THE FUTURE OF DATA ANALYSIS[1]

By John W. Tukey

Princeton University and Bell Telephone Laboratories

Received July 1, 1961.

[1] Prepared in part in connection with research sponsored by the Army Research Office through Contract DA36-034-ORD-2297 with Princeton University. Reproduction in whole or part is permitted for any purpose of the United States Government.

I. GENERAL CONSIDERATIONS

1. Introduction. For a long time I have thought I was a statistician, interested in inferences from the particular to the general. But as I have watched mathematical statistics evolve, I have had cause to wonder and to doubt. And when I have pondered about why such techniques as the spectrum analysis of time series have proved so useful, it has become clear that their "dealing with fluctuations" aspects are, in many circumstances, of lesser importance than the aspects that would already have been required to deal effectively with the simpler case of very extensive data, where fluctuations would no longer be a problem. All in all, I have come to feel that my central interest is in *data analysis*, which I take to include, among other things: procedures for analyzing data, techniques for interpreting the results of such procedures, ways of planning the gathering of data to make its analysis easier, more precise or more accurate, and all the machinery and results of (mathematical) statistics which apply to analyzing data.

Large parts of data analysis are inferential in the sample-to-population sense, but these are only parts, not the whole. Large parts of data analysis are incisive, laying bare indications which we could not perceive by simple and direct examination of the raw data, but these too are only parts, not the whole. Some parts of data analysis, as the term is here stretched beyond its philology, are allocation, in the sense that they guide us in the distribution of effort and other valuable considerations in observation, experimentation, or analysis. Data analysis is a larger and more varied field than inférence, or incisive procedures, or allocation.

Statistics has contributed much to data analysis. In the future it can, and in my view should, contribute much more. For such contributions to exist, and be valuable, it is not necessary that they be direct. They need not provide new techniques, or better tables for old techniques, in order to influence the practice of data analysis. Consider three *examples*:

(1) The work of Mann and Wald (1942) on the asymptotic power of chi-

square goodness-of-fit tests has influenced practice, even though the results they obtained were for impractically large samples.

(2) The development, under Wald's leadership, of a general theory of decision functions has influenced, and will continue to influence data analysis in many ways. (Little of this influence, in my judgment, will come from the use of specific decision procedures; much will come from the forced recognition of the necessity of considering "complete classes" of procedures among which selection must be made by judgment, *perhaps* codified á la Bayes.)

(3) The development of a more effective procedure for determining properties of samples from non-normal distributions by experimental sampling is likely, if the procedure be used wisely and widely, to contribute much to the practice of data analysis.

To the extent that pieces of *mathematical statistics* fail to contribute, or are not intended to contribute, even by a long and tortuous chain, to the practice of data analysis, they must be judged as pieces of *pure* mathematics, and criticized according to its purest standards. Individual parts of mathematical statistics must look for their justification toward either data analysis or pure mathematics. Work which obeys neither master, and there are those who deny the rule of both for their own work, cannot fail to be transient, to be doomed to sink out of sight. And we must be careful that, in its sinking, it does not take with it work of continuing value. Most present techniques of data analysis, statistical or not, have a respectable antiquity. Least squares goes back more than a century and a half (e.g., Gauss, 1803). The comparison of a sum of squares with the value otherwise anticipated goes back more than 80 years (e.g., cp., Bortkiewicz, 1901). The use of higher moments to describe observations and the use of chi-square to assess goodness of fit are both more than 60 years old (Pearson, 1895, 1900). While the last century has seen great developments of regression techniques, and of the comparison of sums of squares, comparison with the development of other sciences suggests that novelty has entered data analysis at a slow, plodding pace.

By and large, the great innovations in statistics have not had correspondingly great effects upon data analysis. The extensive calculation of "sums of squares" is the outstanding exception. (Iterative maximum likelihood, as in the use of working probits, probably comes next, but is not widely enough used to represent a great effect.)

Is it not time to seek out novelty in data analysis?

2. Special growth areas. Can we identify *some* of the areas of data analysis which today offer unusual challenges, unusual possibilities of growth?

The treatment of "spotty data" is an ancient problem, and one about which there has been much discussion, but the development of organized techniques of dealing with "outliers", "wild shots", "blunders", or "large deviations" has lagged far behind both the needs, and the possibilities presently open to us for meeting these needs.

The analysis of multiple-response data has similarly been much discussed, but, with the exception of psychological uses of factor analysis, we see few analyses of multiple-response data today which make essential use of its multiple-response character.

Data of the sort today thought of as generated by a stochastic process is a current challenge which both deserves more attention, and different sorts of attention, than it is now receiving.

Problems of selection often involve multi-stage operations, in which later-stage actions are guided by earlier-stage results. The essential role of data analysis, even when narrowly defined, is clear. (To many of us the whole process is a part of data analysis.) Here we have learned a little, and have far to go.

We have sought out more and more subtle ways of assessing error. The half-normal plot typifies the latest, which will have more extensive repercussions than most of us have dreamed of.

Data which is heterogeneous in precision, or in character of variation, and data that is very incomplete, offer challenges that we have just begun to meet. Beyond this we need to stress flexibility of attack, willingness to iterate, and willingness to study things as they are, rather than as they, hopefully, should be.

Again let me emphasize that these are only *some* of the growth areas, and that their selection and relative emphasis has been affected by both personal ¡diosyncrasies and the importance of making certain general points.

3. How can new data analysis be initiated? How is novelty most likely to begin and grow? Not through work on familiar problems, in terms of familiar frameworks, and starting with the results of applying familiar processes to the observations. Some or all of these familiar constraints must be given up in each piece of work which may contribute novelty.

We should seek out wholly new questions to be answered. This is likely to require a concern with more complexly organized data, though there will be exceptions, as when we are insightful enough to ask new, useful kinds of questions about familiar sorts of data.

We need to tackle old problems in more realistic frameworks. The study of data analysis in the face of fluctuations whose distribution is rather reasonable, but unlikely to be normal, provides many important instances of this. So-called non-parametric methods, valuable though they are as first steps toward more realistic frameworks, are neither typical or ideal examples of where to stop. Their *ultimate* use in data analysis is likely to be concentrated (i) upon situations where relative ranks are really all the data available, and (ii) upon situations where unusually quick or portable procedures are needed. In other situations it will be possible, and often desirable, to analyze the data more thoroughly and effectively by other methods. Thus while non-parametric analysis of two-way tables leading to confidence statements for differences of typical values for rows and columns is quite possible, it is computationally difficult enough to keep me from believing that such techniques will ever be widely used. (The situation for three- and more-way tables is even worse.) As means for defending the results of an analysis against statistical attack on distributional grounds, nonparametric methods will retain a pre-eminent place. And as an outstanding first step in eliminating unacceptable dependence upon normality assumptions they have been of the greatest importance. But these facts do not suffice to make them the methods of continuing choice.

We should seek out unfamiliar summaries of observational material, and establish their useful properties.

But these are not the only ways to break out into the open. The comparison, under suitable or unsuitable assumptions, of different ways of analyzing the same data for the same purpose has been a unifying concept in statistics. Such comparisons have brought great benefits to data analysis. But the habit of making them has given many of us a specific mental rigidity. Many seem to find it essential to begin with a probability model containing a parameter, and then to ask for a good estimate for this parameter (too often, unfortunately, for one that is optimum). Many have forgotten that data analysis can, sometimes quite appropriately, precede probability models, that progress can come from asking what a specified indicator (= a specified function of the data) may reasonably be regarded as estimating. Escape from this constraint can do much to promote novelty.

And still more novelty can come from finding, and evading, still deeper lying constraints.

It can help, throughout this process, to admit that our first concern is "data analysis". I once suggested in discussion at a statistical meeting that it might be well if statisticians looked to see how data was actually analyzed by many sorts of people. A very eminent and senior statistician rose at once to say that this was a novel idea, that it might have merit, but that young statisticians should be careful not to indulge in it too much, since it might distort their ideas. *The ideas of data analysis ought to survive a look at how data is analyzed.* Those who try may even find new techniques evolving, as my colleague Martin Wilk suggests, from studies of the nature of "intuitive generalization".

4. Sciences, mathematics, and the arts. The extreme cases of science and art are clearly distinguished, but, as the case of the student who was eligible for Phi Beta Kappa because mathematics was humanistic and for Sigma Xi because it was scientific shows, the place of mathematics is often far from clear. There should be little surprise that many find the places of statistics and data analysis still less clear.

There are diverse views as to what makes a science, but three constituents will be judged essential by most, viz:

(a1) intellectual content,

(a2) organization into an understandable form,

(a3) reliance upon the test of experience as the ultimate standard of validity.

By these tests, mathematics is not a science, since its ultimate standard of validity is an agreed-upon sort of logical consistency and provability.

As I see it, data analysis passes all three tests, and I would regard it as a science, one defined by a ubiquitous problem rather than by a concrete subject. (Where "statistics" stands is up to "statisticians" and to which ultimate standard—analysis of data or pure mathematics—they adhere. A useful mixed status may be gained by adhering sometimes to one, and sometimes to the other, although it is impossible to adopt both *simultaneously* as *ultimate* standards.)

Data analysis, and the parts of statistics which adhere to it, must then take on the characteristics of a science rather than those of mathematics, specifically:

(b1) Data analysis must seek for scope and usefulness rather than security.

(b2) Data analysis must be willing to err moderately often in order that inadequate evidence shall more often *suggest* the right answer.

(b3) Data analysis must use mathematical argument and mathematical results as bases for judgment rather than as bases for proof or stamps of validity.

These points are meant to be taken seriously. Thus, for example, (b1) does not mean merely "give up security by acting as if 99.9% confidence were certainty", much more importantly it means "give general advice about the use of techniques as soon as there is reasonable ground to think the advice sound; be prepared for a reasonable fraction (not too large) of cases of such advice to be *generally wrong*"!

All sciences have much of art in their makeup. (It is sad that the phrase "the useful and mechanic arts" no longer reminds us of this frequently.) As well as teaching facts and well-established structures, all sciences must teach their apprentices how to think about things in the manner of that particular science, and what are its current beliefs and practices. Data analysis must do the same. Inevitably its task will be harder than that of most sciences. "Physicists" have usually undergone a long and concentrated exposure to those who are already masters of the field. "Data analysts", even if professional statisticians, will have had far less exposure to professional data analysts during their training.

Three reasons for this hold today and can at best be altered slowly:

(c1) Statistics tends to be taught as part of mathematics.

(c2) In learning statistics *per se* there has been limited attention to data analysis.

(c3) The number of years of intimate and vigorous contact with professionals is far less for statistics Ph.D.'s than for physics (or mathematics) Ph.D.'s.

Thus data analysis, and adhering statistics, faces an unusually difficult problem of communicating certain of its essentials, one which cannot presumably be met as well as in most fields by indirect discourse and working side-by-side.

For the present, there is no substitute for making opinions more obvious, for being willing to express opinions and understandings in print (knowing that they may be wrong), for arguing by analogy, for reporting broad conclusions supported upon a variety of moderately tenuous pieces of evidence (and then going on, later, to identify and make critical tests of these conclusions), for emphasing the importance of judgment and illustrating its use, (as Cox (1957) has done with the use of judgment indexes as competitors for covariance), not merely for admitting that a statistician needs to use it. And for, while doing all this, continuing to use statistical techniques for the confirmatory appraisal of observations through such conclusion procedures as confidence statements and tests of significance.

Likewise, we must recognize that, as Martin Wilk has put it, "The hallmark of good science is that it uses models and 'theory' but never believes them."

5. Dangers of optimization. What is needed is progress, and the unlocking of certain of rigidities (ossifications?) which tend to characterize statistics today. Whether we look back over this century, or look into our own crystal ball, there is but one natural chain of growth in dealing with a specific problem of data analysis, viz:

(a1′) recognition of problem,
(a1″) one technique used,
(a2) competing techniques used,
(a3) rough comparisons of efficacy,
(a4) comparison in terms of a precise (and thereby inadequate) criterion,
(a5′) optimization in terms of a precise, and similarly inadequate criterion,
(a5″) comparison in terms of several criteria.

(Number of primes does not indicate relative order.)

If we are to be effective in introducing novelty, we must heed two main commandments in connection with new problems:

(A) Praise and use work which reaches stage (a3), or only stage (a2), or even stage (a1″).

(B) Urge the extension of work from each stage to the next, with special emphasis on the earlier stages.

One of the clear signs of the lassitude of the present cycle of data analysis is the emphasis of many statisticians upon certain of the later stages to the exclusion of the earlier ones. Some, indeed, seem to equate stage (a5′) to statistics —an attitude which if widely adopted is guaranteed to produce a dried-up, encysted field with little chance of real growth.

I must once more quote George Kimball's words of wisdom (1958). "There is a further difficulty with the finding of 'best' solutions. All too frequently when a 'best' solution to a problem has been found, someone comes along and finds a still better solution simply by pointing out the existence of a hitherto unsuspected variable. In my experience when a moderately good solution to a problem has been found, it is seldom worth while to spend much time trying to convert this into the 'best' solution. The time is much better spent in real research $\cdots$." As Kimball says so forcefully, optimizing a simple or easy problem is not as worthwhile as meliorizing a more complex or difficult one. In data analysis we have no difficulty in complicating problems in useful ways. It would, for example, often help to introduce a number of criteria in the place of a single one. It would almost always help to introduce weaker assumptions, such as more flexibility for parent distributions. And so on.

It is true, as in the case of other ossifications, that attacking this ossification is almost sure to reduce the apparent neatness of our subject. But neatness does not accompany rapid growth. Euclidean plane geometry is neat, but it is still Euclidean after two millenia. (Clyde Coombs points out that this neatness made it easier to think of non-Euclidean geometry. If it were generally understood that the great virtue of neatness was that it made it easier to make things complex again, there would be little to say against a desire for neatness.)

6. Why optimization? Why this emphasis on optimization? It is natural, and desirable, for mathematicians to optimize; it focusses attention on a small subset of all possibilities, it often leads to general principles, it encourages sharpening of concepts, particularly when intuitively unsound optima are regarded as reasons for reexamining concepts and criteria (e.g., Cox, (1958, p. 361) and the criterion of power). Danger only comes from mathematical optimizing when the results are taken too seriously. In elementary calculus we all optimize surface-to-volume relations, but no one complains when milk bottles turn out to be neither spherical or cylindrical. It is understood there that such optimum problems are unrealistically oversimplified, that they offer *guidance*, not the *answer*. (Treated similarly, the optimum results of mathematical statistics can be most valuable.)

There is a second reason for emphasis upon the optimum, one based more upon the historical circumstances than upon today's conditions. Let others speak:

"It is a pity, therefore, that the authors have confined their attention to the relatively simple problem of determining the approximate distribution of arbitrary criteria and have failed to produce any sort of justification for the tests they propose. In addition to those functions studied there are an infinity of others, and unless some principle of selection is introduced we have nothing to look forward to but an infinity of test criteria and an infinity of papers in which they are described." (Box, 1956, p. 29.)

"More generally still, one has the feeling that the statistics we are in the habit of calculating from our time series tend to be unduly stereotyped. We are, in a way, in the reverse situation to that which obtained when Fisher wrote about how easy it was to invent a great multiplicity of statistics, and how the problem was to select the good statistics from the bad ones. With time series we could surely benefit from the exploration of the properties of many more statistics than we are in the habit of calculating. We are more sober than in the days of Fechner, Galton, and "K.P."; perhaps we are too sober." (Barnard, 1959a, p. 257.)

The first quotation denies that (a2) above should precede (a3) to (a5′), while the second affirms that (a2) and (a3) should be pusbed, especially in fields that have not been well explored. (The writer would not like to worry about an infinity of methods for attacking a question until he has at least four such which have *not* been shown to have distinguishable behavior.)

7. The absence of judgment. The view that "statistics is optimization" is perhaps but a reflection of the view that "data analysis should not *appear to* be a matter of judgment". Here "appear to" is in italics because many who hold this view would like to suppress these words, even though, when pressed, they would agree that the optimum *does* depend upon the assumptions and criteria, whose selection may, perhaps, even be admitted to involve judgment. It is very helpful to replace the use of judgment by the use of knowledge, but only if the result is the use of *knowledge with judgment.*

Pure mathematics differs from most human endeavor in that assumptions are not criticized because of their relation to something outside, though they are, of course, often criticized as unaesthetic or as unnecessarily strong. This cleavage between pure mathematics and other human activities has been deepening since the introduction of non-Euclidean geometries by Gauss, Bolyai, and Lobachevski about a century and a half ago. Criticism of assumptions on the basis of aesthetics and strength, without regard for external correspondence has proved its value for the development of mathematics. But we dare not use such a wholly internal standard anywhere except in pure mathematics. (For a discussion of its dangers in pure mathematics, see the closing pages of von Neumann, 1947.)

In data analysis we must look to a very heavy emphasis on judgment. At least three different sorts or sources of judgment are likely to be involved in almost every instance:

(a1) judgment based upon the experience of the particular field of subject matter from which the data come,

(a2) judgment based upon a broad experience with how particular techniques of data analysis have worked out in a variety of fields of application,

(a3) judgment based upon abstract results about the properties of particular techniques, whether obtained by mathematical proofs or empirical sampling.

Notice especially the form of (a3). It is consistent with actual practice in every field of science with a theoretical branch. A scientist's actions are *guided*, not determined, by what has been derived from theory or established by experiment, *as is his advice to others*. The judgment with which isolated results are put together to guide action or advice in the usual situation, which is too complex for guidance to be *deduced* from available knowledge, will often be a mixture of individual and collective judgments, but judgment will play a crucial role. Scientists know that they will sometimes be wrong; they try not to err too often, but they accept some insecurity as the price of wider scope. Data analysts must do the same.

One can describe some of the most important steps in the development of mathematical statistics as attempts to save smaller and smaller scraps of certainty (ending with giving up the certainty of using an optimal technique for the certainty of using an admissible one, one that at least cannot be unequivocally shown to be non-optimal). Such attempts must, in large part at least, be attempts to maintain the mathematical character of statistics at the expense of its data-analytical character.

If data analysis is to be well done, much of it must be a matter of judgment, and "theory", whether statistical or non-statistical, will have to guide, not command.

8. The reflection of judgment upon theory. The wise exercise of judgment can hardly help but to stimulate new theory. While the occurrence of this phenomenon may not be in question, its appropriateness is regarded quite differently. Three quotations from the discussion of Kiefer's recent paper before the Research Section of the Royal Statistical Society point up the issue:

"Obviously, if a scientist asks my advice about a complex problem for which I cannot compute a good procedure in the near future, I am not going to tell him to cease his work until such a procedure is found. But when I give him the best that my intuition and reason can now produce, I am not going to be satisfied with it, no matter how clever a procedure it may appear on the surface. The aim of the subject is not the construction of nice looking procedures with intuitive appeal, but the determination of procedures which are proved to be good." (Kiefer (1959, p. 317) in reply to discussion.)

"A major part of Dr. Kiefer's contribution is that he is forcing us to consider very carefully what we want from a design. But it seems to me to be no more reprehensible to start with an intuitively attractive design and then to search for optimality criteria which it satisfies, then to follow the approach of the present paper, starting from (if I may call it so) an intuitively attractive criterion, and then to search for designs which satisfy it. Either way one is liable to be surprised by what comes out; but the two methods are complementary." (Mallows, 1959, p. 307.)

"The essential point of the divergence is concisely stated at the end of section A: 'The rational approach is to state the problem and the optimality criterion and then to find the appropriate design, and not alter the statements of the problem and criterion just to justify the use of the design to which we are wedded by our prejudiced intuition.' I would agree with this statement if the word 'deductive' were inserted in place of the word 'rational'. As a rationalist I feel that the word 'rational' is one which indicates a high element of desirability, and I think it is much broader in its meaning than 'deductive'. In fact what appears to me to be the rational approach is to take designs which are in use already, to see what is achieved by these designs by consideration of the general aims to evaluate such designs, in a provisional way, and then to seek to find designs which improve on existing practice. Having found such designs the cycle should be repeated again. The important thing about this approach is that we are always able to adjust our optimality criteria to designs as well as adjusting our designs to our optimality criteria." (Barnard, 1959b, p. 312).

9. Teaching data analysis. The problems of teaching data analysis have undoubtedly had much to do with these unfortunate rigidities. Teaching data analysis is not easy, and the time allowed is always far from sufficient. But these difficulties have been enhanced by certain views which have been widely adopted, such as those caricatured in:

(a1) "avoidance of cookbookery and growth of understanding come only by mathematical treatment, with emphasis upon proofs".

(a2) "It is really quite intolerable for the teacher then to have to reply, 'I don't know'." (An actual quotation from Stevens (1950, p. 129).)

(a3) "whatever the facts may be, we must keep things simple so that we can teach students more easily".

(a4) "even if we do not know how to treat this problem so as to be either good data analysis or good mathematics, we should treat it somehow, because we must teach the students something".

It would be well for statistics if these points of view were not true to life, were overdrawn, but it is not hard to find them adopted in practice, even among one's friends.

The problem of cookbookery is not peculiar to data analysis. But the solution of concentrating upon mathematics and proof is. The field of biochemistry today contains much more detailed knowledge than does the field of data analysis. The over-all teaching problem is more difficult. Yet the text books strive to tell the facts in as much detail as they can find room for. (Biochemistry was selected for this example because there is a clear and readable account by a coauthor of a leading text of how such a text is revised (Azimov, 1955).)

A teacher of biochemistry does not find it intolerable to say "I don't know". Nor does a physicist. Each spends a fair amount of time explaining what his science does not know, and, consequently, what are some of the challenges it offers the new student. Why should not both data analysts and statisticians do the same?

Surely the simplest problems of data analysis are those of

(b1) location based upon a single set of data,

(b2) relative location based upon two sets of data.

There are various procedures for dealing with each of these problems. Our knowledge about their relative merits under various circumstances is far from negligible. Some of it is a matter of proof, much of the rest can be learned by collecting the results when each member of any class draws a few samples from each of several parent distributions and applies the techniques. Take the one-sample problem as an example. What text, and which teachers, teach the following simple facts about one-sample tests?

(c1) for symmetrical distributions toward the rectangular, the mid-range offers high-efficiency of point-estimation, while the normally-calibrated range-midrange procedure (Walsh, 1949a) offers conservative but still efficient confidence intervals,

(c2) for symmetrical distributions near normality, the mean offers good point estimates and Student's t offers good confidence intervals, but signed-rank (Wilcoxon, 1949; Walsh, 1949b, 1959) confidence intervals are about as good for small and moderate sample sizes,

(c3) for symmetrical distributions with slightly longer tails, like the logistic perhaps, a (trimmed) mean omitting a prechosen number of the smallest and an equal number of the largest observations offers good point estimates, while signed-rank procedures offer good interval estimates,

(c4) for symmetric distributions with really long tails (like the central 99.8% of the Cauchy distribution, perhaps) the median offers good point estimates, and the sign test offers good interval estimates.

