APPLIED STATISTICS

Abdel H. El-Shaarawi
National Water Research Institute, Burlington, Ontario, Canada L7R 4A6

Keywords: Data acquisition, data reduction, exploratory analysis, graphs, tables, sample space, probability, random variable, estimation, models, inference, likelihood, Bayesian methods, experimental design, randomization, replication, control

Contents

Summary

Statistics is a field that has been widely applied to almost all aspects of human endeavors particularly those related to science and technology. This article provides an overview of probability as the foundational bases of statistical methods that justify their use in making inductive inferences based on observed data. In addition, statistical methods used for data exploration, model building and assessment and for data acquisition are briefly presented.

1. Introduction

Observations and experiments produce data that can be used to challenge existing beliefs, make decisions, predict the future and reveal new knowledge. It is not surprising then to see the wide application of statistical methods to almost all types of human activities. National governments periodically collect, analyze and publish demographic, social, economic and environmental statistical data. These are used to document changes, identify causes and make decisions. Many papers published within scientific journals contain some applications of statistical methods. Statistical thinking plays a pivotal role in shaping and guarding the objectivity of scientific methods. In some cases, the interaction of statistics and other disciplines has led to the creation of new branches of sciences with their own journals, societies and educational departments as can be seen in the cases of biostatistics, chemometrics and environmetrics.

One may ask why statistical science is so widely applied. The answer rests on the fact that although data are collected upon a limited number of units, a sample, the objective of statistics is to extract and understand the information contained within the data and to

generalize the results to other potential sampling units that are not included in the sample. It achieves this objective by devising methods for extracting the relevant information contained in the data and by the logical generalization of the results of the sample to un-sampled units.

In this contribution, we shall begin by briefly discussing the probabilistic foundations of applied statistical methods. This is then followed by a broad discussion of the statistical methods for exploratory data analysis, model building, inference, and the design of data collection. We will conclude with some comments about the future practice of applied statistics in a rapidly advancing technological environment.

2. Foundations

The foundations of formal applied statistical methods reside in the theory of probabilities. This arose historically out of discussions of games of chance in the 17th century. From this beginning, it slowly trickled into other areas such as agriculture, official statistics, and now it is applied in many different fields.

Experimental and observational data are subject to a measure of uncertainty. This uncertainty is modeled and analyzed by probabilistic methods. Such methods are mathematically precise and are different from the subjective notion of the word probability used to judge a situation or an event in every-day life.

The mathematical approach begins by the primitive concept of elementary events. All possible such events form a sample space with a subset of these events representing a compound event. For each event, a mathematical measure is attached that represents the probability of its realization.

To clarify these concepts, let us consider a controlled experiment in which three indistinguishable organisms are individually exposed to an equivalent dose of a toxic substance for the same length of time. Let the objective of the experiment be to determine the effects of exposure upon the survival of the organism. If *S* and *D* are used to represent respectively the survival and death of an organism, then an elementary event will correspond to a particular combination of each organism's outcome of *S* or *D*. For example {*S, S, D*} stands for the survival of the first two organisms and the death of the third. The sample space ***S*** consists of the eight elementary events: {*S,S,S*}; {*S,S,D*}; {*S,D,S*}; {*S,D,D*}; {*D,S,S*}; {*D,S,D*}; {*D,D,S*}; {*D,D,D*}.

The next step is to assign a probability to each of these events. For the purpose of illustration, let us assume that the toxic substance has no effect on the organism. Hence that it is logical to view each of these events as equally likely. Thus, $P(S) = P(D) = 1/2$. Therefore, the probability of an elementary event is given as $P\{\textbf{elementary event}\} = (1/2).(1/2).(1/2) = 1/8$.

From this probability measure, we can calculate the probabilities of various compound events by summing the probabilities of all its elementary events. Let *R* be the number of survivors. Then *R* takes on the values 0, 1, 2, 3 with the following probabilities: $P\{R=0\}$

= 1/8; $P\{R=1\}$ = 3/8; $P\{R=2\}$ = 3/8 and $P\{R=3\}$ = 1/8. We note that the sum of the probabilities of R over its entire set of values is one.

R is a compound event. Therefore, each of its possible events is made up of a number of simple events. For instance, the event $R = 2$ corresponds to the three events $\{S,S,D\}$, $\{S,D,S\}$ and $\{D,S,S\}$, so that $P\{R=2\}=3/8$. A variable defined analogously to R is called a random variable. It represents a function defined on the sample space ***S*.** Here R takes on a set of discrete values and for every possible value j of R there exists a non-negative number called the probability of $R=j$, (that is $P\{R=j\}$), such that

$$0 \le P\{R = j\} \le 1.$$

If instead of three organisms, n organisms are used and instead of assuming that the toxic substance has an effect, we assume that $P\{S\} = \theta$ and $P\{D\} = 1 - \theta$, where $0 \le \theta \le 1$, then R is distributed as a binomial distribution with

$$P\{R = j\} = \binom{n}{j} \theta^{j} (1-\theta)^{(n-j)} \quad \text{for } j = 0, 1, \ldots, n$$

This binomial distribution serves as an approximate model to many natural phenomena. There are other probability models for discrete data. These include the multinomial, Poisson, hypergeometric and negative binomial distributions (see Johnson, Kotz and Kemp (1992)). There are also multivariate versions of these distributions that can be found in the book by Johnson, Kotz and Balakrishnan (1997).

