DATA ANALYSIS AND STATISTICS: AN EXPOSITORY OVERVIEW *

J. W. Tukey and M. B. Wilk

Princeton University and
Bell Telephone Laboratories, Inc.
Princeton and Murray Hill, New Jersey

INTRODUCTION

Data analysis is not a new subject. It has accompanied productive experimentation and observation for hundreds of years. At times, as in the work of Kepler, it has produced dramatic results.

As in any other science, what is done in data analysis is very much a product of each day's technology. Every technological development of major relevance—organized tables of functions, knowledge of the mathematical consequences of the Gaussian law of error, desk calculators, stored-program electronic computers, graphical display facilities—has been accompanied by a tendency to rediscover the importance and to reformulate the nature of data analysis.

Today, as in the past, data analysis is usually difficult, cumbersome, and complex—and often very time-consuming, both in man-hours and in elapsed time. It is also of mushrooming importance in business, politics, science and technology.

The basic general intent of data analysis is simply stated: to seek through a body of data for interesting relationships and information and to exhibit the results in such a way as to make them recognizable to the data analyzer and recordable for posterity. Its creative task is to be productively descriptive, with as much attention as possible to previous knowledge, and thus to contribute to the mysterious process called insight.

Four major influences act on data analysis today:

1. The formal theories of statistics.
2. Accelerating developments in computers and display devices.
3. The challenge, in many fields, of more and ever larger bodies of data.
4. The emphasis on quantification in an ever wider variety of disciplines.

The last few decades have seen the rise of formal theories of statistics, "legitimizing" variation by confining it by assumption to random sampling, often assumed to involve tightly specified distributions (in which a bare minimum of adjustable constants deny almost all flexibility) and restoring the appearance of security by emphasizing narrowly optimized techniques and claiming to make statements with "known" probabilities of error. While many of the influences of statistical theory on data analysis have been helpful, some have not.

Exposure, the effective laying open of the data to display the unanticipated, is to us a major portion of data analysis. Formal statistics has given almost no

* Prepared in part in connection with research at Princeton University sponsored by the Army Research Office (Durham).

guidance to exposure; indeed, it is not clear how the informality and flexibility appropriate to the exploratory character of exposure can be fitted into any of the structures of formal statistics so far proposed.

Many bodies of data require routine handling, not data analysis, and can be rightly approached with specific narrow questions. In dealing with them, some facets of formal statistics are more nearly appropriate, and have proved much more useful.

As a discipline, data analysis is a very difficult field. It must adapt itself to what people can and need to do with data. In the sense that biology is more complex than physics, and the behavioral sciences are more complex than either, it is likely that the general problems of data analysis are more complex than those of all three. It is too much to ask for close and effective guidance for data analysis from *any* highly formalized structure, either now or in the near future.

Data analysis can gain much from formal statistics, but only if the connection is kept adequately loose.

The impact on data analysis of the availability and capacity of computer hardware and software has already been substantial, but the full harnessing of this potential has hardly begun. Slowness in development of appropriate computing systems reflects the difficulty of the rethinking and restructuring of the science and art of data analysis which needs to be done.

The revolutionary computer and display developments now taking place will inevitably stimulate major extensions and departures in data analysis, whose beginnings are already visible. Today's first task is not to invent wholly new techniques, though these are needed. Rather we need most vitally to recognize and reorganize the essentials of old techniques, to make easy their assembly in new ways, and to modify their external appearances to fit the new opportunities.

Data has typically been easier to gather than to analyze, though there are outstanding exceptions. Despite the gains in computation and display—perhaps because of them—this is increasingly true. In so many fields, the accumulation of large volumes of data is becoming irresistibly practical and economical. As in the past, much, perhaps most, of even carefully collected data will not be adequately analyzed. In part this is because facts are usually more complex than the hopes which have led to their accumulation; in part because the accumulation of data dampens experimental excitement; in part because collecting data simply serves to keep the experimenter busy while he designs a more adequate experimental setup or develops a needed point of view; but also, in part, because the technology of data analysis is still unsystematized and many of those who could put its tools to good use are unable to do so effectively.

While the data-analysis facilities of the present and the foreseeable future entirely dwarf those of even the near past, it is apparent that the challenge to data analysis is growing instead of receding. Thirty years ago, many thought that good data-analytical techniques were really needed only when data was sparse. Today we also recognize that only the best data analysis will suffice when the data is very extensive. As bodies of data grow in size, the number of essentially different ways of approaching them increases, and the effort involved in their analysis threatens to grow even faster. The analysis of mass data challenges all of our ingenuity and resourcefulness. Meeting this challenge will also help us to handle small amounts of data more effectively and thoroughly.

Increasingly, most disciplines are evolving towards increased quantification and mathematization. Many of them (e.g., medicine) have had long histories of being descriptive and relatively qualitative largely because of the complexities of their phenomena and systems. In these, the demand on data-analysis techniques is often not only that they function in some sort of real time, but even that they perform as well as current expert judgment (which has often been developed over generations and perhaps is based on "data" still unrecognized). The resulting challenges to data analysis are major and increasing.

The wider variety of problems thus brought to data analysis increases the importance of recognizing general elements in diverse problems and untangling these elements from more specific ones. The need is to recognize and understand the similarities and differences of data analyses in nuclear physics, in the physiology of cell nuclei, in cloud-seeding, in the assay of antiviral agents, in chemical engineering, and in opinion polling, to select but a few.

DATA ANALYSIS IS LIKE DOING EXPERIMENTS

Far too many people, in the past and even in the present, have persisted in regarding statistics, and

even data analysis, as a branch of probability theory, nestled deep within modern mathematics. Happily this view is increasingly out of favor. Statistical data analysis is much more appropriately associated with the sciences and with the experimental process in general.

The general purposes of conducting experiments and analyzing data match, point by point. For experimentation, these purposes include (1) more adequate description of experience and quantification of some areas of knowledge; (2) discovery or invention of new phenomena and relations; (3) confirmation, or labeling for change, of previous assumptions, expectations, and hypotheses; (4) generation of ideas for further useful experiments; and (5) keeping the experimenter relatively occupied while he thinks.