(c5) the behavior of the one-sample t-test has been studied by various authors (cp., Gayen, 1949, and references cited by him) with results for asymmetric distributions which can be digested and expressed in understandable form.

These facts are specific, and would not merit the expenditure of a large fraction of a course in data analysis. But they can be said briefly. And, while time spent showing that Student's t is optimum for exactly normal samples may well, on balance, have a negative value, time spent on these points would have a positive one.

These facts are a little complex, and may not prove infinitely easy to teach, but any class can check almost any one of them by doing its own experimental sampling. Is it any better to teach everyone, amateurs and professionals alike, about only a *single* one-sample procedure (or *as is perhaps worse* about the comparative merits of various procedures in sampling from but a *single* shape of parent population) than it would be to teach laymen and doctors alike that the only pill to give is aspirin (or to discuss the merits and demerits of various pills for but a single ailment, mild headache)?

Some might think point (a4) above to be excessively stretched. Let us turn again to the Stevens article referred to in (a2), in which he introduced randomized confidence intervals for binomial populations. In an addendum, Stevens notes an independent proposal and discussion of this technique, saying: "It was there dismissed rather briefly as being unsatisfactory. This may be granted but since...some solution is necessary [because the teacher should not say "I don't know" (J. W. T.)], it seems that this one deserves to be studied and to be used by teachers of statistics until a better one can be found." The solution is admittedly unsatisfactory, and not just the best we have to date, yet it is to be taught, and used!

Not only must data analysis admit that it uses judgment, it must cease to hide its lack of knowledge by teaching answers better left unused. The useful must be separated from the unuseful or antiuseful.

As Egon Pearson pointed out in a Statistical Techniques Research Group discussion where this point was raised, there is a real place in *discussing* randomized confidence limits in advanced classes for statisticians; not because they are useful, not because of aesthetic beauty, but rather because they may stimulate critical thought. As Martin Wilk puts it: "We dare not confine ourselves to emphasizing properties (such as efficiency or robustness) which, although sometimes useful as guides, only hold under classes of assumptions all of which may be wholly unrealistic; we must teach an understanding of *why* certain sorts of techniques (e.g., confidence intervals) are indeed useful."

10. Practicing data analysis. If data analysis is to be helpful and useful, it must be practiced. There are many ways in which it can be used, some good and some evil. Laying aside unethical practices, one of the most dangerous (as I have argued elsewhere (Tukey, 1961b)) is the use of formal data-analytical procedures for sanctification, for the preservation of conclusions from all criticism, for the granting of an *imprimatur*. While statisticians have contributed to this misuse, their share has been small. There is a corresponding danger for data analysis, particularly in its statistical aspects. This is the view that all statisticians *should* treat a given set of data in the same way, just as all British admirals, in the days of sail, maneuvered in accord with the same principles. The admirals could not communicate with one another, and a single basic doctrine was essential to coordinated and effective action. Today, statisticians can communicate with one another, and have more to gain by using special knowledge (subject-matter or methodological) and flexibility of attack than they have to lose by not all behaving alike.

In general, the best account of current statistical thinking and practice is to be found in the printed discussions in the *Journal of the Royal Statistical Society*. While reviewing some of these lately, I was surprised, and a little shocked to find the following:

"I should like to give a word of warning concerning the approach to tests of significance adopted in this paper. It is very easy to devise different tests which, on the average, have similar properties, i.e., they behave satisfactorily when the null hypothesis is true and have approximately the same power of detecting departures from that hypothesis. Two such tests may, however, give very different results when applied to a given set of data. This situation leads to a good deal of contention amongst statisticians and much discredit of the science of statistics. The appalling position can easily arise in which one can get any answer one wants if only one goes around to a large enough number of statisticians." (Yates, 1955, p. 31).

To my mind this quotation, if taken very much more seriously than I presume it to have been meant, nearly typifies a picture of statistics as a monolithic, authoritarian structure designed to produce the "official" results. While the possibility of development in·this direction is a real danger to data analysis, I find it hard to believe that this danger is as great as that posed by over-emphasis on optimization.

11. Facing uncertainty. The most important maxim for data analysis to heed, and one which many statisticians seem to have shunned, is this: "Far better an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise." Data analysis must progress by approximate answers, at best, since its knowledge of what the problem really is will at best be approximate. It would be a mistake not to face up to this fact, for by denying it, we would deny ourselves the use of a great body of approximate knowledge, as well as failing to maintain alertness to the possible importance in each particular instance of particular ways in which our knowledge is incomplete.

II. SPOTTY DATA

12. What is it? The area which is presently most obviously promising as a site of vigorous new growth in data analysis has a long history. The surveyor recognizes a clear distinction between "errors" with which he must live, and "blunders", whose effects he must avoid. This distinction is partly a matter of size of deviation, but more a matter of difference in the character or assignability of causes. Early in the history of formal data analysis this recognition led to work on the "rejection of observations". Until quite recently matters rested there.

One main problem is the excision of the effects of occasional potent causes. The gain from such excision should not be undervalued. Paul Olmstead, for example, who has had extensive experience with such data, maintains that engineering data typically involves 10% of "wild shots" or "stragglers". A ratio of 3 between "wild shot" and "normal" standard deviations is far too low (individual "wild shots" can then hardly be detected). Yet 10% wild shots with standard deviation 3σ contribute a variance equal to that of the remaining 90% of the cases with standard deviation 1σ. Wild shots can easily double, triple, or quadruple a variance, so that really large increases in precision can result from cutting out their effects.

We are proud, and rightly so, of the "robustness" of the analysis of variance. A few "wild shots" sprinkled at random into data taken in a conventional symmetric pattern will, on the average, affect each mean square equally. True, but we usually forget that this provides only "robustness of validity", ensuring that we will not be led by "wild shots" to too many false positives, or to too great security about the precision of our estimates. Conventional analysis of variance procedures offer little "robustness of efficiency", little tendency for the high efficiency provided for normal distributions of fluctuations-and-errors to be preserved for non-normal distributions. A few "wild shots", either spread widely or concentrated in a single cell, can increase all mean squares substantially. (Other spreadings may have different results.) From a hypothesis-testing point of view this decreases our chances of detecting real effects, and increases the number of false negatives, perhaps greatly. From an estimation point of view it increases our variance of estimate and decreases our efficiency, perhaps greatly. Today we are far from adopting an adequately sensitive technique of analysis, even in such simple situations as randomized blocks.

Now one cannot look at a single body of data alone and be sure which are the "wild shots". One can usually find a statistician to argue that some particular observation is not unusual in cause, but is rather "merely a large deviation". When there are many such, he will have to admit that the distribution of deviations (of errors, of fluctuations) has much longer tails than a Gaussian (normal) distribution. And there will be instances where he will be correct in such a view.

It would be unfortunate if the proper treatment of data was seriously different when errors-and-fluctuations have a long-tailed distribution, as compared to the case where occasional causes throw in "wild shots". Fortunately, this appears not to be the case; cures or palliatives for the one seem to be effective against the other. A simple indication that this is likely to be so is furnished by the probability element

$$[(1-\theta)(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}y^2}+\theta[h^{-1}(2\pi)^{-\frac{1}{2}}]e^{-y^2/2h^2}]\,dy,$$

which can be construed in at least three ways:

(a1) as a unified long-tailed distribution which is conveniently manipulable in certain ways,

(a2) as representing a situation in which there is probability θ that an occasional-cause system, which contributes additional variability when it acts, will indeed act,

(a3) as representing a situation in which variability is irregularly nonhomogeneous.

It is convenient to classify together most instances of long-tailed fluctuation-and-error distributions, most instances of occasional causes with large effects, and a substantial number of cases with irregularly non-constant variability; the term "spotty data" is surely appropriate.

13. An appropriate step forward. Clearly the treatment of spotty data is going to force us to abandon normality. And clearly we can go far by studying cases of sampling from long-tailed distributions. How far to go? Which long-tailed distributions to face up to?

To seek complete answers to these questions would be foolish. But a little reflection takes us a surprising distance. We do not wish to take on any more new problems than we must. Accordingly it will be well for us to begin with long-tailed distributions which offer the minimum of doubt as to what should be taken as the "true value". If we stick to symmetrical distributions we can avoid all difficulties of this sort. The center of symmetry is the median, the mean (if this exists), and the $\alpha\%$ trimmed mean for all α. (The $\alpha\%$ trimmed mean is the mean of the part of the distribution between the lower $\alpha\%$ point and the upper $\alpha\%$ point.) No other point on a symmetrical distribution has a particular claim to be considered the "true value". Thus we will do well to *begin* by restricting ourselves to symmetric distributions.

Should we consider all symmetric distributions? This would be a counsel of perfection, and dangerous, since it would offer us far more flexibility than we know how to handle. We can surely manage, even in the beginning, to face a single one-parameter family of shapes of distribution. Indeed we may be able to face a few such families. Which ones will be most effective? Experience to date suggests that we should be interested both in families of shapes in which the tails behave more or less as one would expect from the remainder of the distribution (conforming tails) and in families in which the tails are longer than the remainder suggests (surprising tails). The latter are of course peculiarly applicable to situations involving occasional causes of moderately rare occurrence.

What one-parameter families of distribution shapes with conforming tails should be considered? And will the choice matter much? We may look for candidates in two directions, symmetric distributions proposed for graduation, and symmetric distributions which are easy to manipulate. The leading classical candidate consists of:

(a1) the symmetric Pearson distributions, namely the normal-theory distributions of Student's t and Pearson's r.

The only similar system worth particular mention is:

(a2) N. L. Johnson's [1949] distributions, which, in the symmetric case, are the distributions of tanh N/δ or sinh N/δ, where N follows a unit normal distribution, and δ is an appropriate constant.

While the symmetrical Johnson curves are moderately easy to manipulate in various circumstances, even more facility is frequently offered by what may be called the lambda-distributions, viz:

(a3) the distributions of $P^{\lambda} - (1 - P)^{\lambda}$ where P is uniformly distributed on $(0, 1)$.

(It is possible that it would be desirable to introduce a further system of symmetrical distributions obtained from the logistic distribution by simple transformation.)

While the analytic descriptions of these three conforming systems are quite different, there is, fortunately, little need to be careful in choosing among them, since they are remarkably similar (personal communication from E. S. Pearson for (a2) vs. (a1) and Tukey, 1962 for (a3) vs. (a1)).

The extreme cases of all of these symmetrical families will fail to have certain moments; some will fail to have variances, others will even fail to have means. It is easy to forget that these failures are associated with the last ϵ of cumulative probability in each tail, no matter how small $\epsilon > 0$ may be. If we clip a tail of probability 10^{-40} off each end of a Cauchy distribution (this requires clipping in the vicinity of $x = \pm 10^{20}$), and replace the removed 2.10^{-40} probability at any bounded set of values, the resulting distribution will have finite moments of all orders. But the behavior of samples of sizes less than, say, 10^{30} from the two distributions will be practically indistinguishable. The finiteness of the moments does not matter directly; the extendedness of the tails does. For distributions given in simple analytic form, infiniteness of moments *often* warns of practical extendedness of tails, but, as the infinite moments of

$$dF = [(1 - 10^{-50})(2\pi)^{-\frac{1}{2}}e^{-\frac{1}{2}x^2} + 10^{-50}(4/\pi)(1 + x^2)^{-1}]\,dx$$

show in the other direction, there is no necessary connection. A sure way to drive ourselves away from attaching undue significance to infinite moments, and back to a realistic view is to study some one-parameter families with finite moments which converge to infinite-moment cases. These are easy to find, and include

(b1) the distribution of $\tan X$, where $(2/\pi)X$ is uniform on $(-\theta, +\theta)$ for $0 < \theta < 1$ (as $\theta \rightarrow 1$, this goes to the Cauchy distribution)

(b2) the distribution of $(1 - P)^{-1} - P^{-1}$, where P is uniform on $(\epsilon, 1 - \epsilon)$ for $0 < \epsilon < 0.5$ (as $\epsilon \rightarrow 0$, this goes to a rough approximation to a Cauchy distribution).

(Note that all distributions satisfying (b1) and (b2) have all moments finite.)

Beyond this we shall want in due course to consider some symmetrical distributions with surprising tails (i.e., some which have longer tails than would be expected from the rest of the distribution). Here we might consider, in particular:

(c1) contaminated distributions at scale 3, with probability element

$$\{(1 - \theta)(2\pi)^{-\frac{1}{2}}e^{-y^2/2} + \theta[3^{-1}(2\pi)^{-\frac{1}{2}}]e^{-y^2/18}\}\,dy,$$

(c2) contaminated lambda distributions such as

$$(1 - \theta)[P^{0.2} - (1 - P)^{0.2}] + \theta[\log P - \log(1 - P)]$$

and

$$(1 - \theta)[\log P - \log(1 - P)] + \theta[(1/P) - (1 - P)^{-1}].$$

There is no scarcity of one-shape-parameter families of symmetrical distributions which are reasonably easily manipulated and whose behavior can offer valuable guidance.

Once we have a reasonable command of the symmetrical case, at least in a few problems, it will be time to plan our attack upon the unsymmetrical case.

14. Trimming and Winsorizing samples. In the simplest problems, those involving only the location of one or more simple random samples, it is natural to attempt to eliminate the effects of "wild shots" by *trimming* each sample, by removing equal numbers of the lowest and highest observations, and then proceding as if the trimmed sample were a complete sample. As in all procedures intended to guard against "wild shots" or long-tailed distributions, we must expect, in comparison with procedures tailored to exactly normal distributions:

(a1) some loss in efficiency when the samples do come from a normal distribution,

(a2) increased efficiency when the samples come from a long-tailed distribution.

An adequate study of the effects of trimming naturally proceeds step by step.

It may well start with comparisons of variances of trimmed means with those of untrimmed means, first for normal samples, and then for long-tailed samples. Order-statistic moments for samples from long-tailed distributions are thus required. The next step is to consider alternative modifications of the classical t-statistic with a trimmed mean in the numerator and various denominators. Order statistic moments, first for the normal and rectangular, and then for suitable long-tailed distributions are now used to determine the ratios

$$\frac{\text{variance of numerator}}{\text{average squared denominator}}$$

for various modified t-statistics and various underlying distributions. This enables the choice of a denominator which will make this ratio quite closely constant. (It is, of course, exactly constant for the classical t-statistic.) Detailed behavior for both normal and non-normal distributions is now to be obtained; sophicated experimental sampling is almost certainly necessary and sufficient for this. Close agreement of critical values can now replace close agreement of moment ratios as a basis for selection, and critical values can then be supplemented with both normal and non-normal power functions. At that point we will know rather more about the symmetrical-distribution behavior of such modified t-statistics based upon trimmed samples than we presently do about Student's t itself. (This program is actively under way at the Statistical Techniques Research Group of Princeton University.)

Both this discussion, and the work at S. T. R. G., began with the case of trimmed samples because it is the simplest to think about. But it is not likely to be the most effective for moderately or sub-extremely long-tailed distribution. When I first met Charles P. Winsor in 1941 he had already developed a clear and individual philosophy about the proper treatment of apparent "wild shots". When he found an "outlier" in a sample he did not simply reject it. Rather he changed its value, replacing its original value by the nearest value of an observation not seriously suspect. His philosophy for doing this, which applied to "wild shots", can be supplemented by a philosophy appropriate to long-tailed distributions which leads to the same actions (cp., Anscombe and Tukey, 1962). It seems only appropriate, then, to attach his name to the process of replacing the values of certain of the most extreme observations in a sample by the nearest unaffected values, to speak of Winsorizing or Winsorization.

For normal samples, Winsorized means are more stable than trimmed means (Dixon, 1960, 1957). Consequently there is need to examine the advantages of modified t-statistics based upon Winsorized samples.

The needs and possibilities will not come to an end here. So far we have discussed only the case where a fixed number of observations are trimmed from, or Winsorized at, each end of a sample. But intelligent rejection of observations has always been guided by the configuration of the particular sample considered, more observations being discarded from some samples than from others. Tailoring is required.

Tailored trimming and tailored Winsorizing, respectively, may thus be expected to give better results than fixed trimming or fixed Winsorizing, and will require separate study. (Given a prescribed way of choosing the number to be removed or modified at each end, there will, in particular, be a common factor (for given sample size) by which the length of the corresponding fixed-mode confidence limits can be multiplied in order to obtain tailored-mode confidence limits with the indicated confidence. This is of course only one way to allow for tailoring each sample separately.)

All these procedures need to be studied first for the single sample situation, whose results should provide considerable guidance for those multiple-sample problems where conventional approaches would involve separate estimates of variance. Situations leading classically to pooled variances are well-known to be wisely dealt with by randomization theory so far as robustness of validity is concerned (cp., Pitman, 1937; Welch, 1938). While randomization of trimmed samples could be studied, it would seem to involve unnecessary complications, especially since the pure-randomization part of the randomization theory of Winsorized samples in inevitably the same as for unmodified samples. In particular, robustness of validity is certain to be preserved by Winsorization. And there is every reason to expect robustness of efficiency to be considerably improved in such problems by judicious Winsorization in advance of the application of randomization theory.

Before we leave this general topic we owe the reader a few numbers. Consider the means of a sample of 11, of that sample with 2 observations trimmed from each end, and of that sample with 2 observations Winsorized on each end. The resulting variances, for a unit normal parent, and for a Cauchy parent, are as follows (cp., Dixon, 1960; Sarhan and Greenberg, 1958 for normal variances of symmetrically Winsorized means):

	Normal parent	Cauchy parent
plain mean	0.091	infinite
trimmed mean	0.102	finite
Winsorized mean	0.095	finite*

* Presumably larger than for the corresponding trimmed mean.

The small loss due to Winsorization or trimming in the face of a normal parent contrasts interestingly with the large gain in the face of a Cauchy parent.

15. How soon should such techniques be put into service? The question of what we need to know about a new procedure before we recommend its use, or begin to use it, should not be a difficult one. Yet when one friendly private comment about a new non-parametric procedure (that in Siegel and Tukey, 1960) was that it should not appear in print, and I presume should *a fortiori* not be put into service, even in those places where such a non-parametric procedure is appropriate, until its power function was given, there are differences of opinion of some magnitude. (Especially when the sparsity of examples of non-parametric power functions which offer useful guidance to those who might use the procedure is considered.)

Let us cast our thoughts back to Student's *t*. How much did we know when it was introduced into service? Not even the distribution for a normal parent was known! Some empirical sampling, a few moments, and some Pearson curves produced the tables of critical values. But, since it filled an aching void, it went directly into use.

As time went on, the position of the *single-sample t* procedure became first better, as its distribution, and its optimality, under normal assumptions was established, and then poorer, as more was learned about its weaknesses for non-normal samples, whose frequent occurrence had been stressed by Student himself (1927). (For references to the history of the discovery of the weaknesses see, e.g., Geary, 1947; and Gayen, 1949. For references which tend to support Geary's (1949, p. 241) proposal that all textbooks state that "*Normality is a myth; there never has, and never will be, a normal distribution.*", see Tukey, 1960.)

The case of modified *t*'s based upon trimmed or Winsorized means is somewhat different. There are already well-established procedures with known advantages, viz:

(a1) Student's *t*, optimum for normality

(a2) the signed-rank procedure, robust of validity within symmetry

(a3) the sign-test procedure, robust of validity generally.

How much do we need to know about a competitor before we introduce it to service? For my own part, I expect to begin recommending the first product of the program sketched above, a modified *t* based upon a trimmed mean with fixed degree of trimming, as a standard procedure for most single-sample location problems, as soon as I know:

(b1) that the variance of the numerator for normality is suitably small.

(b2) that the ratio of ave $(\text{denominator})^2$ to var (numerator) is constant to within a few % for a variety of symmetrical parent distributions.

(b3) estimates based upon empirical sampling of the critical values (% points) appropriate for normality.

In bringing such a procedure forward it would be essential to emphasize that it was not the final answer; that newer and better procedures were to be expected, perhaps shortly. (Indeed it would be wrong to feel that we are ever going to completely and finally solve any of the problems of data analysis, even the simple ones.)

Surely the suggested amount of knowledge is not enough for anyone to guarantee either

(c1) that the chance of error, when the procedure is applied to real data, corresponds precisely to the nominal levels of significance or confidence, or

(c2) that the procedure, when applied to real data, will be optimal in any one specific sense.

BUT WE HAVE NEVER BEEN ABLE TO MAKE EITHER OF THESE STATEMENTS ABOUT Student's *t*. Should we therefore never have used Student's *t*?

A judgment of what procedures to introduce, to recommend, to use, must always be a judgment, a judgment of whether the gain is relatively certain, or perhaps only likely, to outweigh the loss. And this judgment can be based upon a combination of the three points above with general information of what longer-tailed distributions do to trimmed and untrimmed means.

III. SPOTTY DATA IN MORE COMPLEX SITUATIONS

16. Modified normal plotting. Spotty data are not confined to problems of simple structure. Good procedures for dealing with spotty data are not always going to be as forthright and direct as those just discussed for simple samples, where one may make a single direct calculation, and where even tailored trimming or Winsorizing need only require carrying out a number of direct calculations and selecting the most favorable. Once we proceed to situations where our "criticism" of an apparent deviation must be based upon apparent deviations corresponding to different conditions, we shall find iterative methods of calculation convenient and almost essential. Iteration will, in fact, take place in "loops within loops" since the details of, e.g., making a particular fit may involve several iterative cycles (as in non-linear least squares, e.g., cp., Deming, 1943, or in probit analysis; e.g., cp., Finney, 1947, 1952), while human examination of the results of a particular fit may lead to refitting all over again, with a different model or different weights. Even such simple problems as linear regression, multiple regression, and cross-classified analyses of variance will require iteration.

The two-way table offers simple instance which is both illuminating and useful. A class of procedures consistent with the philosophies noted in Section 14 [cp., Anscombe and Tukey, 1962] operate in the following way:

(al) if a particular deviation is much "too large" in comparison with the bulk of the other deviations, its effect upon the final estimates is to be made very small

(a2) if a particular deviation is only moderately "too large", its effect is to be decreased, but not made negligible.

The first task in developing such a procedure must be the setting up of a subprocedure which identifies the "too muchness", if any, of the extreme apparent deviations. In the first approximation, this procedure may take no account of the fact that we are concerned with residuals, and may proceed as if it were analyzing a possibly normal sample.

The natural way to attack this problem graphically would be to fit row and column means, calculate residuals (=apparent deviations), order them, plot them against typical values for normal order statistics, draw a straight line through the result, and assess "too muchness" of extreme deviations in terms of their tendency to fall off the line. Such an approach would be moderately effective. Attempts to routinize or automate the procedure would find difficulty in describing just how to fit the straight line, since little weight should be given to apparently discrepant points.

The display can be made more sensitive, and automation easier, if a different plot is used where ordinates correspond to secant slopes on the older plot. More precisely, if we make a *conventional* probability plot on *probability graph paper*, we plot the observed ith smallest value in a set of n over (against) a corresponding fraction on the probability scale. (Different choices are conventionally made for this fraction, such $(i - \frac{1}{2})/n$ or $i/(n + 1)$.) The same plot could be made on ordinary graph paper if we were to plot y_i, the ith smallest, against $a_{i|n}$, the standard normal deviate corresponding to the selected probability.

