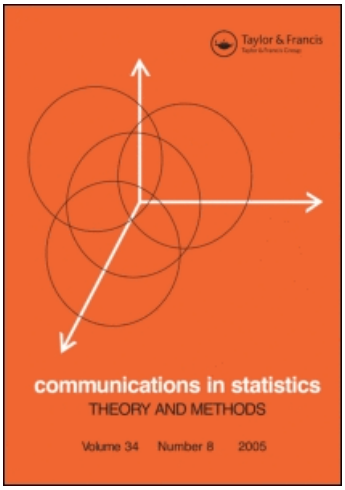


Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713597238

STATISTICS: REFLECTIONS ON THE PAST AND VISIONS FOR THE FUTURE

C. Radhakrishna Rao[a]
[a] Pennsylvania State University, PA, U.S.A.

Online publication date: 30 November 2001

To cite this Article Rao, C. Radhakrishna(2001) 'STATISTICS: REFLECTIONS ON THE PAST AND VISIONS FOR THE FUTURE', Communications in Statistics - Theory and Methods, 30: 11, 2235 — 2257
To link to this Article: DOI: 10.1081/STA-100107683
URL: http://dx.doi.org/10.1081/STA-100107683

COMMUN. STATIST.—THEORY METH., 30(11), 2235–2257 (2001)

STATISTICS: REFLECTIONS ON THE PAST AND VISIONS FOR THE FUTURE

C. Radhakrishna Rao*

Pennsylvania State University, 201 Old Main,
University Park, PA 16802, USA

ABSTRACT

Statistics as data is ancient, but as a discipline of study and research it has a short history. Courses leading to degrees in statistics have been introduced in universities some sixty to seventy years ago. They were not considered to constitute a basic discipline with a subject matter of its own. However, during the last seventy five years, it has developed as a powerful blend of science, technology and art for solving problems in all areas of human endeavor. Now-a-days statistics is used in scientific research, economic development through optimum use of resources, increasing industrial productivity, medical diagnosis, legal practice, disputed authorship, and optimum decision making at individual and institutional levels. What is the future of statistics in the coming millennium dominated by information technology encompassing the whole of communications, interaction with intelligent systems, massive data bases, and complex information processing networks?

*Eberly Professor of Statistics.

The current statistical methodology based on probabilistic models applied on small data sets appears to be inadequate to meet the needs of the society in terms of quick processing of data and making the information available for practical purposes. Adhoc methods are being put forward under the title *Data Mining* by computer scientists and engineers to meet the needs of customers. The paper reviews the current state of the art in statistics and discusses possible future developments considering the availability of large data sets, enormous computing power and efficient optimization techniques using genetic algorithms and neural networks.

Key Words: Bayesian methods; Data mining; Decision theory; Hypothesis testing; Large data sets; Machine learning; Neural networks

1. STATISTICS AS DATA

Statistics as data are ancient, but as a discipline of study and research has a short history. Its origin could be traced to the primitive man who cut notches on trees to keep a record of his personal possessions. The information provided by such a record or *data* probably guided him in his daily activities to gather food and other necessities for himself and his family. The need for systematic collection of data must have arisen when human beings gave up independent nomadic existence and started living in organized communities. They had to assess their needs, pool the available resources and plan for the future on the basis of available information. Then came the establishment of kingdoms ruled by kings, who introduced elaborate systems to collect all sorts of data. It was in their interests to know how many able-bodied men might be mobilized in times of emergency, how many would be needed for essentials of civil life; how numerous or how wealthy were certain minorities who might resent some contemplated changes in the laws of property, or of marriage; what was the taxable capacity of Province, their own and of their neighbors. References to census of people and of agriculture as we carry out today can be found in the old Chinese book, Kuan Tzu (1000 BC), Old Testament (1500 BC) and Arthasastra of Kautilya (300 BC) to mention a few examples. The data relating to a state, to serve as ears and eyes of the government, came to be known as *statistics*, a term coined by the German Scholar Gottfried Achenwall about the middle of the eighteenth century in the wider context of *collection*, *processing and use of data by the*

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

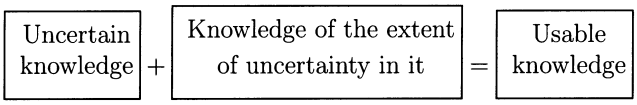
state. [The British were reluctant to use the German word statistics, and they introduced the alternative word *publicistics* which was in use in Britain for some time but soon discontinued.] Statistics acquired special significance, when countries adopted a democratic form of government, as an instrument to comprehend reality through collection of information and to make best possible use of resources for economic development and increasing social welfare. The governments tried to develop national statistical systems to collect data, through administrative channels, special agencies and sample surveys, for use in taking day to day policy decisions, instituting long range plans, and monitoring current projects. Statistical offices were set up in many countries for the purpose of "procuring, arranging and publishing facts calculated to illustrate the conditions and prosperity of the society." [France established the Central Statistical Bureau in 1800, the first one in the world.]

