STATISTICAL METHODS IN LABORATORY AND BASIC SCIENCE RESEARCH

Michael A. Newton
University of Wisconsin, Madison, WI, USA

Keywords: Statistics, Probability, Modeling, Randomization, Molecular biology, Luria Delbrück fluctuation test, Comparative genomic hybridization, Gene expression microarrays

Contents

Summary

To think statistically is to know that the measurements taken in an experiment are subject to systematic and random sources of variation, and that it is beneficial to base methods of data analysis on probabilistic models. Mathematical results from statistical theory indicate certain types of distributions that govern fluctuations in data, and some of these results are reviewed as they bear on statistical analysis in the laboratory and basic sciences. An example of statistical thinking to advance knowledge in molecular biology is described, as are some general strategies for statistical analysis that may be appropriate for a collaborating statistician. Four case studies demonstrate these concepts.

1. Introduction

Only by measurement does the experimentalist record features of the system that he or she is studying, be the system a population of insects growing in the laboratory or the network of biochemical events that cause a cell to divide or a tumor to grow.

Measurements arise in a context within which their naked numerical form acquires the weightier status of information. The process of extracting information from numerical data is a central issue in the field of statistics generally and in applications of statistics to laboratory and basic sciences.

Were it known with certainty the numerical values of measurements that are about to be taken in some experiment, it would seem to be a waste of effort to take the measurements at all! Measurements are unpredictable. Even with a good understanding of the measurement process and the system being studied, one often acknowledges that precise recordings will exhibit unpredictable fluctuations caused by different sources of variation. In spite of these fluctuations, part of the variation may be more systematic, and repeated measurement may elucidate these systematic sources. Statistical methods are ways of processing numerical data for the purpose of drawing inferences about the system: inferences may be to estimate a parameter, test a hypothesis about the parameter, classify an experimental unit into one of several groups, assess the relationship between two factors, predict future measurements, or decide on one of several courses of action in an ongoing experiment.

Statistical methods become enacted during data analysis. The statistical approach to data analysis is founded on the premise that measurements are the realization of a stochastic process. This has a significant effect on the tone of deliberations because emphasis shifts immediately from the particular data in hand to the process by which the data arise. Indeed, many formal discussions distinguish the data which does occur, say x, from the stochastic process X: the function or rule which reveals the actual data $X(\omega) = x$ when the experiment is instantiated as one particular outcome ω amongst a universe of possibilities. This reversion from what actually is measured to what might be measured seems at first to complicate matters, but it is a necessary template for the theory of probability, and it provides a means to make precise quantitative statements about things that are intrinsically unpredictable. Of course it is not to say that the actual data x are ignored – far from it; rather, the significance of particular irregularities in x is gauged in part by the probabilities governing X.

This chapter considers elements of statistical thinking that arise in laboratory and basic sciences. The comments are informed primarily by the experience of being a research statistician who collaborates with biological scientists, and the emphasis is much more on statistics in molecular biology than statistics in the basic sciences generally. Some mathematical results from statistical theory described in the next section are followed by some comments on the role of statistics at different levels of investigation. This is followed by a discussion of data analysis strategies and then a series of four case studies in which statistical thinking has been helpful.

2. Theory: Universal Distributions

There is great diversity in the systems being studied in basic science laboratories. One of the contributions of statistical theory is to identify common structures present in a wide range of experiments — in particular, common features of the variation of certain measurements. The Poisson limit law is a good example. Suppose that the system under consideration is comprised of a large number n of experimental units, and each of these

units provides a binary response to some query. For instance, millions of bacterial cells are growing in culture and one asks whether or not each cell has a particular genetic mutation at one locus in the genome. The total number Y of units which have one of the binary states may be a quantity of some interest as it may affect the experimental design to pinpoint the locus, for example. Under a wide range of conditions on the basic binary variables it is known that fluctuations in Y are well approximated by a Poisson distribution:

