DISCUSSION OF PAPERS BY DOERGE & WEIR, JOHNSON, AND URQUHART & McDONALD, SESSION # 340, JOINT STATISTICAL MEETINGS, SAN FRANCISCO, AUGUST 12

by
Walter T. Federer
Liberty Hyde Professor of Biological Statistics, Emeritus
Cornell University

BU-1214-M* July 1993

ABSTRACT

Three papers by authors cited in the title of this paper were presented at the Joint Statistical Meetings 1993 Session #340 entitled "Biometry: Challenges for the Future". In the main, the authors stated some of the current statistical considerations in mapping genes on chromosomes, trends in statistical consulting in Agriculture of one person at Kansas State University, and several problems encountered in ecology. Statistical problems peculiar to the authors' area were described but little or nothing for the general field of Biometry. To broaden this outlook of the "Future of Biometry", several papers which had considered more general aspects of this topic, were cited, and statistical problems in several areas were discussed. These were in multivariate analyses, laboratory procedures and analyses, modeling and model design, treatment design (specifically fractional replication), sequential experimentation, statistical design (population structure), statistical design for computer experimentation, treatment by sub-population interaction, and biotechnology investigations.

* In the Technical Report Series of the Biometrics Unit, Cornell University, Ithaca, New York 14853

DISCUSSION OF PAPERS BY DOERGE & WEIR, JOHNSON, AND URQUHART & McDONALD

by
Walter T. Federer
Liberty Hyde Professor of Biological Statistics, Emeritus
Cornell University

BU-1214-M July 1993

INTRODUCTION

The theme of the 1993 Joint Statistical Meetings is "Statistics in the Sciences". The "Sciences" is a broad term including all fields of scientific inquiry. Biometry likewise is a broad term which covers the development and application of quantitative (not just statistical) methods to biological phenomenon. Our speakers on this Session entitled "Biometry: Challenges for the Future" have chosen to concentrate on three particular areas, i. e.,

statistical procedures useful for mapping genes on chromosomes,
some aspects of agricultural experimentation, and
an ecological problems associated with sampling and measurement.

Let us briefly summarize what each of their papers discussed.

DOERGE AND WEIR (D&W)

D&W have the following sections in their paper
an introduction to statistical genetics,
a development of notation for statistical genetics using markers for mapping genes,
single marker analysis,
a regression analysis for regressing the trait value on the marker genotype,
a likelihood method for trait values in the marker classes,
many marker analyses (interval mapping, two markers, and many markers),
regression methods (two markers and many markers), and
a discussion of the above.

The paper is a lucid account of current procedures and methods in this area for the most

part but I did not find out what they think the future for Biometry is in their area. Gary Churchill, Cornell University, has effectively used several other procedures for mapping genes on chromosomes. To cite one instance, he (in collaboration with others) found that one of the pooling procedures described in Federer (1991), Chapter 6, was very useful and made a relatively impossible procedure possible. The paper by Wilson (1993), presented at these Meetings, appears to be related to the area covered by these papers.

Neither of the above two authors appeared to present this paper. Instead William Brevis, Pioneer Hybrid Seed Company, presented a talk in their place. No paper was available for review. He discussed the need for more precise methods for use in mapping genes on chromosomes, stated that regression methods appear promising, stated that one challenge of the future is for analytical techniques useful for determining number and position of genes affecting quantitative traits (QTL's), stressed the need to study genotype by environment interactions, and questioned the adequacy of using linear models.

JOHNSON (DEJ)

The main headings of the Johnson paper are
personal notes,
statistical computing (using specific software packages for numerical computing),
some changes in statistical consulting over the last 15 years at KSU,
statistical needs associated with quicker and faster methods for obtaining data,
spatial methods,
on-farm trials,
meta-analysis,
bootstrapping, and
imputation.

DEJ's challenge for the future is for statisticians to learn more about the nature and use of presently available spatial methods.

It would appear that the talk presented by Mihram and Mihram (1993) at these Meetings dwells on topics related to the future of Computing.

URQUHART AND McDONALD (U&M)

U&M state that they will focus on the heritage of modern biometry, its shortcomings, and an area where work is needed. Their outline is

past and current focuses in Biometric literature (primarily analysis),
defining responses outside of agriculture with examples (how Statistics can help), and
needed research.

A copy of their paper was not available but only a copy of their overheads. They find that the very large part of statistical literature deals with how to perform a statistical analysis

with very little relating to designing experiments to maximize information, and almost none on the selection of what measurements, what measuring instruments to use, and what type of responses are required. (See also Federer. 1991, Chapters 2, 5, 7, and 8.) They state that substantial biometric research needs to be directed to

responses underlying the definition of cost effective responses,
definitions and operating criteria of studies targeted toward entire ecosystems, and
implications of spatial correlation for plot design and survey methodology in general.

It should be noted that farming systems such as intercropping in its broadest context fit into the second area above (See Federer, 1993.). A paper presented at these Meetings by Schwager, Robson, Heimbuch, and Wilson (1993) appears to be an answer to one of the questions raised by these authors.

