JOURNAL OF APPLIED BEHAVIOR ANALYSIS 1974, 7, 647-653 NUMBER 4 (WINTER 1974)

STATISTICAL INFERENCE FOR INDIVIDUAL ORGANISM RESEARCH: MIXED BLESSING OR CURSE?[1]

JACK MICHAEL

WESTERN MICHIGAN UNIVERSITY

Descriptive and inferential statistics are described as judgemental aids, stimuli to which the scientist can more easily react than to his raw experimental results. The increasing emphasis on the significance test as the main judgemental aid utilized in experimental psychology is credited with several harmful effects on experimental practice. The area known as "the experimental analysis of behavior" has so far escaped most of these harmful effects, but now we see an increased interest in the development of appropriate significance tests for individual organism research. This interest is based on the view that it is not possible to effect adequate levels of experimental control with much human applied research, and that in such cases a significance test would be quite valuable as a judgemental aid, both of which points are considered to be essentially incorrect, and if accepted, potentially harmful.

Descriptive and Inferential Statistics As Judgemental Aids

The observations resulting from scientific experiments are stimuli that hopefully affect the scientist and his colleagues by producing better practical behavior, more sophisticated follow-up experiments, or better verbal behavior regarding the subject matter. These stimuli, however, may not result in any effective reaction, a fairly common reason being their complexity. Repeated observation of the same experimental condition, for example, may give rise to a set of numbers, all differing considerably from one another. This situation has occurred quite often and methods have been discovered for simplifying it to some degree. Some of the methods generate two-dimensional visual stimuli where the values of each dimension stand in a point-to-point relation to some feature of the data; a frequency polygon is such a stimulus. Another stimulus-simplifying technique results in a smaller set of numbers, each related to some important characteristic of the larger set, such as the mean and range of the raw data.

Using the term "judgement" to refer to any of the various kinds of reactions that a scientist could make to the data of his experiment, it is useful to refer to these stimulus-simplifying techniques and their products as "judgemental aids". In this sense, then, the graphing devices and the measures of central tendency, variability, *etc.* of the field of descriptive statistics are all judgemental aids. In some way they produce a stimulus to which the experimenter can react more easily than to his raw data.

The various judgemental aids do not achieve their simplifying effects without some cost, however. In the first place, they are easier to react to in part because they are abbreviations. Some stimulus aspects of the raw data are simply absent from the aid, and if one's entire reaction is based on the abbreviation, the missing feature cannot affect behavior at all. Further, the scientist must spend some time learning about them, time he might be spending in other activities relevant to his subject matter. Statistics courses displace other topics from the curriculum.

A more complex type of cost consists of the time and effort that must be expended determining the extent to which some particular aid is

[1]This is one in a series of articles available for $1.50 from the Business Manager, *Journal of Applied Behavior Analysis,* Department of Human Development, University of Kansas, Lawrence, Kansas 66045. Ask for Monograph #4.

appropriate to the circumstances and data of a particular experiment. Finally, just as he must accumulate experience with his subject matter by reacting to it in various ways and being affected by the relatively long-term consequences of his reactions, he must now accumulate experience in reacting to the judgemental aid and feel the long-term effects of this behavior.

With such devices as frequency polygons, means, percentages, there seems to be a relatively clear net gain. The time required to learn how to use such techniques and the time spent in determining which one to use in a particular situation is relatively small compared with the simplifying effect achieved. Furthermore, the circumstances where they apply occur often enough that the individual scientist has some chance of acquiring the necessary experience regarding the long-range effects of his reliance on such judgemental aids.

Inferential statistics are also no more nor less than techniques for simplifying a complex stimulus situation. When an experiment results in two sets of numbers, one from a control and another from an experimental condition the comparison may be quite difficult to make. It is usually easier to compare frequency polygons and means, and one's reaction to this state of affairs may be further aided in some way by performing what is called a statistical significance test or computing a confidence interval. The former is the most common inference procedure used in experimental psychology and results in a statement that the probability of such a difference (or a larger one) arising by chance when the population means are actually equal is less than or equal to some specified value.

Significance tests and confidence intervals are more expensive judgemental aids than descriptive procedures. The abbreviation is more extreme, and the time required to learn how to obtain and interpret them is much greater. Determining to what degree the judgemental aid is appropriate to the particular experiment—whether the assumptions underlying the significance test are met—is likely to require reaction to features of the situation that are fully as complex as the features that the aid is supposed to simplify.