With this meaning for the plotting constants $a_{i|n}$, denote the median of the values to be examined by $\dot{y}$ (read "y split"). Then $(0, \dot{y})$ and $(a_{i|n}, y_i)$ are points in the old plot, and the slope of the line segment (secant) joining them is

$$z_i = \frac{y_i - \dot{y}}{a_{i|n}}.$$

A plot of z_i against i is a more revealing plot, and one to which we should like to fit a horizontal line. We have then only to select a typical z, and can avoid great weight on aberrant values by selecting a median of a suitable set of z's.

The choice of $a_{i|n}$ has been discussed at some length (cp., Blom, 1958, pp. 144–146 and references there to Weibull, Gumbel, Bliss, Ipsen and Jerne; cp. also, Chernoff and Lieberman, 1954). The differences between the various choices are probably not very important. The choice $a_{i|n} = \text{Gau}^{-1}[(3i - 1)/(3n + 1)]$, where $P = \text{Gau}(y)$ is the normal cumulative, is simple and surely an adequate approximation to what is claimed to be optimum (also cp., Blom, 1958, pp. 70–71).

17. Automated examination. Some would say that one should not automate such procedures of examination, that one should encourage the study of the data. (Which is somehow discouraged by automation?) To this view there are at least three strong counter-arguments:

(1) Most data analysis is going to be done by people who are not sophisticated data analysts and who have very limited time; if you do not provide them tools the data will be even less studied. Properly automated tools are the easiest to use for a man with a computer.

(2) If sophisticated data analysts are to gain in depth and power, they must have both the time and the stimulation to try out *new* procedures of analysis; hence the *known* procedures must be made easy for them to apply as possible. Again automation is called for.

(3) If we are to study and intercompare procedures, it will be much easier if the procedures have been fully specified, as must happen if the process of being made routine and automatizable.

I find these counterarguments conclusive, and I look forward to the automation of as many standardizable statistical procedures as possible. When these are available, we can teach the man who will have access to them the "why" and the "which", and let the "how" follow along.

18. FUNOP. A specific arithmetic analog of the modified plot of Section 16, which we may call FUNOP (from FUll NOrmal Plot) proceeds as follows:

(b1) Let $a_{i|n}$ be a typical value for the ith ordered observation in a sample of n from a unit normal distribution. (See Section 16.)

(b2) Let $y_1 \leqq y_2 \leqq \cdots \leqq y_n$ be the ordered values to be examined. Let $\dot{y}$ be their median (or let $\overset{\sqcup}{y}$, read "y trimmed", be the mean of the y_i with $\frac{1}{3}n < i \leqq \frac{1}{3}(2n)$.

(b3) For $i \leqq \frac{1}{3}n$ or $> \frac{1}{3}(2n)$ only, let $z_i = (y_i - \dot{y})/a_{i|n}$ (or let $z_i = (y_i - \overset{\sqcup}{y})/a_{i|n}$).

(b4) Let $\dot{z}$ be the median of the z's thus obtained (about $\frac{1}{3}$ $(2n)$ in number).

(b5) Give special attention to z's for which *both* $|y_i - \dot{y}| \geqq A \cdot \dot{z}$ and $z_i \geqq B \cdot \dot{z}$ where A and B are prechosen.

(b5*) Particularly for small n, z_j's with j more extreme than an i for which (b5) selects z_i also deserve special attention (remark of Denis J. Farlie).

Note that if the y's were a sample from a normal population with mean μ and variance σ^2, we should have

$$\text{ave } y_i = \mu + \sigma a^*_{i|n}, \qquad \text{ave } z_i = [(a^*_{i|n}/(a_{i|n})]\sigma \sim \sigma,$$

where $a^*_{i|n}$ is the average value of the ith order statistic in a sample of n from the unit normal distribution. If a few y's are perturbed and thus made larger, the result will be to make a few z's larger, perhaps considerably so, and to shift others by modest amounts. Consequently $\dot{z}$ is a reasonable estimate, with a slight tendency toward inflation, of the σ of the "main normal part of" the distribution from which the y's came.

The requirement that $z_i \geqq B \cdot \dot{z}$ is a requirement that the ith observation is comparatively large for an ith observation. The requirement that $|\, y_i - \dot{y}| \geqq A \cdot \dot{z}$ is roughly that y_i is beyond $\dot{y} \pm A \cdot \sigma$. Some *combination* of these meets many, and I believe most, of the requirements it is reasonable to impose, in different circumstances, upon a procedure for identifying values for special treatment or attention. For the present we must choose A and B mainly on a judgment basis, but we can look forward to future comparisons, probably by experimental sampling, of the effects of various choices under specific circumstances.

The i's with $\frac{1}{3}n < i \leqq \frac{1}{3}(2n)$ are excluded from the formation of z_i's both because the small values of $a_{i|n}$ promote instability and because z_i's for such i's seem unrevealing. The choice of the middle $\frac{1}{3}$ of all i's for omission is again judgment, but reasonable changes here are not likely to affect the behavior of results appreciably. A standard choice is probably worthwhile.

Let us give a simple example with $n = 14$. The needed values of $a_{i|n}$ can be taken as

$$-a_{1|14} = a_{14|14} = \text{Gau}^{-1}\,(2/43) = 1.685$$

$$-a_{2|14} = a_{13|14} = \text{Gau}^{-1}\,(5/43) = 1.194$$

$$-a_{3|14} = a_{12|14} = \text{Gau}^{-1}\,(8/43) = .892$$

$$-a_{4|14} = a_{11|14} = \text{Gau}^{-1}\,(10/43) = .764$$

TABLE 1

Example of a FUNOP *calculation*

As received	Ordered	$y_i - \dot{y}$	z_i	Ordered z's
14	−341	−375	222	
−104	−161	−195	164	
−97	−104	−138	155	
−59	−97	−131	171	135
−161	−59			155
93	14			164
454	22			171
median				
−341	45			176
54	54			222
137	93			260
473	137	103	135	352
45	193	157	176	
193	454	420	352	
22	473	439	260	
	$\dot{y} = 34$			$\dot{z} = 174$

and the calculation proceeds as set out in Table 1. (For $i = 1$, we have $222 = -375/(-1.685)$, for example.) Only $z_{13} = 352$ exceeds even $1.5\dot{z}$, and it barely exceeds $2\dot{z}$. If we were using $A = 0$ and B between 1.5 and 2.0 we would give special attention to $i = 13$, and also to $i = 14$ because it is more extreme (see (b5*) above). These values of i correspond to the original values 454 and 473.

19. FUNOR-FUNOM in a two-way table. Let us consider a specific application of the techniques sketched above. (The general reader may wish to skip the remainder of III at first reading). Because the procedure uses FUNOP and first rejects and then modifies deviations it may be designated FUNOR-FUNOM. (i.e., FUll NOrmal Rejection-FUll NOrmal Modification.)

Suppose we are given values in a r-by-c table, and that we consider that these values are reasonably treated as of the form

$$\text{(general effect)} + \text{(row effect)} + \text{(column effect)} + \text{(deviation)}$$

where the deviation may either (i) come from a long-tailed distribution or (ii) involve "wild shots". The following routine offers a way to avoid most of the evil effects of the tails without compromising real row or column effects:

(a1) Fit row and column means to the original observations and form the residuals

$$y_{jk} = x_{jk} - x_{j.} - x_{.k} + x_{..} \,.$$

(a2) Apply FUNOP to the $n = rc$ residuals, giving special attention to any $y_{jk} = y_i$ with both $|y_i - \dot{y}| \geq A_R \cdot \dot{z}$ and $z_i > B_R \cdot \dot{z}$ where A_R and B_R are prechosen.

(a3) If any such y_{jk} is found, let $y_{jk} = y_i$ be the largest such, and modify x_{jk} as follows:

$$\Delta x_{jk} = z_{jk}\cdot a_{i|n}\cdot \frac{r\cdot c}{(r-1)(c-1)} \qquad x_{jk} \to x_{jk} - \Delta x_{jk}\,.$$

(a4) Repeat steps (a1), (a2) and (a3), using successively modified x's until no y_{jk} deserves special attention. (Note that both the relation of i to (j, k) and the value of $\dot{z}$ will change for each of these FUNOR cycles.)

(a5) When this occurs, give special attention to all $z_{jk} = z_i$ with $z_i \geqq B_M\cdot\dot{z}$, where B_M is also prechosen (A_M is tacitly taken to be zero).

(a6) For each such jk put

$$\Delta x_{jk} = (z_{jk} - B_M\cdot\dot{z})a_{i|n}$$

and modify the corresponding x_{jk} by $x_{jk} \to x_{jk} - \Delta x_{jk}$ (steps (a5) and (a6) constitute the FUNOM cycle).

(a7) Output two two-way tables, one containing the finally modified x_{jk} and the other containing the accumulated modifications.

Certain points in this procedure deserve comment:

(b1) The value of Δx_{jk} in step (a3) includes a factor $rc/(r-1)(c-1)$, because a deviant value affects the corresponding fitted row, column, and grand means. As a result, the residual for a single very large deviation is about $(r-1)(c-1)/rc$ times as large as the deviation. The factor compensates for this. Thus the residual x_{jk} in the *next* cycle will be zero.

(b2) As the FUNOR cycles continue, the residuals for x_{jk} rejected in earlier FUNOR cycles will shift away from zero, but the likely shifts are usually small enough to make resetting them to zero not worthwhile.

(b3) The value of Δx_{jk} in step (a6) is chosen to approximately reduce z_{jk} to $B_M\cdot\dot{z}$. If we neglect changes in fitted means (row, column, and grand) as is more reasonable here, since the Δx_{jk} are considerably smaller, the effect on y_{jk} is

$$y_{jk} \to y_{jk} - (z_{jk} - B_M\cdot\dot{z})a_{i|n}$$

whence if the i corresponding to jk does not alter and $\dot{y}$ is negligibly small, $z_{jk} \to z_{jk} - (z_{jk} - B_M\cdot\dot{z}) = B_M\cdot\dot{z}$.

(b4) Thus each FUNOR cycle effectively rejects one entry, while the FUNOM cycle modifies those remaining entries with surprisingly large values.

The *approximate* overall result of such a procedure is:

(c1) to replace deviations of more than $A_R\cdot\sigma$ (with z's $\geqq B_R\sigma$) by zero, and

(c2) to reduce other deviations which were greater than $B_M\cdot a_{i|n}\cdot\sigma$ to that value.

These are approximately the results which can be supported, as noted above, if either (i) there are wild shots or (ii) the distribution is long-tailed.

TABLE 2

Original x_{rc} values : $1 \leq r \leq 36$, $1 \leq c \leq 15$

Row	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-0.045	0.151	-0.120	0.119	0.097	-0.023	0.022	-0.159	-0.010	-0.218	0.138	-0.099	-0.036	0.329	0.049
2	0.033	-0.210	-0.270	0.006	0.027	0.273	-0.093	-0.113	-0.641	0.124	0.059	-0.060	0.624	0.169	-0.172
3	-0.022	-0.060	-0.238	0.365	0.143	0.156	-0.120	0.094	-0.350	0.118	-0.581	0.120	-0.348	0.188	0.035
4	-0.115	-0.078	0.046	-1.033	-0.053	0.077	0.098	0.624	-0.099	-0.290	0.051	-0.532	-3.330	0.364	-0.241
5	-0.008	-0.301	0.054	1.127	-0.082	0.024	0.045	0.393	-0.066	0.233	0.116	-0.015	0.060	-0.154	0.175
6	0.114	-0.061	-0.007	-16.187	0.122	0.035	0.082	3.832	0.219	0.075	0.417	0.375	-0,902	0.143	0.254
7	0.097	-0.147	0.090	0.871	0.013	0.300	0.061	0.	0.407	0.419	-0.172	-0.168	-0.289	0.057	0.254
8	-0.019	0.005	0.177	0.234	0.085	0.195	0.159	0.757	-0.352	0.062	0.138	0.111	-0.510	0.215	0.161
9	-0.040	0.186	0.160	0.553	0.097	-0.004	0.005	-5.521	0.280	0.044	0.196	0.448	0.218	-0.297	0.015
10	0.116	-0.082	0.256	-0.065	-0.043	-0.038	0.155	0.245	0.000	0.219	0.235	-0.182	0.229	-0.078	0.063
11	0.147	-0.136	0.097	0.282	0.019	0.022	0.199	-0.188	0.375	0.141	0.116	-0.027	0.382	0.227	-0.117
12	-0.097	-0.352	0.193	0.581	0.108	0.070	-0.078	-0.473	0.105	0.781	0.242	0.040	0.183	-0.161	-0.025
13	0.049	0.024	0.078	0.074	-0.013	0.310	-0.032	-0.015	0.437	0.207	-0.021	-0.338	-0.281	0.301	0.035
14	0.095	0.207	0.130	-0.051	-0.031	0.111	-0.021	0.275	0.661	-0.074	0.216	0.049	-0.046	0.053	0.003
15	0.187	0.148	-0.001	0.642	0.050	-0.035	0.411	0.004	0.172	-0.292	0.039	-0.214	-1.191	0.312	0.312
16	-0.021	0.180	0.257	0.397	0.263	-0.273	0.272	0.357	0.054	-0.474	0.415	-0.906	-2.454	0.126	0.192
17	0.114	0.330	-0.180	0.675	0.029	0.095	0.178	0.700	0.037	0.417	0.046	0.410	1.306	0.448	0.026
18	0.088	-0.051	0.016	0.291	0.014	-0.018	0.037	0.350	0.327	0.128	0.056	0.065	-0.004	0.034	0.020
19	0.151	0.101	0.020	-0.612	-0.035	0.057	-0.021	-0.386	-0.207	-0.182	0.053	-0.016	0.150	-0.057	0.026
20	-0.006	0.324	0.022	-0.152	0.084	0.028	-0.085	0.774	0.397	-0.120	0.112	0.005	-0.048	-0.101	-0.087
21	0.053	0.127	-0.026	0.154	0.028	0.190	0.130	0.193	-0.011	0.201	0.213	-0.030	0.261	0.218	-0.035
22	0.170	0.096	-0.021	0.457	0.083	0.083	0.052	0.	-0.196	-0.116	0.201	-0.170	0.195	0.091	-0.230
23	-0.025	-0.072	-0.253	-0.086	0.004	0.004	-0.067	0.998	-0.264	0.300	0.281	0.016	0.555	0.109	0.354
24	0.146	-0.338	0.123	0.087	-0.017	-0.017	0.087	0.172	0.184	0.134	-0.273	0.022	-0.119	0.136	0.095
25	0.106	-0.363	0.187	-0.336	-0.025	-0.025	-0.002	0.499	1.081	0.555	0.426	-0.029	-1.073	0.343	0.636
26	0.070	0.028	0.136	0.058	0.135	0.079	-0.022	0.681	0.261	0.126	0.228	0.147	0.225	0.188	0.053
27	0.087	0.032	0.197	0.348	0.132	0.406	-0.055	0.103	0.198	0.384	0.058	-0.236	0.051	0.248	-0.152
28	0.091	-0.156	0.097	0.085	-0.019	0.108	0.072	0.040	-0.210	0.005	0.209	-0.247	-0.167	-0.134	0.001
29	0.087	0.172	0.129	-0.304	0.145	0.106	-0.542	0.116	0.145	-0.068	0.001	-0.539	-0.238	0.092	0.129
30	0.052	0.067	0.035	0.382	-0.007	0.170	-0.325	0.024	-0.488	0.129	0.055	0.	0.083	0.114	-1.021
31	0.091	-0.053	0.317	-0.202	0.036	0.105	-0.250	0.011	-2.211	0.376	0.169	0.	0.084	0.189	0.298
32	0,061	-0.182	0.054	-2.252	0.267	-0.046	-0.031	-0.166	-0.086	1.112	0.288	0.	-0.738	-0.003	0.134
33	-0.048	0.080	-0.054	-0.406	0.101	-0.120	-0.103	0.498	0.603	-0,909	0.381	-3.378	0.	0.766	0.280
34	0.065	-0.243	-0.027	0.213	0.037	0.053	0.087	1.393	0.195	-0.040	0.324	-0.176	0.	0.179	-0.092
35	-0.113	0.348	0.114	-1.476	-0.028	-0.070	-0.032	0.663	0.199	0.280	0.275	0.198	0.073	0.412	-0.040
36	0.126	0.260	-0.052	-1.430	-0.090	0.014	0.077	0.331	0.217	-0.084	0.271	-0.182	0.259	0.149	0.134

TABLE 3

FUNOR-FUNOM *Steps for data in Table 2*

FUNOR *Steps* ($A_R = 10, B_R = 1.5$)

Cycle	$\dot y$	$\dot z$	$10\dot z$	j, k	Δx_{jk}
1	+.009	0.243	2.435	6, 4	−16.49
2	+.021	0.228	2.277	9, 8	−5.95
3	+.020	0.222	2.222	6, 8	+3.55
4	+.019	0.214	2.136	33, 12	−3.34
5	+.018	0.207	2.067	4, 13	−3.08
6	+.017	0.200	2.004	16, 13	−2.42
7	+.017	0.192	1.923	31, 9	−2.31
8	+.012	0.190	1.901	32, 4	−2.30

FUNOM *Steps* ($A_M = 0, B_M = 1.5$)

i	$\dot y$	$\dot z$	$1.5\dot z$	number of entries modified
9	+.013	0.187	0.280	18 (at − end) 11 (at + end)

20. Example of use of FUNOR-FUNOM. An example of the application of this procedure to a 36 × 15 table consisting of values of a particular multiple regression coefficient in each of 540 = 36 × 15 subgroups (of small to moderate size) may clarify the situation. Table 2 presents the original data. (The 7 instances of "0" correspond to groups involving so few individuals that calculation of the corresponding multiple regression coefficients was not reasonable.) The very wild value in row 6 and column 4 is obvious from a brief scan over the entries.

Table 3 summarizes the 8 FUNOR steps and the concluding FUNOM step. The first FUNOR step directed itself to row 6, column 4 and altered the original x_{64} value to $-16.187 - (-16.49) = +.30$, a value then giving zero residual. The second FUNOR step directed itself to row 9, column 8. Its change of $-(-5.95)$ raised the grand mean by 0.01, which, in particular, lowered the x_{64} residual to $y_{64} = -0.01$. The third FUNOR step directed itself to row 6, column 8. Its change of $-(+3.55)$ altered the fitted mean for row 6 by $-.23$, and the grand mean by about $-.01$. One consequence was to alter the y_{64} residual to $y_{64} = -0.01 - (0.23 - 0.01) = -0.23$. And so on.

At the conclusion of the 8th FUNOR step, no z_{jk} is left with $y_i \geqq 10\dot z$ and $z_i \geqq 1.5\dot z$, so there is no need to "reject" further observations. The FUNOM cycle next modifies $18 + 11 \doteq 29$ further values to bring their z values down, approximately, to $1.5\dot z$. Figure 1 shows, for the 65 lowest values of i, the relation of the rounded z_i to $\dot z$, $2\dot z$, etc. We see $z_1 = z_2 = 2.9\dot z$, $z_3 = z_4 = 2.5\dot z$, $z_5 = 2.4\dot z, \cdots, z_{16} = z_{17} = z_{18} = 1.6\dot z$. The observations corresponding to these 18 values of $i = (j, k)$ are modified in the FUNOM step, along with 11 more for i

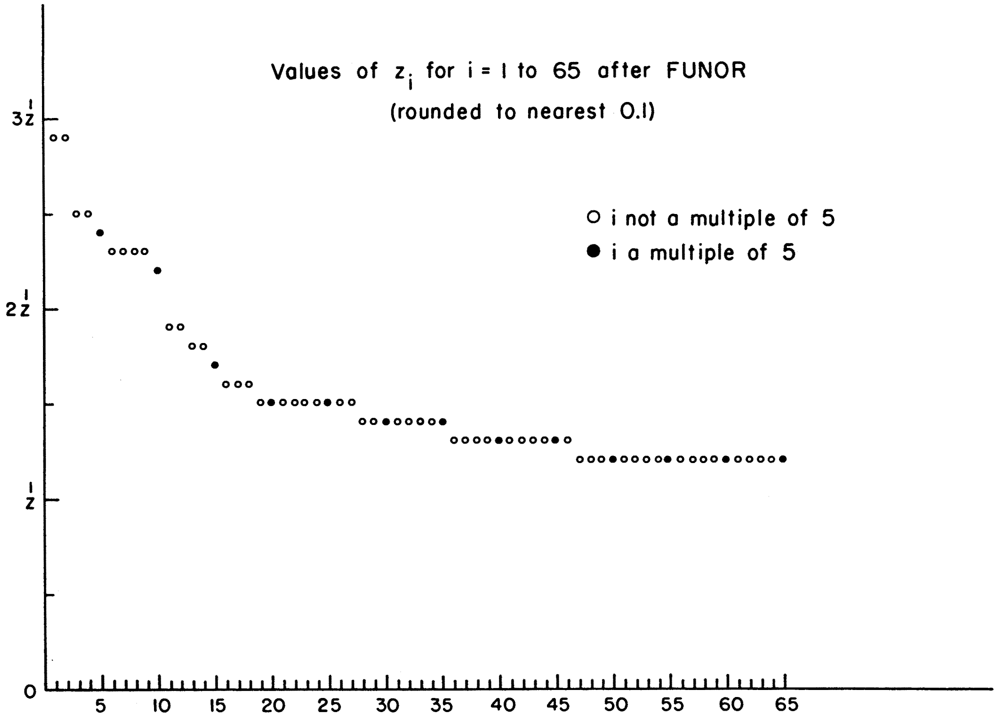

FIG. 1

near 540. The actual (j, k)'s involved can be determined from Table 4 which summarizes the changes for all cycles, which divided into

(1) 8 changes with $|\Delta x| \geqq 2.3$, one made in each FUNOR cycle

(2) 18 changes with $-.374 \leqq \Delta x \leqq -.006$, made in the FUNOM cycle for i near 1,

(3) 11 changes with $.015 \leqq \Delta x \leqq .323$, made in the FUNOM cycle for i near 540.

Table 5 exhibits the residuals for the results of final modification. Eye examination finds little exceptional about this table.

