PROBABILITY AND STATISTICS

Reinhard Viertl
Institute of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria

Keywords: Bayesian statistics, confidence estimates, conventional statistics, data quality, data science, descriptive statistics, estimation, fuzziness, imprecise data, probability, probability distributions, randomness, statistical hypotheses, statistical tests, statistics, statistics and non-precise data, stochastic processes, stochastic quantity, stochastics, uncertainty

Contents

1. Introduction
2. Origin and History
3. Probability
3.1. Continuous Probability Distributions
3.2. Discrete Probability Distributions
3.3. Mixed Probability Distributions
3.4. Mixture Distributions
4. Descriptive Statistics
4.1. Empirical Distributions
4.2. Empirical Distribution Function
4.3. Multivariate Data
4.4. Regression
4.5. Indices
4.6. Time Series
5. Stochastic Models
5.1. Stochastic Vectors
5.2. Correlation and Independence of Stochastic Quantities
5.3. Conditional Distributions
6. Sequences of Stochastic Quantities
6.1. The Law of Large Numbers
6.2. Central Limit Theorem
7. Stochastic Processes
8. From Stochastic Models to Statistical Inference
9. Classical Statistical Inference
9.1. Point Estimates for Parameters and Other Characteristic Values of Probability Distributions
9.2. The Fundamental Theorem of Statistics
9.3. Further Methods in Classical Statistical Inference
10. Bayesian Statistical Inference
11. I nformation and Decision
12. Types of Uncertainty and Data Quality
13. Outlook
Glossary
Bibliography

Biographical Sketch

Summary: Modeling Uncertainty and Descriptive Measures of Reality

Many of the most important quantities in life sciences are uncertain, that is, they are not deterministic. This uncertainty is an inescapable part of life and is called “randomness” or “stochastic variation.” For example the lifetime of a person cannot be predicted in advance.

Probability theory provides powerful tools for the construction and analysis of mathematical models of phenomena with random features. Today probability models are the basis of a great number of applications in modeling different fields of human activity. This especially concerns its use in statistics as well as in modeling different phenomena by stochastic processes for evaluation and prediction in connection with environmental pollution and economic affairs.

Statistics is the science of data and making valid inferences about the characteristics of a group of elements on the basis of information obtained from a random sample of that group. There are two major subdivisions of statistics: “descriptive statistics” and “statistical inference.”

One principal descriptive quantity derived from data is the “mean,” or arithmetic average of the data. It generally is the most reliable single measure of the value of a typical member of the sample. However, if a sample contains some values of a size that have an exaggerated effect on the value of the mean, then the value of a typical member is more accurately represented by the median, which is the value that half the sample values fall below and half above. Measures of the dispersion of values about their mean include the “variance” and the square root of the variance, known as the “standard deviation.” The variance is calculated by determining the mean, subtracting it from each of the sample values, and then calculating the average of the squares of these deviations. The mean and standard deviation provide a complete description of the normal distribution, in which positive and negative deviations from the mean are equally common and small deviations are much more common than large ones.

The theory of statistics is grounded in mathematical probability theory, and has been studied since the seventeenth century. It has been shown that when a sample is drawn at random from a larger population, the composition of the sample is governed by the composition of the population according to well-determined laws of probability. Statistics makes use of these laws through finding ways to infer the composition of the population from that of the sample. These methods belong to statistical inference.

Statistical inferences are of two types: estimation, which is the search for an unknown characteristic value of a population, and hypothesis testing, which involves definitions of one set of possible population values (called hypothesis) and a different set (called alternative). There are many procedures that allow a decision, on the basis of a sample, whether the true population characteristics are more likely to belong to the set of values in the hypothesis or those in the alternative.

Statistics is used in almost every modern human activity; scientific, technological, political, economic, and social. For a very large population, the size of the sample that is needed for standard statistical analysis is generally nearly independent of the size of the underlying population. This allows statisticians to make surprisingly accurate estimates of outcomes on the basis of very small samples.

Probability and statistics can be considered as the two sides of a coin. Probability is the standard mathematical concept to describe stochastic uncertainty. Stochastic uncertainty is the uncertainty concerning some events or a stochastic quantity. An example of a stochastic quantity is the lifetime of a biological system or an engineering system. Since such life times cannot be determined in advance they are called "stochastic quantities," and the "probability distribution" of a stochastic quantity is of interest. The probability distribution of a stochastic quantity gives a number, called probability, to every event connected with the stochastic quantity. This probability distribution is the optimal information concerning a stochastic quantity.