Whether there is net gain, even with the most widely used and simplest inferential procedures depends upon the extent to which the scientist and his colleagues react more effectively with than without such judgemental aids. And although these techniques have been widely used in experimental psychology for over 30 yr it is not at all clear that this particular field is in any way more effective because of them. From an empirical point of view, it would be desirable to have some data comparing the scientific or practical results achieved when significance tests are used with those when judgements are otherwise based. I know of no information of this sort. From a rational point of view, the incorporation of statistical inference into the broader field of decision theory clarifies considerably the possible role of the significance level as a guide to action. When combined with estimates of prior probability values for null and alternative hypotheses, and with quantitative estimates of the utility to the decision maker of correct and incorrect decisions, the significance of a treatment effect may be seen as a part of a very reasonable system for making decisions (as described, for example by Raiffa, 1968, or Schmitt, 1969). There is some hope that developments within this area may eventually prove useful to the psychologist, but at present the assignment of prior probabilities and utilities seems possible in only a few applied research situations and not at all in basic research. From this decision-theoretic point of view, the significance test by itself is a very incomplete basis for any kind of judgement, and there certainly seems to be no rationale for the widespread use of any particular level of significance (0.05 is the most common) as a basis for distinguishing "real" from "chance" effects. If and when these better rationalized inference procedures become available to the psychologist, however, they will be even more expensive in terms of time spent dealing with the details of the judgemental aid itself, and a net gain will be realized only if they produce a

considerable improvement in experimental effectiveness.

Some Detrimental Effects Arising from an Emphasis on Statistical Inference

Although it is not at all clear how statistical inference has helped the field of experimental psychology, it does seem closely linked with some undesirable changes in experimental practice. By the early 1930s, professional statisticians had developed significance test procedures appropriate to experiments of considerable complexity. If one was willing to rely on the result of the significance test as the main basis for reacting to an experiment, it then became possible to "control" statistically for unwanted sources of variation in a dependent variable, especially using the analysis of variance as developed by R. A. Fisher (1925). Before this development, an investigator had to discover techniques for experimentally controlling sources of irrelevant variation before he could even carry out his experiment. In the process he was likely to acquire a very valuable form of knowledge, irrespective of the ultimate value of the specific experiment. He was learning how to control his subject matter—in the case of psychology, the behavior of organisms, and even if the original reason for conducting the experiment was a poor one, something useful was likely to come of it. In addition to the reportable knowledge resulting from the effort to develop experimental control, this activity usually required a good deal of time, and so the experimenter was repeatedly exposed to the relevant contingencies of his problem area. He thus had a chance of being shaped into more effective forms of behavior regarding this subject matter even before his verbal repertoire regarding it was well developed. Also, since most problems concern more than one important independent variable, yet only one could generally be studied at a time, an investigator would usually conduct a series of separate experiments to tease out the various relationships, and was thus further exposed to the contingencies of his problem area.

The possibility of "statistical control" greatly reduced the necessity for developing experimental control. An experimenter could ask his experimental question irrespective of considerable uncontrolled variation in his dependent variable, if he could simply identify the sources of this variation. The study could then be designed in such a way that these sources were "balanced" across the various groups constituting the main comparison, and a satisfactory significance test could be computed with respect to this main comparison. These same methods of experimental design and statistical analysis also made possible the simultaneous investigation of more than one independent variable, thus further reducing experimental time and labor, but also reducing the duration and intensity of the experimenter's contact with his problem area.

At least five harmful effects of this general trend can be discerned.

1. The prolonged and intense interaction with the subject matter undertaken in order to experimentally control irrelevant sources of variation probably constituted a rich source of ideas for further experimentation. The use of statistical control deprives the experimenter of this source and he becomes more dependent upon theory, other researchers' experiments, and a form of commonsense analysis not necessarily related to his problem area as the basis for directing his research.

2. The knowledge developed in order to identify sources of variation and to select subjects in such a way as to "balance" for these sources is considerably less useful to other experimenters or for practical purposes than the knowledge required actually to control such variation.

3. Statistical control in complex experiments is easiest to accomplish by obtaining data from a large number of relatively independent behaving organisms, and such numbers generally preclude prolonged study of any one organism. Experimental situations then, are designed to maximize the efficiency with which they provide exactly the type of information relevant to the particular experimental question being asked,

and become increasingly unlike any other situations, either inside or outside of the laboratory. The results from such experiments are thus less useful for any purpose other than answering the specific question being asked in that experiment, which has the further disadvantage that they are less likely to be verified by another experimenter using the same situation to study a different problem.

4. Reliance on the significance test leading to the extensive use of statistical control and multiple-factor experiments produces an excessive dependency on the significance test, since such experiments cannot be reacted to in any other way. What started out as a supplement to other bases of judgement, has become, in the minds of many researchers an essential aspect of scientific method. Yet, as Skinner points out, "We owe most of our scientific knowledge to methods of inquiry which have never been formally analyzed or expressed in normative rules. (1972, p. 319)"

5. Since extensive preliminary study of an area is seemingly rendered unnecessary if one designs his experiment properly, and since such properly designed experiments cannot be interpreted until all the data are in and the significance tests have been performed, experiments tend to be carried out in a somewhat inflexible manner. In the type of research emphasizing experimental control, and thereby often involving prolonged study of a small number of organisms using relatively simple experimental designs, it is usually possible to change the procedure while the experiment is under way. If it appears that some previously unrecognized source of variation is causing trouble, the main manipulations can be postponed until means for controlling the interferring factor are developed. Or, if some aspect of the incoming results suggests an interesting variation the experiment can be redirected immediately.