Univariate and multivariate probability models are also available for continuous data. Here the sample space is continuous and the probability of realizing a specific value is zero. We speak instead of the probability that the random variable is realized within an interval. Examples of common continuous models include the normal, gamma, uniform, beta, Cauchy, student t, chi-square and Weibull distributions. Most of these models are given in statistics books (see Evans, Hastings and Peacock (2000) or Kendall and Stuart (1976)). Let $f(x,\alpha)$ be the probability density of a continuous random variable X and α is the parameter of the distribution. Note that both X and α may be vectors.

One particular continuous distribution that deserves further discussion is that of the normal distribution. A random variable X is normally distributed if it has the probability density function $\frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma} \right)^2 \right]$. This distribution has been at the cornerstone of statistical theory and applications for decades. Classical regression analysis and the associated analysis of variance methods are based upon this distribution or upon distributions derived from it. These include the student t, the chi-square and the Fisher F distribution. Moreover, the normal distribution (univariate or multivariate) has been derived as the asymptotic limiting distribution for many frequently used test statistics (Van der Vaart (1998)).

Some characteristics of the probability models are frequently of interest in applications. These include the mode, median, quantiles and the moments of the random variables. The mode is the value of the random variable that maximizes its probability density. The median is the value of the random variable that divides the total area under the probability curve into equal parts. Finally the *k*th order moment of a random variable is

$$E(X^k) = \mu_k' = \sum_{i=0}^{\infty} x_i^k P(X = x_i) \quad \text{for a discrete random variable}$$

$$E(X^k) = \mu_k' = \int_{-\infty}^{\infty} x^k f(x, \alpha) dx \quad \text{for a continuous random variable.}$$

Of particular interest to most experimenters, are the first four raw moments. From them, we can derive the expectation E(*X*) and the variance Var(*X*) and the coefficients of skewness γ and kurtosis β:

$$\mathrm{Var}(X) = E(X - \mu_1')^2 = \mu_2' - (\mu_1')^2$$

$$\gamma = \frac{E(X - \mu_1')^3}{(\mathrm{Var}(X))^{3/2}}$$

and

$$\beta = \frac{E(X - \mu_1')^4}{(\mathrm{Var}(X))^2} - 3.$$

The expectation (the first raw moment), the median and the mode are known as the measures of location of the distribution. They have the same value for symmetric distributions. The variance is a measure of the spread of the distribution while the skewness and kurtosis measure the shape of a distribution. The skewness coefficient is zero for a symmetric distribution. The kurtosis determines if the distribution has a heavy or light tail relative to that of the normal distribution.

-
-
-

Bibliography

Anderson, E. (1960). A Semi-graphical Method for the Analysis of Complex Problems. Technometrics, 2, 387-392. [An article on graphical display of multivariate data]

Andrews, D. F. (1972). Plots of High Dimensional Data. Biometrics, 28, 125-136. [An article that suggests the use of curves to represent multivariate data graphically]

Atkinson, Anthony and Riani, Marco (2000). Robust Diagnostic Regression Analysis. Springer-Verlag, New York. [A modern book on regression]

Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, Revised edition, Holden-Day, San Francisco.[A classic basic reference on time series].

Brockwell, P.J. and Davis, R.A. (1988).Time Series: Theory and Methods, 2nd Edition, Springer-Verlag, New York. [A basic reference on time series].

Chernoff, H. (1973). The Use of Faces to Represent Points in k-Dimensional Space Graphically. J. Amer. Statist. Assoc., 68, 361-368. [An article in which the author introduced the human face for the display of multivariate data]

Cleveland, W. S. (1993). Visualizing Data. Hobart Press, Summit, New Jersey. [A basic reference]

Cleveland, W. S. (1994). The Elements of Graphing Data. Hobart Press, Summit, New Jersey. [A well written reference on approaches for the graphical display of data]

Cochran, W. G. and Cox, G. M. (1957). Experimental Designs. New York: Wiley. [A classic book on statistical approaches for the design and analysis of experimental data]

Cressie, N.A.C. (1993). Statistics for Spatial Data, Revised edition, New York: Wiley. [A comprehensive book on spatial statistical methods]

David, H. A. (1981). Order Statistics, second edition. New York: John Wiley & Sons. [A basic book on order statistics].

