ANALYZING DATA:

SANCTIFICATION OR DETECTIVE WORK? [1]

JOHN W. TUKEY [2]

Princeton University and Bell Telephone Laboratories

THERE are many reasons why it is an honor and a pleasure to be here tonight. I first had to deal with masses of data about people's behavior some 27 years ago—other masses have come along from time to time. Psychology, like all science, rests upon good data analysis as one of its foundations. Psychologists are conscious of their data—sometimes, perhaps, too much so. For all these reasons I am glad to try to show you some of the broad aspects of data analysis as I presently see it.

First, three remarks:

1. Every field has data, and a need to analyze them.

2. The problems of different fields are much more alike than their practitioners think, much more alike than different.

3. Data analysis in psychology has a flavor of its own—but one much more due to psychologists than to their science.

Exploratory and Confirmatory Data Analysis

Both exploratory and confirmatory data analysis deserve our attention. Both detection and adjudication play crucial roles—in the progress of science as in the control of crime.

To concentrate on confirmation, to the exclusion or submergence of exploration, is an obvious mistake. Where does new knowledge come from? How can an undetected criminal be put on trial?

Exploration relies greatly on looking around. Indeed, unless practical psychologists produce new ways to receive the message from data, there will continue to be no substitute for visual techniques in exploring data. It may well be true that "plot and eye" is the most diverse channel to the human mind. Not that it transmits more bits per second, but rather that it will transmit a greater variety of messages on unexpected topics easily and rapidly. There really seems to be no substitute for "looking at the data."

When we calculate, rather than looking, we must seek the same flexibility.

We ought to try to calculate what will help us most to understand our data, and their indications. We ought not to be bound by preconceived notions —or preconceived analyses.

In the 1940s, Vannevar Bush and his associates were wise enough to know that (*a*) they should not build differential analyzers without planning what they could be used for; (*b*) they should not expect differential analyzers to be used for what they had planned.

Preplanning data analysis is a little different. One minute of a differential analyzer could be used for only one thing. One body of data can—and usually should—be analyzed in more than one way. Accordingly, it *is* likely to be worthwhile to make the preplanned analysis of your data, but only if you refrain from limiting yourself to the analyses that had been planned in advance.

There is no substitute for examining indications. We want to know what the data *seem* to say, whether or not data mean what they seem—a fortiori whether or not we can *prove* that they mean it.

Flexibility Can Be Dealt With

Bending the data to fit the analysis can be vital—as a saving of capital investment in ways of analysis, as a route that we can take today rather than at some unknown time to come. But bending the question to fit the analysis is to be shunned at all costs.

[1] Prepared in part in connection with research at Princeton University, sponsored by the Army Research Office (Durham).

Invited address to the American Psychological Association, San Francisco, September 1968.

[2] The author would like to thank Robert Abelson, Lee Cronbach, and Lyle Jones for help in clarifying the ideas and wording in the article. Reprints of the article may be requested from John Tukey, Box 824, Murray Hill, New Jersey 07974.

If $\log y$, $\sqrt{y}$, or $(y + 0.7)^{0.37}$ show simpler behavior than y itself, we should hasten to work in the terms with simpler behavior, using as many of our conventional methods of analysis as seem helpful.

If our real question cannot be answered by a correlation coefficient, it can be fatal to insist on using a correlation coefficient to *not answer* our question, whether or not some other question appears to be answered.

Once, not long ago, flexibility in what we compute could have been attacked, perhaps viciously, on an irrelevant point—how are we to do confirmation? Today, the jackknife—a technique about which I shall say more in a moment—offers what is usually an adequately precise approach to confirmation wherever there is appropriate internal evidence about stability. If there is a basis for judging the stability of even the simplest things we might have computed, we can use that same basis to judge the stability of even very complex things—of almost anything we choose to compute.

Confirmation comes from repetition. Any attempt to avoid this statement leads at least to failure and more probably to destruction.

The 131 students in Psychology 101 who were volunteered for an experiment are 131 different repetitions of some things—and one lone repetition of others. No way of twisting the answers can change this. No internal evidence can tell us about how much other repetitions—of those things we actually repeated only once—might differ. External evidence there may be, but rarely will it have been assembled for careful scrutiny by others.

One span of the bridge that leads from data to final judgment is always concerned with how much the things that we only repeated once may differ. Unless the external evidence about possible differences has been carefully pulled together, little but intuition supports this one span.