Statistical Societies were established in Europe and America (e.g., Royal Statistical Society, 1834, American Statistical Association, 1832, and International Statistical Institute, 1855) to discuss problems of data collection ensuring international comparisons.

2. STATISTICS AS A SEPARATE DISCIPLINE

2.1. A Paradigm for Statistical Theory and Methods

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



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

- 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 1), 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, to 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 a guidance to research workers. Referring to the ubiquity of statistics, Sir Ronald Fisher (1), in his Presidential Address to the Royal Statistical Society in 1952, made the

Exhibit 1. Basic Concepts of Statistical Inference

Author	Subject	Year of Introduction
Karl Pearson	Chi-square goodness-of-fit	1900
W. S. Gosset	*t*-test	1908
R. A. Fisher	Exact sampling distributions	1915
	Principles of estimation	1922
	Analysis of variance	1923
	Design of experiments	1926
W. Shewhart	Control charts	1931
J. Neyman	Testing of hypotheses	1933
& E. S. Pearson	Confidence intervals	1938
E. J. G. Pitman	Nonparametric tests	1937
P. C. Mahalanobis & M. Hansen	Sample Surveys	1944
A. Wald	Sequential sampling	1947
	Decision theory	1950

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 the whole gamut of natural and social sciences, engineering and technology, management and economic affairs and art and literature.

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 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 a desired level. 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.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

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 settling 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 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 a 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 (2) 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.

3. FUTURE OF STATISTICS

3.1. 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 if 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

Exhibit 2. Practical Problems Motivating General Statistical Concepts (George Box (3))

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	Ratio 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

by Box (3) 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 isolated 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 a *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, perhaps the logic of all disciplines.

It is a *technology* in the sense that statistical methodology can be built into 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 allowance for uncertainty and thereby maximize the efficiency of individual and institutional efforts.

Statistics is also an *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 (4) who conducted a survey and found that statisticians with different backgrounds used different methods for the analysis of 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.

3.2. Limitation of the Current Statistical Methods

Most of the basic concepts of statistical inference and the related statistical methodology (Exhibit 1) were developed 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 results (closed form solutions to problems) involving minimum computations. Tables of limited percentage points of test statistics were constructed 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 (5) 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"].

In the second half of this century, there was a shift in statistical research from model based to semiparametric and nonparametric methods. 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 (6) as *asymptotitise*.

Recent developments in bootstrap methodology introduced by Efron (7) are currently popular as it is not model based and utilize only computing power. However, its justification is again based on asymptotics and the consequences of bootstrapping in small samples have not been fully examined. Tukey (6) recommended the Jackknife method as an alternative.

In books on theoretical statistics, the statistical paradigm is presented as a scientific method consisting of the following cycle of operations for advancement of natural knowledge:

1. Make a model for phenomena under study. 2. Collect data by an experiment or sample survey. 3. Test the model using the data. 4. Refine or reformulate the model. 5. Go back to 2.

The procedure sounds logical, but how does the current statistical methodology espoused in books and research papers come into the picture. There appears to be no substantial evidence in scientific literature of any *major* discovery being directly attributed to a particular statistical method employed. 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 the birth of theory.
>
> —A. Einstein (1879–1955)

> Einstein was aware of the logical difficulty in establishing a given theory as true. He said:
>
> No amount of experimentation can prove me right; a single experiment can prove me wrong.
>
> —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 do not imply that observational data do not provide any clues to a 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. According to Fisher (8), who placed great emphasis on hypothesis testing,

> tests of significance, when used accurately are capable of rejecting or invalidating hypotheses, in so far as these are contradicted by data; but they are never capable of establishing them as certainly true.

According to the above concept of testing a given 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.

