MODERN BIOMETRY

Susan R. Wilson
School of Mathematical Sciences, Australian National University, Australia

Keywords: Design, sampling, data analysis, biology, medicine, health, agriculture, ecology, biodiversity, mathematics, statistics

Contents

Summary

Biometry is a discipline devoted to the mathematical and statistical aspects of biology. The benefits to mankind of biometrical developments—ranging from their applications in agriculture, and in animal and plant sciences, to those in medical science, and public health—have been enormous.

The interface between biology and mathematics presents many challenges. Fundamental, and important, research problems have surfaced here. This process continues, both because of the explosion of biological data with the continual development of new technologies, and because of the development of ever more powerful computers to organize and analyze the plethora of data. In biology, the challenges range from data at the molecular level, to the biosphere. In mathematics and statistics, the challenges range from application of existing methodologies to the development of new ones, tailored to the biological application, with the aim of giving both broader and deeper insights into biological data and biological systems.

1. Introduction

Biometry is a large and complex field that arises from the application of statistics and mathematics to biology. All phases of research in biology, including design and data collection, analysis, and interpretation of results, depend on statistical principles and statistical methods. Discard any notion that biological statistics is all about hypothesis testing and p-values! The aim of a well-planned biological investigation is to gain insight into questions of scientific, biological interest. A well-developed biomathematical model that accurately describes the data aids in understanding what the data say, and so in making predictions and forming new questions. Unfortunately, many non-scientific, ad hoc, procedures are practiced under the banner “biometrics.”Two reasons have been given for the perpetuation of this state of affairs. Firstly, many biologists and medical researchers are trained without getting any real insight into the methods of science. Secondly, many editors trained in this manner will not accept papers for publication unless they follow these (often well-ingrained) ad hoc procedures.

Underpinning many areas of biometry is the mathematics of probability. In particular, special stochastic models are often developed. Examples include models in genetics, in particular in population genetics, in epidemic theory and predator–prey interactions. The type of model depends on the context. Sir David Cox has drawn rough distinctions between purely empirical models and (at the other extreme) “toy” models; in between lie the intermediate and quasi-realistic models. A purely empirical model has no direct link with the underlying biological process or corresponding interpretation of the parameters. An example of an empirical model is the fitting of a curve to, for example, AIDS incidence data observed over time. In a “toy” model, a highly idealized representation is used to explore the particular circumstances under which a phenotype of interest could be generated from simple starting assumptions. Examples include biological models showing conditions for the extinction or explosion of epidemics, or the extinction of species by competition. An intermediate model is one in which some aspects of a complex biological process are represented, with the objective of obtaining a formulation such that the resulting parameter estimates do have a link with the underlying generating process. An example is the back-calculation procedure used to predict HIV incidence from observed AIDS incidence, assuming a particular formulation for the incubation distribution. Such a model can produce a reasonable fit to the AIDS incidence data. A quasi-realistic model involves complex processes in biological systems. Such models are usually deterministic rather than having an explicit stochastic component.

A natural question arises: when is the introduction of a stochastic element into a model likely to be crucial? For example, in epidemic models the deterministic model gives the corresponding stochastic means, but for small systems such a model may give a poor idea of the behavior of the sample paths. However, the often more biologically realistic, deterministic "toy" models (for which explicit solutions can often be more easily found) can be used with even more realistic and elaborate stochastic models in the interpretation of results from a complex simulation model. For example, the ratio of relevant response variables may be examined in comparison with those predicted by the "toy" model.

For data collection and analysis, the same fundamental principles apply to experiments, to observational studies, and to the secondary analysis of data collected (usually) for another purpose (such as for a disease registry). In essence, the aim of the study is to provide insight by means of numbers, and it is useful to distinguish three broad headings:

- Collection of data.
- Organization of data.
- Drawing conclusions from data.

In all types of study, the key initial questions are:

- What units (individuals) should be included?
- What properties should be measured, and how?
- What interactions should be examined?

Essentially, in the planning stages, one needs to consider how to control the *random* error, and avoid *systematic* error (bias). There is an enormous literature on these issues. Basic principles of design need to be better understood. This is especially true in the laboratory sciences, where a widely-held but erroneous view is that refinement of laboratory technique is preferable to statistical methods for error control. A key element of the scientific method is replication. This needs to be better understood, and applied more in practice.

