J. Ind. Soc. Agril. Statist.
61(2), 2007 : 95-108

Has Statistics a Future? If So in What Form?[1]

C. Radhakrishna Rao
Emeritus Eberly Professor of Statistics, Pennsylvania State University, University Park, PA 16802, USA

SUMMARY

The mathematical foundations of statistics as a separate discipline were laid by Fisher, Neyman and Wald during the second quarter of the last century. Subsequent research in statistics and the courses taught in the universities are mostly based on the guidelines set by these pioneers. Statistics is used in some form or other in all areas of human endeavor from scientific research to optimum use of resources for social welfare, prediction and decision-making. However, there are controversies in statistics, especially in the choice of a model for data, use of prior probabilities and subject-matter judgments by experts. The same data analyzed by different consulting statisticians may lead to different conclusions.

What is the future of statistics in the present millennium dominated by information technology encompassing the whole of communications, interaction with intelligent systems, massive data bases, and complex information processing networks? The current statistical methodology based on simple probabilistic models developed for the analysis of small data sets appears to be inadequate to meet the needs of customers for quick on line processing of data and making the information available for practical use. Some methods are being put forward in the name of data mining for such purposes. A broad review of the current state of the art in statistics, its merits and demerits, and possible future developments will be presented.

Key Words : Bayesian analysis, Cross validation, Data mining, Decision theory, Estimation, Hypothesis testing, Large data sets, Machine learning.

1. STATISTICS AS A SEPARATE DISCIPLINE

1.1 A Paradigm for Statistical Theory and Methods

The word statistics was coined by the German Scholar Gottfried Achenwall about the middle of the eighteenth century in the context of *collection, processing and use of data by government.*

During the nineteenth century, statistics acquired a new meaning as extraction of information from data for decision making. The need arose especially in testing hypotheses or making predictions or forecasts based on information in the observations made on natural phenomena or generated through well designed experiments. It was realized that the information contained in particular data, however well they are ascertained, is subject to some uncertainty and consequently our conclusions based on observed data could be wrong. How then can we acquire new knowledge? We have to evolve a new methodology of data analysis with a view to estimate the amount of uncertainty in extracted information and to formulate rules for making decisions with minimal risk.

The equation

Uncertain knowledge	+	Knowledge of the extent of uncertainty in it	=	Useable knowledge

is used as a *new paradigm* for statistical theory and methods. Thus, statistics acquired the status of a new discipline of study for

[1] *Keynote Address for the International Conference on Statistics and Informatics in Agricultural Research to mark the Diamond Jubilee Celebration of the foundation of Indian Society of Agricultural Statistics held at NASC Complex, New Delhi on 27 December, 2006.*

- acquiring data with maximum possible information for given cost
- processing data to quantify the amount of uncertainty in answering particular questions, and
- making optimal decisions (subject to minimal risk) under uncertainty.

The first systematic efforts for the development of statistical methodology began only in the beginning of the 20-th century, and it is only in the first half of the century the basic concepts of statistical inference were introduced, (Exhibit 3), which enabled rapid developments to take place for possible applications in all areas of human endeavor ranging from natural and social sciences, engineering and technology, management and economic affairs, arts, literature medicine and legal problems. Knowledge of statistics was considered to be essential in all fields of inquiry. Courses in statistics were introduced in the curriculum of social sciences. Specialized books dealing with the applications of statistics in particular areas were written as guidance to research workers. Referring to the ubiquity of statistics, Sir Ronald Fisher (1953), in his Presidential Address to the Royal Statistical Society in 1952, made the optimistic statement

> *I venture to suggest that statistical science is the peculiar aspect of human progress which gives to the twentieth century its special character; and indeed members of my present audience will know from their own personal and professional experience that it is to the statistician that the present age turns for what is most essential in all its more important activities.*

The scope of statistics as it is understood, studied and practiced today extends to all areas of human activity as shown in Exhibit 1.

The *layman* uses statistics (information obtained through data of various kinds and their analyses published in newspapers and consumer reports) for taking decisions in daily life, or making future plans, deciding on wise investments in buying stocks and shares, etc. Some amount of statistical knowledge may be necessary for a proper understanding and utilization of all the available information and to guard oneself against misleading advertisements. The need for

Exhibit 1. Use of statistics in diverse areas of human endeavour

GOVERNMENT		
FRAUD DETECTION Faking currency Notes Stamps Spam blocking	**ADMINISTRATIVE** Policy decisions Economic planning Services (weather etc.) Data compilation and Dissemination Opinion surveys	**NATIONAL SECURITY** Individual identification Bio-terrorism Cyber security Syndromic surveillance
INDUSTRY BUSINESS Efficient management Demand forecasting Quality control Insurance Consumer surveys	**UNIQUITY OF STATISTICS**	**RESEARCH** Hard sciences Soft sciences Arts Literature Archeology Economic history
MEDICINE Diagnosis Prognosis Clinical trials Medical errors	**LAYMAN** Lifetime decisions Investments Daily chores Participation in politics	**LAW** Statistical evidence DNA testing Analysis of crime data Combing evidence

statistical literacy in our modern age dominated by science and technology was foreseen by H.G. Wells

> *Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write.*

For the *government* of a country, statistics is the means by which it can make short and long range plans to achieve specified economic and social goals. Sophisticated statistical techniques are applied to make forecasts of population and the demand for consumer goods and services and to formulate economic plans using appropriate models to achieve a desired rate of progress in social welfare.

In *scientific research*, statistics plays an important role in the collection of data through efficiently designed experiments, in testing hypotheses and estimation of unknown parameters, and in interpretation of results.

In *industry*, extremely simple statistical techniques are used to improve and maintain the quality of

manufactured goods at desired levels. Experiments are conducted in R & D Departments to determine the optimum mix (combination) of factors to increase the yield or give the best possible performance. It is a common experience all over the world that in plants where statistical methods are exploited, production has increased by 10% to 100% without further investment or expansion of plant. In this sense statistical knowledge is considered as a *national resource*. It is not surprising that a recent book on modern inventions lists statistical quality control as one of the great technological inventions of the last century.