$$\text{Prob}(Y = y) = e^{-\lambda}\lambda^{y}/y! \qquad y = 0,1,\ldots$$

where λ is the expected value of Y. Usually this result is presented in the special case where the binary variables are independent and identically distributed Bernoulli random variables. Then the sum Y has a Binomial distribution with parameters n and p, the common expectation of all the Bernoulli variables. With large n and small p, and $\lambda \approx np$, the Poisson approximation becomes valid. The assumptions of independence and common distribution of the binary variables are rather strict, and evidence has mounted that the Poisson approximation may work much more broadly. Indeed the Poisson clumping heuristic theory extends the result significantly; there can be quite complicated forms of dependence amongst the binary variables and still the Poisson limit holds. This is important since in many examples some dependence is expected. For instance, cell lineage effects will cause statistical dependence in the bacterial growth example.

The most important universal distributional result is the central limit theorem which concerns fluctuations in the arithmetic mean of a random sample. It provides conditions under which the sampling distribution is Gaussian (bell-curved) regardless of the nature of fluctuations in the variables which comprise the sample mean. Indeed the theory is really a collection of results dating back to the early work on probability by many including de Moivre, Laplace, and Gauss, and culminating with 20th century work by Polya, Lindeberg, Feller, Levy and others.

Other universal laws receive perhaps less attention but are still very important for making connections between diverse problems. When appropriately centered and scaled, for example, the largest observation in a large random sample must exhibit variations from one of exactly three well-characterized distributional forms, regardless of the sampling distribution of the data (This is sometimes called the extreme value trinity theorem, developed by E. J. Gumbel and others.) The Erdos-Renyi law and extensions of it concern the distribution of long head-rich runs in sequences of coin tosses, and this has found significant application in problems of matching biomolecular sequences. Universal long-range dependence structures have been identified in certain kinds of time-series measurements also. Often distributional forms arise as the stationary distribution of a Markov process characterizing random fluctuations in the system over time. For example the Gamma distribution is the stationary distribution of abundance when a population evolves stochastically according to certain constraints. Knowing these universal laws assists both experimental design and data analysis. They can be used for ‘back-of-the-envelope’ sample size calculations, and they can form the basis for more detailed modeling efforts.

3. The Role of Statistics

3.1. Exceptional Cases

A beautiful illustration of statistics in the service of the basic laboratory sciences is the work by S. E. Luria and M. Delbrück concerning heritable changes in bacteria. Before Luria and Delbrück's work in the 1940's it was well known that a bacterial culture exposed to a certain virus could readily die out, but that periodically there would emerge clones of resistant bacteria. Various explanations presented themselves. Possibly some of the bacterial cells adapt to the invading virus and survive to form a resistant colony. Contrary to this adaptation hypothesis is the mutation hypothesis, which has stood the test of time and which is a central element in modern bacteriology. The mutation hypothesis asserts that bacterial variants (i.e., mutants) arise during normal growth of the colony, and that certain mutants resistant to the virus may by chance exist in the culture prior to viral infection. If so, they emerge for observation simply by the process of selection after the virus has killed the sensitive cells; the virus itself does not effect an adaptation of the bacterial cells.

The ingenious experiment devised by Luria and Delbück to address the problem involved a comparison of the variance of resistant cell counts grown under different conditions. In one condition separate cultures each grew from a very small initial population size; in the control condition a single large colony was separated into a similar number of separate cultures. All cultures then were exposed to the bacteriophage (virus) and counts were made of the number of resistant cells in each culture. Regardless of how bacterial variants arise, one can argue that in cultures grown in the control condition (i.e., from subsets of a large colony) there should be Poisson variation in the number of resistant cells. On the adaptation hypothesis, this same level of variation is expected in cultures of the first type, however the mutation hypothesis predicts extra-Poisson variation. Cultures in which a resistant mutant arises early will present a very large number of resistant cells compared to cultures in which the mutant arises later. It was by comparing the variation in cell counts between these two conditions that evidence favoring the mutation hypothesis was derived.