Neyman and Pearson (9) introduced the concept of the second kind of error which is the probability of accepting a hypothesis when a specified alternative is true, which again does not provide any guidance to a scientist.

Bayesians argue that testing of hypothesis has no logical basis and that one should start with possible alternative hypotheses with known priori probabilities of being true and derive posterior probabilities in the light of observed data. Since hypotheses not considered have zero prior probability, the true hypothesis when it is outside the chosen set of hypotheses (which generally happens) will never be discovered as it has zero posterior probability.

The well known mathematical statistician, J. Wolfowitz (10) 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 is *Decision Theory* as developed by Wald (11), 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 in accepting or rejecting batches of goods produced in a factory, but does not seem to be applicable in scientific research.

Imagine the following scenario of a possible dialogue between Einstein and some contemporary statisticians.

Einstein: I have a new theory for explaining some natural phenomena. Can statisticians help in testing it?
Neyman and Pearson respond: Einstein, you have to do your own experiment, give us your data and also tell us what the possible alternatives are to your theory. We can then tell you the most powerful method of verifying your theory.
Einstein: Alternative theories! There may be, but I do not know.
Fisher responds: I can give you the design of a perfect experiment to perform. The results can reject your theory if it is wrong and cannot confirm if it is true.
Einstein: I am disappointed, you cannot confirm it if it is true.
Wald and Wolfowitz respond: We would like to review your problem in terms of decision theory. Apart from other inputs, we need to know the losses involved in accepting and rejecting your theory.
Einstein: "If my theory is proven successful, Germany will claim me as a German and France will declare that I am a citizen of the world. Should my theory prove untrue France will say that I am a German, and Germany will declare that I am Jew." [This is a true statement made by Einstein in an address at the Sorbonne].

There are other unresolved problems in hypothesis testing such as conditioning a test statistic on ancillary statistics. Since there is no unique way of choosing ancillary statistics, different choices may give different *p*-values to the test statistics. (See Barnard (12) for a discussion of this problems).

How does one select a model for data processing and prediction of future events? There are model selection criteria 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 (13). 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 ((14), (15)). 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 (16), Lehmann (17) and Rao ((18), 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 hypothesis or theory is in explaining the observed data and in predicting future events, and not whether the proposed hypothesis is *true or false*. A working hypothesis is rejected if a better working hypothesis is found. Reference may be made to Inman (19) for a review of the controversy between Karl Pearson and R. A. Fisher on the use of χ^2 in testing a hypothesis. Pearson refers to his chi-square test as (P, χ^2), emphasizing that the P value is used to judge whether a given hypothesis provides a good graduation to data and not for accepting (as true) or rejecting (as false) a hypothesis. Fisher maintains that the purpose of a test is to give the observations a chance to reject a hypothesis if it is not true. (Unfortunately, in many research reports p-values are treated as a measure of departure from the given hypothesis and small values of p are reported as indicating a highly significant difference.)

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.

It may be true that Fisher overstressed the role of tests of significance using 5% and 1% quantiles of test statistics. He did so at a time when research workers were not familiar with statistical methods and needed some rules which can guide them in the interpretation of the experimental

data they generate. The book by Fisher on *Statistical Methods for Research Workers* was written in response to the demand for statistical methods in solving practical problems. It is in the form of a manual for analyzing live data sets through hypothesis testing and estimation, and drawing conclusions.

It is clear, although Fisher did not emphasize, that he regarded tests of significance as of an exploratory nature

- to examine whether the observed data support a specified hypothesis,
- to detect irregularities such as lack of randomness, recording errors or bias in the collection of data (as in Mendel's data and the effect of methods of ascertainment in genetic studies by Fisher ((20), (21))), and
- to possibly provide guidance for further investigation.

As an example of the last alternative reference may be made to the discovery of the Rhesus factor as described by Fisher (22). 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."

3.3. Limitations of Statisticians

In the early days of 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 2 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 directions or sharpening of the existing tools.

Students graduating in statistics learn statistics as a set of rigid rules without acquiring any knowledge of their applications 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 to a statistical consultant.

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 (23), Parzen (24), Rao (18) and other references in these papers.)

It is also relevant to add here what Fisher (25) 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 for 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. Commenting on the articles published in mathematical statistical journals M. G. Kendall deplores:

> Theoretical statistics in the forties could be understood by anybody with moderate mathematical attainment, say at the first year graduate level. I deeply regret to say that the situation has changed so much for the worse that the journals devoted to mathematical statistics are completely unreadable. Most statisticians deplore the fact, but there is not very much that one can do about it.

