FOUNDATIONS OF STATISTICS

Reinhard Viertl
Vienna University of Technology, A-1040 Wien, Austria

Keywords: *a-priori* distribution, *a-priori* information, Bayes risk, Bayes strategy, conditional probability, data, decision, ensembles flous, estimation, fuzzy data, fuzzy probability, information, likelihood, loss, measurement, minimax strategy, philosophical foundations, probability, random sample, risk function, sample, sampling, sampling distributions, sensitivity, statistical inference, statistical population, sufficiency, testing, uncertainty, utility

Contents

Summary

The foundations of statistics are diverse. First the quantitative description of real data is basic for statistical analysis. Second for inference procedures probabilities are fundamental. This includes the different kinds of probability definitions. Also the different kinds of uncertainties which are present in applications are essential for statistical analysis. These are stochastic variation, model uncertainty, parameter uncertainty, and data uncertainty which is different from stochastic uncertainty and is called imprecision or fuzziness. Last but not the least, utility and loss considerations are basic for good decisions. Realistic utility modeling has to take care also of the imprecision in utility modeling. Moreover the most elementary probability distributions in sampling normal distributions are given.

1. Introduction

There is a main difference between probability and statistics. Probability provides mathematical models to describe stochastic (i.e. non-deterministic) phenomena. The question how good a stochastic model (also called probability model) describes a real phenomenon is connected with observations of the phenomenon. These observations are the starting point of statistics.

Contrary to probability theory, which is a deductive mathematical theory, statistics is an inductive science and sometimes also an art. Statistical science is trying to describe data sets in a lucid way, to find structures and dependencies in data, to make conclusions from observed date, to look for stochastic models describing phenomena for which data are available in order to provide scientific support for making well founded decisions, and to analyze decision processes in order to arrive at good or even optimal decisions.

The first step in statistical work is of descriptive nature: To summarize data in order to provide easy-to-grasp information. This can be done in different ways. Most important are characteristic values, measures of dispersion, and empirical distributions. For multivariate data, i.e. the idealized observations are vectors, descriptive measures of dependencies between the different components in the vector-valued observations or dependencies on covariates of observed quantities are considered. Details on these methods are explained in the article *Preliminary Statistical Data Analysis.* These methods are also called *descriptive statistics*.

The foundations of descriptive statistics comprise the study of measurement scales, questions concerning the precision of data (compare the article *Statistical Inference with Non-Precise Data*), the study of the suitability of different descriptive measures, and questions concerning information loss by data compression.

Quite different from descriptive statistics are problems of matching stochastic models to real data sets. Problems of this kind belong to *statistical inference*. In statistical inference the fundamental ideas are *stochastic models* and *populations* from which observations are taken. Based on these observations (also called data) inference is made about the underlying stochastic model. See also the article *Statistical Inference*).

2. Statistical Data

Data in life sciences are of different nature. Ranging from quality of life data to high precision measurements there are different types of data. Corresponding to the different types of data there are different *scales of measurement*. For the so-called *categorical data* (also called nominal data) there is no order in the different possible values. Examples are color, nationality, and religion. Data with a natural order in their values are called *ordinal data*. Examples are quality classes, marks for examinations, personal preferences, and rankings in sports. The next type of more structured data is the so-called *interval data*. For such data the values can be ordered and it is reasonable to define differences. The origin of these scales can be defined arbitrarily. Examples of interval data are time measurements, calendar data, and temperature data. The most structured kind of data are the so-called *ratio scale data*. In addition to the quality of interval data there is an absolute origin. Therefore the quotient (ratio) of two values is independent from the used measurement unit. Examples are length, weight, duration, many physical scales, and income data.

Looking at the *metric data* which are represented by numbers or vectors, different kinds of such quantities are distinguished. In case a real valued quantity can assume only a finite number of different possible values (or an at most countable number of possible values which have no accumulation point) the quantity is called *discrete*. If a quantity can assume all values of an interval (a,b) of real numbers, the quantity is called *continuous*. There are also mixtures of both kinds, called *mixed quantities*.

Another classification of observations of metric data is by the dimensionality of the singe data values. For example for every person in a region the following characteristics can be observed: sex, nationality, age, height, weight. This can be represented by a 5-tuple $(x_1, x_2, x_3, x_4, x_5)$. Such data are called *multivariate data*. It should be noted that the scales of the different components can be different. In case one data point is characterized by a vector $(x_1, \cdots, x_k)$ for $k \geq 2$ the data are called *vector data* (or *k*-dimensional data). Examples are measurements of locations in space $(k = 3)$ and air quality data with *k* types of measurements of concentrations of substances. In such vector data the different components can be from different kind, i.e. one component can be discrete and another component continuous.

