SF 298 MASTER COPY KEEP THIS COPY FOR REPRODUCTION PURPOSES

REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comment regarding this burden estimates or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE March 1997	3. REPORT TYPE AND DATES COVERED Technical - 97-05

4. TITLE AND SUBTITLE

A Cross Disciplinary Approach to Teaching of Statistics

5. FUNDING NUMBERS

DAAH04-96-1-0082

6. AUTHOR(S)

C.R. Rao

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(ES)

Center for Multivariate Analysis
Department of Statistics
417 Thomas Building
Penn State University
University Park, PA 16802

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

ARO 35518.11-MA

11. SUPPLEMENTARY NOTES

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

Historically, the phenomenal growth of statistics as a separate discipline of great ubiquity was motivated by the need to solve practical problems arising in social, biological and natural sciences. Statistical methodology as we practice today involves acquisition of data. extraction of available information and taking optimal decisions under uncertainty. That the role of statistics as the logic and science of solving problems in other disciplines and that continued advances in statistics depend strongly on research stimulated by and directed at problems in other disciplines are not fully reflected in the educational programs for statisticians at the universities. It is suggested that more emphasis should be given to the interface between statistics and other disciplines through data - centered training in statistics.

14. SUBJECT TERMS

Cross disciplinary studies; Cross examination of data; Design of experiments; Exploratory data analysis; Sample surveys; Teaching of statistics.

15. NUMBER IF PAGES

4

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

A CROSS DISCIPLINARY APPROACH TO TEACHING OF STATISTICS

C. Radhakrishna Rao

Technical Report 97-05

March 1997

Center for Multivariate Analysis
417 Thomas Building
Penn State University
University Park, PA 16802

The research work of the author is supported by the Army Research Office under Grant DAAHO4-96-1-0082. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

Typeset by $\mathcal{AMS}$-TEX

A CROSS DISCIPLINARY APPROACH TO TEACHING OF STATISTICS

C. Radhakrishna Rao
Statistics Department
Pennsylvania State University
University Park, PA 16802

Abstract: Historically, the phenomenal growth of statistics as a separate discipline of great ubiquity was motivated by the need to solve practical problems arising in social, biological and natural sciences. Statistical methodology as we practice today involves acquisition of data, extraction of available information and taking optimal decisions under uncertainty. That the role of statistics as the logic and science of solving problems in other disciplines and that continued advances in statistics depend strongly on research stimulated by and directed at problems in other disciplines are not fully reflected in the educational programs for statisticians at the universities. It is suggested that more emphasis should be given to the interface between statistics and other disciplines through data - centered training in statistics.

1. INTRODUCTION

In her book on *Studies in the History of the Statistical Method*, published in 1931, Helen Walker referred to education in statistics as follows:

> Educational statistics is the offspring of a varied ancestry. The greed of ancient kings enumerating their people for taxation; the panic of an English sovereign during London plague; the cupidity of professional gamblers; the scientific ardor of psycho-physicists; the labor of mathematicians and astronomers and physicists and actuaries; the enthusiasm of the students of social phenomena; the disciplined imagination of the biologists; and the vision of the educators planning a new science of education; from these has educational statistics descended.

What is common to all these situations is uncertainty in the available information and the need to take decisions under uncertainty. It was realized in the beginning of the present century that the key to decision making is quantification of uncertainty, which led to the founding of statistics as a separate discipline devoted to the development of methodology for extracting information from data and assessing the amount of uncertainty in it, on the basis of which optimal decisions with minimum risk could be taken.

The fields of application in which the relevance and importance of statistics was first demonstrated were social and biological sciences, but soon statistical reasoning became an integral part of decision making in all fields of human endeavor. Statisticians are regularly employed in large numbers for information gathering and processing for decision making in government, industry and commercial organizations, and the demand for statisticians is likely to extend to other areas. There is also considerable demand for statistical consultants in specialized areas like law, literature and arts. How are we going to educate and provide the necessary expertise to statisticians to work in all these areas involved in data collection and analysis for decision making, to actively collaborate in cross disciplinary research and to do theoretical research for refining existing methodology and developing new tools for data analysis?

If these are the objectives, then the education of a statistician should be more like that of a technologist, broad based and aimed at providing skills which can be used in a variety of situations, and less like that of a scientist specializing in a narrow area with problems of its own and methods of its own in solving them. Unfortunately, the courses given in statistics departments of universities seem to be exercises in mathematics designed to produce statisticians with mathematical skills to solve well posed problems in statistical theory and not statistical skills to formulate and solve problems in other areas. For progress of statistics as a viable subject and for statistical knowledge to have the maximum impact on the welfare of the society, we need specialists of the latter kind.