There is, however, a recent trend to introduce practical oriented courses using real-life data to motivate theory and methods. Experiments are being conducted at the Pennsylvania State University to develop a curriculum of courses at the undergraduate level with emphasis on data collection and analysis.

3.4. Needs of Customers

Who needs statistics? We have seen that scientists use statistics in a marginal way. 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 customers 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 operations research workers to suggest what is claimed to be a different approach to data analysis called *Data Mining* much to the surprise of statisticians.

4. DATA MINING

4.1. What Is Data Mining

Is Data Mining (DM) a form of statistics or a revolutionary concept? Adriaans and Zantinge ((26), 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. An extreme view is as follows:

> Despite . . . somewhat lofty definitions, DM so far has largely been a commercial enterprise. As in most gold rushes of the past, the goal is to "mine the miners". The largest profits are made by selling tools to the miners, rather than doing the actual mining. The concept of DM is used as a device to sell computer hardware and software.
>
> —Friedman

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.

Exhibit 3. 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	Test1	Test2	Num.	Type	Class
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

4.2. 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 3 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 is bound to be rejected when a large data set is available*. Even the computation of test statistics and estimates of parameters such as the simple 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 4.4 without any model assumptions, which cannot be achieved with small data sets.

4.3. 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 (27) 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.

4.4. 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 certain decision rules R_1 based on the discovery of patterns through a *search engine*. The second set S_2 can be used to evaluate the performance of R_1 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 though 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 $R_1, R_2, \ldots$ 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 $R_1, R_2, \ldots$ 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 3 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.).

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

4.5. Data Mining Techniques and Algorithms

4.5.1. 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 ((8), Chapter 2), Tukey (28).) With increase in computing power and possibilities of viewing high dimensional data through parallel coordinates (Wegman (29), Wegman and Luo (30), Wilhelm, Symanzik and Wegman (31)), projections in different directions (Friedman and Tukey (32)), and data reduction by canonical coordinates (Rao (33)), principal components (Rao (34)), correspondence analysis (Benzēcri (35) and Rao (36)), and multidimensional scaling (Kruskal and Wish (37)), graphical analysis is becoming a valuable tool in discovering patterns in data.

4.5.2. 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 (38).

4.5.3. 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.

4.5.4. 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

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

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

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.

4.5.5. Expert System

This is usually defined as a computer system that performs specialized, usually difficult professional tasks at the level (or sometimes beyond the level) of a human expert. The importance of expert systems is highlighted in a statement of Tukey:

> The man or woman with a thoughtful semester of learning how to use half-dozen systems — one for each of a half-dozen areas of the sort we can properly dream about — will be able to do a lot of data analysis more than those who have three semesters of present education.

5. 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. As such, it is a valuable instrument in any field of inquiry. In his paper on foundations of statistics, Fisher (39) stated the three methodological aspects of statistics as

- specification,
- testing of hypothesis, and
- estimation.

During the last century, research in statistics was directed to these areas. As pointed out in Section 3 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 these inputs. 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 (40) 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

1. Fisher, R. A. The expansion of statistics. J. Roy. Statist. Soc., A, **1953**, *116*, 1–6.
2. Rao, C. R. Statistics and Truth: Putting Chance to Work (Second Edition), World Scientific, Singapore, **1997**.
3. Box, G. E. P. Comment (on A Report of the ASA Section on Statistical Education Committee on Training of Statisticians for Industry). The American Statistician, **1980**, *34*, 65–80.
4. Van den Berg, G. 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, **1992**.
5. Buchanan-Wollaston, H. J. Statistical Tests. Nature, **1935**, *136*, 182–183.
6. Tukey, J. W. Major changes for multiple-response (and multiple adjustment) analysis. In *Multivariate Analysis: Future Directions* (Ed. Rao, C. R.), North Holland, **1993**, 401–422.
7. Efron, B. Bootstrap methods: Another look at jackknife. Ann. Stat., **1979**, *7*, 1–26.
8. Fisher, R. A. *Statistical Methods for Research Workers*, Hafner Publishing Company, **1967**.
9. Neyman, J.; Pearson, E. S. On the problem of most efficient tests of statistical hypotheses. Philos. Trans. Roy. Soc. A, **1933**, *231*, 289–337.
10. Wolfowitz, J. Remark on the theory of testing of hypotheses. The New York Statistician, **1967**, *18*, 1–3.
11. Wald, A. Statistical Decision Functions, Wiley, New York, **1950**.