Table 6 exhibits the table of modified observations. Together with Table 4, which exhibits the modifications, Table 6 is the output for further use and study from FUNOR-FUNOM. Notice carefully that changing the original observations according to

$$(\text{observation})_{jk} \to (\text{observation})_{jk} + \Delta\alpha_j + \alpha\beta_k$$

would have had the following effects:

(1) Table 4: entirely unaffected (except for possible changes in roundoff)

(2) Table 6:

$$(\text{modified observation})_{jk} \to (\text{modified observation})_{jk} + \Delta\alpha_j + \Delta\beta_k .$$

TABLE 4

Values of Δx_{jk} : 8 from FUNOR, 29 from FUNOM

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
2	0.	0.	0.	0.	0.	0.	0.	0.	-0.097	0.	0.	0.	0.099	0.	0.
3	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-0.093	0.	0.	0.	0.
4	0.	0.	0.	-0.327	0.	0.	0.	0.	0.	0.	0.	0.	-3.083	0.	0.
5	0.	0.	0.	0.323	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
6	0.	0.	0.	-16.487	0.	0.	0.	3.553	0.	0.	0.	0.	-0.318	0.	0.
7	0.	0.	0.	0.097	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
8	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-0.006	0.	0.
9	0.	0.	0.	0.	0.	0.	0.	-5.949	0.	0.	0.	0.	0.	-0.026	0.
10	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
11	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
12	0.	0.	0.	0.	0.	0.	0.	-0.145	0.	0.054	0.	0.	0.	0.	0.
13	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
14	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
15	0.	0.	0.	0.018	0.	0.	0.	0.	0.	0.	0.	0.	-0.431	0.	0.
16	0.	0.	0.	0.	0.	0.	0.	0.	0.	-0.046	0.	-0.267	-2.419	0.	0.
17	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.331	0.	0.
18	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
19	0.	0.	0.	-0.033	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
20	0.	0.	0.	0.	0.	0.	0.	-0.310	0.	0.	0.	0.	0.	0.	0.
21	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
22	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
23	0.	0.	0.	0.	0.	0.	0.	0.093	0.	0.	0.	0.	0.	0.	0.
24	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
25	0.	0.	0.	0.	0.	0.	0.	0.	0.262	0.	0.	0.	-0.438	0.	0.
26	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
27	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
28	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
29	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
30	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	-0.327
31	0.	0.	0.	0.	0.	0.	0.	0.	-2.314	0.	0.	0.	0.	0.	0.
32	0.	0.	0.	-2.300	0.	0.	0.	0.	0.	0.297	0.	0.	-0.138	0.	0.
33	0.	0.	0.	0.	0.	0.	0.	0.	0.	-0.374	0.	-3.345	0.	0.015	0.
34	0.	0.	0.	0.	0.	0.	0.	0.276	0.	0.	0.	0.	0.	0.	0.
35	0.	0.	0.	-0.730	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.
36	0.	0.	0.	-0.665	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.

TABLE 5

Values of residuals for modified x_{jk}

Row	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-0.049	0.198	-0.120	0.105	0.098	-0.043	0.060	-0.330	-0.036	-0.274	0.048	0.010	0.048	0.237	0.054
2	0.058	-0.132	-0.241	0.021	0.057	0.283	-0.025	-0.255	-0.638	0.098	-0.003	0.078	0.737	0.101	-0.138
3	0.020	0.034	-0.192	0.398	0.190	0.183	-0.035	-0.031	-0.330	0.109	-0.625	0.274	-0.218	0.137	0.087
4	-0.011	0.078	0.154	-0.938	0.056	0.166	0.245	0.560	-0.017	-0.237	0.069	-0.316	-0.054	0.375	-0.128
5	-0.105	-0.347	-0.041	1.020	-0.175	-0.089	-0.010	0.127	-0.186	0.084	-0.069	-0.000	0.051	-0.345	0.086
6	0.027	-0.096	-0.091	0.202	0.039	-0.068	0.037	0.025	0.109	-0.064	0.243	0.400	-0.901	-0.038	0.176
7	-0.014	-0.205	-0.017	0.751	-0.093	0.174	-0.008	-0.278	0.274	0.257	-0.370	-0.167	-0.311	-0.147	0.152
8	-0.104	-0.028	0.095	0.139	0.005	0.094	0.116	0.504	-0.460	-0.075	-0.035	0.138	-0.507	0.036	0.084
9	-0.183	0.095	0.019	0.399	-0.042	-0.163	-0.096	0.117	0.114	-0.151	-0.035	0.416	0.162	-0.534	-0.119
10	0.056	-0.091	0.199	-0.135	-0.098	-0.113	0.137	0.017	-0.074	0.107	0.088	-0.130	0.257	-0.232	0.012
11	0.054	-0.178	0.007	0.178	-0.070	-0.087	0.148	-0.449	0.259	-0.005	-0.065	-0.008	0.376	0.040	-0.201
12	-0.162	-0.365	0.130	0.505	0.047	-0.011	-0.101	-0.706	0.018	0.664	0.090	0.087	0.206	-0.320	-0.082
13	0.006	0.033	0.038	0.021	-0.051	0.251	-0.063	-0.225	0.371	0.112	-0.152	-0.270	-0.236	0.164	0.001
14	0.039	0.203	0.077	-0.117	-0.083	0.039	-0.034	0.051	-0.018	-0.181	0.073	0.105	-0.013	-0.097	-0.044
15	0.159	0.173	-0.025	0.606	0.027	-0.078	0.427	-0.191	0.122	-0.370	-0.075	-0.129	-1.130	0.191	0.294
16	-0.065	0.187	0.216	0.343	0.223	-0.333	0.269	0.145	-0.013	-0.570	0.283	-0.838	0.009	-0.012	0.156
17	-0.186	0.083	-0.477	0.365	-0.266	-0.220	-0.079	0.233	0.286	0.066	-0.341	0.223	1.095	0.055	-0.265
18	0.007	-0.081	-0.062	0.200	-0.062	-0.114	-0.002	0.102	0.224	-0.005	-0.112	0.096	0.003	-0.141	-0.052
19	0.224	0.226	0.096	-0.549	0.043	0.114	0.094	-0.481	-0.157	-0.161	0.039	0.169	0.312	-0.078	0.108
20	0.029	0.411	0.061	-0.126	0.125	0.049	-0.007	-0.906	0.410	-0.136	0.061	0.153	0.077	-0.158	-0.043
21	-0.049	0.077	-0.125	0.042	-0.070	0.072	0.071	-0.076	-0.136	0.047	0.024	-0.019	0.247	0.023	-0.128
22	0.141	0.109	-0.055	0.410	0.050	0.030	0.056	-0.205	-0.256	-0.206	0.076	-0.095	0.246	-0.041	-0.259
23	-0.139	-0.135	-0.369	-0.210	-0.106	-0.126	-0.139	0.716	-0.400	0.134	0.080	0.013	0.529	-0.099	0.249
24	0.127	-0.306	0.107	0.058	-0.031	-0.052	0.110	-0.014	0.143	0.063	-0.379	0.116	-0.050	0.023	0.085
25	-0.017	-0.434	0.068	-0.469	-0.143	-0.164	-0.082	0.209	0.935	0.380	0.216	-0.039	-1.107	0.126	0.522
26	-0.081	-0.071	-0.011	-0.102	-0.011	-0.088	-0.130	0.363	0.088	-0.076	-0.009	0.108	0.163	-0.056	-0.088
27	-0.024	-0.027	0.090	0.227	0.026	0.280	-0.124	-0.176	0.065	0.221	-0.140	-0.235	0.028	0.043	-0.254
28	0.120	-0.075	0.130	0.105	0.105	0.015	0.121	0.144	-0.178	-0.203	-0.017	0.151	-0.050	-0.198	0.040
29	0.134	0.270	0.179	-0.267	0.196	0.137	-0.452	-0.004	0.170	-0.073	-0.039	-0.380	-0.102	0.045	0.185
30	0.110	0.177	0.096	0.430	0.055	0.212	-0.225	-0.086	-0.453	0.135	0.026	0.170	0.229	0.078	-0.954
31	0.015	-0.077	0.244	-0.287	-0.035	0.013	-0.284	-0.232	0.004	0.249	0.006	0.036	0.096	0.019	0.231
32	0.023	-0.169	0.018	0.000	0.233	-0.100	-0.027	-0.372	-0.146	1.022	0.163	0.074	-0.688	-0.136	0.105
33	-0.108	0.072	-0.111	-0.476	0.046	-0.195	-0.121	0.271	0.521	-1.021	0.234	0.019	0.028	0.613	0.229
34	-0.057	-0.314	-0.147	0.081	-0.081	-0.080	0.007	1.103	0.050	-0.214	0.115	-0.186	-0.034	-0.037	-0.205
35	-0.157	0.355	0.072	-1.530	-0.067	-0.130	-0.034	0.451	0.132	0.184	0.143	0.266	0.117	0.274	-0.075
36	0.164	0.350	-0.011	-1.402	-0.047	0.036	0.157	0.201	-0.201	-0.098	0.222	-0.031	0.385	0.093	0.181

TABLE 6

Final Values of Modified x_{jk}

Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	-0.045	0.151	-0.120	0.119	0.097	-0.023	0.022	-0.159	-0.010	-0.218	0.138	-0.099	-0.036	0.329	0.049
2	0.033	-0.210	-0.270	0.006	0.027	0.273	-0.093	-0.113	-0.544	0.124	0.059	-0.060	0.524	0.169	-0.172
3	-0.022	-0.060	-0.238	0.365	0.143	0.156	-0.120	0.094	-0.350	0.118	-0.487	0.120	-0.348	0.188	0.035
4	-0.115	-0.078	0.046	-0.706	-0.053	0.077	0.098	0.624	-0.099	-0.290	0.051	-0.532	-0.247	0.364	-0.241
5	-0.008	-0.301	0.054	0.804	-0.082	0.024	0.045	0.393	-0.066	0.233	0.116	-0.015	0.060	-0.154	0.175
6	0.114	-0.061	-0.007	0.299	0.122	0.035	0.082	0.280	0.219	0.075	0.417	0.375	-0.584	0.143	0.254
7	0.097	-0.147	0.090	0.775	0.013	0.300	0.061	0.	0.407	0.419	-0.172	-0.168	-0.289	0.057	0.254
8	-0.019	0.005	0.177	0.234	0.085	0.195	0.159	0.757	-0.352	0.062	0.138	0.111	-0.504	0.215	0.161
9	-0.040	0.186	0.160	0.553	0.097	-0.004	0.005	0.428	0.280	0.044	0.196	0.448	0.218	-0.271	0.015
10	0.116	-0.082	0.256	-0.065	-0.043	-0.038	0.155	0.245	0.008	0.219	0.235	-0.182	0.229	-0.078	0.063
11	0.147	-0.136	0.097	0.282	0.019	0.022	0.199	-0.188	0.375	0.141	0.116	-0.027	0.382	0.227	-0.117
12	-0.097	-0.352	0.193	0.581	0.108	0.070	-0.078	-0.327	0.105	0.727	0.242	0.040	0.183	-0.161	-0.025
13	0.049	0.024	0.078	0.074	-0.013	0.310	-0.062	-0.015	0.437	0.207	-0.021	-0.338	-0.281	0.301	0.035
14	0.095	0.207	0.130	-0.051	-0.031	0.111	-0.021	0.275	0.061	-0.074	0.216	0.049	-0.046	0.053	0.003
15	0.187	0.148	-0.001	0.624	0.050	-0.035	0.411	0.004	0.172	-0.292	0.039	-0.214	-0.759	0.312	0.312
16	-0.021	0.180	0.257	0.397	0.263	-0.273	0.272	0.357	0.054	-0.427	0.415	-0.639	-0.035	0.126	0.192
17	0.114	0.330	-0.180	0.675	0.029	0.095	0.178	0.700	0.037	0.417	0.046	0.410	0.976	0.448	0.026
18	0.088	-0.051	0.016	0.291	0.014	-0.018	0.037	0.350	0.327	0.128	0.056	0.065	-0.004	0.034	0.020
19	0.151	0.101	0.020	-0.579	-0.035	0.057	-0.021	-0.386	-0.207	-0.182	0.053	-0.016	0.150	-0.057	0.026
20	-0.006	0.324	0.022	-0.152	0.084	0.028	-0.085	-0.465	0.397	-0.120	0.112	0.005	-0.048	-0.101	-0.087
21	0.053	0.127	-0.026	0.154	0.028	0.190	0.130	0.193	-0.011	0.201	0.213	-0.030	0.261	0.218	-0.035
22	0.179	0.096	-0.021	0.457	0.083	0.083	0.052	0.	-0.196	-0.116	0.201	-0.170	0.195	0.091	-0.230
23	-0.025	-0.072	-0.258	-0.086	0.004	0.004	-0.067	0.905	-0.264	0.300	0.281	0.016	0.555	0.109	0.354
24	0.146	-0.338	0.123	0.087	-0.017	-0.017	0.087	0.172	0.184	0.134	-0.273	0.022	-0.119	0.136	0.095
25	0.106	-0.363	0.187	-0.336	-0.025	-0.025	-0.002	0.499	0.819	0.555	0.426	-0.029	-0.635	0.343	0.636
26	0.070	0.028	0.136	0.058	0.135	0.079	-0.022	0.681	0.261	0.126	0.228	0.147	0.225	0.188	0.053
27	0.087	0.032	0.197	0.348	0.132	0.406	-0.055	0.103	0.198	0.384	0.058	-0.236	0.051	0.248	-0.152
28	0.091	-0.156	0.097	0.085	-0.019	0.108	0.072	-0.040	-0.210	0.005	0.209	-0.247	-0.167	-0.134	0.001
29	0.087	0.172	0.129	-0.304	0.145	0.106	-0.542	0.116	0.145	-0.068	0.001	-0.539	-0.238	0.092	0.129
30	0.052	0.067	0.035	0.382	-0.007	0.170	-0.325	0.024	-0.488	0.129	0.055	0.	0.083	0.114	-0.694
31	0.091	-0.053	0.317	-0.202	0.036	0.105	-0.250	0.011	0.103	0.376	0.169	0.	0.084	0.189	0.298
32	0.061	-0.182	0.054	0.048	0.267	-0.046	-0.031	-0.166	-0.086	0.815	0.288	0.	-0.600	-0.003	0.134
33	-0.048	0.080	-0.054	-0.406	0.101	-0.120	-0.103	0.498	0.603	-0.535	0.381	-0.033	0.	0.752	0.280
34	0.065	-0.243	-0.027	0.213	0.037	0.058	0.087	1.117	0.195	-0.040	0.324	-0.176	0.	0.179	-0.092
35	-0.113	0.348	0.114	-0.746	-0.028	-0.070	-0.032	0.663	0.199	0.280	0.275	0.198	0.073	0.412	-0.040
36	0.126	0.260	-0.052	-0.766	-0.090	0.014	0.077	0.331	-0.217	-0.084	0.271	-0.182	0.259	0.149	0.134

Thus the fact that row and column effects were weak in this example is wholly unimportant. If large row and column effects had been superposed, they would have been transferred to Table 6 without change, and the same Table 4 would have emerged.

FUNOR-FUNOM is a procedure which fills a need hitherto not at all satisfied. It would be nice to know more about its properties, its possible competitors, and how the competition comes out. It would be nice to know the consequences of different choices for A_R , B_R , A_M , and B_M . But pending the necessary effort, in which all are invited to participate, it would, in my judgment, be *foolish not to use it.* It is reasonably clear that its use will lead to far more gain than loss, even in rather unreasonable hands.

IV. MULTIPLE-RESPONSE DATA

21. Where are we, and why? Multivariate analysis has been the subject of many pages of papers, and not a few pages of books. Yet when one looks around at the practical analysis of data today one sees few instances where multiple responses are formally analyzed in any way which makes essential use of their multiplicity of response. The most noticeable exception is factor analysis, about which we shall make some more specific remarks below. There is, to be sure, a fairly large amount of data-analysis by multiple regression, some quite sophisticated and incisive, much quite bookish and bumbling. (Cochran's remark of a number of years ago that "regression is the worst taught part of statistics" has lost no validity.) But ordinary multiple regression is overtly not multiple-*response,* no matter how multivariate it may be.

Why is it that so much formalized and useful analysis of single-response data, and of single-response aspects of multiple-response data, is accompanied by so little truly multiple-response analysis? One interesting and reasonable suggestion is that it is because multiple-response procedures have been modeled upon how early single-response procedures were supposed to have been used, rather than upon how they were in fact used.

Single-response techniques started as significance procedures designed in principle to answer the question: "is the evidence strong enough to contradict a hypothesis of no difference?", or perhaps, as became clearer later, the more meaningful question: "is the evidence strong enough to support a belief that the observed difference has the correct sign?" But in early use by the naive these techniques, while sometimes used for the second purpose, were often used merely to answer "should I believe the observed difference?". While this last form is a misinterpretation to a statistician, its wide use and the accompanying growth of the use of such techniques, suggests that it was useful to the practitioner. (Perhaps the spread of confidence procedures will go far to replace it by a combination of technique and interpretation that statisticians can be happy with. Cp., e.g., Roy and Gnanadesikan, 1957, 1958.)

But there was one essential to this process that is often overlooked. Single-response differences are (relatively) simple, and the practitioner found it moderately easy to think, talk, and write about them, either as to sign, or as to sign and amount. Even when fluctuations have negligible influence, multiple-response differences are not simple, are usually not easy to think about, are usually not easy to describe. Better ways of description and understanding of multiple-response differences, and of multiple-response variabilities, may be essential if we are to have wider and deeper use of truly multiple-response techniques of data analysis. While it can be argued that the provision of such ways of description is not a problem of statistics, if the latter be narrowly enough defined, it is surely a central problem of data analysis.

Let me turn to a problem dear to the heart of my friend and mentor, Edgar Anderson, the shape of leaves. It is, I believe, fair to say that we do not at present have a satisfactory way of describing to the *mind* of another person either:

(a1) the nature of the difference in typical pattern (shape and size) of two populations of leaves, or

(a2) the nature of the variability of dimensions for a single population of leaves.

Photographs, or tracings, of shapes and sizes of random samples may convey the information to his *eyes* fairly well, but we can hardly, as data analysts, regard this as satisfactory summarization—better must be possible, but how?

In view of this difficulty of description, it is not surprising that we do not have a good collection of ideal, or prototype multivariate problems and solutions, indeed it is doubtful if we have even one (where many are needed). A better grasp of just what we want from a multivariate situation, and why, could perhaps come without the aid of better description, but only with painful slowness.

We shall treat only one specific instance of the multiple-response aspects of analysis of variance. (This instance, in Section 25 below, combines the boiling-down and factor-identification techniques of factor analysis with the patterned-observation and additive-decomposition techniques of analysis of variance.) A variety of schemes for the analysis of variance of multiple-response data have been put forward (cp., e.g., Bartlett, 1947; Tukey, 1949a; Rao, 1952; Roy, 1958; Rao, 1959), but little data has been analyzed with their aid. A comparative study of various approaches *from the point of view of data analysis* could be very valuable. However, such a study probably requires, as a foundation, a better understanding of what we really want to do with multiple-response data.

It seems highly probable that as such better ways of description are developed and recognized there will be a great clarification in what multiple-response analysis needs in the way of inferential techniques.

22. The case of two samples. It would not be fair to give the impression that there is no use of truly multiple-response techniques except in factor analysis. There are a few other instances, mainly significance testing for two groups and simple discriminant functions.

Hotelling's T^2 serves the two-sample problem well in many cases. Only the need for lengthy arithmetic and the inversion of a matrix stands in its way. The requirement for lengthy arithmetic, both in accumulating sums of products and squares, and in inverting the resulting matrix, which is probably inevitable for a procedure that is affine-invariant (i.e., one whose results are unchanged by the introduction of new responses that are non-singular linear combinations of the given responses), used to stand in the way of its use. More and more of those who have multivariate data to analyze now have access to modern computers, and find the cost of arithmetic low. Those who are not in so favorable a state may benefit substantially from the use of validity-conserving "quick" methods, specifically from applying to more simply obtainable "relatively highly apparent" comparisons, the critical values which would be appropriate to affine-invariant "most apparent" comparison, and which are therefore generally applicable. Such procedures would still be able to establish significance in a large proportion of those cases in which it is warranted. Such "quick" multivariate techniques would not be thorough in establishing lack of significance, but would be far more thorough than would application, at an appropriate error-rate, of a single-response procedure to each component of the multiple response.

Let us sketch how two or three such "quick" multiple-response procedures might be constructed. First, consider calculating means and variances for each component, applying Penrose's method (1947) to obtain an approximate discriminant function, computing values of this discriminant for every observation, calculating Student's t for these values and then referring t^2 to the critical values of Hotelling's T^2 to test the result. A simpler version, under some circumstances, might be to use sample ranges in place of sample variances. A more stringent version would calculate a Penrose discriminant, regress each component (each original response) on this to obtain means and variances of residuals, extract a second Penrose discriminant from the residuals, and combine the two discriminants before testing. Each of these procedures is feasible for high multiplicity of response, since their labor is proportional to only the first power of the number of components, whereas Hotelling's T^2 requires labor in collection of SS and SP proportional to the square of this number, and labor in the matrix inversion proportional to its cube.

For very large numbers of components, as compared to the number of observations, it is necessary to give up affine invariance. Dempster's method (1958, 1960), which involves judgment selection of a measure of size, is applicable in such cases and may well also prove useful in many cases where Hotelling's T^2 would be feasible. There is always a cost to providing flexibility in a method. Affine-invariance implies an ability of the analysis to twist to meet any apparent ellipsoid of variability. Unless the number of cases is large compared to the number of responses, this flexibility requires loose critical values. For moderate numbers of cases, judgment "sizes" may give greater power than would affine invariance. (The observations are still being asked to provide their own error term, thus validity can be essentially the same for the two approaches.)

Many statisticians grant the observer or experimenter the right to use judgment in selecting the response to be studied in a single-response analysis. (Judgment is badly needed in this choice.) It is no far cry to the use of judgment, perhaps by the two parties together, in selecting a "size" for analysis in a multiple-response situation. In this connection, the promising work of Wilk and Gnanadesikan (1962) in using multiple-response sizes to generalize the half-normal-plot analysis for single-response data must be pointed out.

There are many ways to use flexibility in opening up the practice of the analysis of multiple-response data. Few, if any, involve matrices "six feet high and four feet wide", or the solution of very complex maximum likelihood equations.

23. Factor analysis: the two parts. Factor analysis is a procedure which has received a moderate amount of attention from statisticians, and rather more from psychometricians. Issues of principle, of sampling fluctuation, and of computational ease have been confounded. By and large statisticians have been unsatisfied with the result.

Any reasonable account of factor analysis from the data analyst's point of view must separate the process into two quite distinct parts. Every type of factor analysis includes a "boiling-down" operation in which the dimensionality of the data is reduced by introducing new coordinates. In most cases this process is followed, or accompanied, by a "rotation" process in which a set of new coordinates are located which might be believed to approximate a set with some intrinsic meaning. The problems connected with the second phase are rather special, and tend to involve such questions, upon which we should not enter here, as whether simple structure is a hypothesis about the tested, or about the test-makers.

