The Technical Tools of Statistics

John W. Tukey

November, 1964

ABSTRACT

The paper looks at the technical tools of statistics, past, present, and near future. Realistic use of numbers has long been an important instance and has forced statisticians to combine security and insecurity. The ASA is 10/8 century old; about 5/8 ago algebra began to appear in statistician's Journals. Only 2/8 ago came the programmable calculator. Today "software", connecting "hardware" to both data and "brainware", offers statistics far more than most statisticians realize,. They offer us factories, not just machines, and ought to be used flexibly, fluidly, fully and by the side foot (as measured up the side of the stack of printout).

Since, as Chamberlain said, science is the entertaining of hypotheses, data analysis will continue to depend on alternative approaches and (at least until "real time interaction" is real) volume output. Flexibility to interconnect techniques in any way is important, as are such fluid techniques as those which select a monotone transformation from all possible such to match the given data. Graphical output can be of the greatest importance; as yet no one is using it adequately.

In statistical theory, the powers of the computer in classical algebra, combinatorics, experimental sampling, and now algebra are great, and almost untapped.

Adequate languages (or systems, etc.) for data analysis are not yet here, though badly needed. Steps so far taken, though very helpful, are inadequate.

Since the computer is a sharp enough tool to be really useful, you can cut yourself on it.

Most of the technical tools of the future statistician will bear the stamp of the computer manufacture, and will be used in a computer. This will not alter the essential character of these tools, nor will it allow us to escape from (i) using the approximately right, rather than the exactly wrong, ii) making much more frequent use of umbra-penumbra models, (iii) being as happy starting with an estimator as starting with an estimand.

This paper was read at the 125th Anniversary Meeting of the American Statistical Association, Boston, November 1964, and will appear in the April

1965 American Statistician).

MEMORANDUM FOR FILE

We are gathered here to look both forward and back. What have our technical tools been? What are they today? What can we see of what they are to become?

The assessment of the future is always chancy. Who knows this better than a statistician? Yet experience has taught us that it is usually well to extrapolate so long as we go only a modest distance and do what we can to ensure adequate caution.

I am convinced, for reasons that will soon appear, that now is a relatively easy time to forecast the near future of the statistician's technical tools, far easier than would have been the case at our 100th anniversary. Accordingly, I shall focus on the future, saying little about the past, and less about the present.

What have been the technical tools of the statistician through the years, both before and after the founding of our association? The answer to this might logically depend upon what is thought to be statistics. The definitions of statistics are many -- as our recently departed colleague, Walter Willcox, who was only 23 years younger than our association, pointed out nearly 30 years ago. [1]

From the days when "statenkunde" meant the art of summing up in numbers the military strength of a neighboring count, prince, or baron to our own day -- and far into the future -- the two most important technical tools of the statistician remain the same -- and both are attitudes:

1. a willingness to express "it" in numbers and to manipulate the results,
2. a willingness to think hard and realistically, to be "hard-nosed" about the world, the data, and their relationships.

As a result, all statisticians share, to a greater or less degree, a severe stress of a kind that could produce split personalities. They must work with numbers, and usually with other symbols, that are to be manipulated according to clearly established rules; they must be happy with the sorts of (mathematical) manipulations that are probably the most secure things in human life, and are, consequently, also the most inflexible and rule-bound. At the other extreme, they must, at almost the same time, be honest in assessing the uncertainties of their final results. In the latter they cannot be satisfied with allowance for only the likely size of "sampling errors", a task with which routine manipulations can often help them; they must, most particularly and responsibly, make explicit allowance for the likely size of "nonsampling errors", for the extent to which the data given to them was neither what it purported to be nor what it ought to have been. No other profession must support itself over so wide a span from security to insecurity.

Yet if we were to imitate the name of one of our neighbor societies, and call ourselves a "Society of Insecurity Analysts", we should be doing ourselves an injustice. While we must be constantly alert to insecurity, we have an equal responsibility to cut into complex situations and reveal what appears to be true, whether or not we can judge the insecurity of the appearances we there discover.