All in all, it seems possible to argue that what might have been a moderately useful judgemental aid has ultimately had the unfortunate effect of moving psychological research methodology out of the main stream of experimental science.

Statistical Inference for Individual Organism Research: A Weak Solution to an Artificial Problem

Not all areas within experimental psychology have adopted the research methodology deplored above. One that has been relatively unaffected is the area referred to as "the experimental analysis of behavior", "operant conditioning", "Skinnerian psychology", *etc.* The shunning of significance testing by researchers with this orientation may be due, as Gentile *et al.* suggested (1972), to the unavailability of inferential techniques appropriate to typical "single subject" data. On the other hand, this type of individual organism research has been going on for well over 30 yr and it is reasonable to assume that if any strong need for such techniques was felt there would have been some concerted effort to develop them. It seems to me that the relative indifference to statistical inference is more accurately attributable to the strong emphasis on effective experimental control as a major scientific goal and as the main evidence of the scientist's "understanding" of his problem area.[2] The situation where a significance test might seem helpful is typically one involving sufficient uncontrolled variability in the dependent variable that neither the experimenter nor his readers can be sure that there is an interpretable relationship. This is evidence that the relevant behavior is not under good experimental control, a situation calling for more effective experimentation, not a more complex judgemental aid.

In any case, whether by necessity, scientific cunning, or prejudice, operant researchers, basic and applied, have made little use of statistical inference and do not seem to have suffered as a

[2]This emphasis, of course, predisposes investigators toward prolonged study of a small number of organisms, and within-subject comparisons where possible. Between-subject comparisons, however, can also be quite meaningful if behavior is under good experimental control.

result. Increasingly sophisticated methods of experimental control have developed within the area of basic research, and applied researchers have generally been able to make use of the same technology, or develop methods of experimental control appropriate to their own problem areas.

As the applied area expands, however, there seems to be an increasing tendency to present experimental results that are not easily interpreted when simply displayed in graphical form, or as a table of means or per cents. This is said to be due to the practical difficulties that the applied researcher encounters in his efforts to obtain human data in the nonlaboratory environment. It is argued that he does not have the luxury of discontinuing the experiment until he discovers and experimentally controls various sources of irrelevant or confusing variation in his data. He cannot, like the basic researcher, simply discard that pigeon and start over again with another. The opportunity for experimenting may no longer be present, a number of people may have been inconvenienced, a good deal of experimenter time may have been spent, and considerable financial as well as other resources may have been expended. One must, in a sense, make the best of the data as they stand, and this is where the significance test comes in. Faced with data that do not constitute an effective stimulus for judgement, the experimenter and his readers must do whatever is possible, and perhaps they will be able to behave somewhat more effectively if they have the judgemental aid offered by some statistical inference procedure.

This, of course, is what Gentile *et al.* are offering, and although critical of that specific solution, the other authors (Hartmann, 1974; Keselman, 1974; Kratochwill *et al.*, 1974; Thoresen and Elashoff, 1974) offer their own solutions of the same type. It is probably never appropriate to be critical of any valid knowledge-seeking activity *per se,* but one can criticize its rationale. The present interest in obtaining a proper significance test procedure for single-subject data seems based on two faulty premises. First is the belief that applied data are taken under conditions where effective experimental control cannot be expected. While workers in the field of applied behavior analysis have not been as badly affected by the experimental design and statistical significance enthusiasts as some other kinds of psychologists, they may not have escaped entirely. Peaceful coexistence with those who emphasize statistical control and multiple-factor experiments seems to have resulted in an increased tendency to plan, carry out, and then analyze the experiment all as a relatively inflexible unit of behavior—the fifth harmful effect listed earlier. When a dependent variable is not under good control—when there is considerable unexplained variability even though the independent variable being studied is at a constant value—it is not usually necessary to go ahead with the other planned manipulations. Further efforts can be made to obtain a more stable dependent variable, or to discover and eliminate some of the sources of uncontrolled variation.

If these efforts are unsuccessful and if the experiment is an expensive one in terms of time and other resources it is probably wise to abandon it at this point or recognize it as a gamble with a low probability of payoff. There are, of course, a number of "nonscientific" reasons for continuing an apparently unprofitable experiment, such as the necessity of completing a thesis or dissertation requirement, or the belief that if one does not carry out the research project that he spoke so highly of in the grant request he may have trouble getting another grant. That the significance test might be of aid in such situations and could actually further such purposes is certainly no recommendation.