Having a statistical argument central to a major scientific advance is fascinating, especially for people dedicated to the study of statistics; but it seems that such elegant Luria-Delbrük-like case studies are the exception rather than the rule in the application of statistics in the laboratory and basic sciences. Certainly there are wonderful case studies — the formulation of the idea of a tumor suppressor gene was a fundamental advance in cancer research brought about by the work of A.G. Knudson in his statistical analysis of retinoblastoma; the ability to map genes such as those responsible for Huntington's disease and cystic fibrosis is based on statistical properties governing the transmission of DNA during meiosis; the work of Sewall Wright used the variance of phenotypes in different experimental crosses to estimate the number of genetic loci affecting the phenotype. What will be the next great advance?

3.2. Endemic Methods

The ordinary application of statistical thinking in scientific discourse is to characterize imperfect knowledge; it forms one step of many to compile, describe, and report

experimental results. At this level statistical discourse is a basic language for dealing with intrinsic variability; it is endemic in the sense of being regularly occurring, and it affects numerous steps in experimental design and data analysis. As examples of statistical questions consider the following: Upon measuring a cell proliferation rate in two different conditions, are the observed rate differences more than one expects by chance alone? If not, then it may be justified to treat the two conditions as one. When measuring some property of a cell type by preparing cells at different liquid dilutions, how can the measurements be combined efficiently across dilutions? In studying the production of some chemical compound, how can one identify optimal settings of several factors that affect production? Related questions are addressed in the case studies described later.

Supporting the notion that statistical methods are endemic is the fact that basic statistical calculations are embedded in much of the operating software of modern laboratory equipment. For example, a flow cytometer is an important device to determine properties of cells by measuring scattered light and fluorescence of mobilized, fluid-suspended cells. Part of the sophisticated computations built into a cytometer is a statistical discriminant analysis to classify individual cells by features such as cell size or granularity. Statistical discriminant analysis is part of the electronic nose, a device to detect airborne scents for use in food quality testing and other applications.

Statistical calculations are also embedded in the basic protocols of many high-throughput laboratory methods. For example, the technique of comparative genomic hybridization measures DNA copy number variation in cancer cells by processing fluorescence image intensities from labeled tumor and normal DNA that have competitively hybridized to immobilized DNA on a glass slide. Intensity signals from the two sources are measured all along the genome, and statistical signal processing is used to decide when one channel is significantly stronger than the other. Further, DNA microarrays are now widely used to measure simultaneously the level of gene expression of thousands of genes.

A large amount of raw image data constitutes the results of one measurement, and this is processed to create a single record for each gene by a series of statistical manipulations of the image data. For instance, with spotted cDNA microarrays, an algorithm is used to localize each spot, account for local background fluorescence, and normalize measurements across the microarray. Statistical methods are used in these cases, and elsewhere, because they can automatically process large amounts of raw data in a potentially meaningful way.

-
-
-

Bibliography

P. Armitage, 1952. The statistical theory of bacterial populations subject to mutation. *Journal of the Royal Statistical Society. Ser. B (Methodological)*, (1), 1-40. [This is an excellent introduction to the statistical issues surrounding the Luria Dellbruck fluctuation test.]

J. Besag and P. Clifford, 1989, Generalized Monte Carlo significance tests, *Biometrika*, , 633–642. [This paper clearly describes the structure and application of Monte Carlo tests both generally and as applied in several spatial problems.]

G.E.P. Box, W.G. Hunter, and G.S. Hunter. *Statistics for experimenters*, Wiley: New York, 1978. [This book is a classic introduction to statistical issues in both the design and analysis of experiments affected by multiple sources of variation. Emphasis is on linear statistical inference.]