OTHER PAPERS ON THE FUTURE FOR BIOMETRY

What appeared strange was that except for one reference, none of the authors cited previously written articles in statistical literature on future challenges in Statistics and for statisticians such as, e. g., Anderson, (1984), Armitage (1985), Dagnelie (1982), Federer (1984, 1986), National Research Council, CATS (1991), Zelen (1983), and perhaps many others that did not come to my attention under a quick perusal of literature. Several ideas of Federer (1984) are expanded upon in Chapters 5 and 6 of Federer (1991). A recent paper by Finney (1993) discusses many issues related to the future of biometrical problems and research.

DISCUSSION

Some areas of challenge in Biometry which were not mentioned by the speakers are listed below.

Multivariate analyses As shown by Federer and Murty (1987) and Federer (1993), presently developed multivariate procedures, except for the type presented by Pearce and Gilliver (1978, 1979), have little use in analyzing data from intercropping experiments. The same is true for many other areas such as pharmokinetics, ecology, or any investigation dealing with systems and mixtures of treatments. These areas require a multivariate analysis but not of the kind taught in statistics courses. The canonical variables that arise must be such that they can be interpreted by the investigator.

Laboratory procedures and analyses College and university laboratories that I have encountered have little or no statistical quality control procedures. During the course of statistical consulting, it became evident that data could be unreliable because they were sometimes faked but usually because of sloppy procedures in the laboratory. Also, several

things could be done which would greatly improve the efficiency for analyzing samples and greatly reduce the cost to experimenters having their samples analyzed. One such procedure that is very efficient for routine screening samples for a trait is the Dorfmann and Sobel pooling procedures. Several examples of how these procedures have or could have been used are described in Chapter 6 of Federer (1991) as well as several other procedures that would be useful in screening laboratory samples. Pooling procedures alone could result in savings of millions of dollars world-wide as well as getting timely results back to the experimenter. In some studies with which I have been associated, the experimenter had to wait three years before the laboratory returned his analyses and several years in another case. These long waits for laboratory results are unjustifiable when statistical techniques are available which would alleviate the situation.

Modeling and model design Although I predicted that this topic would be one would be a major effort of statisticians some 30-40 years ago, I am confident that it will occupy much of statisticians' time in the first part of the 21st century. Although Box and Hunter (1962) presented several procedural ideas for developing a model and Box and Cox (1964) in their classic paper show how to use curve fitting techniques in selecting a response model, their ideas appear to be slow in being picked up. The recent paper by Atkinson *et al.* (1993) is an exception. These authors consider responses in bioavailability studies such as the area under the concentration curve, the maximum concentration, and time to maximum concentration and find optimum designs for each response and for all combined.

Although the modeling of responses and model selection is gradually receiving more attention from statisticians, the main thrust in this area is by the Operation Research scientists. Their work is mainly on deterministic modeling as is the work by bio-physicists, bio-mathematicians, and bio-engineers. Statisticians need to pursue this topic vigorously , especially stochastic modeling, if they are to keep pace with the needs of 21st century scientists. The nature, causes, and behavior of "error" needs considerable study and discussion. Almost all statistical procedures consider "error" as an add-on to a deterministic model. Especially in the area of stochastic modeling, "error" will need to be considered as a part of the stochastic process.

Treatment design, specifically fractional replication The bulk of the work in this area has been on constructing main effect plans and their properties. When it is desired to construct, say, resolution VII designs for 3^{23}, $3^{14} \times 2^{7}$, etc. with some optimality properties for both saturated and minimal orthogonal designs, we have no procedure for selecting generators, and computers are not the answer. We have some computer programs to aid in this, but none to answer the question. Work is progressing on a method for selecting generators, but the final results are not yet in. There are many other areas of treatment design which need attention.

Sequential experimentation This is an area needing a lot of attention. Perhaps it will be possible to write a Snedecor-Cochran type textbook for sequential procedures by the end of the 20th century. Most research except for, perhaps, graduate student theses have a

sequential nature. We ignore this when we teach statistical methods. By the way, the same thing is true for random samples sizes when the treatment responses may depend upon the sample size configuration. Note that linear model textbooks ignore this fact for the most part. Their procedures are conditional upon the sample pattern present in the data. We also need a Snedecor-Cochran type textbook for this situation.

Statistical design Population structure for all types of statistical design needs to be elucidated and clarified. Statements such as, "This design is a sample of the population of which it is a sample" is meaningless and circular. An investigator must be able to describe the population from which the sample was drawn in order to make meaningful inferences. The population structure for some experiment designs has been described by Federer (1991, 1993) but does not appear in other books purporting to present the topic of statistical (mostly called experimental) design. Survey design statisticians have done a good job with this topic, but the need for describing real world population structures appears to have eluded other statisticians. Inferences to an undescribed and unknown population are academic and useless in the real world.

Statistical design for computer experimentation There more than likely will be textbooks written around the turn of the century which deal exclusively with the planning and execution of experimentation on the computer by subject matter specialists. In fact, many experimenters will be running computer experiments prior to conducting actual experiments in order to make more efficient use of their time and resources. We need many case study examples from real world computer experiments in order to be able to write such textbooks. One example of statisticians doing something in this area is the recent paper by Morris *et al.* (1993) where they were exploring deterministic response functions in response surface prediction.