Good tabular and graphical procedures are invaluable, and have become easier to produce with developments in computer software. Certainly graphical procedures should be more widely used, both for exploratory work and in presenting the conclusions of more elaborate analyses. Also, a current research focus is the area of graphical models that allow interactions between parameters to be clearly shown, and complex structures can now be fitted relatively easily using modern computer power. In statistical analyses that are heading towards the "conclusions" stage of the study, the methods depend, at least in part, on an explicit probabilistic base. When choosing a model, the following should be considered:

- The model should be consistent with previous related studies of the topic.
- If possible, the model should establish a link with underlying substantive knowledge.
- The model should be consistent with, or suggest, a process that might have generated the data.
- The parameters should have clear interpretations, and the error structure should be such that measures of precision are meaningful.
- The fit to the data should be adequate.

These days there is an increasing need to relate the primary conclusions in different studies, including examination of the consistency of conclusions. In medical research this has lead to much interest in meta-analysis. Another fundamental concern for many scientists is the underlying "causal" process. Causality is a "slippery" notion, and a cautious usage is that strong evidence for causality can only come from the synthesis of different kinds of data. Closely related are the complex issues of generalizability and specificity, as well as the importance of any interaction/s that might be present, but has been assumed absent.

Concerning inference, there are many different approaches, ranging from pure likelihood to the Fisherian approach (with its emphasis not only on likelihood but also sufficiency, conditionality and ancillarity), and the Neyman-Pearson approach (with emphasis on power), to the Bayesian approach. The use of highly sophisticated models in the analysis is not always necessary. Both sensible statistics and sensible biology will prescribe the final form of the quantitative model. The underlying assumptions of any model invoked in the analysis must be carefully checked; if any violations occur, the biological significance of the information provided by these models will be greatly reduced. A pointer to the need to examine assumptions will occur when the data disagree markedly with expectations. Moreover, the best quantitative models are useless anyway if they have little relevance to the biological processes they were meant to describe. Certainly analyses should not be an end in themselves; rather they may provide a springboard to as-yet unanswered questions about the biological system being studied. Overall, the selection of an appropriate statistical model for the analysis will be iterative. Both biological and statistical principles are needed to define and refine the quantitative models used to describe the biological processes.

Risk assessment and management are important, and there is extensive discussion of these issues in respect of epidemiology, toxicology, and other topics. The role of judgmental probabilities in such situations is central. In general, individuals have little understanding of extreme probabilities, and their evaluation. Quality control and process improvement methods are also used in various biometric applications. For example, they can be used in multicenter clinical trials to provide quality medical evaluation for the final evaluation.

The problems and questions faced by real-world applied biometry are widespread and far-reaching. How do some birds learn to navigate so well? What factors influence the length of time individuals spend in institutions, like hospitals or nursing homes? Can a particular fish species survive being caught and returned to the ocean? Which histological changes predict cancer at a certain site? No essay, like this one, or set of theme headings, can properly capture the richness, breadth, and depth, of biometrical data and its analysis.

There is no shortage of interesting new ideas and challenging problems in biometry, with many of these stemming from the relatively large datasets that are now proliferating. Collaborations between biologists and biometricians are essential in developing biometrical modeling methods for research in biology. In particular, many current and future challenges are being motivated by questions in molecular biology, genomics, proteomics, and molecular evolution. These will require the development of new techniques and theories.

Much of the following is based on material that is given in the bibliography. This overview is not a comprehensive overview of biometry, but offers a broad-brush picture of the past, present, and future of this fundamental discipline that underpins so much Life Support Systems knowledge and ongoing research.