Another essential aspect of data is their *imprecision*. In case of metric data this means that the result of a measurement procedure of a continuous quantity is never a precise real number but more or less *non-precise*. This is not taken into account in conventional statistics. It is important to note that this concerns the imprecision of one single outcome and not the random variation or errors. The imprecision cannot be described adequately by probability. Since the outcome of one measurement observation is not a precise number a generalization of real numbers is necessary. The most up to date description of non-precise numbers is by the so-called *characterizing functions* $\xi(\cdot)$. A characterizing function of a non-precise number x^* is a generalization of the indicator function of a number. Whereas an indicator function $I_B(\cdot)$ of a subset *B* of $\mathbb{R}$

assumes only the values 0 and 1, i.e.

$$I_B(x) = \left\{ \begin{matrix} 1 & \text{for} & x \in B \\ 0 & \text{for} & x \notin B \end{matrix} \right\} \qquad \text{for} \quad x \in \mathbb{R},$$

a characterizing function can take values from the closed interval $[0,1]$. Moreover a characterizing function $\xi(\cdot)$ of a non-precise number x^* has to obey the following conditions (a) and (b):

(a) There exists a real number x with $\xi(x) = 1$

(b) For every real number $\alpha \in (0,1]$ the so-called $\alpha-$cut $C_\alpha(x^*) = \{x \in \mathbb{R} : \xi(x) \geq \alpha\}$ is a closed finite interval

Therefore a characterizing function can always be represented in the following way: There are two real valued functions $L(x)$ and $R(x)$ such that

$\xi(x) = L(x)$ for all $x \leq a_1$

$\xi(x) = R(x)$ for all $x \geq b_1$

$\xi(x) = 1$ for $x \in [a_1, b_1]$ with $a_1 \leq b_1$

with $L(\cdot)$ is increasing on $(-\infty, a_1]$, $R(\cdot)$ is decreasing on $[b_1, \infty)$.

Remark: Characterizing functions need not being continuous. Therefore also the functions $L(\cdot)$ and $R(\cdot)$ need not to be continuous. In Figure 1 an example of a characterizing function is depicted.

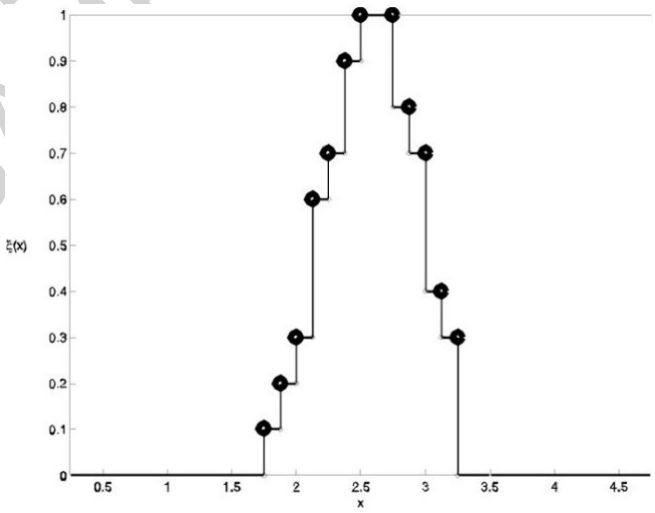

Figure 1: Characterizing function

3. Uncertainty

Observing real phenomena and trying to find a suitable mathematical model to describe uncertainty and to understand dependencies and to make predictions is a central topic of statistics. In this process one is facing different kinds of uncertainty. The first problem is to define a set of basic variables in describing life support systems, so that uncertainties can be treated in a rational way. They are basic in the sense that they are the most fundamental quantities recognized. Examples are properties of materials, environmental loads, quality measurements, and variables describing quality of life. Most quantities that enter into statistical calculations are in reality associated with some uncertainty. Are the magnitudes of all variables bounded or can they be restricted within specified limits? For example limits to resistance are not easily specified. Direct use of limits - if they exist - is extremely uneconomical. Limits imposed by quality control and testing are seldom completely effective. A statistician must be concerned with the nature of the actual variability of physical quantities such as load, life times, material amounts, and others. This variability is also called *physical uncertainty* and can be described in terms of probability distributions or stochastic processes.

The physical variability can only be quantified by examining sample data. But since sample sizes are limited by practical and economical considerations, some uncertainty must remain. This practical limit gives rise to *statistical uncertainty*. Data are collected for the purpose of building a stochastic model of the physical variability. This involves firstly the selection of an appropriate probability distribution type, and then determination of numerical values for its parameters. Probability distributions have usually between one and four parameters which must be estimated on the basis of limited data information with varying data quality.

The parameters of probability distributions are often themselves stochastic quantities and the uncertainty about them will depend on the expert knowledge and the amount of sample data. This uncertainty, unlike physical uncertainty, arises solely as a result of lack of information, and is usually modeled in the Bayesian framework by probability distributions for the parameters. These distributions depend on the available information and therefore it contains also statistical uncertainty.