Whenever we have identifiable repetition, and a specific scheme of calculating a certain indication from data, we can try out the jackknife. It consists of two essentials:

(*a*) a simple scheme to produce pseudovalues corresponding to individual instances or subgroups of what was repeated;

(*b*) use of Student's t for significance or confidence.

For the simplest calculations, such as means, the pseudovalues will coincide with individual values or subgroup means. For less simple calculations, beginning with those regressions where the x's are not fixed in advance, this will not be so. (A souvenir sheet on the jackknife follows as an appendix; a few references have been added.)

Like *all* methods of data analysis, the jackknife is approximate and of only partially known quality. In my personal judgment there are few cases where it is not worth using. Equally, of course, there are few cases where we can be precise about its performance, just as there are few cases, if any, where confirmatory p values obtained by any method you may have chosen deserve complete trust. (For all this, p values still serve us well!)

Flexibility Has To Be Balanced

If in advocating flexibility as necessary—and the jackknife as good—I am thought to be leaning one way, then in advocating explicit and careful attention to problems of multiplicity I am doubtless thought to be leaning in the other.

All of us are probably ready to laugh about the bioassayer of drugs who was telling his friend about how essential new statistical techniques had proved to be in his laboratory. "Why," he cried, "5% of the assays show significant curvature, and 5% show lack of parallelism." We can assign to our elementary statistics classes the problem of what significance level was he using. Many will see the answer. Clearly, he was testing at 5%, and being told, though he did not listen, that his assaying procedure was in fine shape, with no sign at all of curvature or lack of parallelism.

But what of the man who tests 250 predictors against his lone criterion and finds 15 significant at 5%? Dare we do less than make it explicit that the average number expected by chance is 12.5? What attention ought we then give the 15? Is 15 a clear sign that he is finding no meaningful indications? Would finding only 3 be a sign of trouble?

One of my good and respected friends in psychology sent me some of his technical reports. I hope he will not mind if I use some quotations as a "news peg" on which to hang an illustration of how new ways to use confirmatory data analysis might be needed. In discussing the reanalysis of some previously published data he said:

> "My analysis and presentation have been guided by preferences (1) . . . , (2) . . . , (3) for interpreting no relation unless $p < .05$."

"My main analysis tests about 100 effects, and 33 reach the .05 level where the chance expectancy is 5."

By sticking to $p = .05$ he found 33 effects of which some 5 might have come by chance, as opposed to an original analysis, made by another psychologist and involving a larger number of tests, where 27 effects were found of which some 10 might have come by chance. Clearly "about 28 out of 33 meaningful" is better than "about 17 out of 27 meaningful" especially when, as is almost always the case, we do not know which 28 or which 17 are the meaningful ones.

Need we—should we—stick to $p = .05$ if what we seek is a relatively pure list of appearances? No matter where our cutoff comes, we will not be sure of all appearances. Might it not be better to adjust the critical p moderately—say to .03 or .07—whenever such a less standard value seems to offer a greater fraction of presumably real appearances among those significant at the critical p?

We would then use different modifications for different sets of data. No one, to my knowledge, has set himself the twin problems of how to do this and how well doing this in a specific way performs. But someone may well decide to deal with these problems, and his results may be such that we will want to use this novel sort of a confirmatory technique as a guide to our exposition of the results of complex experiments. Flexibility of technique can be helpful; it must, however, be guided by flexible careful thought and supported by empirical trial.

Nothing Learned Is Certain

We learn by taking chances. Every modern learning theorist expects learning to be by trial, with some errors. This is as true for science as for the individual.

Every statistician has the obligation to admit that each particular set of data *could* have come about quite randomly—given the oddest states of nature. He may be able to reassure you by quoting a very small probability. As for the data as a whole, so too for each and every aspect on which it is to throw light. No conclusion or inference becomes knowledge without risk of error.

Only by taking a chance can any of us learn by experience, either as a child or as a scientist. Confirmatory data analysis, done realistically—in particular when problems of multiplicity are faced honestly and rather realistically—can give useful indications of what kinds of chances are being taken.

These indications are themselves based on experience and must therefore be fallible. The rats that were forced to decide between circles and nearly circular ellipses showed behavior changes that were clearly statistically significant; but the psychological significance of these changes has turned out *not* to be what the profession thought for many years was clearly so.