In *business*, statistical methods are employed to forecast future demand for goods, to plan for production, and to evolve efficient management techniques to maximize profit.

In *medicine*, principles of design of experiments are used in screening of drugs and in clinical trials. The information supplied by a large number of biochemical and other tests is statistically assessed for diagnosis and prognosis of disease. The application of statistical techniques has made medical diagnosis more objective by combining the collective wisdom of the best possible experts with the knowledge on distinctions between diseases indicated by tests.

In *literature*, statistical methods are used in quantifying an author's style, which is useful in setling cases of disputed authorship.

In *archaeology*, quantitative assessment of similarity between objects has provided a method of placing ancient artifacts in a chronological order.

In *courts of law*, statistical evidence in the form of probability of occurrence of certain events, such as similarity of DNA is used to supplement the traditional oral and circumstantial evidences in judging cases.

There seems to be no human activity whose value cannot be enhanced by injecting statistical ideas in planning and by using results for feedback and control. It is apodictic to claim: *If there is problem to be solved, seek for statistical advise instead of appointing a committee of experts. Statistics and statistical analysis can throw more light than the collective wisdom of the articulate few.*

In the book on *Statistics and Truth* by Rao (1997b) numerous examples are given in Chapters 5 and 6 of applications of statistical techniques to a variety of problems ranging from disputed authorship, disputed paternity, seriation of Plato's works, foliation of manuscripts, dating of publications and construction of language trees, to weather forecasting, public opinion polls and extra sensory perception.

1.2 What is Statistics?

Is statistics a science, a technology, or an art? Statistics is not a subject like the basic disciplines of mathematics, physics, chemistry or biology. Each of these disciplines has a subject matter of its own and problems of its own which are solved by using the knowledge of the subject. There is nothing like a statistical problem which statistics purports to solve. Statistics is used to solve problems in other disciplines and appropriate methodology is developed for any given situation. The following Exhibit 2 from a paper by Box (1980) shows how most of the important concepts in statistics were motivated by practical problems. In course of time, the subject matter of statistics grew from isolate methods applied to particular problems to the consolidation of different methods under a unified theory based on the concepts of probability. The basic problem of statistics is viewed as quantification of uncertainty, which may be considered as the subject matter of statistics for study and research. As it is practiced today, statistics appears to be a combination of science, technology and art.

It is *science* in the sense that it has an identity of its own with a large repertoire of techniques derived from some basic principles. These techniques cannot be used in a routine way; the user must acquire the necessary expertise to choose the right technique in a given situation and make modifications, if necessary. Further, there are philosophical issues connected with the foundations of statistics - the way uncertainty can be quantified and used - which can be discussed independently of any subject matter. Thus, in a broader sense statistics is a separate discipline.

It is *technology* in the sense that statistical methodology can be built to any operating system to maintain a desired level and stability of performance, as in quality control programs in industrial production. Statistical quality control is described as one of the great technological inventions of the 20th century. Statistical methods can also be used to control, reduce and make

Exhibit 2. Practical problems motivating general statistical concepts (George Box (1980))

Practical problem	Investigator	Derived general concept
Analysis of Asteroid Data. How far is it from Berlin to Potsdam?	Gauss	Least squares
Are planetary orbits randomly distributed?	Daniel Bernoulli	Hypothesis testing
What is the population of France?	Laplace	Radio estimators
How to handle small samples of brewery data?	Gosset	t-test
Improving agricultural practice by using field trials.	Fisher	Design of experiments
Do potato varieties and fertilizers interact?	Fisher	Analysis of variance
Accounting for strange cycles in U.K. wheat prices.	Yule	Parametric time series models
Economic inspection (of ammunition).	Wald Barnard	Sequential tests
Need to perform large numbers of statistical tests in pharmaceutical industry before computers were available.	Wilcoxon	Nonparametric tests
Advanced estimates of agricultural production.	Mahalanobis	Sample surveys*

* not quoted by Box

allowance for uncertainty and thereby maximize the efficiency of individual and institutional efforts.

Statistics is also *art*, because its methodology which depends on inductive reasoning is not fully codified or free from controversies. Different statisticians may arrive at different conclusions working with the same data set. There are frequentists, Bayesians, neo-Bayesians and empirical Bayesians among statisticians each one advocating a different approach to data analysis. (A familiar quote on statistics: *If there are 3 statisticians on a committee, there will be 4 minority reports.* See also Van den Berg (1992) who conducted a survey and found that statisticians with different backgrounds used different methods for the analysis of the same data.) There is usually more information in given data than what can be extracted by available statistical tools. Making figures tell their own story depends on the skill and experience of a statistician, which makes statistics an art. Perhaps, *statistics is more a way of thinking or reasoning than a bunch of prescriptions for beating data to elicit answers.*

While mathematics is the logic of deducing consequences from given premises, statistics may be regarded as a rational approach to learning from experience and the logic of identifying the premises given the consequences, or inductive reasoning as it is called. Both mathematics and statistics are important in all human endeavors whether it is in the advancement of natural knowledge or in the efficient management of our daily chores.

1.3 Fisherian Framework

In his fundamental paper on mathematical foundations of theoretical statistics, Fisher (1922) stated three methodological aspects of statistics:

- *Specification* (choice of a stochastic model for data)
- *Estimation* (of unknown parameters in the chosen model)
- *Testing of hypotheses* (seeking evidence from data for possible rejection of a specified hypothesis or theory)

Fisher's framework has been and still is the basis for the development of statistical methods. However, there are difficulties in using these concepts and methods based on them in statistical analysis of real data.

First is the specification, or the choice of a stochastic model for data: Fisher did not specify any statistical method for model selection. He acknowledged the usefulness of Karl Pearson's chi-square test for specification, only as a method for possible rejection of a given model and not for its acceptance. Reference may be made to Inman (1994) for a review of the controversy between Pearson and Fisher on the role of the chi-square criterion in accepting or rejecting a specified model for data.