Structural analysis of systems is based on mathematical models relating input and output quantities. These models are often deterministic, e.g. linear systems. Such models, by their simplifications, they will therefore add to the uncertainty. This source of uncertainty is termed *model uncertainty* and occurs also by unknown boundary conditions and as a result of the unknown effect of other variables and their interactions which are not included in the model. Especially in large systems like life support systems, model uncertainty has a large effect on structural implications and should not be neglected.

Another important kind of uncertainty is the uncertainty associated with data. Real data can have errors or contain outliers, and often important data are missing. But they are often also *non-precise*. This kind of uncertainty is called *data uncertainty*. Concerning non-precise data this uncertainty is called *imprecision*. It is important to note that this uncertainty is not stochastic in nature but implied by the imprecision of measurements,

which can be modeled by the so-called *non-precise numbers*.

The dominating model for uncertainty in statistics is using probability to describe uncertainty. But it should be noted that probability is not sufficient to model all kinds of uncertainty which appear in real world situations.

4. Probability and Philosophical Foundations

In order to describe random variation, for example the life times of biological units, the most up to date model for that are probability distributions and the basic concepts are *probabilities*.

What are probabilities? There are different opinions about that in the scientific community.

4.1. Classical Probabilities

Historically the first kind of probability was from gambling when only a finite number *m* of possible outcomes is possible. By symmetry argument these outcomes are assumed to be equally probable. Based on the properties of relative frequencies the probability of one of all possible outcomes is defined to be 1. Then the probability of a single outcome is $1/m$. By the additivity property of relative frequencies and their idealizations (probabilities) the probability of a subset *A* of the set of all possible outcomes is given by

$$P(A) = \frac{\text{number of elements in } A}{m}. \qquad (1)$$

This kind of probability is called *classical probability*.

4.2. Geometric Probabilities

If the set of possible outcomes is infinite the classical probability definition does not work. If the set *M* of possible outcomes is a subset of a *k*-dimensional Euclidian space and *M* has finite content $I(M)$, then the probability of a subset *A* of *M* is defined by the ratio of the content of *A* to the content of *M*, i.e.

$$P(A) = \frac{I(A)}{I(M)}. \qquad (2)$$

Remark: For $k = 1$ the content corresponds to length, for $k = 2$ to area, and for $k = 3$ to volume.

Geometric probabilities were developed in the 18th century to model statistical experiments with outcomes corresponding to geometrically measurable quantities. Applications of geometric probabilities are in rendezvous problems.

4.3. Probabilities as Idealized Relative Frequencies

Conducting an arbitrary statistical experiment many times and forming the sequence of relative frequencies $f_n(A)$ for a fixed event *A* of outcomes, this sequence $f_n(A), n \in IN$ converges in probability to a fixed number *p*. This number is called *probability of A*. Compare the law of large numbers in the contribution *Probability and Statistics* and the article *Limit Theorems of Probability Theory*. A mathematically sound theory of this frequentist probability concept is complicated. For that R. v. Mises developed his theory of *collectives*. But this theory was not widely accepted for applications.

4.4. Probability Spaces

In 1933 A. N. Kolmogorov defined a mathematical structure to describe probability distributions for statistical experiments with arbitrary outcome spaces. Let *M* denote the set of possible outcomes from an experiment, then the system $\mathcal{A}$ of subsets of *M* is formed, for which probabilities are of importance. The elements of $\mathcal{A}$ are subsets of *M*, called *events*. Therefore $\mathcal{A}$ is a system of events and in order to define the *probability distribution*, the system $\mathcal{A}$ has to fulfill the following conditions (a) to (c):

(a) M belongs to $\mathcal{A}$

(b) For every event $A \in \mathcal{A}$ also its complement A^c belongs to $\mathcal{A}$

(c) For every countable sequence $A_1, A_2, \cdots$ of events from $\mathcal{A}$ also the union $\bigcup_{n=1}^{\infty} A_n$ belongs to $\mathcal{A}$.

