Computer Science and Statistics:

Proceedings of the Fifteenth Symposium on the Interface

James E. Gentle
editor

North-Holland

COMPUTER SCIENCE AND STATISTICS:

PROCEEDINGS OF THE FIFTEENTH SYMPOSIUM ON THE INTERFACE

Houston, Texas, March 1983

edited by

JAMES E. GENTLE

Manager, Research and Design
IMSL, Inc.
Houston, Texas, U.S.A.

1983

NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM · NEW YORK · OXFORD

ISBN: 0 444 86688 4

Published by:
NORTH-HOLLAND PUBLISHING COMPANY – AMSTERDAM · NEW YORK · OXFORD

Sole distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE
NEW YORK, N.Y. 10017

This work relates to the Department of the Navy Purchase Order N00014-83-M-0041 issued by the Office of Naval Research. The United States Government has a royalty-free license throughout the world in all copyrightable material contained herein.

Library of Congress Cataloging in Publication Data

Symposium on the Interface (15th : 1983 : Houston, Tex.)
Computer science and statistics.

1. Mathematical statistics--Data processing--
Congresses. I. Gentle, James E., 1943-
II. Title.
QA276.4.S95 1983 519.5'028'54 83-11462
ISBN 0-444-86688-4

PRINTED IN THE NETHERLANDS

COMPUTER SCIENCE AND STATISTICS:
Proceedings of the
Fifteenth Symposium on the Interface

PREFACE

The rapidly occurring advances in computing science are making a major impact on the activities of the applied statistician. Conversely, the needs of the statistical data analyst have supplied the computer scientist and the numerical analyst with interesting and challenging problems. This exciting area in the interface of the disciplines has been explored since 1967 in the Symposia on the Interface.

The sessions of the Fifteenth Symposium, held in Houston, generally covered the traditional topics of the interface, such as graphics, numerical algorithms, statistical program packages, and simulation techniques. Reflecting the impact of one of the most significant current trends were the two sessions on smaller computers. The current status of the available software for the desktop computers is rapidly changing. The generally negative reports on this software given at the Symposium did contain a few bright spots, but indicated a need for extensive development of system support software for the micros. There were sessions on computing problems in specific areas of statistics, such as time series and survival analysis, and on general techniques, such as optimization. Other sessions considered such problems as tools for program development and measurement of software complexity.

The format of the Fifteenth Symposium was similar to that of recent Interface Symposia. Following the keynote address in the morning of the first day, there were three concurrent sessions consisting primarily of presentations by invited speakers and audience discussion. Many of the presentations incorporated audio-visual effects that can not be made available through these proceedings; nevertheless the written versions of the papers contained in this volume generally summarize the presentations and perhaps make some points in a more considered manner than the corresponding live presentations. In addition to the audio-visual effects and the spontaneity of oral presentations, there are other aspects of the symposium that the proceedings cannot capture. Much of the real interchange of ideas and information occured in informal contact and in demonstrations. (As I was leaving the hotel around 12:30 a.m. Friday morning, I saw a group animatedly discussing a computing problem in survival analysis. I will not try to describe the problem here!) During the noon break on both days contributed papers were presented in poster sessions. Written versions of these papers also appear in this volume.

There are a number of people who contributed to the success of the Fifteenth Symposium on the Interface and I would like to convey my thanks to all of these. In particular, I wish to thank Doris Ann Moore for help with so many of the details. I would like to thank Cecile Blake, Karen Clark, Bill Evans, Joyce Gentle, Byron Howell, Don Kainer, Tim Leite, Lindsay Reed, Bill Sallas, and Jim Thompson for help with local arrangements. I would like to thank Tom Aird, Tom Boardman, John Dennis, H. E. Dunsmore, Rich Heiberger, Bill Kennedy, Peter Lewis, Bob McGill, Webb Miller, Joe Newton, Michael Steele, Mike Tarter, Harry Wong, and Sid Yakowitz for their help in organizing the program. Tom Glenn and several other members of the Operations Division at IMSL were very helpful in various other aspects of putting on the Symposium, and I wish to thank all of them for their assistance.

Though the Interface Symposium is not sponsored by any separate organization, a number of the professional societies cooperate with the Symposium. I would like to thank the American Statistical Association, particularly its Section on Statistical Computing and its Houston and Southeast Texas Chapters, the Association for Computing Machinery, and the International Association for Statistical Computing for their cooperation.