Comparable objectives in data analysis are (1) to achieve more specific description of what is loosely known or suspected; (2) to find unanticipated aspects in the data, and to suggest unthought-of models for the data's summarization and exposure; (3) to employ the data to assess the (always incomplete) adequacy of a contemplated model; (4) to provide both incentives and guidance for further analysis of the data; and (5) to keep the investigator usefully stimulated while he absorbs the feeling of his data and considers what to do next.

Among the important characteristics shared by data analysis and the experimental process are these:

1. Some prior presumed structure, some guidance, some objectives, in short some ideas of a model, are virtually essential, yet these must not be taken too seriously. Models must be used but must never be believed. As T. C. Chamberlain [1] said, "Science is the holding of multiple working hypotheses."

2. Our approach needs to be multifaceted and open-minded. In data analysis as in experimentation, discovery is usually more exciting and sometimes much more important than confirmation.

3. It is valuable to construct techniques that are likely to reveal such complications as assumptions whose consequences are inappropriate in a specific instance, numerical inaccuracies, or difficulties of interpretation of what is found.

4. In both good data analysis and good experimentation, the findings often appear to be obvious—but generally only after the fact.

5. It is often more productive to begin by obtaining and trying to explain specific findings, rather than by attempting to catalog all possible findings and explanations.

6. While detailed deduction of anticipated consequences is likely to be useful when two or more models are to be compared, it is often more productive to study the results before carrying out these detailed deductions.

7. There is a great need to do obvious things quickly and routinely, but with care and thoroughness.

8. Insightfulness is generally more important than so-called objectivity. Requirements for specifiable probabilities of error must not prevent repeated analysis of data, just as requirements for impossibly perfect controls are not allowed to bring experimentation to a halt.

9. Interaction, feedback, trial and error are all essential; convenience is dramatically helpful.

10. There can be great gains from adding sophistication and ingenuity—subtle concepts, complicated experimental setups, robust models, delicate electronic devices, fast or accurate algorithms—to our kit of tools, just so long as simpler and more obvious approaches are not neglected.

11. Finally, most of the work actually done turns out to be inconsequential, uninteresting, or of no operational value. Yet it is an essential aspect of both processes to recognize and accept this feature, with its momentary embarrassments and disappointments. A broad perspective on objectives and unexpected difficulties is often required to muster the necessary persistence.

In summary, data analysis, like experimentation, must be considered as an open-ended, highly interactive, iterative process, whose actual steps are selected segments of a stubbily branching, tree-like pattern of possible actions.

DATA ANALYSIS IN RELATION TO PEOPLE

The science and art of data analysis concerns the process of learning from quantitative records of experience. By its very nature it exists in relation to people. Thus, the techniques and the technology of data analysis must be harnessed to suit human requirements and talents. Some implications for effective data analysis are: (1) that it is essential to have convenience of interaction of people and interme-

diate results and (2) that at all stages of data analysis the nature and detail of output, both actual and potential, need to be matched to the capabilities of the people who use it and want it.

Data analysis must be iterative to be effective. Human judgment is needed at almost every stage. (We may be able to mechanize an occasional judgment.) Unless this judgment is based on good information about what has been found, and is free to call next for what is now indicated, there cannot be really effective progress.

Nothing—not the careful logic of mathematics, not statistical models and theories, not the awesome arithmetic power of modern computers—nothing can substitute here for the flexibility of the informed human mind. Accordingly, both approaches and techniques need to be structured so as to facilitate human involvement and intervention.

It is insufficient to have results produced; they must be displayed in a manner to satisfy diverse needs of a broad spectrum of individuals. The general goals include (1) using a body of data to answer specific questions, either formulated in advance or developed during the analysis; (2) exposing to people information and structure in the data that is of an unanticipated nature; and (3) compressed summarization for record and communication. All three require communication to people. For each of these purposes, the use of pictures is invaluable, both for guidance at intermediate stages and for final communication or summary.

In many instances, a picture is indeed worth a thousand words. To make this true in more diverse circumstances, much more creative effort is needed to pictorialize the output from data analysis. Naive pictures are often extremely helpful, but more sophisticated pictures can be both simple and even more informative.

For humans, the use of appropriate pictures often offers the possibility of great flexibility all along a scale from broad summary to fine detail, since pictures can be viewed in so many different ways. Moreover, one can change his approach from summary toward detail with great ease and great speed.

A few people will want to understand as much as they can about a given body of data. For them, both numerical detail and many pictures are useful. Others might like to know so much about many bodies of data, that they will have to be content with summarizations, which may often have to be extreme and accordingly somewhat misleading. Data analysis must satisfy needs for both extremes, and, as well, for intermediate degrees of detail. (Formally, and historically, the emphasis has been too much on summarization.)

THE KEY TO EFFECTIVE DATA ANALYSIS

The iterative and interactive interplay of summarizing by fit and exposing by residuals is vital to effective data analysis. Summarizing and exposing are complementary and pervasive.

If certain aspects of the data have been effectively summarized by fitting a straight line, then we can improve exposure by building on this summarization. The plot of y against x is likely to be dominated by what has been summarized by the straight line. A plot of the residuals from the fit against x (or other variables) exposes to view only what has not been summarized, thus avoiding unnecessary distraction and permitting attention to finer details.

Even when used for confirmation alone, data analysis is a process of first summarizing according to the hypothesized model and then exposing what remains, in a cogent way, as a basis for judging the adequacy of this model or the precision of this summary, or both.

Techniques for summarizing are, fortunately, often useful for exposing and vice versa. For example, a half-normal plot [2] of contrasts in a 2^n experiment may serve as an effective summary of the data, with identification of interesting effects. It may also serve as an exposing technique in indicating, for instance, the unanticipated existence of two error terms.

This process of summarizing and exposing is intrinsically iterative. No step is clearly the last before it is taken.

Summarizing data is a process of constrained and partial description—a process that essentially and inevitably corresponds to some sort of fitting, though it need not necessarily involve formal criteria or well-defined computations.

To fit a straight line is to select one of a very restricted family of formal descriptions (a family involving only two constants) and to regard the line, or some coefficients that specify it, as a partial description of the data. To fit row and column means to a 2-way table, or to fit 9 main effects and 5 2-factor interactions to the data of a 2^{9-2} fractional factorial experiment, is to do something entirely similar,

differing mainly in the number of adjustable constants.