One can give a helpful and illuminating analysis of the history of statistics as learning to be certain about less and less. I find both this fact and the details of this analysis helpful and illuminating, but I will not go into details here.

The modern test of significance, before which so many editors of psychological journals are reported to bow down, owes more to R. A. Fisher than to any other man. Yet Sir Ronald's standard of firm knowledge was not one very extremely significant result, but rather the ability to repeatedly get results significant at 5%.

Repetition is the basis for judging variability and significance and confidence. Repetition of results, each significant, is the basis, according to Fisher, of scientific truth.

Certainty is an illusion. We have only to look at physics over the last 100 years to see that this is true for the sciences which have earned the greatest regard. The fact that "all the laws of physics are wrong, though most are extremely good approximations" does not make physics less valuable, either intellectually or practically.

As an illusion, certainty can be wasteful, as well as misleading.

When we classify cases, the philosophers tell us, we should do this unequivocally and without error. Similarly, our grade school teachers tell us to discard digits that are not precisely determined by our arithmetic. Doing either of these things is usually *statistically* wasteful.

If we add up 10,000 numbers, each given to two decimal places, the rounding errors could add up to 50, though the accumulated variance is only about 0.08. The tens digit could be wrong, but throwing it away would be very wasteful, increasing the variance to about 800. Throwing away the units digit would only increase the variance by a factor of 100, while throwing away the first decimal

would only double the variance due to rounding. Statistical efficiency demands keeping both uncertain digits and uncertain classifications.

We can be more specific when the classes are ordered and of comparable size. For here classification is like arithmetic, indeed the same formulas apply. Classifying 10 times as finely is just like keeping another decimal place. The man who must always be "efficient" at all costs—say at least 98% efficient—must then classify so finely that as many reclassifications will depart at least four classes from the result of a prior classification as will agree with it exactly. The man who wishes to avoid serious waste will have to classify nearly as finely as this.

Two caveats go with this statement. First, it refers to reclassification truly independent of classification, something that is harder to obtain than one would think. In practice, seemingly independent reclassification is dependent enough to have much more frequent agreement than would truly independent reclassification. Second, it assumes that using a fine classification does not disturb the classifier, does not lower his performance. Either of these considerations can be relevant enough in a specific situation to make it plausible for classification and reclassification to be seen to agree exactly in, say, 20–50% of the cases. This is still very far from 100% agreement. Certainty of classification is a will-o'-the-wisp hovering over a deep quicksand.

While classification is done by almost every one of us, I have discussed it as much as a symbol as for its own sake. The search for certainty can only lead us astray.

Amount, as well as Direction, Is Vital

Physical scientists are supposed to be a reference group for life and behavioral scientists, including psychologists. Empirically this seems *not* to be really so. The physical sciences have learned much by storing up amounts, not just directions. If, for example, elasticity had been confined to "When you pull on it, it gets longer!" Hooke's law, the elastic limit, plasticity, and many other important topics could not have appeared.

The qualitative properties of *things* have proved much less important than the quantitative ones. Why should this not hold true for people? I believe that just this will prove to be so, but not without much effort. Even if the task is hard, is it not past time to begin, especially in selected, more or less well-understood, subfields?

Bear in mind a simple fact: the great majority of the useful facts that physics has learned—and recorded in numbers—are specific and detailed, not global and general.

Substantial prizes have been well earned by physicists who proved that certain physical numbers were quite closely the same. Are we prepared to look for quantitative constancies in psychology? To reward those that find them? To answer "yes" to such questions, rather than to assume such constancies are natural laws not requiring checking, demands an emphasis on breadth of data, on repetition under different parallel circumstances, of which we have seen far too little.

To move in this direction requires more than a sudden rush into large-scale studies. It requires us to emphasize learning the right things to measure, learning empirically and by small painful steps. Physics made little progress in one of its major fields until Rumford, the Count from New England, separated the notions of heat and temperature.

To move in such a direction means to take an effective attitude to simplicity and complexity, an attitude which is two-sided in an initially uncomfortable way. It means seeking for irremovable complexities, rather than trival ones and choosing the numerical expression of our variables to make things as simple as possible, while grasping greedily at the unsimplicities that remain.

The two-sidedness should not seem strange to us after careful thought. Perhaps the aspect of classical significance test most central to scientific progress is its similarly two-sided nature—its demonstration of difference by trying to make things the same.