Financial support for the Symposium came from the Office of Naval Research, the National Science Foundation, and IMSL, Inc. To all of these I extend my thanks.

James E. Gentle
Houston

Financial Supporters of the Fifteenth Interface Symposium

U. S. Office of Naval Research, U. S. Dept. of the Navy

National Science Foundation

IMSL, Inc.

Cooperating Organizations

American Statistical Association
- Statistical Computing Section
- Southeast Texas Chapter
- Houston Chapter

Association for Computing Machinery

International Association for Statistical Computing

AVAILABILITY OF PROCEEDINGS

13, 14th (1981,82)	Springer-Verlag New York, Inc. 175 Fifth Avenue New York, NY 10010
12th (1979)	Jane F. Gentlemen Dept. of Statistics University of Waterloo Waterloo, Ontario Canada N2L 3G1
11th (1978)	Institute of Statistics North Carolina State Univ. P. O. Box 5457 Raleigh, North Carolina 27650
10th (1977)	David Hogben Statistical Engineering Laboratory Applied Mathematics Division National Bureau of Standards U.S. Dept. of Commerce Washington, D.C. 20234
9th (1976)	Prindle, Weber, and Schmidt, Inc. 20 Newbury St. Boston, Massachusetts 02116
8th (1975)	Health Sciences Computing Facility, AV-111 Center for Health Sciences Univ. of California Los Angeles, California 90024
7th (1974)	Statistical Numerical Analysis and Data Processing Section 117 Snedecor Hall Iowa State Univ. Ames, Iowa 50010
4,5,6th (1971,1972, 1973)	Western Periodicals Company 13000 Raymer Street North Hollywood, California 91605

The Sixteenth Symposium on the Interface will be held March 14-16, 1984, in Atlanta, Georgia. The chair will be Lynne Billard of the University of Georgia.

Contents

KEYNOTE ADDRESS

The Future

R. W. Hamming

COMPUTER SCIENCE AND STATISTICS:
The Interface, James E. Gentle (ed.)

THE FUTURE

R. W. Hamming

Naval Postgraduate School
Monterey, California

This paper predicts that: the past exponential growth in people will not continue, but the publication rate will; computers will have significant effects on the presentation of results; and the field is being saturated by increasing detail. is becoming increasingly personal instead of objective.
Statistics is changing. From compression of data we are going to expansion; from objective results we are going to personal; from a broad concensus of common theories we are going to elaborate specialized theories; and finally the reputation of statistics will continue to fall. You should consider how to alter these predictions by your future behavior.

The purpose of this talk is to persuade you to think about, and plan for, the future, both your own and that of statistics. The future has a habit of coming. It seems to me that because most people do very little long range planning they are forced to accept the future that comes rather than being prepared for, and to some extent making, the future they want.

Such planning implies that: (1) you have some idea of what the future is likely to be; (2) you have some scenarios of the future; (3) you have some desires and goals for the future; and (4) you are willing to plan and work to see that some of these desires are realized.

These are difficult to do. From long observations I believe I am right when I say that most people do not make such extensive plans for the future; they more or less let the the future happen. Among statisticians I do not need to refute the common claim that since the future is uncertain, you therefore cannot be sure that your plans will be realized - hence the futility of planning, so they claim.

Let us look at the future. We have all heard that 90% of the scientists who ever lived are now alive. We know that this is the basis of growth since about the time of Newton (say 1660). We have been doubling knowledge about every 17 years since then. A lot of other observations, from the classical doubling period of library holdings to the growth of the number of employees in Bell Labs, support this doubling period.

Can this growth go on indefinitely? The popular expression, "Science, the endless frontier", suggests that it can, but if it did then in 340 years there would be 20 doublings - that is a million fold growth!

How have we coped with the growth since Newton's time? Essentially by specialization. In my lifetime I have seen an enormous increase in specialization. If you accept the million-fold growth of knowledge, then you must expect something like a million fields of specialization for every field we now have. I doubt you can believe that. Hence, although the past trends in information growth and specialization indicate one thing, common sense indicates another, slower, growth of knowledge in the future.

But there is a closer deterrent. At present there are not the students in school to carry on. Thus the doubling of the number of people must slaken off fairly soon.