A process that is closely similar to these examples is drawing (freehand) a curve that is both increasing and "smooth" and that graduates some data reasonably well. There is, so far as we know, no finite set of adjustable constants that describe all "smooth" curves, but the family of acceptable curves is surely constrained. There is no precisely specified way to conduct the fit, but the result must surely be interpreted as a partial description.

Recognition of the iterative character of the relationship of exposing and summarizing makes it clear that there is usually much value in fitting, even if what is fitted is neither believed nor satisfactorily close. What is left over after the partial description from fitting can often be more effectively approached and structured because there has been some fit, even a poor one.

USING RESIDUALS EFFECTIVELY

When a "fit" has a sufficiently arithmetic character, there is a natural way to express what the fit has not described. Additive residuals defined by

$$\text{observation} = \text{fit} + \text{residual}$$

are widely used.

In other circumstances, it may be appropriate to express residuals in still other ways. After all, we are concerned with appropriate measures of deviation at each of several or many places, and there are many reasons why the appropriate way to measure deviation may vary from place to place, as well as from example to example. Multiplicative residuals or residual factors defined by

$$\text{observation} = \text{fit} \times \text{residual factor}$$

are sometimes useful, but are usually easily reduced to additive residuals by the taking of logarithms. If we are concerned with outlines of leaves, it may be reasonable to measure the deviation of actual outline from fitted outline at each of a number of places, probably at right angles to the fitted outline. And where observation of angles leaves a multiple of 2π undetermined, residuals are often usefully expressed by the sine of the difference between observed and fitted.

Adequate examination of residuals is one of the truly incisive tools of exposure. Perhaps because of the blinding effects of unrealistic optimism about assumptions, perhaps because of the difficulties of computational practice during the era before modern computing evolved, and to an unfortunate degree because computer centers have felt that data output should be compressed, residuals have not received the attention and use they richly deserve.

There is no substitute for examining the collection of detailed individual residuals in diverse ways. It is almost always a sad inadequacy (though far better than nothing) to try to summarize the exposing information in a body of residuals by a mean square error. Even the computation of the individual residuals and their examination as an unstructured mass is not enough. Several graphs of residuals are usually in order. Related numerical analyses, which answer specific questions more stringently but more general questions hardly at all, can also be useful.[3,4]

Kinds of plots of residuals that are very often valuable include (1) plots against fitted (or observed) values; (2) plots against variables which were employed in the summarizing fit; (3) plots against variables not used in the fit (e.g., time); (4) probability plots of ordered residuals, particularly plots of empirical quantiles against quantiles of reference distributions, such as the unit normal.

The first three kinds of plots are often effective in showing what changes in style of fit are needed. Probability plots are particularly helpful in indicating a few peculiar values and in illuminating the overall success of the fit. They provide quick information about location, about spread, about distributional peculiarities, and a palatable summary of individual residual values. In any residual plot it will be helpful to identify each individual residual according to whether or not it comes from an observation which was used in developing the fit. There should be an effort to identify and make evident other important qualitative characteristics of individual residuals.

The usefulness of residuals as a means of exposure depends on the summarizing model having identifiable deficiencies. Accordingly, residuals may fail to reveal deficiencies of a summarizing model when these are too varied and general, or when the data points are few in number or badly distributed.

An example of this is the behavior of residuals from an additive fit in a two-way/classification table. If the departure from additivity is due to the existence of just one highly deviant cell in the table, examining the residuals as a whole will tend to expose this cell. If there are two such deviant values, their

effects can combine to conceal any obvious peculiarities [5] so long as the individual residuals are examined as an unstructured set of numbers.

THE STRATEGY OF DATA ANALYSIS

In addition to the two-pronged use of summarization and exposure, including careful attention to residuals, three of the main strategies of data analysis are:

1. Graphical presentation.
2. Provision of flexibility in viewpoint and in facilities.
3. Intensive search for parsimony and simplicity, including careful reformation of variables and bending the data to fit simple techniques.

Some people are apparently able to absorb broad information from tables of numbers.[6] Most of us can only appreciate matters with full insight by looking at graphical representations. For large-scale data analysis, there is really no alternative to plotting techniques, properly exploited. A picture is not merely worth a thousand words, it is much more likely to be scrutinized than words are to be read. Wisely used, graphical representation can be extremely effective in making large amounts of certain kinds of numerical information rapidly available to people.

Flexibility in viewpoint and in facilities must be built into both the general technology and the individual techniques of data analysis. We must have flexibility in the choice of a model for summarization, in the selection of the data to be employed in computing the summary, in choosing the fitting procedures to be used and in selecting the terms in which the variables are to be expressed. Flexibility in assembly and reassembly of techniques is crucial.

Using human judgment in selection, or cleaning-up, of the data by partial or complete suppression of apparently aberrant values is natural, sensible, and essential. Data is often dirty. Unless the dirt is either removed or decolorized, it can hide much that we would like to learn. Sometimes, it is true, the dirt is blue clay, and contains diamonds in the form of new phenomena and new insights. Whether it is worth much or little to prospect for diamonds among the consequences of a particular set of data, we can do this better after labeling as dirt whatever *appears* from analysis to be such. Moreover, whether or not it is diamond-bearing, clearing out the dirt can do much to help us learn from the cleaner data.

Suppression may be complete and wholly human, as when we decide to exclude a particular set of observations from all computations. Suppression may be partial and wholly automatic, as when we use the median of a set of observations, or the mean of the values in the two center quarters of the empirical distribution. In practice, the processes of selection and suppression are mixed up, rather complex, and for all that quite essential.

Just because values have been suppressed in fitting is no reason for their residuals from the fit to be forgotten. Not every suppression will have suppressed mere dirt. Some will have suppressed clean data, others will have suppressed diamonds. We will never be wholly sure which is which, but calculating, looking at, and thinking about residuals from suppressed data often stimulates the further questions needed to help clear up the situation.

The importance of parsimony in data analysis can hardly be overstated. By parsimony we mean *both* the use of few numerical constants and *also* the avoidance of undue complexity of form in summarizing and displaying. The need for parsimony is both aesthetic and practical.

In general, parsimony and simplicity will be achieved in the summary description either at the price of inadequacy of description or at the price of complexity in the model or in the analysis. Typically those who insist on doing only primitive analyses must often be satisfied with complex—not parsimonious—summaries which often miss important points.