Occam's razor indicates that we should not think things are different if they could be the same. Equally, Occam's razor indicates that we should not think behavior to be complicated if we can make it simple by reexpressing, or reformulating, our variables. Taking this sort of attitude subjects us to cross-pressures and some discomfort. It is by standing up to such pressure that we grow with our science.

My friend Richard Link puts it more vigorously for statisticians, who have to bear even greater cross-pressures than do gatherers and analysts of

data, when he says that they have to be "schizophrenic" to combine the certainty of mathematics with the uncertainties it is one of their prime responsibilities to assess.

Campbellian Measurement Has Scared Us Far Too Long

It is in terms of the variables that most nearly behave simply that physics has made its giant strides. To seek out these variables among all alternatives, while maintaining an even keener interest in any unsimplicities that remain, requires a sort of self-restraint and a kind of quantitative honesty that have been too little practiced. If our basic orientation is to be in such a direction, our data analysis must be forced to help us. Accordingly, our data analysis must stress reexpressing variables for simplicity of behavior.

Norman Robert Campbell was a clear thinker, a man with a deep understanding of classical physics—of classical physics as a static well-developed system, not classical physics as a growing insight. He was able to take over ideas of measurement developed by mathematicians like Hölder and fit them into the physics of his day. His writings were clear and cogent. Yet if physics had had to grow by Campbellian measurement, it would never have reached the state in which Campbell found it. While giving Campbell the respect which his insightful wisdom deserves, it is time to strike off the chains with which response to his writings have bound us, whether through acceptance, through reaction, or through optimism.

Measuring the right things on a communicable scale lets us stockpile information about amounts. Such information can be useful, whether or not the chosen scale is an interval scale. Before the second law of thermodynamics—and there were many decades of progress in physics and chemistry before it appeared—the scale of temperature was not, in any nontrivial sense, an interval scale. Yet these decades of progress would have been impossible had physicists and chemists refused either to record temperatures or to calculate with them.

I hope it is clear that if I had an 1818 physicist as a client I would not let the then scale type of temperature stop me from recommending the calculation and t testing of means of temperatures. With a 1968 psychologist I would take a similar view about many numbers not on interval scales. (In each case I might prefer to use a more resistant summary than the mean, usually one that shifts not as far from the mean as the median, but this is a separate issue.)

Nonparametric statistics for nonparametrism's sake, as we should have expected, has proved to be a side branch on the evolutionary tree. Some "nonparametric" procedures, mainly those that can be formulated in terms of sums and differences, like the two-sample Wilcoxon procedure, have proved to be very good. Since others have not, I credit this to Frank Wilcoxon rather than to nonparametrism. Indeed, we are still learning that order is most useful when combined with a numerical scale, particularly so when we combine the ordered values (on the numerical scale) using addition or subtraction or forming more general linear combinations.

Methods called "nonparametric" that do not make explicit use of a scale are occasionally useful, sometimes because they can show whether or not criticism is carping when it objects to particular analyses of particular sets of data, sometimes because they are handy and portable and contribute to our first-aid kit. For all their uses, such "order only" methods ought not to be taken as either standard or exhaustive.

Data analysis is never going to be basically nonparametric. The costs of nonparametrism usually substantially exceed the advantages. Eliminating one contributory cause of uncertainty is of little value, especially when its contribution is infrequently appreciable.

Today we are hard at work on the next branch up the evolutionary tree, developing methods that keep the p values actually close to the nominal p values throughout a suitably wide variety of underlying distributions while maintaining high efficiency and stringency throughout a similar diversity of circumstances. These newer methods tend to combine ordering and addition, as when we form the mean of the middle half of the sample values. You will hear much more about them in the decade ahead.

Spirit of Conjoint Measurement

The last few years have seen an efflorescence of work on simultaneous conjoint measurement (Luce & Tukey, 1964; Tversky, 1967, and references therein) on a non-Campbellian kind of fundamental measurement in which we measure a response by the way in which this response reacts

to combinations of two or more conditions or stimuli. (In essence, a change from Version A to Version B of one stimulus or condition is defined to produce an equal change in response whatever version of the other stimulus or condition is held fixed. If suitable qualitative axioms hold, responses are then measured in a self-consistent way on a unique interval scale.) So far as I have seen, this work has all been theoretical in the less satisfactory sense of that word—has explored the axiomatization or other mathematical aspects of such "measurement." For this emphasis I am sorry.[3] The major need is still to take conjoint measurement out into the world of observation, experiment, and data, and let it begin to teach us how we ought to express many of the quantities with which we work. Its use will then be a major part of the broad program of learning about the world by seeking out as many empirical constancies as possible.