In recent years, several model selection criteria have been suggested such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), and GIC (General Information Criterion), an extensive review of which can be found in a paper by Rao and Wu (2001). These methods are not directly related to the performance of the estimated models for it is known that

different models may have to be used in the analysis of the same data for different purposes as shown by Rao (1977, 1987). Further, model selection by using AIC, BIC and GIC depends on the sample size; a larger sample size may choose a more complex model.

Recent studies in Chaos theory show that there are difficulties in distinguishing between sequences of observations produced by deterministic and random mechanisms. Attempts are beginning to be made for modeling a sequence of variables such as time series as a combination of deterministic and random components. (See Cox (1990), Lehmann (1990) and Rao (1997b, pp. 26-28).

In real situations scientists are looking for what may be called working hypotheses (which may not be strictly true) which enable prediction of events with reasonable accuracy. So the main question should be to ask how *good* a proposed model or theory is in explaining the observed data and in predicting future events, and not whether the proposed model is *true or false. A working hypothesis is rejected if a better working hypothesis is found. This is how science progresses creating useful knowledge from time to time.*

Fisherian framework provided the basis for the development of theoretical statistics during the first half of the 20th century as shown in Exhibit 3. Neyman and Pearson (1933) developed a theory of testing of hypotheses using the concept of the power function of a test (with respect to possible alternatives to the given hypothesis) for comparing different test criteria and choosing the one with some optimum properties for the power function. The theory provided a justification of some of the test criteria introduced by Fisher on an intuitive basis. Pitman (1937) developed nonparametric tests which do not depend on any stochastic model for data. Wald (1950) formulated estimation and testing of hypothesis in a decision theoretic set up considering a loss function as an input into the problem. Wald (1947) also developed the theory of sequential testing. General asymptotic test criteria, called the *Holy Trinity*, were introduced by Neyman and Pearson (1928), Wald (1941) and Rao (1949).

It is relevant to mention that Fisher introduced the basic concepts of statistical inference and developed the related statistical methodology when computers capable of performing complex computations were not available and there were serious limitations on acquisition of data. Under these restrictions, the statistical methodology developed was *mostly model oriented*, i.e., under the assumption that the observed data is a random sample from a population belonging to a specified family of distributions functions. Often a simple stochastic model was chosen, like the normal distribution, to provide exact

Exhibit 3. Mathematical foundation of theoretical statistics (1900-1950)

T. Bayes	1764	Bayes theorem
K. Pearson	1900	Chi-square test
W.S. Gosset	1908	Students' t-test
R.A. Fisher	1915	Exact distributions of statistics
	1922	Estimation (maximum likelihood)
	1923	Analysis of variance
	1926	Design of experiments
J. Neyman		
E.S. Pearson	1928	*Likelihood ratio test
	1933	Testing of hypotheses
E.J.G. Pitman	1937	Nonparametric tests
H. Jeffreys	1939	Bayesian statistics
A. Wald	1943	*Asymptotic test
P.C. Mahalanobis		
M. Hansen	1944	Sample surveys
A. Wald	1947	Sequential tests
C.R. Rao	1948	*Score test
A. Wald	1950	Decision theory

* The three general asymptotic tests are referred to as the *Holy Trinity*.

results (closed form solutions to problems) involving minimum computations.

Tables of limited percentage points of test statistics were constructed (using desk computers) choosing the normal as the underlying distribution and rules were laid down for rejection of hypotheses at the tabulated levels of significance, usually 5% and 1%. Limitation on the sample size made it difficult to verify model assumptions. [Commenting on the mistrust of British statistical methods by continental statisticians, Buchanan-Wollaston (1935) says, *"The fact that British methods "work" is due to prevalence in Nature of distributions similar to Gaussian rather than to any peculiar value in the methods themselves"].*

Second is the method of estimation: The method of maximum likelihood introduced by Fisher is valuable in the estimation of parameters when the model for data is known and the sample size is not small. However, it is

not robust for slight departures from the specified model and for outliers in data. Robust methods of estimation known as M-estimation and associated tests of hypotheses have received much attention. A variety of procedures have been introduced (without any guidance on what to choose) to eliminate or minimize the influence of outliers or contamination in data. The theory is mostly asymptotic and the performance of M-estimates in small samples has not been adequately examined. The character of research in this area is described by Tukey (1993) as *asymptotitise*.

There are also controversies in expressing the precision of an estimator. Fisher suggested the conditional variance of an estimator given an ancillary statistic as a measure of precision. But, as pointed out by Basu (see Ghosh (1988), pp.3-19), there are difficulties in such a procedure. First there is, in general, no maximal ancillary statistic and different choices of ancillary statistics lead to different expressions for precision, and there is no way of choosing one in preference to the other. Basu also gives examples where the conditional distribution given an ancillary statistic becomes degenerate (Ghosh (1988), pp.161-167), and uninformative about the parameter.

Third is testing of hypotheses: Fisher considered a test of significance using an appropriate criterion as a method which can lead to possible rejection of a given hypothesis, but not for establishing a given hypothesis as certainly true. He used tests of significance in an ingenious way, an example of which is the discovery of the Rhesus factor described in Fisher (1948). It is a brilliant example of how hypothesis testing can be of help "in fitting one scrupulously ascertained fact into another, in building a coherent structure for knowledge and seeing how each gain can be used as a means for further research. Fisher also used tests of significance to detect irregularities in data such as lack of randomness, recording errors or bias in sampling. In this connection, reader is referred to Fisher (1936), where he showed that Mendel's data on his genetic studies are probably faked and to Fisher (1934) where he studied the effects of different methods of ascertaining data in genetic studies of inheritance of diseases.

There is, however, some debate among statisticians on the usefulness of tests of significance. The null hypothesis H_0 as formulated in many problems is known to be wrong and no test of significance is needed. What is of interest is to determine to what extent the true hypothesis can differ from H_0, which is a problem of estimation rather than of testing a hypothesis. Frank Yates, a long time associate of R.A. Fisher, mentioned in the obituary published in *Biographical Memoires of the Royal Society* that Fisher laid too much stress on hypothesis testing. He said that if we are comparing the yields of two varieties of corn, it is useful to ask what the *difference in yields is* rather than whether they *have the same yield*, which is seldom true.