An additional value from parsimony is illustrated in the following idealized example: A cubic in x will always fit the particular numbers that make up our body of data somewhat more closely than a quadratic. This may easily be more a seeming than a truth. If y only differs from a quadratic function of x by fluctuations, independent from one x to another, the fitted quadratic will fit the *average* values of y given x more closely, on the average, than will the fitted cubic.

One further aspect of great strategic importance in data analysis involves the transformation, better called reformation, of variables. Especially insightful choices of modes of expression underlie much of physical science. Changing from raw values to their square roots or logarithms (or other appropriate function) before the data is analyzed is often aston-

ishingly effective. Equally important is the evolution and use of techniques of analysis by which the data itself may be employed to indicate useful transformations.

As a matter of general strategy we may note here that it is almost always easier, and usually better, to "unbend" data to fit known analysis techniques than to bend the techniques to fit the data. If the square root, or logarithm, or reciprocal behaves in a simpler way than the raw form, it is obviously unwise to work with the data in the form where its behavior is more complex. With enough effort we can probably bend any of our techniques of data analysis to work explicitly and effectively on the data in its raw form, but this effort is rarely justified.

FITTING, THE WORKHORSE OF DATA ANALYSIS, HAS VARIED OBJECTIVES

The single most important process of data analysis is fitting. It is helpful in summarizing, exposing and communicating. Each fit (1) gives a summary description, (2) provides a basis for exposure based on the residuals, and (3) may have the parsimony needed for effective communication.

Fitting inevitably raises questions concerning classes of models to be used, selection of criteria of fit, choices of mode of expression for observations, as well as questions of numerical and logical algorithms. The answers to all these questions depend upon the diversity of the objectives of fitting, their character, and the differences amongst them.

These objectives include:

1. *Pure description,* in the sense of drawing, possibly hastily, a curve across the page and saying y appears to depend on x just about this way. If this is our only aim, we do want the curve to fit well, but we do not care at all whether its functional form is more than an accident. Finding, for instance, that a cubic polynomial fits our data well enough is not, at this level, to be thought of as giving any particular support for a cubic "law."

2. *Local prediction,* in the sense that, so long as the situation "remains the same," we should like to do well by substituting x's into the fit and regarding the result as predicting the value of y. This amounts to hoping that our description of the past, however empirical, will continue to be a good description of the future.

3. *Global prediction of local change,* in the sense that we can use our fit to assess the result (averaged over fluctuations) of changing one or more x's moderately, even when both the start and the finish of this change are far from the circumstances for which the fit was developed. If this is to be accomplished successfully, the general situation must be favorable, and theory, or insight, or broad experience must have been responsible for choosing the form of the fit and the nature of the y variables; the data before us can rarely be used to narrow things down enough to provide such good prediction, even of changes, elsewhere.

4. *Global prediction of values,* in the sense that we can use our fit to predict y given x far outside the range of the data on which it was based. Reliance upon outside information (including insight) is now even greater, and the chances of success are correspondingly diminished.

5. Using a fit depending on several mathematical variables (some of which may be functions of the same physical variable) to tell us *which variables have influences and which do not* (which can include telling us about the forms of the dependencies). This is sometimes possible, but nowhere nearly as often as is commonly hoped. Very frequently several alternative sets of variables will each give a satisfactory fit.

6. Using the fit to *estimate coefficients* having the general character of physical constants. Careful descriptions of both what is to be *varied* and what is to be held *constant* are essential before there is any hope of doing this effectively. "Heat capacity," for example, is not an adequate name. Heat capacity at constant volume differs substantially, both in meaning and value, from heat capacity at constant pressure. In most circumstances, indeed, constants to be assessed are not even as simply defined as heat capacity, rather they are only defined in terms of specific, rather complex functional forms.

These six objectives center on what has been fitted rather than on the other essential ingredient of fitting, what remains after the fit. Residuals have two quite distinct sorts of uses. On the one hand, they can be used as an immediate basis for further summarization, as in:

7. Providing *adjusted values for further study,* as when economic series are seasonally adjusted, or when the analysis of covariance is applied.

8. Providing *a basis for immediate further fitting,* as when the residuals from an eye-fitted straight line are fitted by either a further straight line or a quadratic. (So-called stepwise regression procedures operate in this general way, though they tend to omit the calculation of actual residuals.)

The more usual objectives for residuals emphasize exposure and include:

9. Examining and exposing with a view to learning about the *inadequacy of the fit.*

10. Examining and exposing with a view to identifying *peculiar values,* either for study in their own right or for suppression, partial or complete, from further analysis.

We need not assess the relative value and frequency of these specific objectives of fitting—the real need is to identify and distinguish varied objectives (of which these 10 are not all), to recognize their diversity, their tendency to occur one or a few at a time, and the consequent great variety of different demands made upon the fitting process itself and upon such associated procedures as plotting of residuals. No one fitting-and-residuals procedure can serve all our purposes.

REFORMATION OF VARIABLES

When is one expression better than another *for analysis*? Basically, when the data are more simply described, since this implies easier and more familiar manipulations during analysis and, even more to the point, easier and more thorough understanding of the results.

The usual goals of better expression include:

1. Additivity of effects
2. Constancy of variance
3. Normality (Gaussianity) of distribution
4. Linearity of relationship

Widespread and clear understanding of the relative desirability among the first three goals has been impeded by a happy fact—one that only appears accidental: All three tend to occur together. Where a choice has to be made among these three, additivity is to be preferred above all [7] with constancy of variance second.

Linearity of relationship is important for both arithmetic manipulation and graphical presentation. If a response needs to be related to only one factor upon which it depends monotonely, a suitable change of the expression of *either* the factor *or* the response will make the relationship linear. However, when as usual, we have to deal with two or more factors, we may be unable to reach linearity by any such simple device. In most such circumstances, linearity may be achieved if we can attain additivity since appropriate reexpression of *the factors* will then make the dependence of the response on them linear.

Some ways of seeking out desirable expressions are:

1. Explicit trial of various alternatives, whose evaluation is usually best done in terms of corresponding residuals.[3, 8]
2. Use of numerical guides to the next choice.[3, 4]
3. Use of computer iteration to seek that monotone change of expression which produces the greatest amount of additivity.[9]

All of these approaches work. Which one is desirable in a particular case depends upon objectives, on the amount and shape of the data, and on the availability of computing and display equipment.