I shall make a few suggestions based on my experience in designing and formulating courses in statistics at various levels at the Indian Statistical Institute. These courses are meant to provide the basic knowledge and skills needed in statistical practice for collecting or generating data and making effective use of computer software for data analysis. These courses are more suitable for data - oriented and, indeed, will be more effective than mathematics - oriented teaching or training. They also provide the proper background for introducing the theoretical foundations of statistics, and also motivate further research.

2. BASIC COURSES IN STATISTICS

2.1 Sample Surveys

A sample survey is an effective tool for collecting information on a wide variety of aspects of a large population. When properly conducted it would provide a wealth of data, useful for planning and policy purposes, expeditiously, economically and with a reasonable degree of accuracy and at the same time ensuring objectivity of data. Sample surveys has a long history, but it is only in the second quarter of this century the methods were refined and put into regular practice by government agencies and business organizations. A sample survey involves: (1) preparation of schedules for recording information, (2) drafting of a manual, outlining concepts and definitions of measurements or responses involved, and instructions to investigators in the choice of sample units and recording observations, (3) training of investigators, (4) conducting a pilot survey to test draft schedules and to determine sample size to ensure a specified accuracy in sample estimates, (5) choice of a frame for selection of units, (6) efficient designs for random selection of units, (7) planning of field investigation, (8) supervision of field work and use of cross checks, (9) scrutiny of primary data and tabulation and (10) reporting of survey results. The best way of teaching sample surveys is to choose some problem and let the students conduct a survey going through the various stages mentioned above. The course offers an excellent opportunity to discuss various aspects of sampling from a finite population, problems of estimation and statistical inference.

2.2 Design of Experiments

Experiments are often conducted to estimate unknown parameters, to compare alternative treatments as in agricultural or clinical trials, and to determine the optimum mix of factors to produce goods of high quality. Data collected by experimenters are often defective due to lack of controls, insufficient replication and non-random assignment of treatments to experimental units, all of which effect the validity and efficiency of the conclusions drawn. A statistician can play a major role in designing an experiment to provide valid results with minimum cost. R.A. Fisher who introduced and developed the subject of design of experiments stressed its importance in the following terms:

> A competent overhauling of the process of collection, or of experimental design may often increase the yield (precision of results) ten or twelve fold, for the same cost in time and labor. To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.

The basic principles of design of experiments, replication, randomization and local control, could be demonstrated while conducting an actual experiment, and methods of analysis could be discussed with reference to data generated by the experiment. It would also be useful to introduce the concepts of factorial experiments, interaction between factors and optimum mix of factors to improve yield and/or quality.

2.3 Cross Examination of Data

Statisticians are often required to deal with data collected by others, which may be subject to gross recording errors, which may be faked and which may be biased due to editing and some irregularities in sampling. The first and foremost task of a statistician is to cross examine the data, a term coined by Fisher, for reviewing the process by which data are ascertained, to find out inhomogeneities in data due to missing values, clustering and outliers and to decide on a suitable stochastic model for inferential data analysis. A blind application of statistical methods without such exploratory or initial data analysis may lead to misleading inference. There are no routine methods for cross examination of data. The skills and knowledge necessary for this purpose cannot be imparted in a classroom lecture. The only way to train statisticians in this important phase of data analysis is to provide them with experience in handling live data from different disciplines. Some broad principles of such initial data analysis may be formulated while discussing the peculiarities observed in different data sets. There is an excellent paper by Mahalanobis (1933) who used some ingenious methods for rectifying errors in published records. Other publications by Majumdar and Rao (1958) and Rao (1989, pp.36-94) contain extensive accounts of cross examination of data.

2.4 Statistical Inference

Discussion of sample surveys and design of experiments would provide a good opportunity to introduce statistical inference associated with point estimation and tests of significance. While it would be useful to give lectures on general theories of estimation and testing, properties of estimators like consistency, unbiasedness and variance and those of test criteria in terms of power function could also be demonstrated through simulation studies.

2.5 Concepts of Probability

I have discussed in Rao (1970) how discrete probability models can be introduced in a natural way while investigating random phenomena. Data may be generated by drawing cards, throwing coins etc., and the concept of probability could be developed in relation to these random experiments. The law of large numbers, and central limit theorem for discrete variables could also be demonstrated. I have also shown how by comparing a live sequence of male and female births delivered in a hospital with that of heads and tails in a coin tossing experiment, a hypothesis on sex determination could be postulated (see Rao (1970)).