It is not always easy to recognize what the data are saying to us. Over the years, S. S. Stevens has built up a surprisingly large store of numerical information:

(*a*) first about how verbal report of amount is associated with each of many physically measurable quantities: sizes, numerousnesses, strengths of shock, to name but a few;

(*b*) later about how people match comparisons on one of these modalities to comparisons on another.

Only recently is Stevens beginning to point out how this is all so simply expressible by taking logarithms. I forecast rarely, but here I am willing to forecast. I believe that Stevens will, in the decades ahead, be most highly valued for showing that logarithm of verbal report and logarithm of various physically measurable stimuli are the scales that make human behavior in this area simple.

While Stevens' work is not an example of conjoint measurement, its conclusion is in the same spirit, and can be taken as a sign and a portent.

The insights that lead to the great simplicities do not occur in PhD theses, except for an Einstein with Einstein's opportunities, a genius who can build on many earlier stages of simplification. They come after much careful, insightful work that has established facts about amounts.

[3] The author has been told, since the address, that the real situation is better than these words would suggest.

Clarity in the large comes from clarity in the medium scale; clarity in the medium scale comes from clarity in the small. Clarity always comes with difficulty. To seek out clarity in the small by greater flexibility in our data analysis is not to seize a panacea that will lead us to great things at once. If, as we can be sure that it will, it takes away a few stones from our long and winding path, however, it will have served us well. The major messages of conjoint measurement are simple, and like all simple messages not easy to respond to:

(*a*) We need to work with simultaneous changes.

(*b*) We need to shape the expressions of our variables to make the effects of these simultaneous changes simple.

Implications for Research Structure

If conjoint measurement makes our work more difficult, perhaps by requiring essentially more diverse experimental or observational groups, perhaps by requiring tools of measurement of wider application, and almost certainly by calling for larger samples, *let us face the facts*. I am not a practicing psychologist; I do not claim to understand which questions ought to appear of most importance to psychology today. But if we are to study conjoint effects, we are certainly likely to need greater recognized diversity of subjects. We may be able to gain this by sorting out diverse subgroups of the people we now work with, but we are likely to need to sample more widely in terms of geography, or educational attainment, or cultural attitudes. Painful though it may be to gain it, meaningful variation of the subject is going to need to be one of our conjoint circumstances in many cases. With increased diversity, unless we learn to improve our measuring tools rather suddenly, there will have to be increases in sample size. Both effort and expense will have to rise with this.

In every field of science we use the PhD thesis in many ways, including—sometimes unfortunately—as a prototype of how research is to be done. Other sciences have faced the transition from the PhD thesis that stood on its own feet to the PhD thesis that is part of a bigger entity. No PhD builds his own cyclotron as part of his thesis. No PhD orbits his own satellite to get his data.

If, as I would guess, some areas of psychology have reached the point where they need larger,

more diverse bodies of data measured in common ways, there may well be no way out other than cooperation, cooperation both among thesis supervisors and among thesis workers, cooperation over large distances, whether in geography or in a subject's background. Though sometimes necessary, cooperation is never easy; we cannot envy those who must strive to make it work in such a situation as I have suggested.

If subjects must be students in "Psychology 1," they often could be students in various different Psychology 1's. Who believes that the student bodies of all introductory psychology courses are the same? Who knows how much they differ? Or what external features seem to be associated with their differences?

While we are considering the structure of research, we will do well to consider the structure of the PhD. I have never been a psychology PhD candidate, nor worked day by day alongside one. From a distance, however, experimental theses in all fields seem to always be just barely completed in time with the data gathering just getting under the wire. When this is so, the data analysis is under the greatest time pressure of all the steps, and presumably gets the least thought. Where this is true, how can we have failed to educate psychologists that data analysis is:

(*a*) something done hastily.

(*b*) something done to show that one's work is acceptable, rather than something carefully done.

(*c*) surely not an exploratory process turning up clues for future work.