The gains from appropriate reexpressing, transforming—more simply *reforming*—individual variables are likely to be substantial. More major gains (e.g., gains in efficiency by factors of 2, 3, or often much more) come from such efforts than from any other data-analytic step.

SCALING IN DATA ANALYSIS

Measures of similarity (or dissimilarity), even when expressed on a rubber scale, can be used to generate quantitative variables (see Refs. 10–14). Even starting with several responses, each expressed on a well-established numerical scale and thus able to serve as coordinates in a Cartesian space, these (and related) methods can sometimes generate nonlinear transformations of the original responses from "distances" among the points in the initial space.

NEW LINEAR COMBINATIONS FOR OLD

Interest in replacing one set of variables with another made up of linear combinations of the first variables arises:

1. in preparation for dropping some of the new variables;

2. in order to make calculations involving these variables either simpler or more understandable;
3. in a search for a more meaningful or insightful coordinate system.

Canonical Analysis

The computations classically used for calculating canonical correlation coefficients and canonical variates can be used in much more general situations to provide an ordered family of linear combinations of the original working variables, guided by the ability of initial subsets of this family to describe, through linear regression, the behavior of one or more guide variables. Principal components are logically, but probably not computationally, just a particular case.

Orthogonalization

Recognition and elimination of close approximations to linear dependences is almost always important in three ways: computationally, descriptively, and conceptually. Direct quantitative description of amount of dependence upon factors that are substantially correlated with one another still appears almost hopeless, at least so far as communication to the human mind is concerned. Computational difficulties from unrecognized near linear dependencies can be very great. Changes of coordinates to avoid such problems are often very useful.

Complete elimination of correlation—precise orthogonality—is of little consequence to us, so long as our arithmetic processes do not assume it. Still, in practice, we usually seek "orthogonality," mainly because it is specific, clearly defined, and related to simple algorithms. In particular, variables made orthogonal for one set of data are often thoroughly useful in analyzing another, where they are only approximately orthogonal. This happens most frequently, perhaps, when the second set of data is a subset of the first, or, conversely, when the first set is a random sample of the second (as may often be computationally convenient in dealing with large bodies of data).

Orthogonal polynomials in other than the usual order may be useful. Orthogonalization of more general variables is also often both convenient and useful. Notice that canonical variates, pure or modified, (and thus, in particular, principal components) are automatically orthogonal.

Rotation

Rotation of an initial coordinate system to a more useful position can be helpful, but requires quite explicit information about the advantages and disadvantages of specific choices.

CONCENTRATING ON A SUBSET OF THE VARIABLES

Even in the simplest case of naturally or prescriptively ordered variables, working from limited and fallible data, as we always do, offers NO hope of always, or even usually, dividing the variables into ONLY two classes: those it appears we must keep, and those surely of no importance.

Since the "middle class" variables should usually *be carried on to further analysis,* it will almost always NOT be vitally important to be highly precise in just how we select a sequence of linear combinations, the first k of which we are to carry on to further analysis.

In the more general problem of choosing any subset of variables for retention, we face new problems. The 11th, 23rd and 47th variables may, for example, be so highly correlated with each other that any one can deputize for any other without appreciable loss. When this is so, no one can be essential, since it could be replaced; but neither can we be sure that any one of them is *not* important. Also, it is easy to produce a situation where either of two variables alone has negligible descriptive power, yet combined they are very effective.

Taken as a means of selecting reasonably satisfactory subsets and giving some indication of their comparative performance, procedures of stepwise, screening or "steered" regression can prove very helpful in many situations, particularly if the uncertainties and inadequacies of the objectives and the results are clearly recognized.

The basic idea of steered regression is iterative use of the following step: Ask how much can be gained (in the quality of one or more regressions) by adding each of the remaining variables to the currently selected subset, and then add to the subset that variable which gains the most. It is often, perhaps usually, useful to include another process in the iteration. Ask how little it costs to exclude from the subset each of the variables currently in it, and then, if the cost is low enough, remove the least costly variable from the subset. There also needs to be a stopping rule.

Conventionally, gains and costs are assessed in terms of changes in the residual sum of squares.[15–17] More complex measures seem more reasonable in many, if not most, circumstances, however, and the evolution of steered regression is likely to involve more varied and insightful choices of steering functions.

GENERAL CONSIDERATIONS IN GRAPHICAL PRESENTATION

Graphical presentation appears to be at the very heart of insightful data analysis. For most people, graphs convey more of a message than tables and do so more persuasively and attractively. Graphical presentation continues to hold its preeminent place despite both feeble understanding of the reasons for its power and appeal and severe limitations on the variety and character of its techniques, the latter stemming both from past technological limitations and from continuing inadequacies of imagination. Why?

Some reasons are easily found: Graphical displays can be very flexible. The human eye and brain are speedy and proficient in recognizing certain types of geometric configurations. "Smoothness" seems to be very much a geometric concept. The eye seems much more able to comprehend nonunderstood graphs than nonunderstood numbers. Quite large volumes of data can be displayed economically and comprehensibly. And the same graph can transfer effectively either a very compressed summary or an extensive amount of detail, as well as many intermediate packages of information.

Before we discuss some of these reasons in more detail, one key point must be made. While it is often most helpful to "plot the data," this is rarely enough. We need also to "plot the results of analysis" as a routine matter. (There is often more analysis than there was data.)

The innate flexibility of graphical displays is many-sided. It is not merely that we can choose to do many different things graphically. If one expects (or only contemplates the possibility) that y is approximately linear in x, a simple plot will confirm this when it is so and be even more instructive when it is not. If one has no clear anticipation, the same simple plot is likely to reveal whichever one of many alternative structures appears to be present, even though these structures are nowhere collected in a list. (We may, indeed, even doubt whether it is humanly possible to list them all.)

The human eye and brain join easily and speedily in recognizing straightness and certain kinds of smoothness, in assessing amounts of local roughness or local variability (especially when properly aided by a reference "curve"), and in judging the presence or absence of systematic deviation (when the reference curve is neither too steep nor too wiggly). With less precision and more effort, eye and brain can judge symmetry and circularity moderately well, and have a fair chance of recognizing that certain features occur in a roughly periodic way. Further, of basic importance though more difficult to verbalize, the human eye and brain can learn to recognize quite complex and varied configurations with surprising effectiveness.