Clearly I have chosen to talk about the analysis of data, rather than about the design of ways of obtaining data. To do this without due notice would be to slight a large and important area that needs its own treatment.

In this area the key is a combination of experimentation or observation, according to planned patterns that involve replication either open or hidden, with an analysis which takes adequate account of the patterns used, both in assessing typical values -- whether stratum means or main effects and interactions or what have you -- and in assessing the stability of these typical values. The purely observational study, such as the sample survey, has gained steadily in power, incisiveness, and ease of mounting. Such possibilities as studying public reactions to events as they occur are now demonstrated facts [2]. The experimental study patterned for statistical analysis has spread from its early home in agriculture to a very wide variety of applications. The use of these techniques as part of, and as a contributor to the control of, otherwise ongoing processes as in evolutionary operation and certain forms of adaptive control - is just beginning to grow rapidly. To do justice to this area would, require at least a whole further talk.

Similarly, I have not chosen to speak at length about the analysis of data as a part of a larger field of numerical science. We will have things to say about modern computation, but, to name one example, linear programming and its generalizations will not be discussed at all. Discretion must be the better part of valor, so far as our scope is concerned.

Our association has been in existence for ten-eighths of a century. Since we still learn to count on our fingers, by tens, it seems fitting, at this meeting, to count time in eighths -- eighths of a century.

In asking about the growth of our technical tools, then, it is natural to begin by looking at volumes of suitable journals spread across all ten eighths. The Journal of the American Statistical Association and its predecessors cover this range with gaps. The Journal of the Royal Statistical Society covers it all. What do we see when we turn to the past?

Just about five-eighths (of a century) ago there was a sharp discontinuity: algebra began to appear on the pages of the statistician's, journals. The pioneering work of Karl Pearson spread out in many directions; that of Student spread out in many others, including the analysis of economic time series. Since that time the tools of the statistician have been sharpened on the grindstone of algebra and hammered out on the anvil of mathematical models. We learned rapidly that, to a far greater extent than we had dared to think, one could use data to deal both with the uncertainties of nature and with the uncertainties of data-based results.

By three-eighths ago the impact of R. A. Fisher had begun to be widely felt. There were at least three major components to his influence. In two of these -- a much deeper involvement of mathematics, and an emphasis, as a standard to which all men should be obliged to repair, upon optimality under tight specifications -- his influence was joined, successively, by those of Jerzy Neyman and Abraham Wald. The third was far less widely noticeable but equally important and perhaps even longer lasting. This was Fisher's contribution to the rise of new ways of dissecting data to reveal appearances that would otherwise never have been seen. Some may argue that Gauss, Lexis, and von Bortkiewicz foreshadowed the decompositions, and possibly even the interpretations, of the simplest.analyses of variance. But it is to Fisher that we owe the notion of interaction as something that has numerical existence and can be calculated about freely -- and it is to the hand calculator and the patient computress that we owe almost four decades of experience with values of sums of squares of interactions of various, kinds. In rendering honor to Fisher here, however, we must accept responsibility for the things that we ourselves have not done: How often do we look at the interactions themselves, and not merely at the sum of their squares? How much has each of us done to invent or introduce still further new ways of dissecting data?

By two-eighths ago one could see the shadow of the programmable calculator on the outside of the window. Tabulators had plug-boards and there were books on the "Use of the Punched-Card Method in Colleges and Universities."

These machines speeded up computation, but they did not necessarily make it easier, as is still clear- to all those who learned, finally, to make a 601 multiplying punch do some of those things that its instruction manual said it could not be made do -- such as multiplying three numbers together, rather than only two.

There was a war, and a von Neumann; it took only half an eighth to bring the program-self-modifying calculator to reality, and start the rapidly-evolving generations of technical factories that are already at our side. The modern, large-memory, program-self-modifying calculator is far more than a tool, or a kit of tools, or even a room full of kits of tools. It helps to make tools as well as to use them; it is in truth, a factory.