In connection with statistical computing, a Forum on this topic was held by the Committee on Applied and Theoretical Statistics (CATS) of the National Research Council. CATS plans to issue a report on this topic if they haven't already.

Treatment by sub-population interaction Genotype by environment interactions are crucial in making varietal recommendations for a region and studies on this aspect are many. What is needed is an efficient and relatively inexpensive (time and money) procedure for doing this. One such procedure has been proposed by Federer and Scully (1993). Others may be needed in this area as well as in other areas of investigation where treatment by sub-population interactions are important such as in medical drugs for general use.

Biotechnology investigations It appears to me that by and large, biotechnology researchers know little or nothing about statistically designing and analyzing results from investigations. It will be necessary for statisticians to worm their way into the good graces of these researchers and show them how to utilize Statistics, the language of science. One exception appears to be in genetics where statisticians are actively involved in biotechnological genetic research.

Many others For example, what will replace or supersede Gibbs sampling and bootstrapping? How do statisticians respond to the question of whether statistical packages are the tools of the statistician and NOT a replacement for a statistician? How effective are statisticians in acquainting administrators with the need for Statistics as the language of Science, and the need for statisticians to help researchers and to solve their statistical problems? Many more areas are cited in the following references

LITERATURE CITED

Anderson, R. L. (1984). Where are we and where should we be going? *Journal of the American Statistical Association* **79**, 253-258.

Armitage, P, (1985). Biometry and medical statistics. *Biometrics* **41**, 823-833.

Atkinson, A. C., K. Chaloner, and A. M. Herzberg (1993). Optimum experimental designs for properties of a compartmental model. *Biometrics* **49**, 325-337.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Series B, **26**, 211-252.

Box, G. E. P. and W. G. Hunter (1962). A useful method for model building. *Technometrics* **4**, 301-308.

Dagnelie, P. (1982). La biometrie dans les pays du Tiers Monde. In *Actes de la Ileme Confernce Internationale de Biometrie*, Versailles:INRA, pp. 177-182 .

Federer, W. T. (1984). Cutting edges in Biometry. *Biometrics* **40**, 827-829.

Federer, W. T. (1986). Whither Statistics? *In Statistical Design: Theory and Practice. A Conference in Honor of Walter T. Federer*, (Editors: C. E. McCulloch, S. J. Schwager, G. Casella, and S. R. Searle), Biometrics Unit, Cornell University, Ithaca, N. Y. 14853, pp. 211-231.

Federer, W. T. (1991). *Statistics and Society--Data Collection and Interpretation*, second edition, revised and expanded. Marcel Dekker, Inc., New York, xvi + 578 pp.

Federer, W. T. (1993). *Statistical Design and Analysis for Intercropping Experiments. Volume 1. Two Crops.* Springer-Verlag, New York, Berlin, Heidelburg, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest, xx + 299 pp.

Federer, W. T. and B. R. Murty (1987). Uses, limitations, and requirements of multivariate analyses for intercropping experiments. In *Biostatistics* (Editors: I. B. MacNeill and G. J. Umphrey), D. Reidel Publishing Co., Dordrecht, Holland, and Boston, Mass., pp. 269-283.

Federer, W. T. and B. T. Scully (1993). Parsimonious statistical design and breeding procedure for evaluating and selecting desirable characteristics over environments. *Theoretical and Applied Genetics* **86**, 612-620.

Finney, D. J. (1993). Whither Biometry? In *Multiple Comparisons, Selection, and Applications in Biometry (A Festschrift in honor of Charles W. Dunnett.)*, Marcel Dekker, Inc., New York, Basel, Hong Kong, pp. 11-27.

Mihram, A. and D. Mihram (1993). Computers and Statistics: Science in the 21st century. Paper presented at The Joint Statistical Meetings, San Francisco, August 10, Session #173.

Morris, M. D., T. J. Mitchell, and D. Ylvisaker (1993). Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction. *Technometrics* **35**, 243-255.

National Research Council, CATS (1991). The Future of Statistical Software. Proceedings of a Forum, NRC, National Academic Press, Washington, D. C., ix +88 pp.

Pearce, S. C. and B. Gilliver (1978). The statistical analysis of data from intercropping experiments. *J. Agricultural Science, Cambridge*, **91**, 625-632.

Pearce, S. C. and B. Gilliver (1979). Graphical assessment of intercropping methods. *J. Agricultural Science, Cambridge*, **93**, 51-58.

Schwager, S. J., D. S. Robson, D. G. Heimbuch, and H. T. Wilson (1993). A method for detecting long-term trend in environmental monitoring. Paper presented at The Joint Statistical Meetings, San Francisco, August 10, Session # 175.

Wilson, S. (1993). Genomics and Statistics: The challenge of the interface. Paper presented at the Joint Statistical Meetings, San Francisco, August 10, Session #187.

Zelen, M. (1983). Biostatistical science as a discipline: A look into the future (with discussion). *Biometrics* **39**, 827-837.