In Fisher-Neyman framework of testing a null hypothesis, there is some controversy about the level of significance of a test. From the early use of tests of significance by Fisher and the axiomate set up by Neyman, by level of significance is meant "the frequency with which the hypothesis is rejected in repeated sampling of any fixed population allowed by the hypothesis". In his last book, Fisher (1956, p. 91) disassociated himself from such a view by saying: "This intrusive axiom, which is foreign to the reasoning on which tests of significance were in fact based seems to be a real bar to progress". On p. 77, Fisher says: "the population in question is hypothetical, that it could be defined in many ways..., or, that an understanding, of what the information is which the test is to supply, is needed before an appropriate population, if indeed we must express ourselves in this way, can be specified". Fisher has not made explicit how the level of significance can be ascertained given the data.

The well known mathematical statistician, Wolfowitz (1967) reviewing a popular book on testing of hypotheses made the following critical comment.

> *... the history of testing of hypothesis is an example of collaboration between theoreticians and practical statisticians which has resulted in greater obfuscation of important statistical problems and side tracking of much statistical effort.*

Wolfowitz believed that a useful approach to statistical analysis of live data is *Decision Theory* as developed by Wald (1950), which needs inputs such as the class of alternative hypotheses, prior probabilities and losses associated with different possible decisions. Such a procedure of choosing a hypothesis to minimize the expected loss can be implemented in certain situations like acceptance sampling (such as accepting or rejecting batches of goods produced in a factory), but does not seem to be applicable in scientific research.

When a test is applied to test a hypothesis, there are two possible scenarios:

1. The hypothesis is rejected as not being true.
2. The hypothesis is not rejected, but this does not mean that it is accepted as true.

In either case, the scientist has to continue his search for an alternative hypothesis. Does statistics help in the search for an alternative hypothesis? There is no codified statistical methodology for this purpose. Text books on statistics do not discuss either in general terms or through examples *how to elicit clues from data to formulate an alternative hypothesis or theory when a given hypothesis is rejected.*

Fisher (1935) also introduced nonparametric tests based on permutation distributions using the randomization principle, which was hailed as an important contribution by Neyman and others. Unfortunately, the randomization principle is not without logical difficulties (see the paper by Basu on pp. 290-312 in Ghosh (1988)).

A notable contribution to nonparametric testing is Efron's (1979) bootstrap, which has become popular with the enormous computing power we now have. However, its theoretical justification is again based on asymptotics and the consequences of bootstrapping in small samples have not been fully examined. A related method is Jackknife which Tukey (1977) recommended as an alternative. Efron (1979) gave a comparative study of Bootstrap and Jackknife techniques.

1.4 Bayesian Analysis

Some of the inconsistencies in the classical methods described in previous sections of this paper, which depend on properties based on repeated sampling from a population, led several statisticians to use Bayesian methods in data analysis. References to papers emphasizing the need for Bayesian analysis are Berger (2002) and Basu's contributions reproduced in Ghosh (1988). It is argued that in classical statistics of Fisher, Neyman and Wald, statistical methods for drawing inferences on unknown parameters are judged by their (average) performance in repeated sampling from a population. Such a procedure ignores the fact that *all samples are not equally informative* and inference on unknown parameters should be made conditional on the observed sample, which makes the use of Bayesian analysis inevitable.

Other arguments advanced by Bayesians against classical testing procedure refer to the interpretation of p-values and paradoxes associated with it. Lindley (1957) gave an example in which the p-value is fixed at .05, but as the sample size increases, Bayesian posterior probability that the null hypothesis is true approaches unity. It is also shown that for a large class of priors

$$p = p(T > T_{obs} | H_0) \leq p(H_0 | \text{Data})$$

where H_0 is a null hypothesis and T is a test statistic. The above inequality shows that the use of p-values *exaggerates* significance.

Bayesian analysis depends on Bayes theorem

$$p(s|x) = \frac{p(s)p(x|s)}{p(x)}$$

where p(s) is the *prior probability* distribution on the space S of specified family of models, p(x|s) is the probability density of the observation x for specified $s \in S$, p(x) is the overall probability density of x, and p(s|x) is the *posterior probability* of s given x.

An attractive feature of Bayes theorem is that we can make probability statements about the probability models in the light of observed data x using the posterior density p(s|x). But the question is: *Where does p(s), the prior probability come from?* Berger (2002) listed five different approaches to the problem "each of which can be of great value in certain situations and for certain users":

- Objective (non-informative or default priors, maximum entropy and reference priors)
- Subjective and partly subjective and partly objective
- Robust priors
- Frequentist-Bayes
- Quasi-Bayes
- Empirical Bayes (strictly not a Bayesian analysis)

However, it appears that there is no unified approach to Bayesian analysis. How can Bayesian analysis be implemented in the context of a customer, who generates data to throw light on a particular problem, approaches a statistical consultant for analysis of data. Who supplies the input on prior? Surely, not the consultant, but should he (or she) accept the customer's prior? Reference may be made to Cox (2000) for further comments on Bayesian analysis.

1.5 Likelihood Principle

It is generally agreed by statisticians belonging to different schools of thought that the likelihood function introduced by Fisher (1922) plays a pivotal role in statistical inference. The ideal situation is the combined use of the prior distribution and likelihood function to derive the posterior distribution. Attempts have been made in the case of a single parameter θ, to suggest plausible ranges of the parameter, without using priors, using only the ratio $r = L(\hat{\theta}|x) / L(\theta | x)$, where $L(\theta | x)$ is the likelihood function of θ given the observation x and $\hat{\theta}$ is the maximum likelihood estimate of θ. The ranges of θ, classified as very *plausible, somewhat implausible* and *highly implausible*, suggested by Fisher, Jeffreys and Royall are as follows:

	Fisher	Jeffreys	Royall
Very plausible	$r \in (1,2)$	$r \in (1,3)$	$r \in (1,4)$
Somewhat implausible	$r \in (2,5)$	$r \in (3,10)$	$r \in (4,8)$
Highly implausible	$r \in (5,15)$	$r \in (10,00)$	$r \in (8,32)$

It is not clear how these ranges are obtained and how they could be used in practice. When the full likelihood cannot be used due to nuisance parameters, several modified versions have been suggested such as, *partial likelihood, pseudo likelihood, quasi-likelihood, empirical likelihood and a predictive likelihood.* For further details on the likelihood principle, reference may be made to Ghosh (1988, pp. 313-320) and Reid (2002).