2. History

Since the seventeenth century, biological phenomena, like mortality and morbidity, have been the central concern of those who collected and analyzed statistical data. John Graunt (1629–1674) and William Petty (1623–1687) were two pioneers of this time. During this period, the mathematical theory of probability developed from interests in games of chance, and gambling, and major pioneers were Pierre de Fermat (1601–1665), Blaise Pascal (1623–1662), and Jacques Bernoulli (1654–1705). Abraham de Moivre (1667–1754) also was a pioneer in probability theory. He discovered the approximation of the binomial distribution by the normal distribution, as well as investigating mortality statistics. In the eighteenth and nineteenth centuries, the stimulus was astronomy: leading pioneers included Pierre Laplace (1749–1827), and Karl Gauss (1777–1855), who realized the importance of random errors in observations, and developed the method of least squares (that underpins regression). Another astronomer, Adolphe Quetelet (1796–1874), applied statistical methods to problems in biology and medicine. Pioneers in epidemiology also emerged. Louis René Villermé (1782–1863) correlated the variation in mortality he observed in data collected in Paris with variations in environmental factors. William Farr (1807–1883) studied the distribution and determinants of health disorders in English populations: his studies on mortality differences between different occupations helped in understanding industrial hazards. A major epidemiological discovery of this period was John Snow’s 1854 demonstration, using numerical arguments, that cholera was a water-borne disease.

Francis Galton (1822–1911), a cousin of Charles Darwin, made a substantial input to the birth of biometry. Galton found Darwin’s theories on heredity inadequate. Although he did not deduce the principles of heredity, he did lay down some basic foundations for the application of statistics in the biological sciences, inspired by his interests in the analysis of variability, and his developments in the study of correlation and regression of biological measurements (like heights of fathers and their sons). The modern field of mathematical statistics developed out of biometrical problem-solving, and the stimulus is attributed to Galton’s invention of correlation. During this period, Florence Nightingale (1820–1910) pioneered the compilation of relevant vital and medical statistics, accompanied by vivid and revealing graphic representations.

The mathematical work of Karl Pearson (1857–1936), and his colleagues like Raphael Weldon (1860–1906), laid the foundations for modern biometry, and influenced many, including the pioneer medical statistician Major Greenwood (1880–1949). The dominant figure of twentieth-century biometry was Ronald A. Fisher (1890–1962), whose vast contributions included the development of analysis of variance, maximum likelihood methods, and experimental design. Problems in eugenics and in plant breeding motivated Fisher’s statistical work. Work after the Second World War saw a rise in epidemiologic studies focusing on associations between a wide variety of factors and disease, like smoking and lung cancer. In the US, a famous post-war epidemiological investigation has been the Framingham Study of heart disease (from 1947), and an important application of biometrical principles underpinned the 1954 trial of the poliomyelitis vaccine.

Many diverse problems in evolution and genetics have had a fundamental influence on both probability theory and statistics. Galton and Watson (1874) founded the theory of branching processes as a consequence of their investigations of the extinction of human family names. In the mid 1920s, McKendrick and Kermack developed non-linear birth and death processes in answering epidemic theory problems. The work by William Feller (1906–1970) on stochastic processes was partly motivated by population genetics problems.

Counting process models have been developed for studying patterns of arrivals, and interactions of nerve impulses from different neurons. Markov processes have been used in analyzing membrane channel data, studying the kinetic behavior of ionic channels, and understanding DNA damage caused by ionizing radiation.

Stochastic differential equation models have been used for investigating the depolarization of the membrane potential of spatially distributed neurons. The stochastic nature of the measurements has resulted in new developments in stochastic integration and differentiation, and growth of this mathematical field has been stimulated by neurobiology.

In summary, mathematical and statistical techniques have grown in importance over the past century, as has the way in which these methods have been used in biological research and practice. The “green revolution” in agriculture would have been impossible without these tools. Modern medicine and public health practice depend upon carefully designed and interpreted clinical trials, and upon massive observational datasets.

Finally, the rapid increase in computer power in the modern era has seen development and implementation of new ideas that have made a huge impact on biometrical methodology, both for the design of data collection and for analysis. Insightful graphical procedures have become easier to implement too, and good analyses today are accompanied by relevant graphs.

3. Biometric Data Collection and Analysis

3.1. Experimental Design

Experimental design, particularly its historical application to agriculture, has been an important tool in the advancement of biometry. Sir R. A. Fisher, who established the Statistical Laboratory at Rothamsted Experimental Station in the early 1920s, published two articles on crop variation that led to a worldwide revolution in the technique of agricultural trials. This is widely acknowledged as the starting point of research on experimental design.

The basic concepts to increase the accuracy of an experiment were formulated over the next two decades, namely:

1. To increase the size of the experiment.
2. To refine the experimental techniques as much as possible.
3. To select and organize the experimental material to minimize experimental variability.