Hardware alone does not, make a factory -- and hardware alone does not make a computer. The electronics and the magnetics, the tapes and the tape drives, the card readers and the TWX lines, the high-speed printers, card punches, and microfilm devices, all these are far from enough. The hardware has to be connected: to the data, to the printers, and to the minds of those who know what is wanted, minds which ought, perhaps, to be affectionately called "the brainware."

This connection is the function of the software, of the systems programs that are all that allow a reasonably small number of men to make good use of one large computer. These systems programs may be programs that help write individual programs -- when they are often called compilers. They may be programs that control the combined operations of central processors, memories, peripheral equipment, and input-output devices -- when they are often called monitors or executive routines. They may have other functions. But collectively, as the connection between brainware and hardware, they are already more important than the hardware, and are already evolving far more rapidly.

Existence, importance, and rapid evolution are easily overlooked, because the software resides in the computer's memory and seems to most users like a part of the machine -- as in truth, by a wise definition of "machine", it is.

Today, software and hardware together provide far more powerful factories than most Statisticians realize, factories that many of today's most able young people find exciting and worth learning about on their own. Their interest can help us greatly, if statistics starts to make much more nearly adequate use of the computer. However, if we fail to expand our uses, their interest in computers can cost us many of our best recruits, and set us back many years.

We statisticians have not yet begun to use these factories as we should: flexibly, fluidly, fully and by the foot. Each new generation of computers offers us new possibilities, at a time when we are far from using most of the possibilities offered by those already obsolete. Many things that were quite possible an eighth ago are still not common, others that could have been routine a half an eighth ago are not yet started.

When I said that we should be using computers by the foot, I did not mean by the running foot of output paper. I meant, rather, by the 12 inches of thickness of stacks of printout -- by the unit that computer centers that cooperate with statisticians are coming to call the side foot. To call for bulk printout is to deny one of the prime principles of those computer center managers who deal only with classical problems of numerical analysis. For them, the appearance of lots of output means that the programmer did not know what he was doing, did not know what was going on in his program, and that he got lots of output so that he could see where he went wrong.

When such a manager sees even a few side inches of output going to a data-analyzing statistician, he has the same suspicion. There is a germ of truth in this suspicion, but his hasty conclusions are usually wrong. What the statistician often does not know -- is what is going on in the real world situation that generated the data. What the statistician often needs -- is enough different "looks" to have a good chance to learn about this real world. Often he has both a right to, and a need for, the results of many parallel analyses involving a variety of alternative approaches. It will often be sound economy to run each of many analyses through all their many steps instead of planning to come back, and back, and back for more.

It is well for all of us, as statisticians, to recall how Professor Chamberlain, a great geologist, defined science more than half a century ago. To him, science was the entertaining of multiple working hypotheses. If we are to be scientific in data analysis, it is just this that we must do.

There are two changes in the mechanics of coupling brainware to the computer and back again which may alter the equation

data analysis = volume output

to an uncertain extent. When the remote console is a commonplace, and the supersize fast memory allows supermultiprogramming, the statistician may begin to find it easy enough to get individual questions answered immediately to ask them one at a time. Conversely, when microcard output, as we all should hope, arrives to supplement microfilm output (now often used to make enlargements), and is more and more used for massive and tabular results -- when one can put the equivalent of today's side foot of printout in a coat pocket -- bulk of output will be only an information retrieval problem, not a matter of three porters to move paper for one 7094.

The impact of such changes are hard to assess. What will the over-all economics of data analysis be in an era of many remote consoles? Won't it always be easier to carry the, answers with you from room to room, on the subway, or on the turnpike? Whether or not the side foot turns into the Pocket inch, the usefulness of alternative, diversified detailed output <u>in being</u> seems likely to remain for a long time. Volume of output, much of it obtained on a contingency basis, different parts of it never examined in different instances, seems likely to remain a hallmark of the analyst of data, the distinctive sign of exploration and inference.

Not only by the foot, but flexibly -- where flexibility means just what it seems to say: freedom to run down long lists of alternative approaches and choose any sublist for application to each specific set of data; freedom to stack up techniques in any order -- as in Cuthbert Daniel's use of fractional factorial designs in reducing the computation needed to understand a messy multiple regression problem; freedom to introduce new approaches; freedom, in a word, to be a journeyman carpenter of data-analytical tools.