It is surprising that Fisher, who in his early research work emphasized tests of significance based on a test statistic and the p-value in the tails of its distribution, recommended in his last book (Fisher (1956)) the use of likelihood ratio without any reference to its probability distribution.

There is another problem about the use of likelihood which is often referred to as Rao's paradox in sample surveys (see Rao (1971), Cox (1997) and Smith (1997)). Consider a finite population defined by the set $\{Y_i, X_i\}$, $i = 1,..., N\}$, where Y_i is a label identifying the i-th member and X_i is, in a general set up, a random variable with probability density $f_i(X, \theta_i)$ depending on an unknown parameter θ_i. A sample is a selection of n pairs $(y_1, x_1), ..., (y_n, x_n)$ drawn from $\{Y_i, X_i\}$, where y_i takes one of the values $(Y_1, ... Y_n)$ and x_i is an observation on X_i. The problem is that of estimating a function of $\theta_1,..., \theta_N$. The probability density at the observed values $\{(y_i, x_i)\}$, i.e., the likelihood of the parameters based on the sample is

$$\prod_{i=1}^{n} \frac{1}{N} f_{r_i}(x_i, \theta_{r_i})$$

where the i-th pair in the sample corresponds to the unit labeled r_i with parameter θ_i. The above likelihood function contains only n of the parameters $(\theta_1,..., \theta_N)$ and has thus no information on the rest of the N – n parameters. In such a case the likelihood approach based on the entire sample is not applicable.

However, if we disregard the labels and retain only $(x_1,...,x_n)$, then the likelihood based on $(x_1,...,x_n)$ under random sampling is

$$\prod_{i=1}^{n} \frac{1}{N} \sum_{r=1}^{n} f_r(x_i, \theta_r)$$

which contains all the unknown parameters. If one wants to use the likelihood principle, *it may be necessary to throw away part of the data*! This raised new problems on the choice of statistics (functions of the sample) for setting up the likelihood function without loss of information.

2. STATISTICS IN THE INFORMATION AGE

2.1 Limitations of the Current Statistical Methods

Mostly model oriented. Where does the stochastic model for data come from? This question has often be debated. From whoever it may come, either the customer who may have some knowledge of the data and the mechanism generating the data, or the statistical consultant from his previous experience of similar data, there is no reasonable statistical procedure for validating it for use on current data. The reader is referred to a recent paper by Breiman (2001) and the discussion by Cox, Efron, Parzen and others. The author mentions two cultures, data modeling (practiced by 98% of all statisticians) and algorithmic modeling (practiced by 2%) and makes a strong case for model free analysis using techniques such as *neural networks, genetic algorithms, machine learning* (support vector machines of Vipnik).

In the author's opinion, a combination of both cultures will be ideal. If one can succeed in getting a model, at least an approximate one, characterizing the source and the mechanism generating data, it may contribute to expansion of natural knowledge. One way of achieving this is to use a model, wherever it may come from, and validating it by algorithmic modeling. In this connection, it is interesting to note what Fisher (1956,1960, Sec. 21.1 at the end of Chapter 3) said about nonparametric tests which Fisher himself introduced round about 1935:

> *The utility of such nonparametric tests consists in their being able to supply confirmation whenever, rightly or, more often, wrongly it is suspected that the simpler tests have been appreciably injured by departures from normality.*

Lack of firm basis for measurement of uncertainty. There are various methods of expressing uncertainty such as the variance of estimators conditional on a certain configuration of the sample, confidence limits, fiducial limits, plausible limits based on the likelihood function or posterior probability and so on, which are all subject to debate.

Lack of methodology for distinguished between random noise and chaos. The reader is referred to examples given in Rao (1997b, pp. 26-28).

Methodology based on asymptotics. Current contributions to statistical theory are based on asymptotic behavior of estimators and tests (as the sample size tends to infinity) without examination of their usefulness in small samples.

2.2 Limitations of Statisticians

In the early days of the development of statistics as a method of extracting information from data and taking decisions, research in statistics was motivated by practical problems in biological and natural sciences, as indicated in Exhibit 2. Methods developed for use in one area found applications in other areas with minor modifications. Gradually, statistics came to be adopted as an inevitable instrument in all investigations scientific or otherwise as discussed in Section 1 of this paper. Then the need arose for training professionals in statistics to help the government and research organizations in the collection and analysis of data. Statistics was introduced as a compulsory subject in the curriculum of courses in some scientific and technological disciplines.

Gradually, universities started separate departments of statistics where statistical theory and methodology is taught without any serious focus on applications. Venues of interaction between faculty members in statistics and other departments have gradually closed, and the lack of contact with live problems has impeded the expansion of statistics in desired direction or sharpening of the existing tools.

Students graduating in statistics learn statistics as a set of rigid rules without acquiring any knowledge of their application to practical problems. The students are not made aware that statistics is a dynamic and evolving discipline and fertile research in statistics can result only by collaborative work with researchers in other sciences.

Statistics Departments in the universities generally tend to produce statisticians as a separate breed of scientists, which is detrimental to their usefulness as professionals helping research workers in natural and social sciences in data collection and its analysis. They teach statistics as a deductive discipline of deriving consequences from given premises. *The need for examining the premises, which is important for practical applications of results of data analysis, is seldom emphasized.*

It is also surprising that in many universities courses in design of experiments and sample surveys are not given, or listed as optional. Knowledge of these two methodological aspects of data collection is extremely important in all investigations.

Further, students specializing in statistics do not acquire in-depth knowledge of any basic discipline and are, therefore, unable to collaborate with scientists in research work. There has been some thinking on the education and training of statisticians, but no attempts have been made to change the present system (see Kettenring (1995), Parzen (1997), Rao (1997a) and other references in these papers.)