There is merit in having a PhD thesis encompass all the admitted steps of the research process. Once we recognize that research is a continuing, more or less cyclic, process, however, we see that we can segment it in many places. Why should not at least a fair proportion of theses start with a reasonably careful analysis of previously collected and presumably already lightly analyzed data, a process usefully spread out over considerable time. Instant data analysis is—and will remain—an illusion. From this analysis would come the appearances, clues, and questions that ought to play a part in the next effort of data collection. Students who gained their PhD in this way would take data analysis seriously and, I am convinced, would contribute more throughout their research life. Like the present pattern, their theses would cover a full cycle of the research process, which they would regard much more clearly as an ongoing process, rather than a set of disconnected papers.

Correlation Coefficients Are a Symptom

Like the late Charles P. Winsor, a statistician far ahead of his time, I find the use of a correlation coefficient a dangerous symptom. It is an enemy of generalization, a focuser on the "here and now" to the exclusion of the "there and then." Any influence that exerts selection on one variable and not the other will shift the correlation coefficient. What usually remains constant under such circumstances is one of the regression coefficients. If we wish to seek for constancies, then, regression coefficients are much more likely to serve us than correlation coefficients.

Why then are correlation coefficients so attractive? Only bad reasons seem to come to mind. Worst of all, probably, is the absence of any need to think about units for either variable. Given two perfectly meaningless variables, one is reminded of their meaninglessness when a regression coefficient is given, since one wonders how to interpret its value. A correlation coefficient is less likely to bring up the unpleasant truth—we *think* we know what $r = -.7$ means. *Do we?* How often? Sweeping things under the rug is the enemy of good data analysis. Often, using the correlation coefficient is "sweeping under the rug" with a vengeance. Being so disinterested in our variables that we do not care about their units can hardly be desirable.

In statistical theory, the correlation coefficient seems to be mildly convenient in discussing the behavior of estimates and other statistics. This is very different from calculating a correlation coefficient from data.

When we are calculating from data, regression formulas, covariances, and even variance components answer meaningful questions better.

What I Should Have Done Tonight

In preparation for coming here to talk to you, I talked to a few of my very good friends among your profession and asked them what I should say. One of them—a man I have always respected, and now respect more—wrote me a letter that I am going to share with you. This letter gives his view of what I ought to have done by this point in the evening.

I share in his feelings as to where we should come out. I leave it to you to judge whether he is right as to where we started the evening. If there be truth in his starting point, anything I have done to take us a step or two along the way is very much for the good. Clearly, however, no one could dream of producing—in one evening—any substantial part of the change that he feels is needed.

Let me then read, preserving his anonymity so that you must judge the words for themselves, without any specific psychological authority:

> I have the feeling that Psychology currently is without a dominant viewpoint concerning a model for data analysis. In the forties and early fifties, a hypothetico-deductive framework was popular, and our mentors were keen on urging the design of "crucial" experiments for the refutation of specific predictions made from one or another theory. Inductive empiricism was said to be disorderly and inefficient. You and I knew then, as we know now, that no one approach is uniformly most powerful. We adapt easily to the uncertainties inherent in a world where the choice of methods for experimental design and analysis are relative to the purposes of an investigation. Such tolerance for ambiguity, however, doesn't come easily for many of our friends, who strive to be told the "rules." (Unfortunately, most textbooks are obliging; investigators in psychology are well-known for citing texts in psychological statistics as authority for the methods they have selected.) I attribute the stultifying uniformity of statistical usage less to journal editors than to the bulk of investigators (as I attribute counting of publications as criteria for academic advancement less to deans than to faculty).
>
> Following a rule book for research seems to stimulate the attack on trivial problems. The great challenge is to teach investigators to formulate questions that have a chance of leading somewhere, not to be too tightly bound in the *formulation* by a preconceived model of research design. Only after the formulation (but before empirical study) need there be attention to the procedures to be adopted for collecting and evaluating evidence—not right away changing the questions to fit a standard procedure, but (hopefully) selecting and/or adapting procedures which are suitable.
>
> Your task, then, is simple. Provide security to psychological researchers by presenting a rule structure that contains few fixed rules. Divert them from depending upon the "authority" of standard textbook solutions, but without being able to substitute a second religion for the first. Stimulate intelligent problem formulation, without being able to say quite how this is done. Demand high standards of statistical reasoning, but without specifying a single model of statistics which might serve as a criterion for quality of reasoning. And achieve all of these things on members of an audience, each of whom already is convinced that his limited conception of scientific method is the only right one.
>
> Sympathetically yours,

I must warn you of one thing: the view you ought to take of data analysis and the way you ought to practice it does not depend on what you think of this letter. I know of no thoughtful account of scientific research as a dynamic process—I assume psychology is not yet seeking stasis—according to which data analysis should fail to be flexible, or should fail to be a handmaiden rather than a high priestess.