The "boiling-down" process is something else again. If we strip it of such extraneous considerations as:

(a1) a requirement to extract as much useful information as possible with as *few* coordinates as possible,

(a2) a requirement to try to go far enough but not one bit further,

we find that it is an essentially simple problem of summarization, one with which most, if not all, statisticians can be content (so long as not too much is claimed for it). If we start with 40 responses, and really need only 8 coordinates to express the 'meat' of the matter, then finding 10, 12, or even 15 coordinates which are almost sure to contain what is most meaningful is a very useful process, and one not hard to carry out. (The early stages of almost any method of factor analysis will serve, provided we do not stop them too soon.) And, before we go onward beyond such rough summarization, it may be very much worth while to investigate how changes in mode of expression, either of the original coordinate or the new ones, may lead to a further reduction in dimensionality. Adequate packaging, and broad understanding, of effective methods of "boiling down, but not too far" which are either simple or computationally convenient could contribute much to the analysis of highly multiple-response data.

24. Factor analysis: regression. Whatever the conclusion as to the subject-matter relevance and importance of the "rotation" part of the process, we may, as statisticians, be sure of its demand for more and more data. For it is essentially based, at best, upon the detailed behavior of multivariate variance components. It is commonplace that variance-component techniques applied to single responses will require large samples. When they are applied to multiple responses, their demands for large samples cannot help being more extensive. Consequently there must be a constant attempt to recognize situations and ways in which techniques basically dependent upon regression rather than variance components can be introduced into a problem, since regression techniques always offer hopes of learning more from less data than do variance-component techniques.

When we have repeated measurements, we can, if we wish, make a multivariate analysis of variance, and factor-analyze a variance component for individuals (or tests) rather than factor-analyzing the corresponding mean square (or sum of squares). It has already been pointed out that it is better to turn to the variance component (Tukey, 1951) in such instances.

A somewhat related situation arises when the multiple responses are accompanied by a (small) number of descriptive variables. We clearly can replace the original responses by residuals after regression on descriptive variables before proceeding to factor analysis. Should we?

If this question has a general answer, it must be that we should. For a description of the variation of the multiple responses in terms of a mixture of

(a1) regression upon descriptive variables, and

(a2) variation of residual factors,

will convey much more meaning than the results of a simple factor analysis, if only because of the lower sensitivity of the regression coefficients to fluctuations. And the residual factor analysis will often be far more directly meaningful than the raw factor analysis. Consider a study of detailed aspects of children's personalities, as revealed in their use of words. Suppose the sexes of the individual children known. Elimination of the additive effect of sex would almost surely lead to more meaningful "factors", and elimination of reading speed as well, if available for elimination, might lead us to even closer grips with essentials.

25. Factor analysis: the middle lines. An instance which illustrates this point, and is related to the earlier one, is provided by the Management Attitude Survey which has been given on a completely anonymous basis to tens of thousands of Bell System supervisors. Within the limits of anonymity it is possible to obtain data upon sex, department, level of supervision, and (geographical) operating area for each respondent. Sex and department are interrelated, sometimes strongly, so that it seems advisable to treat department, sex, and job level as a single composite factor, of which, for some analyses, it is reasonable to consider only the 15 most popular such combinations as the versions, while some 36 (partly consolidated) geographical zones serve as versions of a second factor.

If we had only a single response, we should want to make an analysis of variance according to the skeleton

Item	df
Department, Sex, Job level	14
Geographical zones	35
Interaction	490
Within	tens of thousands

If we are to study the multiple responses by a factor analysis, we have strong reasons for excluding the additive effect of sex, and likely the additive effects assignable to combinations of department, job level, and geographical location. Thus we should be careful to omit the corresponding upper variance components from the analysis. The argument about the tens of thousand of degrees of freedom for differences among people identified alike is more subtle, depending in part (but only in part) on whether the purpose of the analysis is to get at basic psychology or to learn about Bell System supervisors. Granted that it is the latter, it is appropriate to avoid the variance component within supervisors identified alike, and to concentrate our attention upon the middle line(s) of the analysis of variance. If we can face the computational labor, we should factor analyze the multivariate *variance component* corresponding to the 490 degrees of freedom for the interaction line; otherwise we should factor analyze the corresponding *mean square*, in which the variance component which we should like to factor is combined with a fraction of the within component. (Before saying that such an analysis is wholly unsatisfactory, we must note how much less of the within component is combined in the middle-line mean square than had we analysed a random subsample of individuals, or had separated out only the "upper lines" before analysis.)

In practice factoring the mean square would mean forming the mean of each response for each of the $15 \times 36 = 540$ groups of supervisors identified alike, and then forming residuals by adjusting for Department-Sex-Job-level means and Geographic zone means. The residuals, one for each response for each group of supervisors identified alike, would then be subjected to some form of factor analysis.

This example has been mentioned not so much for its intrinsic importance, but rather because it illustrates a joining of attitudes and techniques some of which are traditionally associated with single-response analysis while others are limited to multiple-response analysis. The difficulties in dealing with this set of data are not matters of significance, confidence, estimation, or of statistics narrowly defined. Rather they are matters of how to convert a mass of data into possibly meaningful numbers, which is surely the first, though not the only, task of data analysis.

Another way to relate factor analysis, as practiced with psychometric batteries, to single-response techniques more familiar to the statistician has been pointed out by Creasy (1957). Eysenck (1950, 1952) has proposed, under the name of "criterion analysis", a different sort of interplay between external variables and factor analysis, which is intended to aid in improving the definition of the external variables.

26. Taxonomy; classification; incomplete data. Problems which can be fitted under the broad heading of taxonomy (or, if you must, nosology) are truly multiple response problems, whether plants, retrievable information, or states of ill-health are to be classified. It is already clear that there are a variety of types of problems here. In plants, for example, where the separation of species does not usually call for formal data analysis, the work of Edgar Anderson (e.g., 1949, 1957) has shown how even simple multiple response techniques (e.g., Anderson, 1957) can be used to attack such more complex questions as the presence and nature of inter-specific hybridization.

The rise of the large, fast, and (on a per-operation basis) cheap electronic computer has opened the way to wholly mechanized methods of species-finding. Here biologists have had a prime role in attacking the problem (cp., Sneath, 1957a, 1957b; Michener and Sokal, 1957; Sneath and Cowan, 1958; Rogers and Tanimoto, 1960).

Formal and semiformal techniques for identifying "species" from multiple-response data are certain to prove important, both for what they will help us learn, and for the stimulus their development will give to data analysis.

Once "species" are identified, the problem of assigning individuals to them is present. There is a reasonable body of work on the classical approaches to discrimination, but much remains to be done. The possibilities of using the Penrose technique repeatedly (see Section 22 above) have neither been investigated or exploited. And the assessment, as an end product, for each individual of its probabilities of belonging to each "species", in place of the forcible assignment of each individual to a single species, has, to my knowledge, only appeared in connection with so-called *probability weather forecasts.* (For an example of assessment as an expression of the state of the evidence, see Mosteller and Wallace, 1962.)

When we realize that classification includes medical diagnosis, and we recognize the spread of affect, from naive optimism to well-informed hope for slow gains as a result of extensive and coordinated effort, with which the application of electronic computer to medical diagnosis is presently being regarded by those who are beginning to work in this field, we cannot find the classification problem simple or adequately treated.

Once either taxonomy or classification is an issue, the problem of incomplete data arises. This is particularly true in medical diagnosis, where no patient may have had all the tests. There is a small body of literature on the analysis of incomplete data. Unfortunately it is mostly directed to the problems of using incomplete data to estimate population parameters. Such results and techniques can hardly be used in either taxonomy or classification. Great progress will be required of us here also.

V. SOME OTHER PROMISING AREAS

27. Stochastic-process data. The number of papers about the probability theory of stochastic processes has grown until it is substantial, the number of papers about statistical inference in stochastic processes is seeking to follow the same pattern, yet, with a few outstanding exceptions, there does not seem to be anything like a comparable amount of attention to *data analysis* for stochastic processes. About the cause of this there can be quite varied views. A tempting possibility is that we see the ill effects of having the empirical analysis of data wait upon theory, rather than leading theory a merry chase.

Techniques related to circular frequencies, cosines, or complex exponentials, and to linear, time-invariant black boxes are one of the outstanding exceptions. The estimation of power spectra has proved a very useful tool in the hands of many who work in diverse fields. Its relation to variance components had been discussed quite recently (Tukey, 1961a). Its more powerful brother, the analysis of cross-spectra, which possesses the strength of methods based upon regression, is on its way to what are likely to be greater successes. All these techniques involve quadratic or bilinear expressions, and their variability involves fourth-degree expressions. Their use is now beginning to be supplemented by quite promising techniques associated with (individual, pairs of, or complex-valued) linear expressions, such as (approximate) Hilbert transforms and the results of complex demodulation.

The history of this whole group of techniques is quite illuminating as to a plausible sequence of development which may appear in other fields:

(al) the basic ideas, in an imprecise form, of distribution of energy over frequency became a commonplace of physics and engineering,

(a2) a belief in the possibility of exact cycles, of the concentration of energy into lines, led to the development of such techniques as Schuster's periodograms and the various techniques for testing significance in harmonic analysis (e.g., Fisher, 1929),

(a3) an abstract theory of generalized harmonic analysis of considerable mathematical subtlety was developed by Wiener (e.g., 1930).

(a4) various pieces of data were analyzed in various ways, mostly variously unsatisfactory,

(a5) a Bell Telephone Laboratories engineer wanted to show a slide with a curve which represented the rough dependence of energy upon frequency for certain radar, and had some data analyzed,

(a6) examination showed a curious alternation of the estimates above and below a reasonable curve, and R. W. Hamming suggested (0.25, 0.50, 0.25) smoothing. (The striking success of this smoothing seized the writer's atten·tion and led to his involvement in the succeeding steps.)

(a7) so the 4th degree theory of the variability and covariability of the estimates was worked out for the Gaussian case, to considerable profit,

(a8) gradually the simpler 2nd degree theory for the average values of spectral estimates, which does not involve distributional assumptions, came to be recognized as of greater and greater use,

(a9) more recently we have learned that still simpler 1st degree theory, especially of complex demodulation (cp., Tukey, 1961a, pp. 200–201), offers new promise,

(a10) and the next step will call for investigation of the theory, roughly of 2nd degree, of the variability of the results of such techniques.

In this history note that:

(b1) decades were lost because over-simple probability models in which there was a hope of estimating everything were introduced and taken seriously (in step a2),

(b2) the demands of actual data analysis have driven theoretical understanding rather than vice versa (e.g., cp., steps a6, a7, a9, a10),

(b3) by and large, the most useful simpler theory was a consequence of more complex theory, although it could have been more easily found separately (e.g., cp., steps a7 and a8).

There seems to be no reason why we should not expect (b1) (b2) and (b3) to hold in other areas of stochastic-process data analysis.

If I were actively concerned with the analysis of data from stochastic processes (other than as related to spectra), I believe that I should try to seek out techniques of data processing which were not too closely tied to individual models, which might be likely to be unexpectedly revealing, and which were being pushed by the needs of actual data analysis.

28. Selection and screening problems. Data analysis, as used here, includes planning for the acquisition of data as well as working with data already obtained. Both aspects are combined in multistage selection and screening problems, where a pool of candidates are tested to differing extents, and the basic questions are those of policy. How many stages of testing shall there be? How much effort shall be spent on each? How is the number of candidates passed on from one stage to another to be determined?

This subject has been studied, but the results so far obtained, though quite helpful, leave many questions open. (Cp., Dunnett, 1960, for some aspects; Falconer, 1960, and Cochran, 1951, for others.) This is in part due directly to the analytical difficulty of the problems, many of whose solutions are going to require either wholly new methods, or experimental simulation. An indirect effect of analytical difficulty is that available solutions refer to criteria, such as "mean advance" which do not fit all applications.

Not only do problems of this class occur in widely different fields—selecting scholarship candidates, breeding new plant varieties, screening molds for antibiotic production are but three—but the proper criterion varies within a field of application. The criterion for selecting a single national scholarship winner should be different from that used to select 10,000 scholarship holders; the criterion for an antibiotic screen of earth samples (in search of new species of antibiotic producers) should be different from that for an antibiotic screen of radiation-induced mutants (in search of a step toward more efficient producers); and so on.

Practical selection (cp., e.g., page 223 in Lerner, 1954) has long made use of selection for the next stage, not of a fixed number, nor of a fixed fraction, but of those whose indicated quality is above some apparent gap in indicated quality. It is reasonable to believe (as I for one do) that such flexible selection is more efficient than any fixed % methods. But, to my knowledge, we have no evidence from either analytic techniques or empirical simulation as to whether this is indeed the case. This is but one of many open questions.

29. External, internal, and confounded estimates of error. The distinction between external and internal estimates of error is a tradition in physical measurement, where external estimates may come from comparisons between the work of different investigators, or may even be regarded as requiring comparisons of measurements of the same quantity by different methods. A similar distinction is of course involved in the extensive discussions of "the proper error term" in modern experimental statistics. No one can consider these questions to be of minor importance.

But there is another scale of kinds of error estimate whose importance, at least in a narrower field, is at least as great; a scale which can be regarded as a scale of subtlety or a scale of confusion. The first substantial step on this scale may well have been the use of "hidden replication" (cp., Fisher, 1935) in a factorial experiment as a basis of assessing variability appropriate as a measure of error. This can be regarded, from one standpoint, as merely the use of residuals from an additive fit to assess the stability of the fitted coefficients. As such it would not be essentially different from the use of residuals from a fitted straight line. But if, in this latter case, the fluctuations vary about a crooked line, we know of no sense in which the residuals, which include both variability and crookedness, are a specifically appropriate error term. In the factorial case, however, the presence of arbitrary interactions does not affect the correctness of the error term, provided the versions of each factor are randomly sampled from a population of versions (e.g., Cornfield and Tukey, 1956) whose size is allowed for. Thus "hidden replication" may reasonably be regarded as a substantial step in subtlety.

The recovery of inter-block information in incomplete block experiments (including those in lattices) is another stage, which has been extended to the recovery of inter-variety information (Bucher, 1957).

But the latest step promisės much. Until it was taken, we all felt that error should be assessed in one place, effects in another, even if different places were merely different *pre-chosen* linear combinations of the observations. The introduction of the half-normal plot (cp., Daniel, 1959) has shown us that this need not be the case, that, under the usual circumstance where many effects are small, while a few are large enough to be detected, we may confound *effects* with *error* and still obtain reasonable analyses of the result.

30. The consequences of half-normal plotting. The importance of the introduction of half-normal plotting is probably not quite large enough to be regarded as marking a whole new cycle of data analysis, although this remains to be seen. The half-normal plot itself is important, but the developments leading out of it are likely to be many and varied.

The work of Wilk and Gnanadesikan (1962) on multivariate extensions of the half-normal plot offers real promise of "shaking up" the analysis of multiple-response data. And it may well come in due course to modify our original treatment of the half-normal-plot situation itself.

A substantial part of the stimulus for the FUNOP, FUNOR, FUNOM approach to spotty data illustrated in Sections 18 to 20 came from the existence and use of the half normal plot.

The extension of half-normal technique to mean squares with varied numbers of degrees of freedom is inevitable, and in progress. What is not clear are the likely consequences of its by-products.

The introduction of random balance experimentation (cp., Various Authors, 1959) has made some attention to "super-saturated" experimental patterns inevitable. Adequately incisive analysis, in terms of what is possible, must turn to something like the half-normal plot in the sense that "error" will have to be estimated from contrasts which may contain contributions from real effects. (Some such procedures, such as that suggested by Beale and Mallows, 1958, are far from being graphical.) Note that not only situations supersaturated with main effects are in question here, but also situations with main effects reasonably clear of one another where the supersaturation is associated with two-factor interactions.

Undoubtedly still other important chains of growth will spring from the half normal plot.

These techniques, like the half-normal plot itself, will begin with indication, and will only later pass on to significance, confidence, or other explicitly probabilistic types of inference. When they do pass on, the resulting probabilistic treatments are relatively certain to be admittedly approximate. The hypothesis that fluctuations-and errors are exactly normally distributed, though demonstrably contrary to fact, is much more likely to be accepted without explicit question than is the hypothesis that a specific finite set of effects behave exactly like a sample from a normal distribution. Yet the latter is the *type* of assumption needed to make an exact probabilistic treatment of the half-normal plot, or of its progeny, exact rather than approximate. At this juncture the everpresent approximateness of data analysis must be more completely faced.

31. Heterogeneous data. Some sort of homogeneity of fluctuation-and-error behavior is a highly conventional, even stylized, part of the development of most procedures of data analysis. In a few cases we have come to recognize the extent to which it is needed and the role it plays. In balanced analyses of variance, for example, homogeneity of variance plays no role in determining the average values of mean squares, but can substantially affect their variability, and thus affect the exactness of F-tests of ratios of mean squares. The validity of the analysis-of-variance table as an indicator, which is usually its greatest importance, is unaffected by heterogeneity of variance, though conventional statements of significance and confidence may become only approximate in its presence.

There are at least two reasons why our attention, which we have often comfortably kept away from problems of heterogeneity of variation, is going to be brought into focus upon them:

(a1) the rise of procedures for dealing with spotty data, and

(a2) the rise of half-normal-plot-like techniques.

The removal, actual or effective, of observations which appear to be "wild shots" cannot be the whole result of an effective procedure for dealing with spotty data. *For the most useful information is not infrequently found to reside in the apparent wild shots themselves.* Consequently, good procedures for spotty data will provide two separate and distinct outputs, one consisting of cleaned-up observations, or of results based on such, while the other describes the apparent wildnesses.

We noted above that, besides the extreme deviations occurring rarely in any event, apparent "wild shots" can come from such varied sources as occasionally-acting causes, long-tailed distributions of deviations and errors, *and* inhomogeneity of variance. Which of these was the actual source for a specific "wild shot" will be a matter of concern in many instances. In seeking answers we shall have to face up to many more situations of inhomogeneous variability.

The case of half-normal-plot-like techniques is not too different. They are specifically directed toward the detection of those "occasionally-acting causes which are associated with namable effects." They cannot avoid bringing up occasional instances of purely accidental fluctuations-and-errors of extreme size. And they have to be more sensitive to situations of heterogeneous variability than do more omnibus methods. Their use will also bring us more frequently face to face with heterogeneity of variability.

32. Two samples with unequal variability. Why is the treatment of heterogeneous variability in such poor shape? In part, perhaps, because we are stuck at the early stage of the Behrens-Fisher problem. Analytical difficulties are undoubtedly a contributing cause, but it is doubtful that they would have been allowed to hold us up if we had reached a clear conclusion as to what do with the two-sample problem when the variances are to be separately estimated with appreciable uncertainty.

While it has been the fashion to act as though we should solve this problem in terms of high principle, either a high principle of "making use of all the information" or a high principle of "making exact probability statements", it now seems likely that we may, in due course, decide to settle this question in a far more realistic way. Both the highly-principled approaches before us require *precise* normality of the two parent distributions. We know this condition is not met in practice, but we have not asked how much the case of non-normal populations can tell us about what technique can reasonably be applied, or whether either of the well-known proposals is reasonably robust. We should ask, and think about the answer.

Martin\`Wilk is vigorous in pointing out that solving the two-sample *location* problem will not solve the problem of comparing two samples whose variances are likely to differ. He considers that the crux of the issue lies beyond, where both difference in variance and difference in mean are equally interesting to describe and test. Discussion of this problem, like that of so many left untouched here, would open up further interesting and rewarding issues for which we have no space.

While no clear general line has yet been identified, along which major progress in meeting heterogeneity of variability is likely to be made, there are a number of clear starting points.

VI. FLEXIBILITY OF ATTACK

33. Choice of modes of expression. Clearly our whole discussion speaks for greater flexibility of attack in data analysis. Much of what has been said could be used to provide detailed examples. But there are a few broader issues which deserve specific mention.

Shall the signal strength be measured in volts, in volts^2 (often equivalent to "watts"), in $\sqrt{\text{volts}}$, or in log volts (often described as "in decibels")? As a question about statement of final results this would not be for data analysis and statisticians to answer alone, though they would be able to assist in its answering. But, as a question about how an analysis is to be conducted, it is their responsibility, though they may receive considerable help from subject-matter experts.

This is an instance of a choice of a mode of expression for a single response, a question about which we now know much more than we did once upon a time. Clarification of what data characteristics make standard techniques of analysis more effective, and of the nature of measurement, as it has been developed in science, has made it possible to set down goals, and to arrange them in order of usual importance. Moreover, there are reasonably effective techniques for asking sufficiently large bodies of data about the mode in which their analysis should go forward. Today, our first need here is to assemble and purvey available knowledge.

The related question of expressing multiple-response data in a desirable way is almost unattacked. It is substantially more complex than the single-response case, where the order of the different responses is usually prescribed, so that all that remains is to assign numerical values to these responses in a reasonable way.

First and last, a lot is said about changes of coordinates, both in pure mathematics, and in many fields of mathematical application. But little space is spent emphasizing what "a coordinate" is. From our present point of view, a coordinate is the combination of two things:

(a1) a set of level surfaces, namely a classification of the points, objects, responses, or response-lists into sets such that any two members of a set are equivalent so far as that particular coordinate is concerned (as when a 200 lb. woman aged 53 has the same height as a 110 lb. man aged 22), and

(a2) an assignment of numerical values, one to each of these level surfaces or equivalence classes.

In dealing with modes of expression for single-response data, part (a1) is usually wholly settled, and we have to deal only with part (a2).

In dealing with multiple-response data, the least we are likely to face is a need to deal with part (a2) for each response, at least individually but often somewhat jointly. Actually, this simplest case, where the identification of the qualitative, (a1), aspects of all coordinates is not at all at the analyst's disposal, is rather rare. A much more frequent situation is one in which coupled changes in (a1) and (a2) are to be contemplated, as when any linearly independent set of linear combinations of initially given and quantitatively expressed coordinates make up an equally plausible system of coordinates for analysis. In fact, because of its limited but real simplicity, we often assume this situation to hold with sometimes quite limited regard for whether this simplicity, and the consequent emphasis on affine invariance of techniques, is appropriate in the actual instance at hand.

In many instances of multiple-response data we can appropriately consider almost any system of coordinates, giving at most limited attention to the qualitative and quantitative relations of the coordinates for analysis to the initial coordinates. Our knowledge of how to proceed in such circumstances is limited, and poorly organized. We need to learn much more; but we also need to make better use of what we do know. All too often we approach problems in terms of conventional coordinates although we know that the effect to be studied is heavily influenced by an *approximately-known combination* of conventional coordinates. Time and temperature in the stability testing of foods, chemicals, or drugs is one instance. (The combination of time and temperature corresponding to 30 kcal/mole activation energy can be profitably substituted for time in most problems. A little information will allow us to do even better.) Temperature and voltage in the reliability testing of electronic equipment is another.

As in so many of the instances we have described above, *the main need in this field* is for techniques of indication, for ways to allow the data to express their *apparent* character. The need for significance and confidence procedures will only begin to arise as respectable indication procedures come into steady use.