It is also relevant to add here what Fisher (1938) said on who should teach statistics:

> *I want to insist on the important moral that the responsibility for the teaching of statistical methods in our universities must be entrusted, certainly to highly trained mathematicians, but only to such mathematicians as have had sufficiently prolonged experience of practical research and of responsibility for drawing*

conclusions from actual data, upon which practical action is to be taken. Mathematical acuteness is not enough.

This is generally disregarded in the recruitment of faculty members of Statistics Departments at the Universities.

Regarding research work in statistics published in Journals, S.C. Pearce says:

In many fields of statistics numerous techniques have been published with little to guide the practical man as to their spheres of influence.

Current research in statistics should be directed to and made available for immediate use in problems waiting to be solved "rather than getting published in archival Journals", as the editors of the newly started Journal *Biostatistics* put it.

2.3 Needs of Customers

Who needs statistics? The scientists use statistics in a marginal way. Current technology enables scientists to make measurements with a high degree of accuracy and generate large amounts of data under identical conditions. In such a situation, it *may* not be necessary to use sophisticated methods of data analysis. There appears to be no substantial evidence in scientific literature of any major discovery being directly attributed to results or insight provided by statistical analysis. Let us look at the following quotations:

If your experiment needs statistics, you ought to have done a better experiment.

- Lord Rutherford (1871-1931)

A theory can be proved by an experiment but no path leads from experiment to theory.

- A. Einstein (1879-1955)

It is safe to say that no discovery of some importance would have been missed by lack of statistical knowledge.

- F.N. David (1909-1993)

All these statements do not imply that observational data cannot provide clues to scientific discovery. Perhaps, lack of interest in using statistical methods in scientific research may be due to the limited role of hypothesis testing as formulated by statisticians in knowledge discovery. The aim of statistical analysis should be not only to *answer specific questions* but also to *raise new questions* and indicate what further investigations are needed to answer them.

Perhaps, the greatest beneficiaries of statistics are the national governments (responsible for socio-economic development, optimum utilization of national resources, protecting the environment and providing essential public services), industry (in maintaining quality of manufactured goods, increasing productivity) and business (in efficient management and working out optimal strategies). Is the current statistical methodology adequate to meet the demands of customer in these areas?

With computerization of all activities in science, commerce and government, we will have access to unprecedented quantity and variety of data. We also have enormous computing power. These provide us an opportunity to meet the customer's demands for timely and useful information on a wide variety of issues.

There is need to develop new statistical methods for managing large data sets, on-line automatic processing of data (OLAP) to judge the performance of existing practices (working hypotheses), extracting *new* information useful to customers rather than to answer specific questions, decision making and assessing the risks involved and making automatic adjustments for missing or contaminated data. The limitations of the current statistical methods in handling large data sets for extracting useful information have led computer scientists, engineers and operational research workers to suggest what is claimed to be a different approach to data analysis called *Data Mining* much to the surprise of statisticians.

3. DATA MINING

3.1 What is Data Mining

Is Data Mining (DM) a form of statistics or a revolutionary concept? Adriaans and Zantinge (1996, p. 5) describe DM or a more general concept known as KDD (Knowledge Discovery in Databases) as

the non-trivial extraction of implicit, previously unknown and potentially useful knowledge from data.

It is conceived as a multidisciplinary field of research involving machine learning, database technology, statistics, expert systems and visualization.

Some statisticians think that the concepts and methods of DM have their basis in statistics or already subsumed under current statistical methodology. We shall review the current literature on DM, examine to what extent they meet the needs of customers compared to the available statistical methodology, and comment on possible developments in the future.

3.2 Massive Data Sets

What motivated and made DM popular is the availability of large data sets which are automatically generated, stored and easily retrievable for analysis. They are high dimensional in terms of features, cases and classes. The stochastic model for the observations is generally not fully known. There may be some missing values and contaminated data. (See Exhibit 4 for examples of such data sets). Generally, data relating to business transactions, services provided by the government and even scientific programs like the genome mapping and sky surveys in astronomy run into multi-gigabytes.

Conventional statistical methods of testing of hypotheses and building models for prediction may not be suitable. *Every conceivable hypothesis or model in bound to be rejected when a large data set is available*. Even the computation of test statistics and estimates of parameters such as the sample median may pose difficulties. What can we do when large data sets are available?

The characteristics of a large sample are, by asymptotic consistency theorems, close to that of the population on which observations are made. As such, inferences drawn from a sample will have a low degree of uncertainty. Further, the amount of uncertainty itself can be estimated with a high degree of precision by double cross validation (revalidation) as explained in Section 3.4 without any model assumptions, which cannot be achieved with small data sets.

3.3 Data Mining versus Traditional Data base Queries

Using traditional data analytic methods, we can estimate certain parameters of interest and examine the performance of certain decisions (or hypotheses) formulated on the basis of previous studies or some theoretical considerations. Such an analysis is often called on-line analytical processing (OLAP) or providing answers to certain queries.

In DM, through the use of specific algorithms or *search engines* as they are called, attempts are made to discover *previously unknown patterns and trends* of interest in the data and take decisions based on them. We shall examine some of the methods reported in the literature on DM which is described by Wegman (1998) as

> *exploratory data analysis with little or no human intervention using computationally feasible techniques, i.e., the attempt to find interesting structures unknown a priori.*

3.4 Cross Validation and Revalidation

When a large data set, say with S cases, is available, we can divide it into subsets with S_1 and S_2 cases which are also sufficiently large. We can use the subset S_1 to formulate a certain decision rule R based on the discovery of patterns through a search engine. The second set S_2 can be used to evaluate the performance of R through some loss function. In view of the largeness of S_2, we expect to get a precise estimate of the average loss. This procedure known as cross validation is well known in statistical literature, but its application in small samples through methods such as LOO (leave one out) may not be effective.

There are other possibilities when a large sample is available, especially when the search engine suggests several possible rules R_1, R_2,... based on the subset S_1 of cases. We then divide S_2 into two subsets S_{21} and S_{22}, and use cross validation of rules R_1, R_2,... on S_{21} and choose the rule R^* with the minimum loss. Now, we can compute the loss in applying R^* on the second subset S_{22}. We thus have an unbiased estimate of loss in using the rule R^*. This method may be described as revalidation. (See Exhibit 4 where different divisions of the available cases as Train (S_1), Test 1 (S_{21}) and Test 2 (S_{22}) are given in some real large data sets.)