12. Barnard, G. A. Scientific practice and statistical inference. *Symposium on the Foundations of Statistical Inference, with Special Emphasis on Applications in honor of D.A. Sprott*, **1996**, 1–9.
13. Rao, C. R.; Wu, Y. On model selection. IMS Lecture Notes (in press), **2000**.
14. Rao, C. R. Prediction of future observations with special reference to linear models. J. Multivariate Analysis, **1977**, *4*, 193–208.
15. Rao, C. R. Prediction of future observations in growth curve type models (with discussion). J. Statistical Science, **1987**, *2*, 434–471.
16. Cox, D. R. Role of models in statistical analysis. Statistical Science, **1990**, *5*, 169–174.
17. Lehmann, E.L. Model specification: The views of Fisher and Neyman and later developments. Statistical Science, **1990**, *5*, 160–168.
18. Rao, C. R. A cross disciplinary approach to teaching of statistics. Proc. 51st session of the Int. Statist. Institute, Istanbul, **1997**.
19. Inman, H. F. Karl Pearson and R. A. Fisher on statistical tests: A 1935 exchange from Nature. The American Statistician, **1994**, *48*, 2–11.
20. Fisher, R. A. The effect of methods of ascertainment upon the estimation of frequencies. Ann. Eugenics, **1934**, *6*, 13–25.
21. Fisher, R. A. Has Mendel's work been rediscovered? Annals of Science, **1936**, *1*, 115–137.
22. Fisher, R. A. The Rhesus factor: A study in scientific research. Am. Sc., **1948**, *35*, 95–102, 113.
23. Kettenring, J. What industry needs. The American Statistician, **1995**, *49*, 2–4.
24. Parzen, E. Data mining, statistical methods mining and history of statistics. Tech. Rept. Texas A & M University, **1997**.
25. Fisher, R. A. Presidential Address: Indian Statistical Conference. Sankhyā, **1938**, *4*, 14–17.
26. Adriaans, P.; Zantinge, D. Data Mining. Addison Wesley, **1996**.
27. Wegman, E. J. Visions: The evolution of statistics. Keynote talk at the Conference, New Techniques and Technologies in Statistics, Sorrento, Italy, **1998**.
28. Tukey, J. W. *Exploratory Data Analysis*. Reading, Mass: Addison-Wesley, **1977**.
29. Wegman, E. J. Hyperdimensional data analysis using parallel coordinates. J. Am. Statist. Assoc., **1990**, *68*, 664–675.
30. Wegman, E. J.; Luo, Q. High dimensional clustering using parallel coordinates and grand tour. Computing Science and Statistics, **1997**, *28*, 352–360.

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

31. Wilhelm, A.; Symanzik, J.; Wegman, E. J. Visual clustering and classification. The oronsay particle size data revisited. Computational Statistics, **1999**, *14*, 109–146.
32. Friedman, J.; Tukey, J. W. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers, **1974**, *23*, 881–889.
33. Rao, C. R. The utilization of multiple measurements in problems of biological classification. J. Roy. Statist. Soc. B, **1948**, *9*, 128–140.
34. Rao, C. R. The use and interpretation of principal components analysis in applied research. Sankhyā A, **1964**, *26*, 329–358.
35. Benzēcri, J. P. *Correspondence Analysis Handbook*, Marcel Dekker Inc., New York, **1992**.
36. Rao, C. R. A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. Qüestiio, **1995**, *9*, 23–63.
37. Kruskal, J. B.; Wish, M. *Multidimensional Scaling*. Sage Publications, **1978**.
38. Agrawal, R.; Imielinski, T.; Swami, A. Mining association rules between sets of items in large databases. *Proc. Int. Conf.*, ASMSIGMOD, Washington, D.C., **1993**, 207–216.
39. Fisher, R. A. On the mathematical foundations of theoretical statistics. Philos. Trans. Roy. Soc., **1922**, *A222*, 309–368.
40. Weiss, S. M.; Indurkhya, N. *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers, Inc., **1998**.

Received November 2000
Revised April 2001

Request Permission or Order Reprints Instantly!