A natural question is how to obtain the probability distribution of a stochastic quantity from observed data. This is one of the main problems of statistical inference.
But when facing data on a stochastic phenomenon, the first task is doing some descriptive statistics with the data. This gives some empirical evidence and first hints concerning the underlying probability distribution of a stochastic quantity. After that more detailed analyses and decisions have to be made in order to find a suitable probability model for the real phenomenon under consideration.

A topic of growing importance is "data quality," this is particularly true in environmental science when describing amounts of certain materials released to the environment. The question of the precision of data is fundamental also in relation to the Rio declaration, the Kyoto protocol, and the den Haag conference on environmental pollution in connection with sustainable development. Moreover the results of measurements are always non-precise. This imprecision is different from errors and is called "fuzziness" which arises due to vagueness or linguistic descriptions such as "large," "small," "old," "young" and so on. Generalizations of statistics for non-precise data are necessary and related results exist.

Probability and statistics are the methods for modeling uncertainty and measuring real phenomena. Today many important political, health, and economic decisions are based on statistics.

1. Introduction

What is probability?

Some scientists think probability is the only mathematical concept to model uncertainty. This is questionable because the uncertainty of a single measurement of a continuous quantity is not only of stochastic nature. Measurement results are not precise real numbers but they are intrinsically also "non-precise." Still probabilities are very important to describe the uncertainty of many real phenomena in different areas of life. One of the most interesting quantities is the lifetime of a biological system. Such life

times are uncertain before the system dies and in general it is not possible to predict the lifetime in the sense of deterministic calculations. Therefore the lifetime is called a "stochastic quantity" (also called random quantity or stochastic variable or random variable).

Stochastic quantities are denoted by one of the symbols $X,Y,Z,T,\tilde{\theta}$ in this article. They are describing parts of the results of "statistical experiments." These are experiments, which can be conducted repeatedly under "identical" conditions. The set of all possible outcomes of a statistical experiment is called the "outcome space" and is denoted by Ω. In the case of lifetimes, the related statistical experiment is the reporting of the lifetime of a unit. In this statistical experiment the outcome space is the set of all non-negative real numbers, i.e. $\Omega = [0, \infty)$.

In survival analysis for different time periods $[t_1, t_2]$, for example a year, it is important to know if the life time falls into this time interval or not. The optimal information, which is possible about this, is the "probability" of obtaining a lifetime from the subset $[t_1, t_2]$ of Ω. Such probabilities are real numbers between 0 and 1 that describe quantitatively the degree of belief that the lifetime will fall into the interval. Therefore certain subsets A of the outcome space Ω are of interest.

For a general statistical experiment it is important to have quantitative information concerning the occurrence of certain subsets A of Ω. Such subsets of the outcome space are called "events." In this set up probabilities are numbers to describe the uncertainty about obtaining an outcome of the statistical experiment, which belongs to an event A. They are denoted by $P(A)$. To construct suitable mathematical models for statistical experiments is the subject of "probability theory." Basic explanations on this are given in section 3.

In probability models usually constants θ appear which give the possibility to adapt a probability model to given empirical evidence. These constants are called "parameters" of the probability model.

For applications the main problem is to find appropriate probability models to describe the variability of quantities based on given data. This is achieved by finding an "estimator" $\hat{\theta}$ for the parameter θ. This is a problem of "statistical inference."

Another problem of statistical inference is to decide if certain assumptions about a probability model are justified in face of empirical given data. The formulation of such assumptions is called "statistical hypothesis." Formal "decision rules" to make a decision if a hypothesis is acceptable or has to be rejected are called "statistical tests."

What is statistics?

A different viewpoint concerning variation is that of "descriptive statistics" which has to describe real world phenomena. Here the starting points are observations (also called data) from surveys, experiments, or records in databases. Topics of importance in descriptive statistics are numerical summaries of data sets, time series, that is the

display of the time dependence of certain quantities in environmental analysis, the description of the dependence between different quantities, calculation of indices related to certain phenomena, measures of concentration, empirical distributions, and others.

The first step in statistical data analysis is to obtain an overview and a first classification of given data concerning a real phenomenon. The methods for that belong to "descriptive statistics." Descriptive statistics is the major methodology to describe issues of global change and sustainability. All kind of assessments needs enormous amounts of data that carry variability and uncertainty. The scientific method to survey, collect and analyze them is statistics.

Since there are different kinds of data, last but not least questions of how to measure real phenomena as well as problems of data quality are important. Especially in environmental science the quantitative description of data quality and precision is an important topic (see "Statistical inference with non-precise data," EOLSS on-line, 2002). Descriptive statistics uses no probability theory. The methods of statistical data analysis using probability concepts belong to statistical inference.

Related to statistical experiments for stochastic quantities "statistical inference" provides objective methods for the fitting of suitable stochastic models—also called *probability models*—to empirically given data of a stochastic quantity. Details on this are given in "Foundations of statistics," "Applied statistics," and "Statistical inference," EOLSS on-line 2002.