Almost all graphical techniques correspond to one or more natural reference situations. Graphs are most effective when these conceptually simple situations produce simple configurations, *above all when reference situations produce straight lines.*

"Smoothness" seems to be an essentially geometrical concept for which we do not yet seem to have a reasonable analytical approximation. "Smooth" extrapolation and interpolation, especially with irregularly spaced data, continues to be easier and more persuasive when conducted and exhibited graphically rather than numerically.

One great virtue of good graphical representation is that it can serve to display clearly and effectively a message carried by quantities whose calculation or observation is far from simple. Many kinds of spectra, analog and numerical, illustrate this principle.

A scatter diagram with 100 or 500 points need not be more difficult to scan than one with 10 or 50. The same is often true with 1000 to 5000 points (possibly with 10,000 or 50,000). Tables of numbers simply cannot be expanded comparably without tremendous increases in difficulty of examination and understanding.

Similar problems and advantages occur in the even simpler case of an unstructured collection of single-number data. As the volume of data increases, it becomes very difficult to appreciate from a table even the most elementary properties of the collection, such as location, range or gaps. Yet simple graphical representations, as empirical cumulative distribution plots [18] or, perhaps, even as sensibly constructed histograms, provide rapid, easy and insightful indication of many properties, both sum-

mary and detailed, and do so as conveniently for large bodies of data as for smaller samples.

A common and extremely effective human response to a scatter plot of y versus x is often to draft in a smooth "freehand" or "eyeball" curve as an aid to judging the data. (Doing this is a natural step toward summarizing and exposing—toward the complementary processes of fitting and inspection of residuals.) Despite the apparent ease—and substantial agreement—with which humans can do this, there does not yet exist any automatic procedure that does it at all satisfactorily.

The issues and problems of graphical presentation in data analysis need and deserve attention from many different angles, ranging from profound psychological questions to narrow technological ones. These challenges will be deepened by the evolution of facilities for graphical real-time interactions.

ONE-VARIABLE GRAPHS

Graphical portrayal of frequency distributions by bar charts and histograms can be improved in various directions. For comparing with a fitted curve, in particular, the use of hanging (or suspended) rootograms, in which heights are proportional to the square root of frequency and blocks are attached to the fitted curve (not the base line), can be a considerable improvement.[19]

KINDS OF TWO-VARIABLE GRAPHS

Point plots, linked plots, and curve plots are only three of several distinct styles for two-variable graphs. Point clouds, scatter displays and progressive patterns are associated with useful distinction among purposes. Regularity of spacing, frequency of wild observations, absence of changes in variability, and correlations among fluctuations are also important considerations.

LINKING-UP POINTS AND RELATED ISSUES

Whether or not the points of a progressive pattern are linked together by line (or curve) segments can significantly influence the usefulness of such a graph. Linking up is not likely to help unless the points are reasonably close together (in x) and reasonably uniformly spaced (in x) and the corresponding "spectrum" is not too flat.

THE NEED FOR VARIOUS MENTAL APPROACHES

Given an appropriate plot, we still need an appropriate attitude or "set" for the brain to take toward its message. In this regard, the degree and character of correlations among the fluctuations of the various points are particularly important.

THREE-VARIABLE GRAPHS

Visual presentation of $z = f(x,y)$ is far from easy, yet badly needed. Of three classes of possibilities—contours, families of cross sections, and isometric views—the first seems, so far, most likely to be effective, though direct-interaction graphical consoles may offer other possibilities.

GENERAL CHARACTERISTICS OF DATA ANALYSIS

In productive data analysis:

1. *Those who seek are more likely to find.*

Tight frameworks of probable inference demand advance specifications of both a model and a list of questions to be asked, followed by data collection and then by analysis intended to answer only the prechosen questions. A man who lived strictly by this paradigm would have a hard time in learning anything new.

Some may be uncomfortable in not having tight global probability-like measures to calibrate their optimism and pessimism, yet in thinking about this difficulty it is vital to remember that science has not required independent confirmation without reason. To have clear evidence that something was not chance in one single circumstance is feeble proof that it happens in general. The price of losing a crisp evaluation of the results for a single circumstance is thus never great.

The price of not looking around, on the other hand, is the loss of opportunity to have the data suggest new things. What price would be greater?

In a strictly confirmatory experiment, there is a clear place for a relatively narrow and constricted analysis. But even there, there is likely to be a basic need and responsibility for accompanying such an analysis with a careful look around for new suggestions.

2. *Flexibility in viewpoint and in facilities is essential for good data analysis.*

Data analysis is very much a bootstrap exercise. Our facilities and our attitudes must encourage flexibility: use of alternate models, choice of subsets of the data, choice of subsets of auxiliary or associated variables, choice of forms of expression of these variables, and of the data, choice of alternate criteria, both in fitting and in evaluating.

3. *Both exploration and description are major objectives of data analysis; for both reasons data analysis is intrinsically iterative.*

Both the search for insight and for the unanticipated require that the available information be displayed. *Description as a preparation to display and insight is, in a certain sense, the main business of data analysis.* But, equally, adequate insight, however informal or intuitive, is a necessary precursor for incisive description of the anticipated. Accordingly, insightful exploration and description require an iterative, interactive, complementary process involving both summarization and exposure.

SIMPLICITY

Simplicity is in the mind, and it is valuable because it lets the mind work better. *Simplicity is often learned and comes in many forms.* To a man who understands only straight lines, a parabola may seem complex. As a conic section, however, it has a simplicity that has attracted men's minds since the days of the Greeks.

But, in data analysis, just as in experimentation, *attaining the simple is often a complex task.* Head-on collision between two high-velocity atoms or ions is simple in concept, yet a colliding-rings particle accelerator is a real complexity. Producing a simply-portrayed description of a body of data may require a similar complexity both in arithmetic approach and in display.

Progress in science is a curious mixture: The simple becomes complex as we learn about the ifs and buts; the complex becomes simple as new generations, using new concepts, learn to regard new things as simple.

Thus, in data analysis, *useful results, including useful techniques, need to be made simple.* This requires a broad spectrum of appropriate concepts.