Not only flexibly but fluidly -- where fluidity means that we are prepared to use structures of analysis that can flow rather freely, superparametrically rather than nonparametrically, to fit the apparent desires of the data. All today's examples of fluidity depend on a key fact: a computer can be told to approach a desired result by successive steps and to stop when nearly enough there. It can do this effectively, even in cases where it would be impossible to give a fixed set of arithmetic steps that would lead to the goal. Without iterative calculation we could not follow such paths as that, blazed by Roger Shepard [3,4] and Joseph Kruskal [5,6], of seeking out that particular increasing function of a raw response that leads to the seemingly most insightful analysis of each particular set of data. Such a fluid approach well deserves the honorable title of <u>superparametric</u>, since it involves an unlimited number of parameters. Indeed this approach shows its fluidity in more ways than one, for when one watches a movie of the progress of such an iteration something the computer and the microfilm plotter can easily combine to produce -- one truly sees the points flowing back and forth.

Not only fluidly, but fully -- where full use of our computer has three more essentials: adequate use of its capabilities for graphical output, intensive use of its computing power in the development of new statistical methodology, and learning how to fit it out with a programming system adequate for -- and suited to -- the needs of data analysis.

Computer-prepared "pictures" of data have been accessible for at least two eighths. If a classical punched-card tabulator had a digit selector, as many did, it could be used to produce 20-by-M scatter diagrams or contingency tables, either as conventional printout or directly on ditto masters ready for reproduction. How many of us did this?

Today microfilm output is being put on more and more large computers. It offers an extremely wide variety of displays, and provides them rapidly, cheaply, and effectively. Hand-drawing of graphs, except perhaps for reproduction in books and in some journals, is now economically wasteful, slow, and on the way out. Computer-drawn graphs are cheaper, arrive quicker, and can be afforded in great numbers. In exploration they are going to be the data analyst's greatest single resource. In presentation, and in education, they are going to play a new role. Here are several new kits of tools. How many of us are using them? Which equipment manufacturers are actively educating their customers?

The figure shows you something of what is now easily possible. (This is a very simple example of what can be done.) Except for the obviously printed caption and for photographic enlargement, everything there is just as it came from the 402O microfilm printer. Points, lines, curves, grids, scales, and legends were all produced by the 4020 according to instructions produced by the 7094. In Wilk and Gnanadesikan's Annals paper, their figure was redrawn for aesthetic advantages, but one can easily see everything on the microfilm output that one can see in the redrawn figure. We should look to the computer for our graphs, and do this more often.

At what may seem today to be an extreme-of graphics, the making of animated movies on technical subjects, computer production is very substantially cheaper than hand methods (once an appropriate "language" is available for programming) and many mathematical and statistical topics can be approached with surprising ease.

As yet I know of no person or group that is taking nearly adequate,advantage of the graphical potentialities of the computer.

Next, the uses of the computer in statistical theory. Much could be said about them; we shall take time here for only the main points:

1. Since, for example, almost anything that can be reduced to a single integration is trivial on a modern computer, and ordinary differential equations are little more trouble than explicit quadratures, the making of many kinds of tables is no longer difficult: the question becomes "Do I want a table, or do I want a short computer program to give me the particular value that I want when I want it?", to which the answer is likely to be: "Both".

2. Quite complex combinatorial problems can be handled in reasonable time on the computer, accordingly there are new opportunities for the study of techniques based upon order relations among the observations.

3. Experimental sampling runs smoothly and rapidly on Modern computers, wherever it falls from the most naive sampling (which first gave us Student's t) to the most subtle forms of Monte Carlo (which can often gain us speed by large factors -- sometimes in the millions -- over naive experimental sampling).

4. Procedures for efficiently doing algebra (on polynomials in several variables and on ratios of same) are already in use [8,9,10] and will soon be widely available. This is going to revolutionize many aspects of computation. If a power series, for example, has machine-producible coefficients, especially if these are rational numbers, the use of 50 terms --or even 500 -- can be easy, cheap, and effective. And we can calculate algebraic answers to problems far beyond our pencil-and-paper reach.