An important but less respected topic is the interpretation of results from statistical work in a way suitable for decision-makers.

The whole field of probability and statistics consists of the following areas:

- data collection and data modeling
- descriptive statistics
- probability theory
- statistical inference
- theoretical statistics
- applied statistics.

These areas are overlapping but give a first orientation in the field of statistical data science.

2. Origin and History

Historically there is a big difference between statistics and probability. Statistics has a much longer tradition than probability. Statistical work has existed for about 5,000 years, whereas the mathematical treatment of probabilities started almost 500 years ago.

Censuses are mentioned in ancient Egypt, ancient China, and the Bible. Such surveys gave statistical data but no systematic efforts were made to analyze them, and to draw formal conclusions.

The development of probability calculations started around the year 1520 by Girolamo Cardano, and in the middle of the sixteenth century by Galileo Galilei.

In the seventeenth century Blaise Pascal and Pierre de Fermat solved problems from gambling, which is known from the correspondence between them. The first printed book on probability is *De Rationiis in Ludo Aleae*, published by Christiaan Huygens in 1657. Jakob Bernoulli who extended it to economic problems read this book. This work was published after his death by his nephew Nikolaus Bernoulli in 1713 with the title *Ars conjectandi.*

In 1662 the English salesman John Graunt published a study on death statistics in London and stated also the cause of death. This paper is the first well-known example of a statistical analysis of population data. Graunt discovered that in the crowded city, even in years of no epidemic, the number of deaths was higher than the number of births, a result that was contrary to rural areas, where it was just the opposite.

In the eighteenth century the Berlin priest Johann Süßmilch published a book: *Die göttliche Ordnung in den Veränderungen des menschlichen Geschlechts, aus der Geburt, dem Tode und der Fortpflanzung desselben erwiesen*. (This means "The divine order in the changes of human species as demonstrated by birth, death, and propagation.") He collected all the facts on population statistics which were available to him at that time, and investigated causes and consequences and generated new scientific findings. Also in Belgium, England, and France many statisticians, including Adolphe Quetelet and Sir Francis Galton worked on similar problems.

Also at the beginning of the eighteenth century, Abraham de Moivre recognised the limiting relationship between the Binomial distribution and the Normal distribution. In the same century Thomas Bayes wrote his well-known treatise on probabilities of hypotheses.

At the beginning nineteenth century Pierre Simon de Laplace published a survey on probability theory and a philosophical treatise on probabilities. Siméon Denis Poisson found the law of large numbers, and Johann Friedrich Carl Gauss found the normal distribution as limiting distribution and developed the method of least sum of squares. Later John Venn published the book *Logic of Chance* and Andrey A. Markov made basic contributions to the formal description of dependent variables in stochastic processes.

Francis Galton, who developed the correlation coefficient, and Francis Ysidro Edgeworth were responsible for the most important developments in statistical research during the nineteenth century. At the turning point of the centuries George Udny Yule should be mentioned, and the last important contribution to formal statistical inference in the nineteenth century was the publication of the χ^2- goodness of fit test by Karl Pearson in 1900, and this year provided that there was a good basis for further research in statistics and probability.

During the last part of the nineteenth ccentury and throughout the twentieth century mathematically based statistics developed rapidly.

The twentieth century brought a tremendous development of statistics and probability. In 1901 Alexander Michailowitsch Ljapunov proved the central limit theorem. One of the milestones is the dissertation *Intégrale, Longueur, Aire* by Henri Lebesgue in 1902. The Lebesgue integral is one of the basic mathematical concepts in modern probability theory. In 1908 William Sealy Gosset found the *t*-distribution. The most influential scientist in statistics in this century was Sir Ronald Aylmer Fisher. He laid the foundations of modern statistical inference (see "Foundations of statistics" and "Statistical inference," EOLSS on-line, 2002). Other outstanding statisticians in the twentieth century were Abraham Wald, Bruno de Finetti, Leonard Jimmy Savage, William Cochran, Henry Scheffé, Norbert Wiener, Jerzy Neyman, Egon S. Pearson, Lucien LeCam, and John Tukey.

Modern statistical inference is based on probability theory that was axiomatically founded by the Russian mathematician A. N. Kolmogorov in his famous booklet in German language *Grundbegriffe der Wahrscheinlichkeitsrechnung* (this means "Foundations of the calculus of probability") published in 1933.

There were three other books, which influenced the development of probability theory and theoretical statistics tremendously. The first was *Mathematical Statistics* by Harald Cramér, published in 1946. The second, *An introduction to Probability Theory and its Applications* by William Feller in 1950 made the up to date methods of probability theory available to students all over the world. The third was a basic text for theoretical statistics, *Testing Statistical Hypotheses* by Erich L. Lehmann, published in 1968.