2.6 Graphical Techniques

The book on *Statistical Methods for Research Workers* by the late Sir Ronald Fisher has a remarkable opening chapter entitled, Diagrams, which I think has not received due recognition by the statisticians. Other serious books on statistics do not mention diagrams and graphs. Graphical presentation of data must be regarded as an integral part of data analysis and I hope future books on statistics give more attention to it. Besides being of value "in suggesting statistical tests and in explaining the conclusion based on them", (as mentioned by Fisher), graphical presentation of data provides an insight into the problem under investigation and sometimes enable us to draw conclusions without or with a little further analysis. A graph provides a better summary than averages and indices and is, therefore, a valuable tool for preliminary examination of data leading to the choice of an appropriate model and possible transformation of the variables to simplify the statistical analysis. A graph brings out the outliers clearly and is thus a useful device in scrutiny of data and in looking for contamination in data.

Graphical methods have universal applicability and they can be taught without any advance preparation of students in statistical methodology or mathematics. For example, the concept of a control chart could be introduced as a graphical techniques and its usefulness in industrial production explained.

2.7 Statistical Software

Statistical computing has become increasingly comprehensive and user friendly as computer software has advanced from special - purpose programs to statistical packages and to interactive computing, some with extensive graphic capabilities. Attempts are being made to produce more intelligent software, or what is called expert systems, for (a) guiding the user to existing sources of information, (b) helping to define the statistical problem, screen the data, and select an appropriate model, (c) suggesting appropriate methods of analyzing data, (d) automatically performing data analysis, (e) explaining the meaning of results, and (f) pointing out patterns and peculiarities in data and suggesting further evaluations. Such a facility, if properly utilized, will be of immense help in data - centered teaching and will be an asset to statisticians in actual practice. Some questions have been raised about the misuse of statistical software, but this refers only to untrained statisticians. Talking about expert systems, Tukey (1985) says:

> 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 now properly dream about - will be able to do a lot of effective data analysis more than those who have three semesters of present education.

3. WHO SHOULD TEACH STATISTICS?

To sum up statistical knowledge is an asset in all investigations, scientific or otherwise. There is a growing demand for statisticians to work on practical problems in many areas of human endeavor. The greatest theoretical advances in statistics have come as a direct result of important applications, and in many cases advances in substantive knowledge and in statistical theory are inseparable.

In view of this a restructuring of the courses for the education and training of statisticians currently offered at the universities seems to be necessary. It is suggested that in the courses for statisticians greater emphasis should be given to data acquisition techniques like sample surveys and design of experiments, and data analytic techniques like exploratory data analysis including graphics, and knowledge of expert systems. Such courses offered through data - oriented teaching will go a long way in meeting the demand for statistical practioners in different areas, as well as research workers needed for the advancement of useful statistical theory.

Then, a question arises as to who should teach statistics. There is considerable debate on the subject, but I believe in what Fisher has said:

> I want to insist on the important moral that the responsibility of 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 of drawing conclusions from actual data, upon which practical action is to be taken. Mathematical acuteness is not enough.

Acknowledgement: The work was supported by a Grant from the Army Research Office: DAAH04-96-1-0082.

4. REFERENCES

Mahalanobis, P.C. (1933). A revision of Risley's anthropometric data relating to the tribes and castes of Bengal. *Sankhyā* 1, 76-105.

Majumdar, D.N. and Rao, C.R. (1958). Bengal anthropometric survey, 1945 - A statistical study. *Sankhyā* 19, 201-408.

Rao, C.R. (1969). A multidisciplinary approach for teaching statistics and probability. *Sankhyā* B, 321-340.

Rao, C.R. (1974). Teaching of statistics at the secondary level - an interdisciplinary approach. In *Statistics at the School Level*, Almquist and Wiksell Int., Amsterdam, 121-140.

Rao, C.R. (1989). *Statistics and Truth: Putting Chance to Work.* Available in English, and Japanese, Spanish, Polish, German and Chinese translations.

Tukey, J. (1985). Comment on a paper. *American Statistician* 39, 12-14.

Résumé: Le phenomène de la croissance des statistiques comme une discipline omniprésente était historiquement motivè par le besoin de résoudre les problèmes pratiques qui se sont soulevés dans les sciences sociales, biologiques et naturelles. La méthodologie statistique, comme nous la pratiquons aujourd' hui, implique l'acquisition de données, la soustraction d'informations disponibles et la prise de décisions optimales dans l'incertitude. Le role des statistiques, en tant que la logique et la science servant à résoudre les problèmes au niveau de d'autres disciplines, et tendant vers l'avancement continu des statistiques, dépend fortement sur la recherche encouragée par, er dirrigée vers les problèmes dans d'autres disciplines. Pour les statisticiens, ces problèmes ne sont pas entièrement soulevés dans les programmes d'éducation au niveau universitaire. La suggestion ici faite est que plus d'intérêt soit porté au lien existant entre les statistiques et les autres disciplines, par le biais d'un entrainement basé sur les données de statistiques.