Detective Work versus Sanctification

Let us return, then, to the original question: What ought to be the nature of data analysis?

Data analysis needs to be both exploratory and confirmatory. In exploratory data analysis there can be no substitute for flexibility, for adapting what is calculated—and, we hope, plotted—both to the needs of the situation and the clues that the data have already provided. In this mode, data analysis is detective work—almost an ideal example of seeking what might be relevant.

Confirmatory data analysis has its place, too. Well used, its importance may even equal that of exploratory data analysis. We dare not, however, let it be an imprimatur or a testimony of infallibility. "Not a high priestess but a handmaiden" must be our demand. Confirmatory data analysis must be the means by which we adjust optimism and pessimism, not only ours but those of our readers. To do this is not easy and may require new approaches and unfamiliar ways of thinking.

The Roman Catholic Church is a long-lived and careful institution. It has long held that sanctification was only for the dead—indeed only for those already dead for an appropriate period. I believe, and I urge you to feel, that sanctification of data is equally only for dead data—data that are only of historical importance, like Newton's apple. If we could all live by this precept, we might have to think more—painful though that might be—but we would, by the same token, accomplish more.

Data analysis has its major uses. They are detective work and guidance counseling. Let us all try to act accordingly.

REFERENCES

LUCE, R. D., & TUKEY, J. W. Simultaneous conjoint measurement: A new type of fundamental measurement. *Journal of Mathematical Psychology,* 1964, 1, 1–27.

TVERSKY, A. A general theory of polynomial conjoint measurement. *Journal of Mathematical Psychology,* 1967, 4, 1–20.

APPENDIX

SOUVENIR SHEET ON THE JACKKNIFE

Given:

1. Data divided into r pieces which we are willing to treat as equally important.

2. Agreement that the differences among these pieces reflect the sources of variation for which we expect to allow.

3. Some chosen way to calculate a number from data that apply whenever we have enough data of the proper kind.

Sought:

1. Significance and confidence statements about the result of the calculation, including some allowance for bias depending on amount of data.

Basics:

1. Let y_{all} be the number calculated from all the data.

2. Let $y_{(i)}$, read "y skip i," be the number calculated from all $r-1$ pieces of data except the ith.

3. Let y_{*i}, the ith pseudovalue, be given by $y_{*i} = r \cdot y_{\text{all}} - (r-1) \cdot y_{(i)}$.

Procedure:

1. Treat the pseudovalues as if they were a sample and apply Student's t.

Cautions:

1. Need to avoid cases where one piece is likely to dominate all others, whether present or absent. (When something is sold by Macy's, Sears Roebuck, and small country hardware stores we must stratify.)

2. Need to avoid too small pieces when this unduly constrains the variation of the pseudovalues. (When y is the median of a sample, using single observations as pieces is to be avoided.)

REFERENCES TO APPENDIX

AVERSEN, J. Jackknifing variances. Technical Report No. 18. United Public Health Service Grant 2T1 GM25-11. Stanford University, 1968.

BRILLINGER, D. R. The asymptotic behaviour of Tukey's general method of setting approximate confidence limits (The Jackknife) when applied to maximum likelihood estimates. *Review of the International Statistical Institute,* 1964, **32**, 202–206.

MILLER, R. D., JR. A trustworthy jackknife. *Annals of Mathematical Statistics,* 1964, **35**, 1594–1605.

MILLER, R. G. Jackknifing variances. *Annals of Mathematical Statistics,* 1968, **39**, 567–582.

MOSTELLER, F., & TUKEY, J. W. Data analysis, including statistics. In G. Lindzey & E. Aronson (Eds.), *Handbook of social psychology.* (2nd ed.) Reading, Mass.: Addison-Wesley, 1968.

QUENOUILLE, M. H. Notes on bias in estimation. *Biometrika,* 1956, **32**, 353–360.

TUKEY, J. W. Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics,* 1958, **29**, 614.

TVERSKY, A. Additivity, utility, and subjective probability. *Journal of Mathematical Psychology,* 1967, **4**, 175–201.