In the first forty years of the twentieth century, statistics was not considered to be a serious mathematical subject in the mathematical community of the German-speaking region. There the acceptance of mathematical statistics as a sound mathematical theory was relatively late during and after the Second World War.

Important developments are ongoing and have to be evaluated at a later time.

-
-
-

Bibliography

Andersen, P. K.; Borgan, O.; Gill, R. D.; Keiding, N. 1993. *Statistical Models Based on Counting Processes.* New York, Springer. [Modern statistical methodology for time dependent data based on stochastic process models.]

Ash, R. B.; Gardner, M. F. 1975. *Topics in Stochastic Processes.* New York, Academic Press [Excellent textbook for mathematical aspects of stochastic processes.]

Bhattacharyya, G. K.; Johnson, R. A. 1977. *Statistical Concepts and Methods.* New York: Wiley. [Application oriented text on mathematical methods of statistics.]

Everitt, B. S. 1998. *The Cambridge Dictionary of Statistics*, Cambridge, Cambridge University Press. [Up to date dictionary of statistical terms.]

Gibbons, J. S.; Chakraborti 1992. *Nonparametric Statistical Inference.* New York, Marcel Dekker. [Comprehensive text on nonparametric methods of statistics.]

Lehmann, E. L. 1986. *Testing Statistical Hypotheses*, New York, Wiley. [Standard text on mathematical aspects of statistical tests.]

Mood, A. M.; Graybill, F. A.; Boes, D. C. 1974. *Introduction to the Theory of Statistics*. New York, McGraw-Hill. [Very good text on the principles of probability and mathematical statistics.]

Mukhopadhyay, N. 2000. *Probability and Statistical Inference*, New York, Marcel Dekker. [Standard text on mathematical aspects of probability and statistics.]

Parthasarathy, K. R. 1977. *Introduction to Probability and Measure.* Delhi, Macmillan. [Excellent text on the mathematical foundations of probability theory.]

Prabhu, N. U. 1965. *Stochastic Processes: Basic Theory and Its Applications.* New York, Macmillan. [Excellent introduction to stochastic processes containing also Markov chains and renewal processes.]

Press, S. J. 1989. *Bayesian Statistics: Principles, Models, and Applications.* New York, Wiley. [Basic textbook on the Bayesian approach to statistics.]

Ross, S. M. 1997. *Introduction to Probability Models.* San Diego, Academic Press. [Well-written and comprehensive introductory text to stochastic models.]

Thiessen, H. 1997. *Measuring the Real World: A Textbook on Applied Statistical Methods.* Chichester, Wiley. [Basic text on the application of descriptive methods of Statistics.]

Viertl, R. 1996. *Statistical Methods for Non-Precise Data.* Boca Raton, CRC Press. [Basic text on the description and statistical analysis of non-precise data.]

Biographical Sketch

Reinhard Viertl was born on March 25 1946, at Hall in Tyrol, Austria. He studied civil engineering and engineering mathematics at the Technische Hochschule Wien and received a Dipl.-Ing. degree in engineering mathematics in 1972. He obtained a Doctor of engineering science degree in 1974. He joined as assistant at the Technische Hochschule Wien and became University Docent in 1979. He was a research fellow and visiting lecturer at the University of California, Berkeley, from 1980 to 1981, and visiting Docent at the University of Klagenfurt, Austria in winter 1982. Since 1982, he is a full professor of applied statistics at the Department of Statistics, Vienna University of Technology. He was a visiting professor at the Department of Statistics, University of Innsbruck, Austria from 1991 to 1993, and is invited to be a visiting professor at the University of Calgary in summer 2002.

He is a fellow of the Royal Statistical Society, London, held the Max Kade fellowship in 1980, and is founder of the Austrian Bayes Society, member of the International Statistical Institute, and president of the Austrian Statistical Society from 1987 to 1995. He was invited to membership in the New York Academy of Sciences.

He authored the books *Statistical Methods in Accelerated Life Testing* (1988), *Introduction to Stochastics* in German language (1990), *Statistical Methods for Non-Precise Data* (1996). He edited the books *Probability and Bayesian Statistics* (1987), *Contributions to Environmental Statistics* in German language (1992). He is co-editor of a book titled *Mathematical and Statistical Methods in Artificial Intelligence* (1995), and co-editor of two special volumes of journals. He authored over eighty scientific papers in algebra, probability theory, accelerated life testing, regional statistics, and statistics with non-precise data.

He serves on the editorial board of scientific journals, is editor of the publication series of the Vienna University of Technology, and has organized different scientific conferences.