Davison, A. C. and Hinkley, D. V. (1998). Bootstrap Methods and their Applications. Cambridge University Press. [An excellent book with many applications]

Draper, N.R. and Smith, H. (1981). Applied Regression Analysis, 2nd edition, New York: Wiley. [A basic book on regression].

Efron, B. (1979). Bootstrap methods, Another Look at the Jacknife, Ann. Statist. 7, 1-26. [Original research paper by the founder of the bootstrapping as a method for statistical inference]

Efron, B. and Tibshirani, R. J. (1993). An Introduction to the Bootstrap. London: Chapman & Hall. [An elementary reference on bootstrapping. It presents the basic fundamentals].

Evans, Merran, Hastings, Nicholas, and Peacock, Brian (2000). Statistical Distributions, Third edition, New York: John Wiley & Sons. [An easy to use reference on commonly used frequency distributions]

Gnanadesikan, R. (1977). Methods for Statistical Data Analysis of Multivariate Observations. John Wiley, New York. [An excellent applied book on multivariate analysis methods]

Hall, P. (1992). The Bootstrap and Edgeworth Expansion, Springer-Verlag. [The theory of bootstrap is presented in this book].

Härdle, Wolfgang (1990). Smoothing Techniques with Implementation in S, Springer-Verlag. [A good reference on kernel smoothing with software routines for their implementations]

Hoaglin, D.C., Mosteller, F., and Tukey, J. W. (1985). Exploring Data Tables, Trends and Shapes. New York: Wiley. [A fundamental book on exploratory data analysis methods]

Huber, P. (1981). Robust Statistics, New York: John Wiley & Sons. [A systematic theory for robust statistical inference is presented in this book].

Jeffreys, H. (1939). The Theory of Probability. Oxford: Oxford University Press. [A classic book on the development of theory of the objective specification of non-informative prior probabilities].

Johnson, N. L., Kotz, S., and Balakrishnan. N. (1997). Discrete Multivariate Distributions, New York: John Wiley & Sons. [A basic book on multivariate discrete distributions]

Johnson, N. L., Kotz, S., and Kemp, A. W. (1992). Univariate Discrete Distributions, Second edition, New York: John Wiley & Sons. [Provides a summary of known distributions along with their main properties]

Kendall, M.G. and Stuart, A. (1976). The Advanced Theory of Statistics, Vol. 1, Griffin, London. [A classical book for common distributions and their characteristics].

Krzanowski, W.J. (1988). Principles of Multivariate Analysis. Oxford University Press. [A comprehensive reference for various aspects of multivariate analysis. It contains applications to real life applications].

Plackett, R. L. (1949). " A Historical Note on the Method of Least Squares", Biometrika, vol. 36, part 3 and 4, p. 458. [It gives the history of the discovery of the celebrated method of least squares].

Tukey, J. W. (1960). A Survey of Sampling from Contaminated Distributions. In Contributions to Probability and Statistics, pp. 448-485, eds. Olkin, I. et al. Stanford University Press.[This paper shows the effects of minor errors upon the performance of classical methods].

Tukey, J.W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading. [A fundamental reference containing ideas about various ways to extract signals from messy data]

Van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press. [Derivation of asymptotic distributions of test statistics and their use in practical applications are the focus of the book. It is well written and not too difficult to study].

Venables, W.N. and Ripley, B.D. (1997). Modern Applied Statistics with S-PLUS (2nd ed.). New York: Springer-Verlag. [This book is directed mainly to the users of SPLUS routines by providing examples of their use in applications].

Viertl, R. (1996). Statistical Methods for Non-Precise Data. CRC Press, Boca Raton, Florida. [A fundamental reference on the treatment of non-precise data]

Viertl, R. (1997). Statistical Inference for Non-Precise Data. Environmetrics, 8, 541-568. [A research article describing the steps involved in making inferences from non-precise data].

Biographical Sketch

Abdel El-Shaarawi was born in Egypt. He received his B.Sc. and M.Sc. degrees from Cairo University and his Ph.D. in Statistics in 1972 from University of Waterloo. He is a research scientist at the Canada Centre for Inland Waters in Burlington, Ontario and a professor in the Department of Mathematics and Statistics, McMaster University and is currently a visiting professor at the University of Genova. He is founding Editor of the journal Environmetrics and founding President of The International Environmetrics Society. He is an elected member of the International Statistical Institute and a Fellow of the Royal Statistical Society (UK), the American Statistical Association and the Modelling and Simulation Society of Australia and New Zealand. He received several Awards including the Distinguished Achievement Medal of the ASA Section on Statistics and the Environment and the Citation of Excellence Award from the Government of Canada.