The second faulty premise is that the significance test is an especially helpful judgemental aid, and therefore worth a good deal of time and inconvenience. When experimental control is emphasized and results can be portrayed in relatively simple graphical form, the probability of those results or more extreme ones given the

null hypothesis is a very crude form of information, compared with the other stimulus features available to the experimenter, and is likely to be ignored if it is not consistent with the interpretation arrived at otherwise.[3] In the typical multiple-factor experiment relying heavily on statistical control, the significance value is no more informative in an absolute sense, but since the results cannot generally be reacted to in any other way, it seems more useful. This means only that one should avoid experimenting in such a way that he is forced to rely on such a weak tool.

An overvaluation of the significance test by itself is a relatively harmless misunderstanding, but it is likely to cause other changes in experimental practice that are more serious. If the significance test is valued above all other judgemental aids, experimenters are likely to try to design their experiments so that a significance test can be computed, an obvious loss in terms of experimental flexibility. Note in this connection Hartmann's (1974) suggestion that ". . . at least 12 and preferably more stable data points should be available for each condition."

Another undesirable possibility is that a good deal of time will be spent in learning about and interacting with the judgemental aid, rather than in contact with the experimental area itself. In the operant area we already have a powerful source of distraction from our primary "target", in that many experimenters often find it at least temporarily more satisfying to experiment with their behavior control equipment—electromechanical, solid state, and more recently on-line computer—than to experiment with behavior. In the case of the autoregressive techniques that seem to be "just around the corner", their understanding will surely require a good deal of graduate instruction time and their proper usage could easily become a main concern from the point of view of data analysis—clearly a case of the tail wagging the dog.[4]

If the decreased experimental flexibility and the distraction from our primary subject of interest is not sufficient reason to be unenthusiastic about this development, there is the further distinct possibility that editors confronted with results that are in an obvious sense relatively meaningless may be induced to foist these results off on the readers if they are accompanied by an appropriate significance test that reaches the 5% value.

What Gentile, Roden, and Klein, and the other authors as well, are offering researchers in the area of behavior analysis is an opportunity to adopt a practice that has had a 30-yr trial period and is still of uncertain value. It is a practice, furthermore, that seems historically almost incompatible with the emphasis on experimental control that has characterized the operant research orientation. This would seem to be an offer we can afford to refuse.

REFERENCES

Fisher, R. A. *Statistical methods for research workers.* Edinburgh: Oliver and Boyd, 1925.

Gentile, R. R., Roden, A. H., and Klein, R. D. An analysis-of-variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis,* 1972, **5**, 193-198.

[3]It is possible that there are still some researchers who overvalue the significance test because they believe that the significance value reached in any particular test is equivalent to the probability that the null hypothesis is true. We cannot blame the professional statisticians for this misinterpretation, however, except that in warning us to avoid this error they have not often substituted a plausible alternative.

[4]It is often pointed out that the time spent dealing with the statistical judgemental aid can now be minimized by utilizing computer programs developed for this purpose. The experimenter can simply "plug in" his data and read out the significance value as well as some indication of the appropriateness of the particular technique to those data. This would seem to represent even further dependency upon an expertise which is beyond one's own critical scrutiny, an essentially undesirable direction to take. It can be argued, of course, that we all depend upon experts in other areas—an example is the biologist's dependency upon the optical specialists who design and construct his microscopes. We do it, however, on the basis of earned confidence, and the statisticians' contribution to experimental psychology seems quite uncertain when compared with the optical specialists' contribution to biology.

Hartmann, D. P. Forcing square pegs into round holes: some comments on 'An analysis-of-variance model for the intrasubject replication design.' *Journal of Applied Behavior Analysis,* 1974, **7,** 635-638.

Keselman, H. J. Concerning the statistical procedures enumerated by Gentile *et al.:* another perspective. *Journal of Applied Behavior Analysis,* 1974, **7,** 643-645.

Kratochwill, T., Alden, K., Demuth, D., Dawson, D., Panicucci, C., Arnston, P., McMurray, N., Hempstead, J. and Levin, J. A further consideration in the application of an analysis of variance model for the intrasubject replication design. *Journal of Applied Behavior Analysis,* 1974, **7,** 629-633.

Raiffa, H. *Decision analysis.* Reading, Mass.: Addison-Wesley, 1968.

Schmitt, S. A. *Measuring uncertainty.* Reading, Mass.: Addison-Wesley, 1969.

Skinner, B. F. *Cumulative record.* 3d ed. New York: Appleton-Century-Crofts, 1972.

Thoresen, C. E. and Elashoff, J. D. 'An analysis-of-variance model for intrasubject replication design:' some additional comments. *Journal of Applied Behavior Analysis,* 1974, **7,** 639-641.

Received 20 August 1974.
(Published without revision.)