Exhibit 4. The number and type of features, classes and cases used for training, cross validation (Test 1) and revalidation (Test 2) in seven data sets

	Cases			Features		
Dataset	Train	Test 1	Test 2	Num.	Type	Classes
Medical	2079	501	522	33	Num+Binary	2
Telecom	62414	34922	34592	23	Num+Binary	2
Media	7133	3512	3672	87	Num	2
Control	2061	685	685	22	Num	Real
Sales	10779	3591	6156	127	Num+Binary	3
Service	4826	2409	2412	215	Binary	2
Noise	20000	5000	5000	100	Num	2

As new data come in, we have a chance to evaluate the performance of rules in current practice and update if necessary.

3.5 Data Mining Techniques and Algorithms

3.5.1 Visualization

The use of graphs in exploratory data analysis (for understanding the nature of observations and choosing an appropriate model), and in reporting the results of statistical analysis is well known in statistical literature. (See Fisher (1967, Chapter 2), Tukey (1977)). With increase in computing power and possibilities of viewing high dimensional data through parallel coordinates (Wegman (1990), Wegman and Luo (1997), Wilhelm, Symanzik and Wegman (1999)), projections in different directions (Friedman and Tukey (1974), and data reduction by canonical coordinates (Rao (1948a)), principal components (Rao (1964)), correspondence analysis (Benzecri (1992) and Rao (1995)) and multi-dimensional scaling (Kruskal and Wish (1978)), graphical analysis is becoming a valuable tool in discovering patterns in data.

3.5.2 Finding associations

A typical problem is that of finding association between items purchased by customers in a grocery shop (e.g., those who purchase bread also buy butter). In the abstract, the problem may be stated as follows. We have a set of vectors with zeros and ones such as (10010...), where 1 denotes the presence of a specific characteristic (such as purchase of an item) and 0 otherwise. The object is to find whether there is a high percentage of vectors with all 1's in certain positions. A fast algorithm for this purpose was developed by Agrawal, Imielinski and Swami (1993).

3.5.3 Clustering, pattern recognition and decision trees

These methods first introduced in statistical literature and developed by computer scientists and engineers for specific purposes are extensively used in data mining.

3.5.4 Machine learning, neural networks and genetic algorithms

Suppose the problem is that of predicting a target (or class) variable y using a concomitant vector variable x (called features). In statistics, we generally start with a probability model for the variables (x,y) and estimate the conditional distribution of y given x, on the basis of observed samples $(x_1, y_1), \ldots, (x_n, y_n)$. We can then use the conditional distribution of y given x to predict y. In machine learning, we do not explicitly use any probability model. We use an algorithm to find a function $f(\cdot)$ such that

$$\sum_{i=1}^{m} \phi[y_i - f(x_i)]$$

is minimized, where ϕ is a given loss function and $m(< n)$ is the number of samples set apart for learning. This is done by specifying a wide class of functions for f and using a search method like neural networks or genetic algorithms. The efficiency of an estimated function $\hat{f}$ is judged by cross validation, i.e., applying it on the remaining $(n - m)$ samples and computing the average loss

$$(n-m)^{-1} \sum_{m+1}^{n} \phi\left[y_i - \hat{f}(x_i) \right]$$

If the computed loss is large, we alter the class of functions f and search for an optimal solution. The final solution is obtained by a series of iterations.

4. SOME FINAL THOUGHTS

We view this pile of data as an asset to be learned from. The bigger the pile, the better - if you have the tools to analyze it, to synthesize it and make yourself more and more creative.

- Britt Mayo
Director of Information Technology

Statistics is a broad based scientific discipline with theory and methods developed through the calculus of probability for taking optimal decisions under uncertainty. During the last century, research in statistics was directed to the concepts laid down by Fisher, Neyman and Wald. As pointed out in Section 1 of this paper, there are difficulties in formulating the problems to be solved and in applying these concepts to practical problems. [There has been an uncharitable criticism that statisticians are providing exact solutions to the wrong problems, where as in practice, what is needed is an approximate solution to the right problem]. Current statistical methodology has no satisfactory rules governing the choice of the inputs needed such as data modeling, prior probabilities, and expression of uncertainty in decision making. Data mining methods, applied on large datasets, seem to bypass stochastic considerations, and derive decision rules using "machine learning" methods and evaluate their performance through cross validation. The techniques used in data mining problems such as pattern recognition, decision trees, clustering and cross validation have their roots in statistics, but perhaps not actively pursued by statisticians. We may agree with what Weiss and Indurkhya (1998) say:

> *Statistical models are competitive with those developed by computer scientists and may overlap in concept. Still, classical statistics may be saddled with a timidity that is not up to the speed of modern computers.*

In conclusion, I believe DM is a form of much needed statistics neglected by statisticians.

REFERENCES

Adriaans, P. and Zantinge, D. (1996). *Data Mining.* Addison Wesley.

Agrawal, R., Imielinski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. *Proc. Int. Conf.*, ASMSIGMOD, Washington, D.C., 207-216.

Barnard, G.A. (1996). Scientific practice and statistical inference. *Symposium on the Foundation of Statistical Inference, with Special Emphasis on Applications in honor of D.A. Sprott,* 1-9.

Benzecri, J.P. (1992). *Correspondence Analysis Handbook.* Marcel Dekker Inc., New York.

Berger, J.O. (2002). Bayesian analysis: A look at today and thoughts of tomorrow. In : *Statistics in 21st Century* (Editors: Raftery, Tanner and Wells), Chapman and Hall/ CRC, 275-291.

Box, G.E.P. (1980). Comment (on A Report of the ASA Section on Statistical Education Committee on Training of Statisticians for Industry). *Amer. Stat.*, **34**, 65-80.

Breiman, L. (2001). Statistical modeling: Two cultures. *Statist. Sci.*, **16**, 199-231.