34. Sizes, nomination, budgeting. The choicc of qualitative and quantitative aspects of coordinates is not the only way in which to approximately exercise judgment in approaching the analysis of multiple response data. The work of Dempster (1958, 1960) and of Wilk and Gnanadesikan (1962) points the way toward what seems likely to prove an extensive use of judgment-selected measures of "size" for differences of multiple response. The considerations which should be involved in such choices have not yet been carefully identified, discussed, and compared. Still, it is, I believe, clear that one should not limit oneself to information and judgment about the actual variability, individual and joint, of the several responses (or of more useful coordinates introduced to describe these responses). It will also be wise and proper to give attention to what sorts (=what directions) of real effects seem more likely to occur, and to what sorts of effects, if real, it is more likely to be important to detect or assess.

The problems which arise in trying to guide the wise choice of "size" are new, but not wholly isolated. The practice at East Malling, where experiments take a major fraction of a scientific lifetime, of *nominating* (cp., Pearce, 1953) certain comparisons, appears to have been the first step toward what seems to be an inevitable end, the budgeting of error rates in complex experiments. We consider it appropriate to combine subject-matter wisdom with statistical knowledge in planning what factors shall enter a complex experiment, at how many versions each shall appear, and which these versions shall be. This granted, how can there be objection to using this same combination of wisdom and knowledge to determine, in advance of the data, what level of significance shall be used at each of the lines of the *initial* analysis of variance. If wisdom and knowledge suffice to determine whether or not a line is to *appear* in the initial analysis, surely they suffice to determine whether 5%, 1%, or 0.1% is to be the basis for immediate attention. Yet budgeting of error rate does not seem to have yet been done to any substantial extent.

35. A caveat about indications. It may be that the central problem of complex experimentation may come to be recognized as a psychological one, as the problem of becoming used to a separation between indication and conclusion. The physical sciences are used to "praying over" their data, examining the same data from a variety of points of view. This process has been very rewarding, and has led to many extremely valuable insights. Without this sort of flexibility, progress in physical science would have been much slower. Flexibility in analysis is often only to be had honestly at the price of a willingness not to demand that what has *already* been observed shall establish, or prove, what analysis *suggests*. In physical science generally, the results of praying over the data are thought of as something to be put to further test in another experiment, as indications rather than conclusions.

If complex experiment is to serve us well, we shall need to import this freedom to reexamine, rearrange, and reanalyze a body of data into all fields of application. But we shall need to bring it in *alongside*, and not in place of, preplanned analyses where we can assign known significance or confidence to conclusions about preselected questions. We must be prepared to have indications go far beyond conclusions, or even to have them suggest that what was concluded about was better not considered. The development of adequate psychological flexibility may not be easy, but without we shall slow down our progress.

This particular warning about the place of indication besides conclusion applies in many places and situations. It appears at this particular point because sizes, nomination, and error-rate budgeting are all directed toward the attainment of conclusions with valid error-rates of reasonable overall size. It would have been dangerous to have left the impression that suggestive behavior which did not appear conclusive because of the wise use of such somewhat conservative techniques should, consequently, not be taken as an indication, should not even be considered as a candidate for following up in a later study. Indications are not to be judged as if they were conclusions.

36. FUNOP as an aid to group comparison. The most standard portion of the output of an analysis of variance is a set of means, and an estimate of their variance. As we have learned, more than one thing can be done at this point. Classically, an F-test was used to answer the question "Does the data provide firm evidence that the means are not all the same?". More recently, we have had a variety of multiple comparison methods leading to significance, confidence, and decision statements about more detailed comparisons.

But these approaches, which concentrate upon "Have I proved it?" or "What should I do next?", are not the only ones possible. And there are other sorts of questions. Suppose that the various means are so well determined as to make every comparison unequivocally significant, and suppose, even, that we have put confidence limits for all these comparisons on record, what then?

There is still a real need for guidance in how to think of the real differences among the means. Does one differ *unusually* from the others, do a few? (If the answer to this question is "yes", it will *not* imply that the remaining means are the same, for *all* comparisons were assumed significant, it will merely *suggest* that the *character* of the differences between the one (or few) and the rest are likely to be different from the *character* of the differences among the rest.) Clearly FUNOP can be used, in a variety of possible ways, to provide answers to this question. The simplest ways involve special attention for observations with $z_i \geqq B_A \cdot \dot{z}$ (and for y's more extreme than such y_i).

The choice of B_A is going to be a matter of judgment. And will ever remain so, as far as *statistical theory* and *probability calculations* go. For it is not selected to deal with uncertainties due to sampling. Instead it is to be used to classify situations in the real world in terms of how it is "a good risk" for us to *think* about them. It is quite conceivable that empirical study of many actual situations could help us to choose B_A , but we must remember that the best B_A would be different in different real worlds.

For the present, I propose to experiment with $B_A = 2$. Experience may lead to changes.

Let us then consider an example involving numbers which will arise below as part of a more complex situation. Table 7 sets out the relevant values: column numbers which serve as identification, the observed means in both raw and linerly coded forms, values of i when the means are ordered, values of z_i , and of $\dot{z}$ and $2\dot{z}$. The z value for column 4 exceeds $2\dot{z}$ so that we are led to give this mean special attention. The z value for column 8 comes close to 2 , and might

TABLE 7

Use of FUNOP *to guide thinking about groups of determinations*

Col.	Observed Column Mean		i	FUNOP z†	
	Raw*	Coded			
1	.042	14	7	—	$\dot{z} = 215$
2	.022	−104	4	177	$2\dot{z} = 430$
3	.023	−97	5	230	
4	−.081	−723	1	482	
5	.030	−59	6	—	
6	.013	−161	3	199	
7	.055	93	11	159	
8	.115	454	14	381	
9	−.017	−341	2	310	
10	.049	54	10	—	
11	.062	137	12	180	
12	.118	473	15	298	
13	.047	45	9	—	
14	.072	193	13	201	
15	.043	22	8	—	

* Rounded off.

† Applicable to either raw or coded values; —'s from middle third.

well be considered a case for special attention on this ground, since we have little basis for setting B_A at precisely 2. If we do give column 8, and thus $i = 14$, special attention, we should do the same for $i = 15$, namely column 12, which is even more extreme.

Actually, if we repeat FUNOP on the values remaining after excluding "column 4" we obtain the figures set out above in Table 1 (Section 18), and find that columns 8 and 12 now exceed 2 . For the present purpose it is likely not to be necessary to do the repeat calculation; the application of judgment to the first cycle will usually suffice.

In any event, we come to the following suggestion:

special attention: column 4
some special attention: columns 8, 12
undistinguished: columns 1, 2, 3, 5, 6, 7, 9, 10, 11, 13, 14, 15.

If the error variance associated with these means were small, the proper label on the third group above would be "undistinguished but significantly different", and any further discussion would have to involve both the actual names of the columns and subject-matter insight.

37. Continuation. But suppose, as was actually the case in the situation in which the means above were generated, that the error variance was not very small. It may be that only a few true means are different, and that, at least for all our evidence tells us, the others may be all equal. How are we to assess the reasonableness of such a possibility?

There is certainly ground for considering carefully just how we should proceed in marginal cases, but there is little doubt that we should at least look at the corresponding mean squares. For the instance at hand, these mean squares (expressed for the coded values) are (rounded):

	DF	MS
All columns	14	85010
Omit column 4	13	48430
Omit cols. 4, 8, 12	11	21200
Possible error term	490	56960
Better error term	442	46400

The qualitatively reasonable indications are clear:

(a) Except for column 4 there is no strong indication of differences between true means.

(b) Except for columns 4, 8, 12 there is no visible indication of differences between true means.

The reader may find it instructive to consider what the qualitatively reasonably indications would have been had the error mean square been 5000, 10000, 20000, or 80000.

VII. A SPECIFIC SORT OF FLEXIBILITY

38. The vacuum cleaner. In connection with stochastic process data we noted the advantages of techniques which were revealing in terms of many different models. This is a topic which deserves special attention.

If one technique of data analysis were to be exalted above all others for its ability to be revealing to the mind in connection with each of many different models, there is little doubt which one would be chosen. The simple graph has brought more information to the data analyst's mind than any other device. It specializes in providing indications of unexpected phenomena. So long as we have to deal with the relation of a single response to a single stimulus, we can express almost everything qualitative, and much that is quantitative, by a graph. We may have to plot the differences between observed response and a function of the stimulus against another function of stimulus; we may have to re-express the response, but the meat of the matter can usually be set out in a graph.

So long as we think of direct graphing of stimulus against response, we tend to think of graphing as a way to avoid computation. But when we consider the nature and value of indirect graphs, such as those mentioned above, we come to realize that a graph is often the way in which the results of a substantial computational effort is made manifest.

We need to seek out other tools of data analysis showing high flexibility of effectiveness in other situations. Like the simple graph they will offer us much. We should not expect them to be free of a substantial foundation of computation, and we should not expect their results to be necessarily graphical. Our aim should be flexibility.

In the area of time-series-like phenomena, as we have already noted, spectrum analysis and, more particularly, cross-spectrum analysis (and their extensions) offer such tools.

For data set out in a two-way table, exhibiting the combined relation of two factors to a response, we have long had a moderately flexible approach: the fitting of row, column, and grand means, *and the calculation of residuals*. We have had much profit from the fitting of row, column, and grand means, though not as much as if we had usually gone on to calculate individual residuals, rather than stopping with calculation of the sum of their squares (of the "sum of squares for interaction" or "for discrepancy" etc. in the corresponding analysis of variance). But it is easy to write down two-way tables of quite distinct structure where the fitting of row, column, and grand means fail to exhaust the bulk of this structure. We need to have general techniques that go farther than any just mentioned.

A first step was practiced under such titles as "the linear-by-linear interaction", and was later formalized as "one degree of freedom for non-additivity" (Tukey, 1949b). By isolating a further single (numerical) aspect of the tabulated values, it became possible to ask the data just one more question, and to retrieve just one more number as an answer.

How does one ask still more questions of a two-way table in such a way as to detect as much orderly behavior as possible? The answer here must depend upon the nature of the two factors. If one or both is quantitative, or naturally ordered, we will have access to techniques not otherwise available. Let us be Spartan, and suppose that neither factor has natural quantitative expression or natural order.

If we are to ask reasonably specific questions we must plan to be guided by the table itself in choosing which ones we ask. (This is true to the extent that general prescriptions can be given. Subject-matter knowledge and insight can, and of course should, guide us in specific instances.) If the start is to be routine, prechosen, there is little chance that the fitting of row, column, and grand means can be replaced by some other first step that is both equally simple and better. And the question becomes, once these have been fitted, and residuals formed, how are we to be guided to a next step? Only the fitted means offer us new guidance.

The table of coefficients whose entries are products of "row mean minus grand mean" with "column mean minus grand mean" designates the one degree of freedom for non-additivity. To go further we should perhaps make use of the individual factors rather than their product. When we seek for ways to applv "row mean minus grand mean", for example, we see that we can apply these entries separately in each column, obtaining one regression coefficient per column, a whole row of regression coefficients in all; and so on. This procedure generates what, since its initial use was to produce "vacuum cleaned" residuals (residuals more completely free of systematic effect than those obtained by mean fitting), is conveniently called the basic *vacuum cleaner*. (Those who feel that FUNOR-FUNOM really removes the dirt, may wish to adopt Denis Farlie's suggestion that the present procedure be called FILLET, (i) because it is like taking the bones out of a fish (a herring, perhaps), (ii) from the acronym "Freeing Interaction Line from the Error Term".)

39. Vacuum cleaning: the subprocedure. Again we revert to detail; this time the general reader may wish to skip to VIII, page 60 on first reading.

Our first task is to describe formally a regression procedure which is equivalent to regressing (the values in) each row of a two-way table on (the values in) a separately given row, regressing (the values in) each column of the table on (the values in) a separately given column, regressing the whole table on the two-way array consisting of all products of an entry in the separate row with an entry in the separately given column, and then subtracting this last regression from each of the other two. The result is a four-part breakdown:

(original values) = (dual regression)
+ (deviations of row regression from dual regression)
+ (deviations of column regression from dual regression-
+ (residuals)

This subprocedure begins with a two-way array of entries $\{y_{rc}\}$, where $1 \leqq r \leqq R$, $1 \leqq c \leqq C$, and two conformable one-way arrays, $\{a_r\}$ for $1 \leqq r \leqq R$ and $\{b_c\}$ for $1 \leqq c \leqq C$. (Clearly "r" stands for "row" and "c" for column.) The regression coefficient of column c of the y_{rc} upon $\{a_r\}$ will be denoted

$$[y/a]_c = \frac{\sum_r a_r y_{rc}}{\sum_r a_r^2},$$

and provides the decomposition

$$y_{rc} = a_r[y/a]_c + \{y_{rc} - a_r[y/a]_c\}$$

into an array of rank one involving $\{a_r\}$ and an array of residuals, each column of which yields zero regression on $\{a_r\}$. In the particular case $a_r \equiv 1$, we have $[y/a]_c = y_{\cdot c}$, the mean of the cth column, and the decomposition is that corresponding to the removal of column means.

We can clearly also calculate the regression coefficient

$$[y/b]_r = \frac{\sum_c y_{rc} b_c}{\sum_c b_c^2}$$

of each row upon the given row $\{b_c\}$. Doing both together, we are led, by the familiar procedure of removing row, column, and *grand* means first to notice that

$$\frac{\sum_c [y/a]_c\, b_c}{\sum_c b_c^2} \equiv \frac{\sum\sum a_r\, y_{rc}\, b_c}{\sum\sum a_r^2\, b_c^2} \equiv \frac{\sum_r a_r [y/a]_r}{\sum a_r^2} = [y/ab]$$

where the notation $[y/ab]$ is by analogy with $[y/a]_c$ and $[y/a]_r$, and then to write down the 4-part decomposition

$$y_{rc} = a_r[y/ab]b_c + \{a_r[y/a]_c - a_r[y/ab]b_c\} + \{[y/b]_r b_c - a_r[y/ab]b_c\} \\ + \{y_{rc} - a_r[y/a]_c - [y/b]_r b_c + a_r[y/ab]b_c\}$$

whose successive terms represent (i) dual regression, (ii) regression within column differing from dual regression, (iii) regression within row differing from dual regression, (iv) residuals, each row or column of which yields no regression. In the special case $a_r \equiv 1 \equiv b_c$ these four parts of course reduce to (i) grand mean, (ii) deviation of column mean from grand mean, (iii) deviation of row mean from grand mean, (iv) residuals, each row or column of which yields zero mean. Thus the breakdown provided by our subprocedure is a generalization of a familiar one.

As we may expect, there is a corresponding breakdown of the sum of squares:

$$\sum\sum y_{rc}^2 \equiv ([y/ab])^2 \cdot \sum a_r^2 \cdot \sum b_c^2 + \sum ([y/a]_c - [y/ab]b_c)^2 \cdot \sum a_r^2 \\ + \sum([y/b]_r - a_r[y/ab])^2 \cdot \sum b_c^2 + \sum\sum (y_{rc} - a_r[y/a]_c - [y/b]_r b_c \\ + a_r[y/ab]b_c)^2$$

here the entries in parentheses are natural results of the breakdown, viz: (i) dual regression coefficient, (ii) and (iii) deviation regression coefficients, (iv) residuals.

Simplicity of computation is enhanced if we apply the constraint $\sum a_r^2 \equiv 1 \equiv \sum b_c^2$, which (a) eliminates divisors in calculating regression coefficients, and (b) eliminates factors in calculating the breakdown into sums of squares. The resulting subprocedure, then;

(a1) accepts a column of numbers, a row of numbers, and a two-way table $\{y_{rc}\}$ of numbers,

(a2) multiplies the entries in the column by an appropriate constant to attain $\sum a_r^2 = 1$, and those in the rw by another to attain $\sum b_c^2 = 1$. The new row and the new column are conveniently called *carriers*, (or perhaps *regressors* cp., Hannan, 1958),

(a3) calculates a two-way table of *residuals*; finding (a row of, a column of, and an individual) *coefficients* on the way.

(a4) accumulates the corresponding analysis-of-variance sums of squares,

(a5) outputs the carriers, the residuals, the coefficients, and an analysis of variance.

40. The basic vacuum cleaner, and its attachments. In what is called the basic vacuum cleaner the subprocedure just described is applied twice:

(b1) in the first application each carrier consists of identical entries, and *the subprocedure removes row, column, and grand means,*

(b2) in the second application, the carriers are constructed (by a2) from the coefficients obtained in the first stage, and *the subprocedure extracts row-by-row regression upon "column mean minus grand mean" and column-by-column regression on "row mean minus grand mean".*

The respective single degrees of freedom (corresponding to dual regression) are those (b1) for the grand mean, and (b2) for one degree of freedom for nonadditivity.

Since the first application removes $R + C - 1$ degrees of freedom and the second application removes $R + C - 3$ degrees of freedom, the table has to be large enough for $RC > 2R + 2\text{C} - 4$, which requires $R, C \geqq 3$.

In early trials, the vacuum cleaner was applied to relatively raw data containing some quite wild values. As a result, the vacuum cleaner spent its (b2) effort in trying to account for the wildest values and had none left over to look for more dispersed structure. FUNOR-FUNOM was actually developed so that it could be applied, on a routine basis, ahead of the vacuum cleaner. Some similar sort of clean-up procedure will usually have to precede application of the vacuum cleaner.

There are a number of ways to continue the basic vacuum cleaner, a number of "attachments" which can be added. Clearly the stage b2 coefficients could be used to define stage b3 carriers, as could the results of orthogonalizing the squares of the stage b1 coefficients to the stage b1 coefficients themselves. It would also be possible to proceed in a way suggested by expression of the residual table in terms of eigen-vectors. We shall not follow up any of these possibilities here in any detail.

The problem of determining the eigen-vectors and eigen-values associated with a two-way array is beginning to appear in diverse places in data analysis. Tucker (1958) has been led to it in connection with factor analysis; it is surely one of the natural ways to continue the basic vacuum cleaner; and we may expect the same problem to arise in connection with a number of quite distinct problems.

To date, work on eigen-values and eigen-vectors computation has concentrated upon taking a matrix as given error-free and finding vectors and values to an assigned accuracy. The effort involved increases moderately rapidly with the size of the array. Large arrays are going to require, and smaller arrays can be wisely analyzed through, the use of approximate solutions. We shall need to know how the precision required to make the errors of approximation small with respect to sampling fluctuations depends on the proper analog of sample size. And we shall have to investigate such methods of seeking the largest eigen-value, and its associated vectors, as beginning with k randomly chosen vectors and repeatedly multiplying each by the array or its transpose. (With or without changes in the k-dimensional coordinate system after each multiplication.)

TABLE 8

Original Values for Second Example

ROWS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	MEAN
1	0.126	-0.092	0.268	-0.284	0.101	-0.002	0.045	-0.305	-0.604	-0.201	-0.493	-0.023	0.340	-0.102	0.042	-0.079
2	0.031	0.050	0.076	-0.058	-0.020	-0.075	0.076	0.065	0.173	0.081	0.287	0.045	0.198	-0.044	0.108	0.066
3	0.045	0.386	0.015	0.127	0.030	0.009	0.117	-0.103	-0.033	-0.027	-0.134	0.288	0.327	0.277	0.035	0.091
4	0.113	0.068	-0.009	-0.043	0.050	-0.083	0.069	0.586	-0.047	0.233	0.298	-0.514	0.664	-0.263	0.191	0.088
5	0.154	0.398	-0.027	-0.361	-0.019	-0.114	0.062	0.467	0.055	-0.056	-0.008	0.224	0.512	-0.106	-0.090	0.073
6	0.068	-0.072	-0.210	-0.015	0.025	0.138	0.071	1.113	0.115	0.040	0.305	0.714	0.251	0.131	-0.043	0.175
7	-0.009	-0.097	0.044	-0.788	0.145	-0.106	0.168	0.	-0.034	-0.420	0.277	-0.356	-0.464	0.512	-0.000	-0.075
8	0.044	0.177	-0.031	0.366	0.008	-0.138	0.081	0.960	0.128	0.161	-0.065	0.885	1.039	-0.140	0.302	0.252
9	0.218	-0.153	0.112	-0.824	0.045	0.141	-0.036	-0.	0.001	-0.243	0.032	0.102	-0.717	-0.007	0.050	-0.085
10	-0.039	0.028	0.021	0.066	0.030	0.177	0.035	0.240	0.054	-0.038	-0.057	0.415	-0.039	-0.126	0.014	0.052
11	-0.004	0.096	0.028	0.184	0.020	0.162	-0.052	0.084	0.096	0.030	0.015	-0.109	-0.129	-0.178	-0.054	0.013
12	0.078	-0.298	0.062	-0.155	-0.108	-0.029	0.111	-0.265	0.147	0.264	-0.020	-0.130	-0.157	0.041	0.078	-0.025
13	-0.006	-0.126	-0.052	0.014	0.106	-0.110	0.055	-0.000	-0.172	-0.039	-0.005	0.114	-0.178	0.224	-0.010	-0.012
14	0.038	-0.043	0.076	0.051	-0.099	0.084	0.117	0.161	0.157	0.147	0.122	-0.043	-0.073	-0.011	-0.010	0.045
15	-0.056	0.096	-0.340	0.193	0.198	0.127	-0.035	0.425	0.128	-0.103	0.217	-0.107	0.740	0.486	-0.158	0.121
16	0.042	0.220	0.119	0.133	0.002	-0.367	0.035	-0.264	-0.012	0.763	0.483	-0.202	-0.394	0.824	0.157	0.103
17	0.097	0.001	0.162	-0.168	-0.044	0.209	0.006	-0.108	-0.007	0.692	0.516	0.556	-0.527	0.789	0.071	0.150
18	0.061	0.115	0.064	-0.052	0.092	0.145	0.236	0.057	-0.040	-0.006	0.215	-0.026	0.211	0.055	0.035	0.077
19	-0.005	-0.261	0.057	0.233	0.018	0.106	0.004	-0.321	0.202	-0.095	-0.012	0.027	-0.093	-0.090	-0.198	-0.029
20	0.163	0.176	-0.032	-0.008	0.270	0.103	0.109	0.924	-0.449	-0.226	-0.193	0.636	0.401	-0.201	0.085	0.117
21	-0.014	0.034	-0.004	0.185	0.024	0.066	0.040	-0.020	0.104	-0.030	-0.063	0.059	-0.174	0.048	0.088	0.023
22	-0.017	0.137	-0.070	0.037	0.165	0.165	0.135	0.	0.391	0.034	-0.112	0.090	0.130	-0.124	0.089	0.070
23	0.071	0.038	0.086	0.247	0.150	0.150	0.184	-0.323	-0.037	-0.100	-0.070	0.101	-0.450	-0.141	0.023	0.009
24	-0.009	0.061	0.052	0.348	0.043	0.043	-0.031	0.380	0.036	0.028	0.141	-0.165	0.145	-0.721	-0.246	0.007
25	-0.005	0.142	-0.113	0.005	0.031	0.031	-0.084	-0.352	0.302	-0.022	0.228	-0.235	0.115	-0.381	0.052	-0.019
26	0.032	-0.225	-0.006	-0.085	0.001	0.003	0.110	-0.108	-0.225	0.111	0.170	0.174	0.063	0.048	-0.135	-0.005
27	-0.009	-0.010	0.042	-0.004	-0.113	-0.176	0.084	-0.799	0.265	-0.057	0.234	0.262	-0.203	0.114	-0.056	-0.028
28	0.037	-0.068	0.013	0.067	0.115	-0.005	0.110	-0.011	-0.258	0.075	-0.230	0.153	0.454	0.196	-0.002	0.043
29	-0.047	0.142	0.058	0.303	-0.202	-0.077	0.130	-0.059	-0.325	-0.009	0.024	-0.124	-0.086	0.056	-0.003	-0.015
30	0.066	0.142	0.045	-0.337	0.040	-0.026	0.225	0.072	0.220	0.307	0.165	0.	0.010	0.073	0.681	0.112
31	-0.010	0.035	-0.030	-0.215	-0.039	0.046	0.094	0.665	0.020	0.176	-0.075	0.	0.190	0.226	-0.059	0.068
32	0.084	0.039	0.141	-0.231	-0.224	0.113	0.417	-0.136	-0.490	-0.166	0.143	0.	-0.026	0.316	0.515	0.033
33	0.146	-0.336	0.280	-0.637	0.038	-0.180	-0.037	-0.053	-0.509	0.342	-0.040	0.432	0.	0.007	0.084	-0.031
34	0.079	0.072	0.088	-0.062	-0.147	-0.266	-0.413	0.683	0.116	-0.214	0.194	0.326	0.	0.654	-0.006	0.073
35	-0.020	-0.013	-0.026	-0.371	0.115	0.003	-0.031	0.285	0.325	0.081	-0.102	0.381	-0.038	-0.504	-0.052	0.002
36	-0.027	-0.053	-0.118	-0.788	0.225	0.190	-0.221	0.210	-0.416	0.038	-0.142	0.314	-0.342	0.647	-0.018	-0.033
MEAN	0.042	0.022	0.023	-0.081	0.030	0.013	0.055	0.115	-0.017	0.049	0.062	0.118	0.047	0.072	0.043	0.040

41. The vacuum cleaner: an example. Table 8 presents $540 = 36 \times 15$ values of a multiple regression coefficient as obtained in 540 small to moderate groups and modified (i) by entering "0" for 8 groups containing too few individuals to make calculation of the corresponding multiple regression coefficients reasonable, and (ii) applying FUNOR-FUNOM. (Neither variate in this example is the same as in the example of Section 19.) For our present purposes, this is just another moderately cleaned-up two-way table.