The key elements of this last concept include the blocking of experimental units into groups that are as homogeneous as possible, and the use of covariates. Technically, improved experimental design involves development of increasingly sophisticated experimental layouts, along with corresponding methods of analysis. Good experimental design requires an understanding of the objectives, and the nature, of the experimental units. Randomization in the assignment of treatments to experimental units is fundamental to reducing possible biases from other sources of variation, that are either unrecognized or of no importance to the question/s under investigation. Increased computational power in recent times has seen the development of more complex designs, along with computational algorithms. As well as in crop research, experimental designs are used in animal research (for example, into dairy animals and pigs), and in evaluating drugs and other medical treatments. Classical designs for experiments include factorial experiments, fractional factorial designs, balanced incomplete block designs, Latin square designs, and Lattice designs.

For the basic analysis of such designs, Fisher introduced what is termed the analysis of variance (ANOVA). This provides a worktable for evaluation of the null hypothesis of all levels of categories of treatment having the same effect on the (continuous) outcome measured. Under the assumption of normality, relevant ratios of the mean squares (sums of squares divided by their degrees of freedom) can be shown to follow well-known distributions, against which values from the relevant test of the null hypothesis can be evaluated. There are strong links between these models and linear regression. The approach is readily extendable to accommodated concomitant variables (ANCOVA: analysis of covariance). These days the data can be clustered, and random effects models, as opposed to fixed effects models, can be utilized. Robust approaches to guard against inappropriate assumptions have also become increasingly popular.

Models of experimental data are often for prediction purposes: for personal ozone exposure assessment, for example, or for estimating seedling responses to environmental effects. When making inferences over some population of conditions, it is important that the increased uncertainty that naturally accompanies the broader inference is reflected in the associated measures of confidence. Assuming, say, that a fixed effects model generally overstates the level of confidence in the estimates, it is important that the experiments have randomly sampled the reference populations of conditions. This can be a problem if, for example, research sites are chosen for specific reasons, such as convenience. Environments sampled can almost never be regarded as truly random, and one hopes that the environments encountered over the experiment's timespan are reasonably representative. To the extent that the sample of environments is not representative, there will be bias in the estimates.

Good design of experiments, and corresponding good data analysis, remain fundamental to good research. The number of important areas that benefit from good design are always growing. Today they include topics from diverse research areas: evaluations of chemical pollution; of effects in social experiments, such as offering economic incentives to see if there are effects on lengths of stay in nursing homes; of the feeding behavior of birds, to gain insight into their spatial association learning; and research on the ozone level. Often, cost considerations can be accommodated. For example, recent research has shown diagnostic tests for disease prevalence on pools of serum samples can, when properly designed, reduce cost and yet increase precision. As computer power has been increasing, so we can contemplate greater complexities in our modeling and data collection, and can better combine fragmentary data while simultaneously analyzing multivariate responses.

-
-
-

Bibliography

Anderson, R. M.; May, R. M. 1991. *Infectious Diseases of Humans: Dynamics and Control*. Oxford, Oxford University Press.

Armitage, P.; Colton, T. (eds.) 1998. *Encyclopedia of Biostatistics*. Chichester, UK, Wiley. 4913 pp. [Major complementary source.]

Armitage, P.; David, H. A. (eds.) 1996. *Advances in Biometry*. New York, Wiley. 473 pp. [Collection of articles to show the aims and achievements of the International Biometric Society and its journal Biometrics.]

Biometrics. http://stat.tamu.edu/Biometrics/ [Scientific journal of the International Biometric Society, containing the latest biometrical research.]

Bookstein, F. L. 1997. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge, Cambridge University Press. 435 pp.

Bürger, R. 2000. *The Mathematical Theory of Selection, Recombination, and Mutation*. Chichester, John Wiley. 409 pp.

Carlin, B. P.; Louis, T. A. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. 2nd edn. Boca Raton, Chapman and Hall/CRC. 399 pp. [Contains many biometrical applications.]

Diggle, P. 1990. *Time Series: A Biometrical Introduction*. New York, Oxford University Press. 257 pp.

Ewens, W. J.; Grant, G. R. 2001. *Statistical Methods in Bioinformatics: An Introduction*. New York, Springer-Verlag. 476 pp. [Introductory account of the probability theory, statistics, and stochastic process theory appropriate to computational biology and bioinformatics.]