Remark: From the above conditions (a) to (c) by the de Morgan's laws it follows that the countable intersection of events is also an event. Moreover finite unions and intersections of events are also events, i.e. they belong to $\mathcal{A}$. This follows from

$$\bigcup_{i=1}^{n} A_i = \bigcup_{i=1}^{\infty} A_i \quad \text{with} \quad A_{n+k} = \varnothing \quad \text{for all } k \geq 1.$$

The following equations for set theoretic operations for subsets of a fixed set are called de Morgan's laws:

$$\left(\bigcup_{i \in I} A_i \right)^c = \bigcap_{i=I} A_i^c \tag{3}$$

And

$$\left(\bigcap_{i=I} A_i\right)^c = \bigcup_{i=I} A_i^c \tag{4}$$

Kolmogorov defined a probability distribution P on $\mathcal{A}$ in the following way:

(P1) $P : \mathcal{A} \to [0,1]$

(P2) $P(M) = 1$

(P3) For every countable sequence A_n, $n \in \mathbb{N}$ of pairwise disjoint events A_n the following must hold: $P\left(\bigcup_{i=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n)$

Remark: Condition (P3) is called *countable additivity*. From this the finite additivity follows immediately. This means for every finite family $A_1, \cdots, A_n$ of pairwise disjoint events the following holds:

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} (A_i) \tag{5}$$

A triplet $(M, \mathcal{A}, P)$ with the foregoing properties is called *probability space*.
The concept probability space is basic for most of contemporary stochastic modeling.

-
-
-

Bibliography

W. H. Beyer (1988): *CRC Handbook of Tables for Probability and Statistics*, 2nd Edition, CRC Press, Boca Raton, Florida. [Comprehensive collection of tables for statistics and probability]

A. Birnbaum (1992): On the Foundations of Statistical Inference. In: S. Kotz, N. L. Johnson (eds.): *Breakthroughs in Statistics, Vol, I, Foundations and Basic Theory*, Springer Verlag, New York. [Reprint of a paper presented at a special discussion meeting of the American Statistical Association in 1961]

B. de Finetti (1974): *Theory of Probability, Volumes 1 and 2*, Wiley, London. [This is a classical and important book on the foundations of probability modeling and statistics]

R. A. Fisher (1922): On the Mathematical Foundations of Theoretical Statistics, reprinted in S. Kotz, N. L. Johnson (Eds.): Breakthroughs in Statistics, Vol. I, Foundations and Basic Theory, Springer-Verlag, New York. [One of the most influential papers on the foundations of statistics in the 20th century]

P. J. Huber (1980): *Robust Statistics*, Wiley, New York. [Important monograph for robustness studies concerned with influence functions in statistical inference]

R. Launer and G. Wilkinson (1979): *Robustness in Statistics*, Academic Press, New York. [This contains different aspects of robustness in statistical work.]

A. M. Mood, F. A. Graybill and D. C. Boes (1974): *Introduction to the Theory of Statistics*, Third Edition, McGraw-Hill, New York. [Well written text on mathematical aspects basic for statistics, including the necessary background in probability]

K. R. Parthasarathy (1980): *Introduction to Probability and Measure*. London: Macmillan Press. [This is an excellent book on probability foundations of statistics]

L. J. Savage (1972): *The Foundations of Statistics*, Dover Publ., New York. [Early monograph on a decision oriented approach to statistical foundations]

G. Tintner (1949): Foundations of probability and statistical inference, *Journal of the Royal Statistical Society*, Ser. A, 62. [Discussion on foundations of statistics from different viewpoints]

R. Viertl (1987): Is it necessary to develop a Fuzzy Bayesian inference. In: *Probability and Bayesian Statistics*, Plenum Press, New York. [In this paper fundamental problems of Bayesian inference based on fuzzy data are stated]

R. Viertl (1999): Non-Precise Data. In: *Encyclopedia of Statistical Sciences, Update Volume 3*, Wiley, New York. [Foundations for the formal and quantitative description of non-precise data and their statistical analysis]

Biographical Sketch

Reinhard Viertl born March 25, 1946, at Hall in Tyrol, Austria. Studies in civil engineering and engineering mathematics at the Technische Hochschule Wien. Receiving a Dipl.-Ing. degree in engineering mathematics in 1972. Dissertation in mathematics and Doctor of engineering science degree in 1974. Appointed assistant at the Technische Hochschule Wien and promotion to University Docent in 1979. 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 1981 - 1982. Since 1982 full professor of applied statistics at the Department of Statistics, Vienna University of Technology. Visiting professor at the Department of Statistics, University of Innsbruck, Austria from 1991 to 1993. 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, president of the Austrian Statistical Society from 1987 to 1995. Invitation to membership in the New York Academy of Sciences in 1998. Author of the books *Statistical Methods in Accelerated Life Testing* (1988), *Introduction to Stochastics* in German language (1990), *Statistical Methods for Non-Precise Data* (1996). Editor of the books *Probability and Bayesian Statistics* (1987), *Contributions to Environmental Statistics* in German language (1992). Co-editor of a book titled *Mathematical and Statistical Methods in Artificial Intelligence* (1995), and co-editor of two special volumes of journals. Author of over 70 scientific papers in algebra, probability theory, accelerated life testing, regional statistics, and statistics with non-precise data.

Editor of the publication series of the Vienna University of Technology, member of the editorial board of scientific journals, organiser of different scientific conferences.