A very large share of the advances of statistical theory during the next eighth are going to depend upon the computer in an essential way.

Long before the end of that eighth, we are going to reach a position we should have reached long ago. We are going, if I have to build it myself, to have a programming system -- a "language" if you like -- with all that that implies, suited to the needs of data analysis. This will be planned to handle numbers in organized patterns of very different shapes, to apply a wide variety of data-analytical operations to make new patterns from old, to carry out the oddest sequences of apparently unrelated operations, to provide a wide variety of outputs, to automatically store all time-expensive intermediate results "on disk" until the user decides whether or not he will want to do something else with them, and to do all this and much more easily.

We have already begun to move much further toward such a situation than most of us realize. Here and there, groups are pulling together sets of programs for data analysis, sometimes for geophysical time series [11], sometimes for psychological-psychometric problems [12], sometimes for rather general data analysis [13]. So far, these "systems" are valiant -- and extremely useful -- efforts to do what can be done within patterns established with a view to quite other uses.

Some of my friends felt that I should be very explicit in warning you of how much time and money can be wasted on computing, how much clarity and insight can be lost in great stacks of computer output. In fact, I ask you to remember only two points:

1. The tool that is so dull that you cannot cut yourself on it is not likely to be sharp enough to be either useful or helpful.

2. Most uses of the classical tools of statistics have been, are, and will be, made by those who know not what they do.

Most of the technical tools of the future statistician will bear the stamp of computer manufacture, and will be used in a computer. We will be remiss in our duty to our students if we do not see that they learn to use the computer more easily, flexibly, and thoroughly than we liver have; we will be remiss in our duties to ourselves if we do not try to improve and broaden our own uses.

This does not mean that we shall have to continue to teach our students the elements of computer programming; most of the class of '70 is going to learn that as freshmen or sophomores. Nor does it mean that each student will write his own program for analysis of variance or for seasonal adjustment, this would be a waste. It ought to mean far less time using hand calculators, over one of which R. A. Fisher once said [14] "he had learned all he knew". It must mean learning to put together, effectively and easily -- on a program-self-modifying computer and by means of the most helpful software then available -- data analytical steps appropriate to the need, whether this is to uncover an anticipated specific appearance or to explore some broad area for unanticipated, illuminating appearances, or, as is more likely, to do both.

As the computer revolution finally penetrates into the technical tools of statistics, it will not change the essential characteristics of these tools, no matter how much it changes their appearance, scope, appositeness and economy. We can only look for:

1. more of the essential erector-set character of data analysis techniques, in which a kit of pieces are available for assembly into any of a multitude of analytical schemes,

2. an increasing swing toward a greater emphasis on graphicality and informality of inference,

3. a greater and greater role for, graphical techniques as aids to exploration and incisiveness,

4. steadily increasing emphasis on flexibility and on fluidity,

5. wider and deeper use of empirical inquiry, of actual trials on potentially interesting data, as a way to discover new analytic techniques,

6. greater emphasis on parsimony of representation and inquiry, on the focussing, in each individual analysis, of most of our attention on relatively specific questions, usually in combination with a broader spreading of the remainder of our attention to the exploration of more diverse possibilities.

In order that our tools, and their uses, develop effectively:

- we shall have to give still more attention to doing the approximately right, rather than the exactly wrong, in particular by giving up, as far as we are able, both tight specifications and reliance on likelihood functions and fiducial arguments, at least as we now know them.

- we shall have to make very frequent use of umbrapenumbra models where, for instance, we require validity over a range of circumstances far wider than the one over which we seek optimality.

- we shall have to be as happy starting from an estimator and inquiring what it estimates, as we are starting from something to be estimated, and inquiring what reasonably estimates it.

My extrapolation of our near future is before you, at least in rough outline. If statistics evolves as rapidly over the next few eighths of a century as it has over the last ten, the technical tools of our middle future will go far beyond all of today's dreams.