Johnson, N. L., Kotz, S. (eds.) 1988. *Encyclopedia of Statistical Sciences*. New York, Wiley. 5 vols. [Major complementary source.]

Journal of Agricultural, Biological and Environmental Statistics (JABES). http://www.tibs.org/jabes/ index.html. [Major scientific journal.]

Lange, N.; Ryan, L.; Billard, L.; Brillinger, D.; Conquest, L.; Greenhouse, J. 1994. *Case Studies in Biometry*. New York, Wiley. 496 pp. [Twenty-one biometric studies at the intersection of theory and applications.]

Levin, S. A. (ed.). *Mathematics and Biology: The Interface Challenges and Opportunities*. http://www.bis.med.jhmi.edu/ Dan/mathbio/T.html. [Report on opportunities at the interface between biology and mathematics.]

Maindonald, J. 2000. *The Design of Research Studies: A Statistical Perspective*. http://www.anu.edu.au/graduate/pubs/ occasional-papers/gs00_2.pdf. [Excellent notes addressing broad planning principles that apply to many research areas; source of references to various biometrical areas, including experimental design and evidence-based medicine.]

Manly, B. F. 1991. *Randomization and Monte Carlo Methods in Biology*. New York, Chapman and Hall. 281 pp.

McCullagh, P.; Nelder, J. A. 1989. *Generalized Linear Models*. 2nd edn. London, Chapman and Hall. 511 pp. [Deals with an important and useful class of statistical models.]

Methodology Group, NHS R&D Health Technology Assessment Programme. http://www.hta.nhsweb.nhs .uk/. [Downloadable monographs and review papers on health technology.]

Rasch, D.; Tiku, M. L.; Sumpf, D. 1994. *Elsevier's Dictionary of Biometry*. Amsterdam, Elsevier. 887 pp.

Sokal, R. R.; Rohlf, F. J. 1995. *Biometry: The Principles and Practice of Statistics in Biological Research*. 3rd edn. San Francisco, Freeman. 859 pp. [Popular introductory text.]

Statistics in Medicine. http://www.interscience.wiley. com/. [Major scientific journal covering biostatistics.]

Statistical Methods in Medical Research. http:// www.smmrjournal.com. [Major scientific journal covering biostatistics.]

Taylor, H. M.; Karlin, S. 1998. *An Introduction to Stochastic Modeling*, 3rd edn. San Diego, Academic Press. 631 pp. [Popular text.]

Welsh, A. H. 1996. *Aspects of Statistical Inference*. New York, Wiley. 451 pp. [Introduces modern inferential procedures motivated by real biological problems.]

Biographical Sketch

Susan Wilson is Professor and Head, Statistical Science Program, Centre for Mathematics and its Applications, School of Mathematical Sciences, and Co-Director, Centre for Bioinformation Science (joint with John Curtin School of Medical Research), at the Australian National University (ANU). She obtained her B.Sc. from the University of Sydney, followed by her Ph.D. from the ANU in 1972. Sue then spent two years as a Lecturer in the Department of Probability and Statistics at Sheffield University. She returned to ANU towards the end of 1974 and has since held a variety of positions there, both in some of the Statistical groupings, as well as at the National Centre for Epidemiology and Population Health.

Susan has over 150 publications in biometry and applied statistics, with a particular emphasis on statistical genetics/genomics. These papers have arisen from her extensive consulting experience in the biological, social, and medical sciences, leading to statistical modeling developments to answer substantive research questions in these disciplines. She is currently involved in the establishment of a bioinformatics research facility at ANU.

Susan is an elected member of the International Statistical Institute, a Fellow of the American Statistical Association and a Fellow of the Institute of Mathematical Statistics (IMS). She was President, International Biometric Society, 1999–2000 (Vice President, IBS, 1998, 2001). Currently she is Associate Editor, *Annals of Human Genetics*; Associate Editor, *Computational Statistics and Data Analysis*; Member, Editorial Board, *Statistical Methods in Medical Research*; Member, Conference Advisory Committee, IBS; Member, IMS Committee on Memorials; Member, IMS Nominations Committee; Member, Editorial Committee for the 6th edition of the ISI's *Dictionary of Statistical Terms*.