W.F. Dove, R.T. Cormier, K.A. Gould, R.B. Halberg, A.J. Merritt, M.A. Newton, and A.R. Shoemaker (1998). The intestinal epithelium and its neoplasms: genetic, cellular, and tissue interactions. *Phil. Trans. R. Soc. Lond. B*, , 915-923. [This review paper summarizes the state of knowledge regarding the cellular dynamics of the intestinal epithelium; emphasis is on mouse experiments to understand the factors affecting cancer growth in this tissue.]

M.M. Fisher, J.L. Klug, G. Lauster, M.A. Newton, and E.W. Triplett, 2000. Effects of resources and trophic interactions on freshwater bacterioplankton. *Microbial Ecology*, 40:125-138. [This research article describes Dr. Fisher's work characterizing bacterial biodiversity in a certain freshwater environment.]

M.A. Newton, C.M. Kendziorski, C.R. Richmond, F.R. Blattner, and K.W. Tsui, 2001. On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology*. (1), 37-52. [This research article introduces the empirical Bayes statistical methodology for the analysis of high-throughput -gene expression data.]

M. A. Newton, T. Yeager, C.A. Reznikoff, 1999. A statistical analysis of cancer genome variation. In, *Statistics in Genetics, IMA Volumes in Mathematics and its Applications*, M.E. Halloran and S. Geisser (eds), , 223-236, New York:Springer. [This article reviews the two-step instability-selection probability model as it applies to data measuring genomic aberrations in cancer. Emphasis is on data from chromosome-based comparative genomic hybridizations taken in a bladder cancer study.]

J. Pickands III, 1975. Statistical inference using extreme order statistics. *The Annals of Statistics*, , 119-131. [This is an example of work which considers the extreme-value trinity theorem.]

S. Sarkar, 1991. Haldane's solution of the Luria-Delbrück distribution. Reprinted in, *Perspectives on Genetics*, pp 199-203, J.F. Crow and W.F. Dove editors. The University of Wisconsin Press, 2000. [This excellent review of research on the Luria-Delbruck problem emphasizes both statistical and biological aspects of the problem.]

T.R. Yeager, S. DeVries, D.F. Jarrard, C. Kao, S.Y. Nakada, T.D. Moon, R. Bruskewitz, W.M. Stadler, L.F. Meisner, K.W. Gilchrist, M.A. Newton, F.M. Waldman, and C.A. Reznikoff (1998). Overcoming cellular senescence in human cancer pathogenesis. *Genes and Development*, , 163–174. [This research article describes experiments with human cancer cell lines and manipulations aiming to understand rate-limiting factors affecting cell division.]

Biographical Sketch

Michael Abbott Newton received his Bachelor degree in Mathematics and Statistics from Dalhousie University in Halifax in 1986, Nova Scotia, Canada, after primary schooling on Cape Breton Island. Dr. Newton did graduate work in the Statistics Department at the University of Washington, Seattle, where he earned a Masters degree in 1988 and a PhD degree under the supervision of Adrian Raftery in 1991. Since then he has been on the faculty at the University of Wisconsin at Madison. He is currently professor in the Departments of Statistics and of Biostatistics and Medical Informatics, member of the University of Wisconsin Comprehensive Cancer Center, and member of the Genome Center of

Wisconsin, associate editor at both Biometrics and the Journal of the American Statistical Association, and member of the Genome Study Section at the National Institutes of Health. Dr. Newton's research concerns computational statistical methodology, especially as it arises in the biological sciences. For example, he has studied statistical procedures for assessing estimation uncertainty in evolutionary tree reconstruction, contributing both to the theory of bootstrap resampling and to the earliest Markov chain sampling procedures. He has made various contributions to Bayesian analysis including one of the earliest applications to identify quantitative trait loci in genetics. He has contributed new methodology for characterizing genomic aberrations presented by cancer cells and for analyzing transcriptional fluctuations measured by gene-expression microarrays.