Let us next look at publications. So long as we have the pattern of "publish or perish" we will have the glut of publication. The situation has positive feedback! I have noted with interest that top management in the schools, because they are less and less able to judge the quality of a candidate for promotion, are insisting more and more on volume of publication. They are measuring quantity, not quality. But is there any other solution for them? I cannot see it coming soon. It is less apparent in industry, but as specialization increases they will have, in their turn, to depend on outside evidence, and this tends to be quantity not quality. We ourselves are guilty! We will seldom express our opinions clearly about the trash that appears monthly lest we in our turn be critized; we only complain. Since we will not, then I say management cannot do other than increasingly use quantity in place of quality.

As a curious aside, I note that JASA appears to reject every first submission of a paper. Apparently no one, including the reviewers themselves, can write a proper paper the first time. I look at the delays indicated, and wonder if they are worth the

"improvements" that come from the recycling. A statistical study of the dates of first submission and final acceptance would reveal a number of startling things, and raise serious questions about editorial policy.

The tricks for building up a big bibliography are well known. You take an idea, publish it in three papers rather than digest it first (risking prior publication by someone else) and have one good paper. You need energy to resubmit the paper after the first refereeing, since apparently (see above JASA remarks) no reviewer feels he is doing his job unless he asks for some alterations. Hence you must plan to send it back promptly with the changes asked, (even if in your eyes and others the changes are bad); you wear the reviewer down. It is a matter of having more energy than he has! It is also well known that you publish the same thing under different appearing titles in different places - the excuse is that it is of wide importance and needs wide circulation! I think that almost anyone can build up a bibliography in these days of specialization by becoming the recognized expert in their pinheaded field of specialization.

With the falling off of the number of scientists, can the growth of publication continue? Yes! The computer is the solution. Merely using a word processor will greatly ease the burden of rewriting the paper to satisfy the reviewer. It is also a great aid in preparing alterate versions for publication in various journals.

Many of you think that the computer will save us from being swamped by publications. You think that, by suitable cross referencing the papers, the computer can find the relevant literature. I doubt that it will do a lot, although there are some possibilities in this direction. The real truth of the matter is that a major problem in publication is to find a problem that has not yet been solved! That is often more work than solving the problem when you have found it.

Some of you may think that the computer can be programmed to search the product space of subclassification labels to find an empty hole that has surrounding papers already published. Then the computer merely names the missing entry, prints out the relevant papers along the lines of cross references, and you have on your desk all the material ready to whip out the missing theorems and proofs!

It is too glib for me to believe, but then I am not an ardent fan of artificial intelligence. If you believe that it can be done then maybe you can get the machine to write much of the paper too. In principle the machine has already delivered all the relevant papers needed to find the proof. But if you believe that you can write a program that can separate important information from junk, then I suggest you connect it to a source of random noise, and let it find important information. After all, a random source will, sooner or later, deliver every finite piece of information, and you claim your program can filter out the unimportant results! If you doubt the practicality of this proposal, then you need to back down and begin to consider, as I have, the impracticality of the earlier wishful thinking that computers will make significant improvements in our ability to do important research - note the qualifying words!

Is there no limit in the growth of knowledge and the concomitant specialization? As I earlier said, I do not believe in the "Science - the endless frontier". I believe, instead, in diminishing returns. When, finally, you have a massive three volume work on Maximum Likelihood Estimators, written by a select committee of experts, will you want to continue in the field? I suggest that most humans will find further effort repulsive! And when you have a five volume work, each 1000 pages or more, on Maximum Entropy, will that not be enough? How far do you believe people will be willing to push this endless frontier? Not how far can it be pushed!

I have already indicated a limit on the growth of activity in science because of the lack of students now in school to carry on. I have just indicated that I believe that the economic theory of diminishing returns also applies to research - the same amount of effort no longer produces in your own eyes the same value. There is a third limiting factor. It is the sense of creativity. Creativity is not specialized as a talent. One has a general urge to create. And if the apprenticeship in one field requires many long years before you can do real creations of importance in your own eyes, then you are likely to turn to other fields where the frontier is closer and the hope of creative activity is much higher. The creative person tends to go where there is a chance to be creative soon - not after 20 years of post-graduate study to find out what is known and only then begin to create.

I believe that this stage of diminishing returns is now here in statistics. The period of most activity often occurs after the period of diminishing returns starts. The inertia, the "publish or perish" syndrome, and the working habits of the people already in the field, all tend to maintain the frantic explosion of papers. How many of you believe that great, important progress is now going on in Statistics? I doubt that many of you do. What is going on is more filling in of missing pieces, finding variations similar to earlier parts, rewriting of things more

abstractly that in fact adds little new knowledge. So far as I can see, the important papers, per person working and publishing, seems to be way down from what it was in earlier years, and I predict a further falling off, percentagewise.