ACKNOWLEDGMENTS

This manuscript was prepared in part in connection with research at Princeton University sponsored by the Army Research 0ffice (Durham). Reproduction for any purpose of the U. S. Government encouraged. Thanks are due to: Christopher Bingham, Keith Hastings, Oscar Kempthorne, William Kruskal, Richard Link, Frederick Mosteller, and Milton Terry, for valuable comments on various drafts of this talk. They bear no responsibility for the wording finally adopted.

REFERENCES

1. Walter F. Willcox 1935. Definitions of Statistics. Revue de l'Inst. Int. de Statist., Z, 388-399. (For earlier history note also V. John 1883. The term "STATISTICS", J. Statist. Soc. (London), 46, 656-679, (translated from a German separate) and G. Udny Yule 1905, The'intrQduction of the words "statistics". ft statistical" into the English Language. J. Roy. Statist. Soc., 88, 391-396. I owe these references to William Kruskal.) (Who knows the references to the paper that raised Willcox's 116 definitions to 531?)

2. Paul B. Sheattley and Jacob J. Feldman 1964. The assassination of President Kennedy: A preliminary report on public reactions and behavior. Public Opinibon Quarterly, 28, 189-215.

3. R. N. Shepard 1962. -The analysis of proximities: Multidimensional scaling with an unknown distance function. I. Psychometrika, ZZ, 125-140. II. Psychometrika, ZZ, 219-246.

4. R. N. Shepard 1963. Analysis of proximities as a technique for the study of information processing in man. Human Factors, 5, 33-48.

5. J. B. Kruskal 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika, 29, 1-27.

6. J. B. Kruskal 1964. Nonmetric multidimensional scaling: A numerical method. Psychometrika, 29, 115-129.

7. M. B. Wilk and R. Gnanadesik-an 1964. Graphical methods for internal comparisons in multiresponse experiments. Annals of Math. Statist., 35, 613-631.

8. W. S. Brown 1963. The ALPAK System for Nonnumerical Algebra on a Digital Computer I: Polynomials in Several Variables and Truncated Power Series with Polynomial Coefficients. Bell System Technical Journal, 42, 2081-2119.

9. W. S. Brown, J. P. Hyde and B. A. Tague 1964. The ALPAK System for Nonnumerical.Algebralon a Digital Computer II,; Rational Functions of Several Variables and Truncated Power Series with Rational-Function Coefficients. Bell System Technical Journal, 43, 785-804.

10. J. P. Hyde 1964. The ALPAK System for Nonnumerical Algebra on-a Digital Computer III: Systems of Linear Equations and a Class of Side Relations. T3ell System Technical Journal. 43 1547-1562.

11. E. C. Bullard, F. Oglebay, W. H. Munk, and G. Miller 1964. A User's Guide to BOMM, A system of programs for the analysis of time series. Institute of C-eophysics and Planetary Physics, La Jolla (University of California at San Diego) and Cambridge, England. Reproduced, lll pp. (Unpublished).

12. Roald Buhler 1964. P-STAT; a system of statistical programs for the-7090/7094. Princeton University Computer Center (Programming Notes No. 12).

13. W. J. Dixon (Editor) 1964. BMD Biomedical ComputerPrograms, 585 pp. Available from UCLA Student Store, 308 Westwood-Boulevard, Los Angeles 24, California.

14. Besse Day Maiiss 1964. Personal communication.

Note: Other "systems" approaches to data analysis are discussed or referred to in

- Vladimir V. Almendinger 1963. SPAN Reference Manua I. Technical Memorandum TM1563/000/00 System Development Corporation.

- Rolf E. Bargmann and Michael W. Browne lq63. Generation and Analysis of Data for Multivariate Analysis of Covariance. Thomas J. Watson

Research Center (IBA).

- William W. Cooley and Paul R. Lohnes lq62. Multivariate Procedures for the Behavioral Sciences. Wiley 1962.
- Brian E. Cooper 1964. Designing the data presentation of statistical programs for the experimentalist. Bull Inst. Statist. 40, 567-585.
- Kenneth J. Jones 1964. The Multivariate Statistical Analyzer'. Harvard Computing Center.