Buchanan-Wollaston, H.J. (1935). Statistical Tests. *Nature*, **136**, 182-183.

Cox, D.R. (1990). Role of models in statistical analysis. *Statist. Sci.*, **5**, 169-174.

Cox, D.R. (1997). *The International Statistical Review.* **65**, 261-276.

Cox, D.R. (2000). The five faces of Bayesian statistics. *Cal. Stat. Assoc. Bull.*, **50**, 127-136.

Efron, B. (1979). Bootstrap methods: Another look at jackknife. *Ann. Statist.*, 7, 1-26.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Phil. Trans. Roy. Soc.*, **A222**, 309-368.

Fisher, R.A. (1934). The effect of methods of ascertainment upon the estimation of frequencies. *Ann. Eugen.*, **6**, 13-25.

Fisher, R.A. (1935, 1960, 7th edition). *The Design of Experiment.* Oliver and Boyd, Edinburgh.

Fisher, R.A. (1936). Has Mendel's work been rediscovered? *Ann. Sci.*, **1**, 115-137.

Fisher, R.A. (1948). The Rhesus factor: A study in scientific research. *Amer. Stat.*, **35**, 95-102, 113.

Fisher, R.A. (1953). The expansion of stastistics. *J. Roy. Statist. Soc.*, **A116**, 1-6.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.

Fisher, R.A. (1967). *Statistical Methods for Research Workers.* Hafner Publishing Company.

Friedman, J. and Tukey, J.W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers,* **23**, 881-889.

Ghosh, J. (1988). *Statistical Information and Likelihood: A Collection of Critical Essays by D. Basu.* Lecture Notes in Statistics, **45**, Springer Verlag.

Inman, H.F. (1994). Karl Pearson and R.A. Fisher on statistical tests: A 1935 exchange from Nature. *Amer. Stat.*, **48**, 2-11.

Kettenring, J. (1995). What industry needs? *Amer. Stat.*, **49**, 2-4.

Kruskal, J.B. and Wish, M. (1978). *Multidimensional Scaling.* Sage Publications.

Lehmann, E.L. (1990). Model specification: The view of Fisher and Neyman and later developments. *Statist. Sci.*, **5**, 160-168.

Lindley, D.V. (1957). A statistical paradox. *Biometrika*, **44**, 187-192.

Neyman, J. and Pearson, E.S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika,* **20A**, 175-240 and 263-294.

Neyman, J. and Pearson, E.S. (1933). On the problem of most efficient tests of statistical hypotheses. *Phil. Trans. Roy. Soc.,* **A231**, 289-337.

Parzen, E. (1997). Data mining, statistical methods mining and history of statistics. Tech. Rept. Texas A&M University.

Pitman, E.J.G. (1937). Significance tests which may be applied to samples from any population. *Suppl. J. Roy. Statist. Soc.*, 199-130, 225-232 and *Biometrika,* **29**, 322-335.

Rao, C.R. (1948a). The utilization of multiple measurements in problems of biological classification. *J. Roy. Statist. Soc.,* **B9**, 128-140.

Rao, C.R. (1948b). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proc. Comb. Phil. Soc.*, **44**, 50-57.

Rao, C.R. (1964). The use and interpretation of principal components analysis in applied research. *Sankhya,* **A26**, 329-358.

Rao, C.R. (1971). Some aspects of statistical inference in problems of sampling from finite population. In: *Foundations of Statistical Inference,* pp.177-202, Holt, Rinehart and Winston, Canada.

Rao, C.R. (1977). Prediction of future observations with special reference to linear models. *J. Multivariate Analysis,* **4**, 193-208.

Rao, C.R. (1987). Prediction of future observations in growth curve type models (with discussion). *Statist. Sci.,* **2**, 434-471.

Rao, C.R. (1995). A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Qüestiio*, **9**, 23-63.

Rao, C.R. (1997a). A cross disciplinary approach to teaching of statistics. *Proc 51st session of the Int. Statist. Institute,* Istanbul.

Rao, C.R. (1997b). *Statistics and Truth: Putting Chance to Work* (Second Edition). World Scientific, Singapore.

Rao, C.R. and Wu, Y. (2001). On model selection. IMS Lecture Notes - Monograph Series, **38**, 1-64.

Reid, N. (2002). Likelihood. In: *Statistics in 21st Century* (Eds: Raftery, Tanner and Wells), Chapman and Hall/CRC, ϕ419-431.

Smith, T.M.F. (1997). Social surveys and social science. *Canad. J. Statist.,* **25**, 23-44.

Tukey, J.W. (1993). Major changes for multiple-response (and multiple adjustment) analysis. In *: Multivariate Analysis : Future Directions* (Ed. C.R. Rao), North Holland, 401-422.

Tukey, J.W. (1977). *Exploratory Data Analysis.* Reading, Mass: Addison-Wesley.

Van den Berg, G. (1992). Choosing an analysis method: An empirical study of statistician's ideas in view of the design of computerized support. Ph. D. Thesis, University of Leiden.

Wald, A. (1941). Asymptotically most powerful tests of statistical hypotheses. *Ann. Math. Statist.,* **12**, 1-19.

Wald, A. (1947). *Sequential Analysis.* Wiley, New York.

Wald, A. (1950). *Statistical Decision Functions.* Wiley, New York.

Weiss, S.M. and Indurkhya, N. (1998). *Predictive Data Mining: A Practical Guide.* Morgan Kaufmann Publishers, Inc.

Wegman, E.J. (1990). Hyperdimensional data analysis using parallel coordinates. *J. Amer. Statist. Assoc.*, **68**, 664-675.

Wegman, E.J. (1998). Visions: The evolution of statistics. Keynote talk at the conference, "New Techniques and Technologies is Statistics" Sorrento, Italy.

Wegman, E.J. and Leo, Q. (1997). High dimensional clustering using parallel coordinates and grand tour. *Comput. Statist.*, **28**, 352-360.

Wilhelm, A., Symanzik, J. and Wegman, E.J. (1999). Visual clustering and classification: The oronsay particle size data revisited. *Comput. Statist.,* **14**, 109-146.