You should be wary of doing things because they can be done. Without a sound basis, in mathematics and probability, a sound superstructure is very hard to build. There seems currently to be a growth in "personal statistics". Not that personal statistics may not get better results (without defining what I mean) than can objective statistics, but it is hardly a way of convincing others to commit their actions to your personal vision. For better or worse, the results you find must be communicable to others and win their assent before they can be translated into large scale actions. By "personal statistics" I mean the following: give the same data to 10 competent statisticians and you could get up to 20 different answers! The essence of science used to be reproducibility.

No doubt the connection between statistics and computing will become closer, but the theory of computer science seems to me to be going in directions that have little interest to statisticians at present. Of course some of the fallout from computer science may benefit you, but don't expect too much. Like statistics, computer science has become rather inwardly directed, as judged by the publication (or perish!) record.

Back to my opening remarks. What will happen, in the long run, to statistics and yourself? As to statistics I hope that there will be some serious examination of the foundations, both theoretical, which is happening, and experimentally which is not. I have been preaching for many years that 90% of the time the next independent measurement will fall outside of the previously announced 90% error bounds. I came to the idea from Youden's paper Enduring Values, and have seen mountains of evidence since then which support the claim - approximately. So I wonder why all the fancy statistics is studied and taught when the simplest parts fly in the face of reality. If, statistically speaking, statistics gives wildly wrong predictions, then what is wrong, and why will statisticians not look? The usual expert syndrome, I suspect.

I have meditated on this 90% effect for many years, and have come up with one small factor to partially explain it. When the experimentalist finally gets the equipment working, then the final stages of adjustment of the equipment begins. What criterion is used? Reproducibility, of course! Run-to-run should give consistent answers. But what is reproducibility except low variance? The equipment is adjusted to get low variance, and it is exactly this data that the statistician, sitting in his ivory tower, uses to estimate the probable error of the data.

Now we come down to predicting your future in statistics. I cannot say what your desires should be, nor what you should want, but I can say that if you look carefully and wonder about where you will be, and what statistics will probably be at the time of your mid-and end-career, then you are in a position to plan to have happen what you wish would happen.

I have omitted the topic of dramatic new discoveries that completely change the field. I have given some thought about their possibilities, but the results are so meager that I am ashamed to tell you. Hence, I will take cover under the standard remark, "barring sudden and unexpected developments in the world."

I have talked about only one aspect of the effects of computers on statistics. We need to examine the topic in more detail in our projections of the future. Perhaps the most dramatic advance will be in the general area of display in three or more colors. It is bound to be widely available and cheap in the near future. The ability to look at a set of data, displayed in many ways, and from many angles, with a rich battery of transformations to apply on demand, will encourage the art of prospecting in the raw data for the nuggets of value. Furthermore, the cheapness of computing will encourage many different summarizations of the data - until the summarizations exceed the original data in volume. We are entering the era of the explosion of the data. And the automation of the collection of data will exascerbate the problem many-fold. Unless we demand a mathematical basis for the processing we will face a deluge of personal data summarizations and presentations with no "degrees of freedom" decrease to compensate for the selection mechanism. I still remember the internal memorandum at Bell Telephone Labs that had less than 5 pages of text and over 100 pages of the output graphs. There is almost no upper limit on the ability to take a small amount of data and blow it up to an overwhelming amount.

In summary, I see changes I do not like. I was raised on the definition that statistics was the boiling down of the mass of data; I predict the regular creation of a mass of output from a small amount of input data. I used to believe that the purpose of statistics was to gain insight; tomorrow I see a volume of output that by its very mass can only prevent insight. I used to think statistics strove for objectivity; it seems

headed toward personal data exploration. I see the continuation of mathematical statistics (how else do you get publications and theses?) to the neglect of the obvious current defects in prediction. I see the elaboration of every detail without end until nausea sets in; I do not see the statisticians daring to express value judgements on what is worthwhile. I see the reputation of statisticians matching that, in the public eye, of lawyers - people totally unconcerned with truth, justice, etc., but instead devoted to winning the case. In the future it will still be said with justice, "There are liars, damned liars, and statistics." Presumably I am an old fogey. If you do not want this future, now is the time for you to do something about it!