Table 9 shows the first application of the subprocedure, corresponding to the removal of row and column means. The entries in the body of the table are residuals after this fit. Thus, the upper left entry in Table 9 is

$$0.126 - (0.014)(0.258) - (-.459)(0.167) - (0.921)(0.258)(0.167) \doteq 0.20.$$

which would have been calculated in terms of row and column means (see Table 8 for same) as

$$0.126 - 0.042 - (-0.079) + (0.040) \doteq 0.20.$$

And so on and on. (Clearly, if all we wanted to do was to fit row, column, and grand means, we would have saved some arithmetic to do this directly.) The result of the first application, so far as the upper left cell is concerned, is to dissect the value 0.126 into four parts:

$$0.126 \doteq 0.20 + (0.014)(0.258) + (-0.459)(0.167) + (0.921)(0.258)(0.167)$$

to set 0.014, -0.459, and 0.921 aside for later consideration, and to carry 0.20 on for further analysis.

The appended analysis of variance indicates the presence of (small) row and column contributions, since the row and column mean squares are each about 1.5 times the error mean square. If all degrees of freedom are to be taken at face value, the pooled mean square for rows and columns is significant at 5%.

Table 10 shows the second application of the same subprocedure. The carriers are now normalized forms of the coefficients obtained in the first application. (Note ratios of about 0.9 for the row vectors and about 0.6 for the column vectors.) Again the body of the table contains residuals after removing the indicated regressions on the carriers.

The mean squares for row slopes and column slopes are larger than were the mean squares for row means and column means. If fitting row and column means to the original table was worthwhile, continuing the fitting this further stage was more worthwhile. The individual slope mean squares are all significant at 5%; indeed the ratio of pooled mean squares for slopes to pooled mean squares for means is significant at 5% (if we dare trust a two-sample-like F test).

The second stage reduced the residual mean square by something more than 15% below its post-first-stage value. Whether or not such a reduction is important will depend very much upon the purpose for which the residuals are to be used. If, for example, they were to be correlated with residuals from another similar table, failing to remove this amount of systematic structure could easily

TABLE 9

Results of First Application of Subprocedure

Rows	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	CAR.	COEF
1	00.20	0.00	0.36	-0.08	0.19	0.10	0.11	-0.30	-0.47	-0.13	-0.44	-0.02	0.41	-0.05	0.12	0.167	-0.459
2	-0.04	0.00	0.03	-0.00	-0.08	-0.11	-0.01	-0.08	0.16	0.01	0.20	-0.10	0.12	-0.14	0.04	0.167	0.103
3	-0.05	0.31	-0.06	0.16	-0.05	-0.05	0.01	-0.27	-0.07	-0.13	-0.25	0.12	0.23	0.15	-0.06	0.167	0.197
4	0.02	-0.00	-0.08	-0.01	-0.03	-0.14	-0.03	0.42	-0.08	0.14	0.19	-0.68	0.57	-0.38	0.10	0.167	0.186
5	0.08	0.34	-0.08	-0.31	-0.08	-0.16	-0.03	0.32	0.04	-0.14	-0.10	0.07	0.43	-0.21	-0.17	0.167	0.128
6	-0.11	-0.28	-0.37	-0.07	-0.14	-0.01	-0.12	0.86	-0.00	-0.14	0.11	0.46	0.07	-0.08	-0.22	0.167	0.526
7	0.06	-0.00	0.14	-0.59	0.23	-0.00	0.23	-0.00	0.10	-0.35	0.33	-0.36	-0.40	0.56	0.07	0.167	-0.445
8	-0.21	-0.06	-0.27	0.23	-0.23	-0.36	-0.19	0.63	-0.07	-0.10	-0.34	0.55	0.78	-0.42	0.05	0.167	0.822
9	0.30	-0.05	0.21	-0.62	0.14	0.25	0.03	0.01	0.14	-0.17	0.10	0.11	-0.64	0.05	0.13	0.167	-0.484
10	-0.09	-0.01	-0.01	0.13	-0.01	0.15	-0.03	0.11	0.06	-0.10	-0.13	0.28	-0.10	-0.21	-0.04	0.167	0.049
11	-0.02	0.10	0.03	0.29	0.02	0.18	-0.08	-0.00	0.14	0.01	-0.02	-0.20	-0.15	-0.22	-0.07	0.167	-0.105
12	0.10	-0.26	0.10	-0.01	-0.07	0.02	0.12	-0.32	0.23	0.28	-0.02	-0.18	-0.14	0.03	0.10	0.167	-0.252
13	0.00	-0.10	-0.02	0.15	0.13	-0.07	0.05	-0.06	-0.10	-0.04	-0.02	0.05	-0.17	0.20	-0.00	0.167	-0.202
14	-0.01	-0.07	0.05	0.13	-0.13	0.07	0.06	0.04	0.17	0.09	0.05	-0.17	-0.13	-0.09	-0.06	0.167	0.021
15	-0.18	-0.01	-0.44	0.19	0.09	0.03	-0.17	0.23	0.06	-0.23	0.07	-0.31	0.61	0.33	-0.28	0.167	0.314
16	-0.06	0.13	0.03	0.15	-0.09	-0.44	-0.08	-0.44	-0.06	0.65	0.36	-0.38	-0.50	0.69	0.05	0.167	0.244
17	-0.06	-0.13	0.03	-0.20	-0.18	0.09	-0.16	-0.33	-0.10	0.53	0.34	0.33	-0.68	0.61	-0.08	0.167	0.426
18	-0.02	0.05	0.00	-0.01	0.02	0.09	0.14	-0.10	-0.06	-0.09	0.11	-0.18	0.13	-0.05	-0.05	0.167	0.146
19	0.02	-0.21	0.10	0.38	0.06	0.16	0.02	-0.37	0.29	-0.08	-0.01	-0.02	-0.07	-0.09	-0.17	0.167	-0.264
20	0.04	0.08	-0.13	-0.00	0.16	0.01	-0.02	0.73	-0.51	-0.35	-0.33	0.44	0.28	-0.35	-0.04	0.167	0.301
21	-0.04	0.03	-0.01	0.28	0.01	0.07	0.00	-0.12	0.14	-0.06	-0.11	-0.04	-0.20	-0.01	0.06	0.167	-0.065
22	-0.09	0.08	-0.12	0.09	0.10	0.12	0.05	-0.15	0.38	-0.05	-0.20	-0.06	0.05	-0.23	0.02	0.167	0.118
23	0.06	0.05	0.09	0.36	0.15	0.17	0.16	-0.41	0.01	0.08	-0.10	0.01	-0.47	-0.18	0.01	0.167	-0.120
24	-0.02	0.07	0.06	0.46	0.05	0.06	-0.05	0.30	0.09	0.01	0.11	-0.25	0.13	-0.76	-0.26	0.167	-0.126
25	0.01	0.18	-0.08	0.14	0.06	0.08	-0.08	-0.41	0.38	-0.01	0.22	-0.29	0.13	-0.39	0.07	0.167	-0.228
26	0.03	-0.20	0.02	0.04	0.02	0.03	0.10	-0.18	-0.16	0.11	0.15	0.10	0.06	0.02	-0.13	0.167	-0.172
27	0.02	0.04	0.09	0.15	-0.07	-0.12	0.10	-0.85	0.35	-0.04	0.24	0.21	-0.18	0.11	-0.03	0.167	-0.263
28	-0.01	-0.09	-0.01	0.14	0.08	-0.02	0.05	-0.13	-0.24	0.02	-0.30	0.03	0.40	0.12	-0.05	0.167	0.014
29	-0.04	0.17	0.09	0.44	-0.18	-0.04	0.13	-0.12	-0.25	-0.00	0.02	-0.19	-0.08	0.04	0.01	0.167	-0.210
30	-0.05	0.05	-0.05	-0.33	-0.06	-0.11	0.10	-0.12	0.16	0.19	0.03	-0.19	-0.11	-0.07	0.57	0.167	0.282
31	-0.08	-0.02	-0.08	-0.16	-0.10	0.00	0.01	0.52	0.01	0.10	-0.17	-0.15	0.11	0.13	-0.13	0.167	0.111
32	0.05	0.02	0.12	-0.14	-0.25	0.11	0.37	-0.24	-0.47	-0.21	0.09	-0.11	-0.07	0.25	0.48	0.167	-0.025
33	0.17	-0.29	0.33	-0.49	0.08	-0.12	-0.02	-0.10	-0.42	0.36	-0.03	0.38	0.02	0.01	0.11	0.167	-0.273
34	0.00	0.02	0.03	-0.02	-0.21	-0.31	-0.50	0.53	0.10	-0.30	0.10	0.17	0.08	0.55	-0.08	0.167	0.131
35	-0.02	0.00	-0.01	-0.25	0.12	0.03	-0.05	0.21	0.38	0.07	-0.13	0.30	-0.05	-0.54	-0.06	0.167	-0.145
36	0.00	-0.00	-0.07	-0.62	0.27	0.25	-0.20	0.17	-0.33	0.06	-0.13	0.27	-0.32	0.65	0.01	0.167	-0.280
CAR.	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	0.258	PROD,	
COEF	0.014	-0.104	-0.097	-0.724	-0.059	-0.161	0.093	0.454	-0.341	0.054	0.137	0.473	0.045	0.193	0.022		0.92102

ANOVA

Isolations				Pools and Residuals			
df	Site	Sum of Squares	Mean Square	df	Site	Sum of Squares	Mean Square
1	Grand Mean	0.849	0.848				
14	Col Main	1.190077	0.085006	49	Main	4.052967	0.082714
35	Row Main	2.862890	0.081797	490	Res After Main	27.86270	0.05686

TABLE 10

Results of Second Application of Subprocedure

Row	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	CAR.	COEF.
1	0.09	0.03	0.19	0.08	0.08	0.00	0.00	-0.06	-0.46	-0.11	-0.47	0.03	0.62	-0.11	0.08	-0.271	0.110
2	-0.01	-0.02	0.06	-0.11	-0.06	-0.11	0.03	-0.09	0.13	0.01	0.22	-0.07	0.08	-0.11	0.05	0.061	-0.133
3	0.00	0.28	-0.00	-0.03	-0.01	-0.04	0.07	-0.30	-0.13	-0.13	-0.21	0.17	0.15	0.21	-0.04	0.117	-0.220
4	0.07	-0.02	-0.02	-0.13	0.01	-0.11	0.02	0.36	-0.11	0.13	0.21	-0.67	0.49	-0.35	0.12	0.110	-0.130
5	0.11	0.36	-0.01	-0.15	-0.03	-0.08	-0.02	0.12	0.13	-0.16	-0.13	-0.08	0.36	-0.25	-0.16	0.076	0.279
6	0.01	-0.21	-0.12	0.15	0.02	0.20	-0.05	0.33	0.17	-0.20	0,07	0.13	-0.20	-0.12	-0.19	0.311	0.487
7	-0.05	0.06	-0.00	-0.21	0.14	-0.05	0.10	0.10	0.22	-0.35	0.25	-0.45	-0.21	0.44	0.03	-0.263	0.447
8	-0.01	-0.08	0.07	0.14	-0.02	-0.13	-0.02	0.07	0.00	-0.15	-0.31	0.33	0.40	-0.38	0.10	0.486	0.104
9	0.17	0.02	0.07	-0.19	0.04	0.20	-0.11	0.11	0.28	-0.16	0.01	-0.00	-0.44	-0.08	0.09	-0.286	0.501
10	-0.08	-0.01	0.00	0.10	-0.00	0.16	-0.02	0.10	0.05	-0.10	-0.12	0.29	-0.12	-0.20	-0.04	0.029	-0.040
11	-0.04	0.07	-0.04	0.05	-0.03	0.09	-0.07	0.22	0.01	0.03	0.02	-0.01	-0.08	-0.16	-0.07	-0.062	-0.390
12	0.04	-0.26	-0.01	-0.06	-0.15	-0.07	0.08	-0.09	0.17	0.30	-0.01	-0.06	-0.02	0.04	0.08	-0.149	-0.158
13	-0.05	-0.08	-0.10	0.21	0.08	-0.12	0.01	0.05	-0.10	-0.03	-0.03	0.08	-0.08	0.18	-0.02	-0.119	0.035
14	-0.00	-0.09	0.04	-0.01	-0.14	0.04	0.08	0.11	0.11	0.10	0.08	-0.08	-0.13	-0.05	-0.05	0.012	-0.205
15	-0.10	-0.04	-0.34	0.00	0.16	0.09	-0.09	0.11	0.02	-0.24	0.11	-0.29	0.47	0.39	-0.26	0.186	-0.192
16	0.00	0.10	0.10	-0.08	-0.04	-0.42	-0.01	-0.48	-0.13	0.65	0.40	-0.32	-0.60	0.76	0.07	0.144	-0.281
17	0.05	-0.13	0,21	-0.17	-0.06	0.22	-0.09	-0.68	-0.03	0.50	0.34	0.16	-0.89	0.61	-0.06	0.252	0.174
18	0.02	0.03	0.05	-0.13	0.05	0.11	0.19	-0.13	-0.10	-0.09	0.14	-0.15	0.06	-0.02	-0.03	0.087	-0.143
19	-0.04	-0.25	-0.05	0.12	-0.04	0.02	0.00	-0.01	0.13	-0.04	0.04	0.24	0.07	-0.03	-0.18	-0.156	-0.468
20	0.11	0.11	0.03	0.23	0.26	0.16	-0.00	0.35	-0.36	-0.39	-0.37	0.18	0.12	-0.41	-0.02	0.178	0.447
21	-0.05	0.00	-0.06	0.09	-0.02	0.01	0.02	0.05	0.04	-0.04	-0.07	0.11	-0.16	0.04	0.06	-0.039	-0.317
22	-0.06	0.04	-0.11	-0.19	0.11	0.10	0.11	-0.06	0.26	-0.03	-0.15	0.08	0.01	-0.15	0.03	0.070	-0.382
23	0.04	0.01	0.00	0.08	0.10	0.07	0.17	-0.14	-0.14	0.11	-0.05	0.24	-0.39	-0.11	0.01	-0.071	-0.460
24	-0.04	0.04	-0.03	0.20	-0.01	-0.03	-0.04	0.56	-0.05	0.04	0.16	-0.04	0.21	-0.69	-0.26	-0.075	-0.429
25	-0.04	0.14	-0.21	-0.15	-0.03	-0.06	-0.08	-0.05	0.21	0.03	0.28	-0.02	0.25	-0.32	0.06	-0.135	-0.509
26	-0.01	-0.19	-0.04	0,14	-0.02	0.00	0.06	-0.11	-0.14	0.11	0.13	0.10	0.14	-0.01	-0.15	-0.101	0.095
27	-0.04	0.02	-0.05	-0.02	-0.16	-0.24	0.07	-0.54	0.24	-0.01	0.27	0.41	-0.05	0.15	-0.04	-0.156	-0.327
28	-0.00	-0.10	-0.01	0.11	0.08	-0.02	0.06	-0.12	-0.26	0.02	-0.29	0.05	0.40	0.13	-0.05	0.008	-0.049
29	-0.08	0.16	-0.02	0.30	-0.25	-0.13	0.11	0.13	-0.35	0.02	0.04	-0.02	0.03	0.07	-0.00	-0.124	-0.272
30	0.02	0.03	0.06	-0.39	0.01	-0.04	0.16	-0.29	0.17	0.17	0.05	-0.25	-0.24	-0.05	0.59	0.166	-0.016
31	-0.06	0.00	-0.02	-0.02	-0.06	0.07	0.01	0.35	0.09	0.08	-0.19	-0.28	0.05	0.09	-0.13	0.066	0.246
32	0.04	0.04	0.13	-0.03	-0.24	0.13	0.35	-0.30	-0.42	-0.22	0.06	-0.18	-0,06	0.22	0.47	-0.015	0.167
33	0.10	-0.21	0.28	0.03	0.05	-0.09	-0.14	-0.21	-0.22	0.34	-0.13	0.14	0.12	-0.14	0.08	-0.162	0.694
34	0.03	0.04	0.11	0.18	-0.16	-0.23	-0.50	0.31	0.21	-0.32	0.06	-0.00	-0.16	0.50	-0.08	0.077	0.336
35	-0.06	0.03	-0.05	-0.10	0.09	0.02	-0.09	0.22	0.43	0.07	-0.16	0.25	0.01	-0.58	-0.07	-0.086	0.183
36	-0.08	0.09	-0.11	-0.03	0.24	0.30	-0.33	0.01	-0.08	0.04	-0.25	-0.02	-0.22	0.48	-0.03	-0.166	0.814
CAR.	0.013	-0.095	-0.089	-0.663	-0.054	-0.148	0.085	0.416	-0.313	0.049	0.125	0.434	0.042	0.177	0.020	PROD:	
COEF.	-0.422	0.125	-0.627	0.678	-0.409	-0.365	-0.394	0.856	0.084	0.066	-0.151	0.150	0.759	-0.219	-0.131		0.51902

Anova

Isolations				Pools and Residuals			
df	Site	Sum of Squares	Mean Square	df	Site	Sum of Squares	Mean Square
1	Lin. by Lin.	0.269379	0.269379				
13	Col Slopes*	2.933377	0.225644	47	Slopes	7.084908	0.150743
34	Row Slopes	4.151531	0.122104	442	Res after Slopes	20.50842	0.04640

*That is, variation from column to column in regression on carrier presented as a column vector (and proportional to deviations of original row means from the grand mean.)

lead to avoidable errors as serious as ascribing the wrong sign to the resulting correlation or regression coefficients.

Note that FUNOR-FUNOM was applied before the vacuum cleaner. If this had not been done, and if, for instance, the original table had contained a single unusually deviant observation, the second stage would have devoted most of its effort to accounting for the presence of this "wild shot", and would thus have been unable to attend to removing the organized structure which was actually removed. Unless "wild shots" are surely absent, preliminary FUNOR-FUNOM is likely to be essential in making effective use of the basic vacuum cleaner.

42. The example continued. The production of residuals freer of systematic structure than those obtained by merely fitting rows and columns is a central purpose of using the vacuum cleaner. To this end, the 540 residuals in Table 10 are one of the main results of the technique, and would normally be subject to further analysis or examination.

But the results of the vacuum cleaner can be applied to other purposes. The usual purposes of fitting row and column means include a summarization of (some of) the systematic appearances of the data. These purposes are more fully met by giving the row and column slopes as well as the row and column means. The 50 constants required to specify grand, row, and column means describe the differences between the values of Table 8 and the first residuals of Table 9. The 48 further constants required to specify linear-by-linear (or dual), row, and column slopes, when combined with the first 50, describe the differences between the original values of Table 8 and the second residuals of Table 10. For some aspects of the purpose of describing the apparently systematic behavior it would suffice to stop with these 50 + 48 values. For others further analysis may be helpful.

Having obtained the various coefficient vectors, we can try to understand them. The first question to ask is: "Is one entry, or are a few entries, of overwhelming importance?" As we have seen above (Section 36) FUNOP can offer guidance in this problem. Table 11 sets out the relevant detail for the various row vectors in the example. The entries for column 4 turn out to be comparatively large. As comparison with Table 7 shows, the entries for the "main" coefficients, which are linearly-coded sample means, were already analyzed in Section 36, with the result that there is rather clear ground for giving special attention to column 4, and possible ground for looking at columns 8 and 12, while the others seem to need no attention.

Turning to the slope coefficients, no z_i exceeds $2\dot{z} = 880$ but z_1, corresponding to column 4 nearly reaches this level. In view of (a) our uncertainty that $2.00\dot{z}$ is the right dividing line, and (b) the appearance of column 4 in the detailed analysis of the mean vector, it seems reasonable to give the column 4 slope special attention. If we do this, we should do the same for the more extreme entries for columns 8 and 13. Analysis of the remaining 12 entries yields a mean square of only about 1.5 times the error mean square, and FUNOP gives no strong indication of further need for special attention.

TABLE 11

Detailed behavior of row vectors in example
(Entries 1000 × those above)

Col.	(Carrier)		Coefficient		FUNOP	
	(Main)	(Slope)	Main	Slope	Main	Slope
1	(258)	(13)	14	−422	—	306
2	(258)	(−95)	−104	125	177	399
3	(258)	(−90)	−97	−627	230	362
4	(258)	(−663)	−723†	678	482†	843
5	(258)	(−54)	−59	−409	—	382
6	(258)	(−148)	−161	−365	199	604
7	(258)	(86)	93	−394	159	481
8	(258)	(416)	454	856	381	603
9	(258)	(−313)	−341	84	310	—
10	(258)	(49)	54	66	—	—
11	(258)	(125)	137	−151	180	—
12	(258)	(434)	473	150	298	326
13	(258)	(42)	45	759	—	721
14	(258)	(177)	193	−219	201	—
15	(258)	(20)	22	−131	—	—
					$\dot{z}$ = 215	440

* FUNOP z's for coefficient entries.
† FUNOP $z > 2\dot{z}$.