THE LIMITATIONS OF DATA ANALYSIS

Data analysis cannot make knowledge grow out of nothing or out of mere numbers, nor can it salvage or sanctify poor work. *It can only bring to our attention a combination of the content of the data with the knowledge and insight about its background which we must supply. Accordingly, validity and objectivity in data analysis is a dangerous myth.*

Developing models of one sort to aid in appreciating or assessing the performance of models of another sort (perhaps describing methods of analysis) may indeed be useful to the discipline of data analysis. Such theories of inference must, however, be taken only as a guidance, and kept from becoming impediments. Assumptions and theory are indispensable, but, in use, the focus of data-analysis techniques must be on the data and the analysis, with the theory aiding insight by providing alternative backdrops.

It seems too easy for some to believe that detailed assumptions can make the data tell much more, either qualitatively or quantitatively, than would otherwise be the case. But if these assumptions are unwarranted their consequences may be misleading.

For example, combining an unexamined assumption of additivity in a two-way table with a classical test for main effects may indicate the absence of statistical significance, yet an elementary examination of residuals or interactions may reveal important information.

Both the guidance and the conduct of data analysis demand approximation. The combination of individually useful approximations often fails to be useful, either because errors accumulate or because ranges of adequacy fail to overlap. Thus, when *simple, individually useful models or data-analytic steps are linked together,* it is essential, as scientists have long realized, to think about the scientific problem as a whole and to make empirical tests of the worth of the combined chain.

GUIDANCE AND MODELS

Data analysis cannot be effectively conducted without guidance: implicit and vague or detailed and explicit. Contemplation of raw observations with an empty mind, even when it is possible, is often hardly more beneficial than not studying them at all.

In the sense in which we here use the word "model"—a means of guidance without implication of belief or reality—*all the structures that guide data analysis,* however weak and nonspecific, *are models* —even when they are not explicitly mathematical. Without them we are almost certainly lost (and surely completely primitive); were we to accept them unquestioningly we would be equally lost in a different morass; taking them as limited guidance, we may, however, succeed in finding some of what the data conceals. As Francis Bacon so well said, "Truth arises more easily from error than from confusion."

Definiteness in detailing objectives and assumptions in a formal model can simplify mathematical problems and increase the simplicity and impact of the results reached. But tightness of detail usually forces such a formal model unnecessarily far away from the realities of the data-gathering situation, obscuring possibly important phenomena. Looser structures can often do as well in simplicity and clarity of results while retaining robustness and breadth. *Both for guidance and the encouragement of exploration, it is most desirable that models be loose and noncommittal,* thus encouraging diverse alternative working hypotheses.

Even as simple a problem as comparing the location of two samples illustrates these points. When our model includes assumptions of approximate equality of variance, absence of seriously aberrant observations, and close normality (Gaussianity) of distribution, we are likely to calculate Student's t and halt. If we admit that any or all of these assumptions may be false, we will know that we need to do much better. (So far as the narrow objective of location comparison is concerned, it may suffice to use a more robust modification of Student's t.)

In most circumstances, we will gain by broadening our interests, and supplementing either t-value by at least inquiring what the sample has to say about inequality of variances, presence of aberrant observations, or nonnormality of distribution. Exhibiting the two samples on a single normal probability plot will surely open our eyes in these directions and will even, occasionally, direct our attention to less anticipated phenomena. The gain from doing this is likely to be great.

Trying to answer questions concerning the adequacy of a model by the use of data may be interesting and valuable. *In data analysis, however, models and techniques are to be thought of and developed as assisting tools with the focus on the data.* The models need not fit perfectly or even adequately to prove usefully insightful. We must never believe so deeply in any model as to constrain our insight.

Thus, for example, although the use of half-normal plotting [2] in analysis of 2^n experiments was suggested by a model combining equally distributed normal errors, simple factorial effects, and the "null" hypothesis that these latter effects vanish, such plots merely use these assumptions to provide a "backdrop" for exposure and remain both descriptive and instructive in most circumstances when the assumptions fail, even badly. Indeed, the plot itself may indicate or reveal the inappropriateness of the assumptions.

BRIEF COMMENTS ABOUT SOME CLASSICAL STATISTICAL PROCEDURES

Particularly in textbooks, statistical procedures are usually described as operational wholes (e.g., multiple linear regression), too often in terms of a fragmentary list of formal objectives (e.g., to estimate regression coefficients and/or test hypotheses about narrow aspects of the model). The main emphasis in statistical theory and in textbook presentations of methods has been on confirmation and summarization (e.g., the basing of multiple regression methods on linear hypothesis theory).

The development of techniques and concepts *useful for exposure* has had very little guidance from formal statistical theory. In actual practice, statistical methods embodied in such categories as experimental design, analysis of variance, multivariate analysis, time series analysis, goodness of fit, etc., are employed *often and productively* for purposes typically very distinct from those used in their textbook derivation and justification. (For example, multiple regression is typically useful as a generator of residuals and a producer of empirical analytical descriptions and summarizations.)

THE TECHNOLOGY OF DATA ANALYSIS

The basic purpose of a technology is to provide and organize tools and techniques to meet relatively well-specified, but often very broad, objectives. Well-organized technologies are usually associated with better-organized and more basic bodies of knowledge, conveniently referred to as the corre-

sponding sciences. Objectives, science, and technology can only evolve and develop together—and by means of active mutual interaction.

By the term data analysis we mean to encompass the techniques, attitudes, interests and concepts which are relevant to the process of learning from organized records of experience. This area has always been of fundamental importance. It is quite apparent that, currently and in the near future, widespread harnessing of the explosive potential of organized data analysis depends upon active development of its technology. The progress of that development suffers from the fragmentary understanding of both the science and the proper objectives of data analysis. Moreover, it seems to be true that both the mathematical developments of modern statistical theory and the glamor of computer and display hardware have, for different reasons and different persons, provided a diversion from the socially and scientifically important challenges of statistical data analysis. Still, the mushrooming opportunities of modern computing and display provide a major stimulus.

In thinking about data analysis technology, the antithesis between hardware and software (between machinery and organized know-how) is important, not only in the conventional uses of these terms within a computer system, but also in data analysis itself. Specific techniques, such as a three-way analysis of variance, considered as parts of data analysis, are really data-analytic hardware. The mystery and art of when and why to use such techniques make up the data-analytic software, which is today very soft indeed.