So far as row vectors are concerned, the best-defined part of the systematic appearance will be covered if we specify the entries for columns 4, 8, 12, and 13, leaving the others zero, and adjusting the dual regressions (grand mean and linear-by-linear) accordingly. Thus the 14 + 13 = 27 row vector constants have been boiled down to 4 + 4 = 8 numerical values and the selection of 4 columns for special attention. (A similar process could be applied to the column vectors.)

This boiling down of apparent structures does not oblige us to alter our residuals. The purpose of using the vacuum cleaner to generate residuals was to provide residuals clear of likely effects. This it did. A *three-part* decomposition of the observations,

$$\begin{aligned}(\text{observed value}) &= (\text{apparently systematic part})\\ &+ (\text{possibly systematic part})\\ &+ (\text{hopefully residual part})\end{aligned}$$

is often more desirable than a two-part decomposition in which everything is forced to be either apparently systematic or hopefully residual.

When special attention is given to these four columns it is easy, in this particular instance, to learn what is going on. As noted at the beginning of the example the 540 original values are regression coefficients, one for each of 540 groups of individuals, these groups, being cross-classified into 36 and 15 classes respectively. Each group can provide not only an apparent regression coefficient, but also an estimate of the variance of this apparent regression coefficient. When these estimates of variance are examined, they are found to be systematically large in the four special-attention columns.

Thus the apparent significance of row means and slopes can plausibly be ascribed to inhomogeneity of variance. Two remarks are relevant:

(1) As noted elsewhere (Section 31) the use of more incisive tools is more likely to reveal *both* what is being sought for *and* what may, perhaps uncomfortably, be present (such as non-constant variability).

(2) The effects of non-constant variability already gave rise to a detectable, and nominally significant, effect in the *conventional* stage of the analysis, the fitting of row and column means.

VIII. HOW SHALL WE PROCEED?

43. What are the necessary tools? If we are to make progress in data analysis, as it is important that we should, we need to pay attention to our tools and our attitudes. If these are adequate, our goals will take care of themselves.

We dare not neglect any of the tools that have proved useful in the past. But equally we dare not find ourselves confined to their use. If algebra and analysis cannot help us, we must press on just the same, making as good use of intuition and originality as we know how.

In particular we must give very much more attention to what specific techniques and procedures do when the hypotheses on which they are customarily developed do not hold. And in doing this we must take a positive attitude, not a negative one. It is not sufficient to start with what it is supposed to be desired to estimate, and to study how well an estimator succeeds in doing this. We must give even more attention to starting with an estimator and discovering what is a reasonable estimand, to discovering what is it reasonable to think of the estimator as estimating. To those who hold the (ossified) view that "statistics is optimization" such a study is hindside before, but to those who believe that "the purpose of data analysis is to analyze data better" it is clearly wise to learn what a procedure really seems to be telling us about. It would be hard to overemphasize the importance of this approach as a tool in clarifying situations.

Procedures of diagnosis, and procedures to extract indications rather than conclusions, will have to play a large part in the future of data analyses. Graphical techniques offer great possibilities in both areas, and deserve far more extensive discussion than they were given here. Graphs will certainly be increasingly "drawn" by the computer without being touched by hands. More and more, too, as better procedures of diagnosis and indication are automated, graphs, and other forms of expository output, will, in many instances, cease to be the vehicle through which a man diagnoses or seeks indications, becoming, instead, the vehicle through which the man supervises, and approves or disapproves, the diagnoses and indications already found by the machine.

44. The role of empirical sampling. Numerical answers about the absolute or comparative performance of data analysis procedures will continue to be of importance. Approximate answers will almost always serve as well as exact ones, provided the quality of the approximation is matched to the problem. There will, in my judgment, be no escape from a very much more extensive use of experimental sampling (empirical sampling, Monte Carlo, etc.) in establishing these approximate answers. And while a little of this experimental sampling can be of a naive sort, where samples are drawn directly from the situation of concern, the great majority of it will have to be of a more sophisticated nature, truly deserving the name of Monte Carlo. (Cp., Kahn, 1956, for an introduction.)

It is, incidentally, both surprising and unfortunate that those concerned with statistical theory and statistical mathematics have had so little contact with the recent developments of sophisticated procedures of empirical sampling. The basic techniques and insights are fully interchangeable with those of survey sampling, the only difference being that many more "handles" are easily available for treating a problem of statistical theory than are generally available for treating a problem about a human population or about an aggregation of business establishments. (cp., Tukey, 1957, for an instance of interchangeability.)

As one comes to make use of all that he knows, both in replacing the original problem by one with an equivalent answer, and in being more subtle in analysis of results, one finds that no more than a few hundred samples suffice to answer most questions with adequate accuracy. (The same modifications tend to reduce the demands for extreme high quality of the underlying "random numbers".) And with fast electronic machines such numbers of samples do not represent great expenditures of time or money. (Programming time is likely to be the bottleneck.)

45. What are the necessary attitudes? Almost all the most vital attitudes can be described in a type form: *willingness to face up to X*. Granted that facing up can be uncomfortable, history suggests it is possible.

We need to face up to *more realistic problems*. The fact that normal theory, for instance, may offer the only framework in which some problem can be tackled simply or algebraically may be a very good reason for *starting* with the normal case, but never can be a good reason for STOPPING there. We must expect to tackle more realistic problems than our teachers did, and expect our successors to tackle problems which are more realistic than those we ourselves dared to take on.

We need to face up to the *necessarily approximate nature of useful results in data analysis*. Our formal hypotheses and assumptions will never be broad enough to encompass actual situations. Even results that pretend to be precise in derivation will be approximate in application. Consequently we are likely to find that results which are approximate in derivation or calculation will prove no more approximate in application than those that *pretend* to be precise, and even that some admittedly approximate results will prove to be *closer* to fact in application than some supposedly exact results.

We need to face up to the need for *collecting the results of actual experience with specific data-analytic techniques.* Mathematical or empirical-sampling studies of the behavior of techniques in idealized situations have very great value, but they cannot replace experience with the behaviour of techniques in real situations.

We need to face up to the *need for iterative procedures in data analysis.* It is nice to plan to make but a single analysis, to avoid finding that the results of one analysis have led to a requirement for making a different one. It is also nice to be able to carry out an individual analysis in a single straightforward step, to avoid iteration and repeated computation. But it is not realistic to believe that good data analysis is consistent with either of these niceties. As we learn how to do better data analysis, computation will get more extensive, rather than simpler, and reanalysis will become much more nearly the custom.

We need to face up to the need for *both indication and conclusion in the same analysis.* Appearances which are not established as of definite sign, for example, are not all of a muchness. Some are so weak as to be better forgotten, others approach the borders of establishment so closely as warrant immediate and active following up. And the gap between what is required for an interesting indication and for a conclusion widens as the structure of the data becomes more complex.

We need to face up to the need for *a free use of* ad hoc *and informal procedures in seeking indications.* At those times when our purpose is to ask the data what it suggests or indicates it would be foolish to be bound by formalities, or by any rules or principles beyond those shown by empirical experience to be helpful in such situations.

We need to face up to the fact that, as we enter into new fields or study new kinds of procedures, *it is natural for indication procedures to grow up before the corresponding conclusion procedures do so.* In breaking new ground (new from the point of view of data analysis), then, we must plan to learn to ask first of the data what it suggests, leaving for later consideration the question of what it establishes. This means that almost all considerations which explicitly involve probability will enter at the later stage.

We must face up to the need for a *double standard in dealing with error rates, whether significance levels or lacks of confidence.* As *students* and *developers* of data analysis, we may find it worth while to be concerned about small difference among error rates, perhaps with the fact that a nominal 5% is really 4% or 6%, or even with so trivial a difference as from 5% to 4.5% or 5.5%. But as *practitioners* of data analysis we must take a much coarser attitude toward error rates, one which may sometimes have difficulty distinguishing 1% from 5%, one which is hardly ever able to distinguish more than one intermediate value between these conventional levels. To be useful, a conclusion procedure need not be precise. As working data analysts we need to recognize that this is so.

We must face up to the fact that, in any experimental science, *our certainty about what will happen in a particular situation does not usually come from directly applicable experiments or theory*, but rather comes mainly through analogy between situations which are *not known* to behave similarly. Data analysis has, of necessity, to be an experimental science, and needs therefore to adopt the attitudes of experimental science. As a consequence our choices of analytical approach will usually be guided by what is known about simpler or similar situations, rather than by what is known about the situation at hand.

Finally, we need to give up the vain hope that data analysis can be founded upon a logico-deductive system like Euclidean plane geometry (or some form of the propositional calculus) and to face up to the fact that *data analysis is intrinsically an empirical science*. Some may feel let down by this, may feel that if data analysis cannot be a logico-deductive system, it inevitably falls to the state of a crass technology. With them I cannot agree. It will still be true that there will be aspects of data analysis well called technology, but there will also be the hallmarks of stimulating science: intellectual adventure, demanding calls upon insight, and a need to find out "how things really are" by investigation and the confrontation of insights with experience.

46. How might data analysis be taught? If we carry the point of view set forth here to its logical conclusion, we would teach data analysis in a very different way from any that I know to have been tried. We would teach it like biochemistry, with emphasis on what we have learned, with some class discussion of how such things were learned perhaps, but with relegation of all question of detailed methods to the "laboratory work". If we carried through the analogy to the end, all study of detailed proofs, as well as all trials of empirical sampling or comparisons of ways of presentation would belong in "the laboratory" rather than "in class". Moreover, practice in the use of data analysis techniques would be left to other courses in which problems arose, just as applications of biochemistry are left to other courses.

It seems likely, but not certain, that this would prove to be too great a switch to consider putting into immediate effect. Even if it is too much for one step, what about taking it in two or three steps?

I can hear the war cry "cookbookery" being raised against such a proposal. If raised it would fail, because the proposal is really to go in the opposite direction from cookbookery; to teach not "what to do", nor "how we learned what to do", but rather "what we have learned". This last is at the opposite pole from "cookbookery", goes beyond "the conduct of taste-testing panels", and is concerned with "the art of cookery". Dare we adventure?

47. The impact of the computer. How vital, and how important, to the matters we have discussed is the rise of the stored-program electronic computer? In many instances the answer may surprise many by being "important but not vital", although in others there is no doubt but what the computer has been "vital".

The situations where the computer is important but not vital are frequently those where the computer has stimulated the development of a method which then turns out to be quite applicable without it. FUNOP for small or moderate sized sets of values is an example. Using pen, paper, and slide rule, I find that I can FUNOP a set of 36 values in, say, twice or thrice the time it would take me to run up sums and sums of squares, and find s^2 on a desk computer. And I observe:

(1) I learn at least two or three times as much from FUNOP as from $\bar{x}$ and s^2.

(2) Hand FUNOP is faster than hand calculation of conventional measures of non-normality.

(3) It is easier to carry a slide rule than a desk computer, to say nothing of a large computer.

This is but one instance, but it is unlikely to be the only one.

On the other hand, there are situation where the computer makes feasible what would have been wholly unfeasible. Analysis of highly incomplete medical records is almost sure to prove an outstanding example.

In the middle ground stand techniques which could be done by hand on small data sets, but where speed and economy of delivery of answer make the computer essential for large data sets and very valuable for small sets. The combination of FUNOR-FUNOM and the basic vacuum cleaner (with FUNOP on the coefficient vectors) will tear down a two-way table more thoroughly than statisticians were prepared to do, even by interspersing many man hours of careful study between spells of computation, only a few years ago. With a few trimmings, such as estimation of separate variances for individual rows and columns, such a procedure, teamed with a competent statistician who could spot and follow up clues in the print-out, could greatly deepen our routine insight into two-way tables.

48. What of the future? The future of data analysis can involve great progress, the overcoming of real difficulties, and the provision of a great service to all fields of science and technology. Will it? That remains to us, to our willingness to take up the rocky road of real problems in preference to the smooth road of unreal assumptions, arbitrary criteria, and abstract results without real attachments. Who is for the challenge?

REFERENCES

ANDERSON, EDGAR (1949). *Introgressive Hybridization.* Wiley, New York.

ANDERSON, EDGAR (1957). A semigraphical method for the analysis of complex problems. *Proc. Nat. Acad. Sci. USA* **13** 923–927. (Reprinted with appended note: *Technometrics* **2** 387–391).

ANSCOMBE, F. J. and TUKEY, J. W. (1962) The examination and analysis of residuals (submitted to *Technometrics*).

AZIMOV, ISAAC (1955). The sound of panting. *Astounding Science Fiction* **55** June 104–113.

Barnard, G. A. (1959a). Control charts and stochastic processes. *J. Roy. Statist. Soc.* Ser. B. **21** 239–257 and reply to discussion, 269–271.

Barnard, G. A. (1959b). (Discussion of Kiefer, 1959). *J. Roy. Statist. Soc.* Ser. B. **21** 311–312.

Bartlett, M. S. (1947). Multivariate analysis. *J. Roy. Statist. Soc.* Suppl. **9** 176–197.

Beale, E. M. L. and Mallows, C. L. (1958). On the analysis of screening experiments. Princeton Univ. Stat. Tech. Res. Grp. Tech. Rpt. 20.

Benard, A. and Bos-Levenbach, E. C. (1953). The plotting of observations on probability paper. (Dutch) *Statistica* (Rijkswijk) **7** 163–173.

Blom, Gunnar (1958). *Statistical Estimates and Transformed Beta-Variables.* Wiley, New York.

Bortkiewicz, Ladislaus von (1901). Andwendungen der Wahrscheinlichkeitsrechnung auf Statistik. *Encyklopadie der Math. Wissenschaften* (Teil 2, 1900–1904) 822–851. Leipzig, Teubner. (Especially p. 831).

Box, G. E. P. (1956). (Discussion) *J. Roy. Statist. Soc.* Ser. B. **18** 28–29.

Bucher, Bradley D. (1957). The recovery of intervariety information in incomplete block designs. Ph.D. Thesis, Princeton, 108 pp.

Chernoff, Herman and Lieberman, Gerald J. (1954). Use of normal probability paper. *J. Amer. Statist. Assoc.* **49** 778–785.

Chernoff, Herman and Lieberman, Gerald J. (1956). The use of generalized probability paper for continuous distributions. *Ann. Math. Statist.* **27** 806–818.

Cochran, W. G. (1951). Improvement by means of selection. *Proc. Second Berkeley Symp. Math. Stat. Prob.* 449–470. Univ. of California Press.

Cornfield, Jerome and Tukey, John W. (1956). Average values of mean squares in factorials. *Ann. Math. Statist.* **27** 907–949.

Cox, D. R. (1957). The use of a concomitant variable in selecting an experimental design. *Biometrika* **44** 150–158 and 534.

Eysenck, H. J. (1950). Criterion analysis. An application of the hypotheticodeductive method to factor analysis. *Psychol. Bull.* **57** 38–53.

Eysenck, H. J. (1952). *The Scientific Study of Personality.* Routledge and Kegan, London.

Finney, David J. (1947), (1952). *Probit Analysis; a Statistical Treatment of the Sigmoid Response Curve.* Cambridge Univ. Press.

Fisher, R. A. (1929). Tests of significance in harmonic analysis. *Proc. Roy. Soc. London* Ser. A. **125** 54–59.

Fisher, R. A. (1935), (1937, 1942, 1947, ···). *The Design of Experiments.* Oliver and Boyd, Edinburgh and London. (especially Sec. 41).

Gauss, Carl Friedrich (1803). Disquisitiones de elementis ellipticis pallidis. *Werke* **6** 20–24 (French translation by J. Bertrand, 1855. English translation by Hale F. Trotter. Princeton Univ. Stat. Tech. Res. Grp. Tech. Rpt. 5.)

Gayen, A. K. (1949). The distribution of 'Student's' t in random samples of any size drawn from non-normal universes. *Biometrika* **36** 353–369 (and references therein.)

Geary, R. C. (1947). Testing for normality. *Biometrika* **34** 209–242.

Hannan, E. J. (1958). The estimation of spectral density after trend removal. *J. Roy. Statist. Soc.* Ser. B. **20** 323–333.

Johnson, N. L. (1949). Systems of frequency curves generated by methods of translation. *Biometrika* **36** 149–176.

Kahn, Herman (1956). Use of different Monte Carlo sampling techniques. *Symposium on Monte Carlo Methods* 146–190 (edited by Herbert A. Meyer). Wiley, New York.

Kiefer, J. (1959). Optimum experimental designs. *J. Roy. Statist. Soc.* Ser. B. **21** 272–304 and reply to discussion 315–319.

Kimball, George E. (1958). A critique of operations research. *J. Wash. Acad. Sci.* **48** 33–37 (especially p. 35).

Lerner, I. Michael (1950). *Population Genetics and Animal Improvement.* Cambridge Univ. Press.

MALLOWS, C. L. (1959). (Discussion of Kiefer, 1959). *J. Roy. Statist. Soc.* Ser. B. **21** 307–308.

MANN, H. B. and WALD, A. (1942). On the choice of the number of intervals in application of the chi square test. *Ann. Math. Statist.* **13** 306–317.

MICHENER, C. D. and SOKAL, R. R. (1957). A quantitative approach to a problem in classification. *Evolution* **11** 130–162.

MOSTELLER, FREDERICK and WALLACE, DAVID L. (1962). Notes on an authorship problem. *Proc. Harvard Symp. Digital Computers and their Applications*. Harvard Univ. Press (in press).

PEARCE, S. C. (1953). Field experimentation with fruit trees and other perennial plants. Commonwealth Bur. Horticulture and Plantation Crops. Tech. Comm. 23.

PEARSON, KARL (1895). Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philos. Trans. Roy. Soc. London* Ser. A. **186** 343–414.

PEARSON, KARL (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Phil. Mag.* (5) **50** 157–175.

PENROSE, L. S. (1947). Some notes on discrimination. *Ann. Eugenics* **13,** 228–237.

PITMAN, E. J. G. (1938). Significance tests which may be applied to samples from any populations. III. The analysis of variance test. *Biometrika* **29** 322–335.

RAO, C. RADHAKRISHNA (1952). *Advanced Statistical Methods in Biometric Research*. Wiley, New York.

RAO, C. RADHAKRISHNA (1959). Some problems involving linear hypotheses in multivariate analysis. *Biometrika* **46** 49–58.

ROGERS, DAVID J. and TANIMOTO, TAFFEE T. (1960). A computer-program for classifying plants. *Science* **132** 1115–1118.

ROY, J. (1958). Step-down procedure in multivariate analysis. *Ann. Math. Statist.* **29** 1177–1187.

ROY, S. N. and GNANADESIKAN, R. (1957). Further contributions to multivariate confidence bounds. *Biometrika* **44** 399–410.

ROY, S. N. AND GNANADESIKAN, R. (1958). A note on "Further contributions to multivariate confidence bounds". *Biometrika* **45** 581.

SARHAN, ARMED E. and GREENBERG, BERNARD G. (1958). Estimation of location and scale parameters by order statistics from singly and doubly censored samples. *Ann. Math. Statist.* **29** 79–105.

SIEGEL, SIDNEY and TUKEY, JOHN W. (1960). A non-parametric sum of ranks procedure for relative spread in unpaired samples. *J. Amer. Statist. Assoc.* **55** 429–445.

SNEATH, P. H. A. (1957a). Some thoughts on bacterial classifications. *J. Gen. Microbiol.* **17** 184–200.

SNEATH, P. H. A. (1957b) The application of computers to taxonomy. *J. Gen. Microbiol.* **17** 201–226.

SNEATH, P H. A. and COWAN, S. T. (1958). An electrotaxonomic survey of bacteria. *J. Gen. Microbiol.* **19** 551–565.

STEVENS, W. L. (1950). Fiducial limits of the parameter of a discontinuous distribution. *Biometrika* **37** 117–129.

STUDENT (William Sealy Gosset) (1927). Errors of routine analysis. *Biometrika* **19** 151–164. (Reprinted as pages 135–149 of "Student's" *Collected Papers* (edited by E. S. Pearson and John Wishart) Biometrika Office, Univ. Coll. London 1942 (1947).

TUCKER, LEDYARD R. (1958). An interbattery method of factor analysis. *Psychometrika* **23** 111–136.

TUKEY, JOHN W. (1949a). Dyadic anova—An analysis of variance for vectors. *Human Biology* **21** 65–110.

TUKEY, JOHN W. (1949b). One degree of freedom for non-additivity. *Biometrics* **5** 232–242.

TUKEY, JOHN W. (1951). Components in regression. *Biometrics* **7** 33–69. (especially pp 61–64.)

TUKEY, JOHN W. (1957). Antithesis or regression. *Proc. Camb. Philos. Soc.* **53** 923–924.

TUKEY, JOHN W. (1960). A survey of sampling from contaminated distributions. Paper 39 (pp. 448–485) in *Contributions to Probability and Statistics* (edited by I. Olkin *et al*). Stanford Univ. Press.

TUKEY, JOHN W. (1961a). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics* **3** 191–219.

TUKEY, JOHN W. (1961b). Statistical and quantitative methodology. In *Trends in Social Sciences*, (edited by Donald P. Ray) Philosophical Library, New York.

TUKEY, JOHN W. (1962). The symmetrical λ-distributions (in preparation).

VARIOUS AUTHORS (1959). Various titles. *Technometrics* **1** 111–209.

VON NEUMANN, JOHN (1947). The mathematician, pp. 180–196 of the *Works of the Mind.* (edited by R. B. Heywood). Chicago Univ. Press. (Reprinted as pp. 2053–2069 of the *World of Mathematics* (edited by James R. Newman). Simon and Schuster, New York, 1956).

WALSH, JOHN E. (1949a). On the range-midrange test and some tests with bounded significance levels. *Ann. Math. Statist.* **20** 257–267.

WALSH, JOHN E. (1949b). Some significance tests for the median which are valid under very general conditions. *Ann. Math. Statist.* **20** 64–81.

WALSH, JOHN E. (1959). Comments on "The simplest signed-rank tests". *J. Amer. Statist. Assoc.* **54** 213–224.

WELCH, B. L. (1937). On the z-test in randomized blocks and latin squares. *Biometrika* **29** 21–52.

WIENER, NORBERT (1930). Generalized harmonic analysis. *Acta Math.* **55** 118–258.

WILK, M. B. AND GNANADESIKAN, R. (1961). Graphical analysis of multiple response experimental data using ordered distances. *Proc. Nat. Acad. Sci. USA* **47** 1209–1212.

YATES, F. (1955). (Discussion) *J. Roy. Statist. Soc.* Ser. B. **17** 31.