Our difficulties are twice compounded; we must develop data-analytic software to harness the power of our data-analytic hardware and, at the same time, develop computer software for data analysis that adequately harnesses the power of our computer and display hardware.

Flexibility in our objectives must be combined with easy and effective iteration and combination. Flexibility, easy iteration, and efficiency all demand the identification and use of functional components that can be assembled in many ways. Existing techniques need to be decomposed into appropriate components; new components need to be recognized and created. Among these functional components we shall find such diverse things as data structures, output formats, algorithms, logical operations, and even components for the construction of other components. Their selection and description needs improved guidance from an appropriate appreciation and structuring of objectives. By the same token, their definition and creation will stimulate the evolution of broad objectives and sophisticated techniques.

As far as numerically carrying out classical statistical procedures goes, little has been done, until very recently indeed,[20] to attempt to recognize the common arithmetic and logical operations which underlie a great many of the techniques of data analysis, and which may serve as the lowest-level functional components in a more organized approach to it. The recognition of these functional components, and the facility to combine them freely and flexibly, will greatly increase the power and scope of data-analytic methods. Side benefits would also accrue in husbanding programmer effort.

Two considerations are important in implementing data analysis: First, that the process of analysis usually involves a volume of output much greater than the original body of data. Second, that there is no clear barrier between output and input in the overall process of data analysis. The input for analysis is always the output from something else (whether from a previous analysis or a data source). The output from a step of analysis—as for instance an array of residuals or a covariance matrix—is likely to be the input to another phase of analysis. The resulting requirements upon the technology of computation for ease and compatibility are of major importance.

Current facilities for computing, display, and real-time interaction have developed substantially beyond our understanding of how to use them effectively in data analysis. Current limitations in data analysis technology are mainly in explicating and organizing the science of data analysis and in defining and implementing the necessary associated computer software.

From the statistical side of the discipline must come: broader, more permissive, empirically oriented concepts and theories; more inclusive and realistic classifications of objectives; more effective and coherent classifications of useful techniques; research toward more empirically informative techniques that will provide both exposure and summarization; more understanding and research on techniques of reforming and reexpressing variables; deeper insight into the psychology of graphs, pictures and output formats in general, both for inter-

action and for communication; progress toward standardized data structures of great flexibility and comprehensiveness.

From the computing side of the discipline is required software to provide: convenience with flexibility, simple and effective bookkeeping and history keeping, adequate editing, effective means for treating output as input, more flexible and general graphical presentations, and a variety of means to facilitate real-time interaction.

Though some progress is being made on many of these needs, the technology of data analysis is still in its infancy.

ACKNOWLEDGMENTS

We would like to thank G. A. Barnard, D. R. Cox, R. Gnanadesikan, C. L. Mallows, F. Mosteller, H. O. Pollak, and D. R. Wallace for their useful comments on related treatments of the topics considered in this paper.

REFERENCES

1. T. C. Chamberlain, "The Method of Multiple Working Hypotheses," reprint of 1890 version, *Science,* vol. 148, pp. 754–59 (1965).

2. Cuthbert Daniel, "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments," *Technometrics,* vol. 1, pp. 311–42 (1959).

3. F. J. Anscombe and J. W. Tukey, "The Examination and Analysis of Residuals," *Technometrics,* vol. 5, pp. 141–60 (1963).

4. G. E. P. Box and D. R. Cox, "An Analysis of Transformations," *Jour. Roy. Stat. Soc.,* vol. 26, pp. 211–43 (1964).

5. Jane F. Munk and M. B. Wilk, "Detecting Outliers in a Two-Way Table," unpublished manuscript (1966).

6. E. S. Pearson, "Some Aspects of the Geometry of Statistics: The Use of Visual Presentation in Understanding the Theory and Application of Mathematical Statistics," *Jour. Roy. Stat. Soc.* (A), vol. 119, pp. 125–49 (1956).

7. R. Duncan Luce and John W. Tukey, "Simultaneous Conjoint Measurement: A New Type of Fundamental Measurement," *Jour. Math. Psych.* vol. 1, pp. 1–27 (1964).

8. Peter G. Moore and John W. Tukey, "Answer to Query 112," *Biometrika,* vol. 10, pp. 562–68 (1954).

9. J. B. Kruskal, "Analysis of Factorial Experiments by Estimating Monotone Transformations of the Data," *Jour. Roy. Stat. Soc.* (B), vol. 27, pp. 251–63 (1965).

10. R. N. Shepard, "The Analysis of Proximities: Multidimensional Scaling with an Unknown Distance Function, I," *Psychometrika,* vol. 29, pp. 125–40 (1962).

11. ——, "Analysis of Proximities," pt. II, ibid, pp. 219–46.

12. ——, "Analysis of Proximities as a Technique for the Study of Information Processing in Man," *Human Factors,* vol. 5, pp. 33–48 (1963).

13. J. B. Kruskal, "Multidimensional Scaling by Optimizing Goodness of Fit to a Non-Metric Hypothesis," *Psychometrika,* vol. 29, pp. 1–27 (1964).

14. ——, "Non-Metric Multidimensional Scaling: A Numerical Method," ibid, pp. 115–29.

15. M. A. Efroymson, "Multiple Regression Analysis," in *Mathematical Models for Digital Computers* (Anthony Ralston and Herbert S. Wilf, eds.), Wiley and Sons, New York, 1960, pp. 191–203.

16. Robert G. Miller, "Statistical Prediction by Discriminant Analysis," *Meteorological Monographs* (Boston, American Meteorological Society), vol. 4, no. 25 (1962).

17. Norman J. MacDonald and Fred Ward, "The Prediction of Geomagnetic Disturbance Indices: 1. The Elimination of Internally Predictable Variations," *J. Geophys. Res.,* vol. 68, pp. 3351–73 (1963).

18. M. B. Wilk and R. Gnanadesikan, "Probability Plotting Methods for the Analysis of Data," unpublished manuscript (1966).

19. John W. Tukey, "The Future of Processes of Data Analysis," *Proceedings of the 10th Conference on the Design of Experiments in Army Research, Development and Testing,* U. S. Army Research Office (Durham), 1965, pp. 691–729.

20. Albert G. Beaton, "The Use of Special Matrix Operations in Statistical Calculus," Ed.D. thesis, Grad. School of Education, Harvard University 1964.