Psychological Bulletin
1961, Vol. 58, No. 4, 305–316

SCALES AND STATISTICS: PARAMETRIC AND NONPARAMETRIC[1]

NORMAN H. ANDERSON
University of California, Los Angeles

The recent rise of interest in the use of nonparametric tests stems from two main sources. One is the concern about the use of parametric tests when the underlying assumptions are not met. The other is the problem of whether or not the measurement scale is suitable for application of parametric procedures. On both counts parametric tests are generally more in danger than nonparametric tests. Because of this, and because of a natural enthusiasm for a new technique, there has been a sometimes uncritical acceptance of nonparametric procedures. By now a certain degree of agreement concerning the more practical aspects involved in the choice of tests appears to have been reached. However, the measurement theoretical issue has been less clearly resolved. The principal purpose of this article is to discuss this latter issue further. For the sake of completeness, a brief overview of practical statistical considerations will also be included.

A few preliminary comments are needed in order to circumscribe the subsequent discussion. In the first place, it is assumed throughout that the data at hand arise from some sort of measuring scale which gives numerical results. This restriction is implicit in the proposal to compare parametric and nonparametric tests since the former do not apply to strictly categorical data (but see Cochran, 1954). Second, parametric tests will mean tests of significance which assume equinormality, i.e., normality and some form of homogeneity of variance. For convenience, parametric test, F test, and analysis of variance will be used synonymously. Although this usage is not strictly correct, it should be noted that the t test and regression analysis may be considered as special applications of F. Nonparametric tests will refer to significance tests which make considerably weaker distributional assumptions as exemplified by rank order tests such as the Wilcoxon T, the Kruskal-Wallis H, and by the various median-type tests. Third, the main focus of the article is on tests of significance with a lesser emphasis on descriptive statistics. Problems of estimation are touched on only slightly although such problems are becoming increasingly important.

Finally, a word of caution is in order. It will be concluded that parametric procedures constitute the everyday tools of psychological statistics, but it should be realized that any area of investigation has its own statistical peculiarities and that general statements must always be adapted to the prevailing practical situation. In many cases, as in pilot work, for instance, or in situations in which data are cheap and plentiful, nonparametric tests, shortcut parametric tests (Tate & Clelland, 1957), or tests by visual inspection may well be the most efficient.

[1] An earlier version of this paper was presented at the April 1959 meetings of the Western Psychological Association. The author's thanks are due F. N. Jones and J. B. Sidowski for their helpful comments.

Practical Statistical Considerations

The three main points of comparison between parametric and nonparametric tests are significance level, power, and versatility. Most of the relevant considerations have been treated adequately by others and only a brief summary will be given here. For more detailed discussion, the articles of Cochran (1947), Savage (1957), Sawrey (1958), Gaito (1959), and Boneau (1960) are especially recommended.

Significance level. The effects of lack of equinormality on the significance level of parametric tests have received considerable study. The two handiest sources for the psychologist are Lindquist's (1953) citation of Norton's work, and the recent article of Boneau (1960) which summarizes much of the earlier work. The main conclusion of the various investigators is that lack of equinormality has remarkably little effect although two exceptions are noted: one-tailed tests and tests with considerably disparate cell n's may be rather severely affected by unequal variances.[2]

A somewhat different source of perturbation of significance level should also be mentioned. An over-all test of several conditions may show that something is significant but will not localize the effects. As is well known, the common practice of t testing pairs of means tends to inflate the significance level even when the over-all F is significant. An analogous inflation occurs with nonparametric tests. There are parametric multiple comparison procedures which are rigorously applicable in many such situations (Duncan, 1955; Federer, 1955) but analogous nonparametric techniques have as yet been developed in only a few cases.

Power. As Dixon and Massey (1957) note, rank order tests are nearly as powerful as parametric tests under equinormality. Consequently, there would seem to be no pressing reason in most investigations to use parametric techniques for reasons of power *if* an appropriate rank order test is available (but see Snedecor, 1956, p. 120). Of course, the loss of power involved in dichotomizing the data for a median-type test is considerable.

Although it might thus be argued that rank order tests should be generally used where applicable, it is to be suspected that such a practice would produce negative transfer to the use of the more incisive experimental designs which need parametric analyses. The logic and computing rules for the analysis of variance, however, follow a uniform pattern in all situations and thus provide maximal positive transfer from the simple to the more complex experiments.

There is also another aspect of power which needs mention. Not infrequently, it is possible to use existing data to get a rough idea of the chances of success in a further related experiment, or to estimate the N required for a given desired probability of success (Dixon & Massey, 1957, Ch. 14). Routine methods are available for these purposes when parametric statistics are employed but similar procedures are available only for certain nonparametric tests such as chi square.

[2] The split-plot designs (e.g., Lindquist, 1953) commonly used for the analysis of repeated or correlated observations have been subject to some criticism (Cotton, 1959; Greenhouse & Geisser, 1959) because of the additional assumption of equal correlation which is made. However, tests are available which do not require this assumption (Cotton, 1959; Greenhouse & Geisser, 1959; Rao, 1952).

Versatility. One of the most remarkable features of the analysis of variance is the breadth of its applicability, a point which has been emphasized by Gaito (1959). For present purposes, the ordinary factorial design will serve to exemplify the issue. Although factorial designs are widely employed, their uses in the investigation and control of minor variables have not been fully exploited. Thus, Feldt (1958) has noted the general superiority of the factorial design in matching or equating groups, an important problem which is but poorly handled in current research (Anderson, 1959). Similarly, the use of replications as a factor in the design makes it possible to test and partially control for drift or shift in apparatus, procedure, or subject population during the course of an experiment. In the same way, taking experimenters or stimulus materials as a factor allows tests which bear on the adequacy of standardization of the experimental procedures and on the generalizability of the results.

An analogous argument could be given for latin squares, largely rehabilitated by the work of Wilk and Kempthorne (1955), which are useful when subjects are given successive treatments; for orthogonal polynomials and trend tests for correlated scores (Grant, 1956) which give the most sensitive tests when the independent variable is scaled; as well as for the multivariate analysis of variance (Rao, 1952) which is applicable to correlated dependent variables measured on incommensurable scales.

The point to these examples and to the more extensive treatment by Gaito is straightforward. Their analysis is more or less routine when parametric procedures are used. However, they are handled inadequately or not at all by current nonparametric methods.

It thus seems fair to conclude that parametric tests constitute the standard tools of psychological statistics. In respect of significance level and power, one might claim a fairly even match. However, the versatility of parametric procedures is quite unmatched and this is decisive. Unless and until nonparametric tests are developed to the point where they meet the routine needs of the researcher as exemplified by the above designs, they cannot realistically be considered as competitors to parametric tests. Until that day, nonparametric tests may best be considered as useful minor techniques in the analysis of numerical data.

Too promiscuous a use of F is, of course, not to be condoned since there will be many situations in which the data are distributed quite wildly. Although there is no easy rule with which to draw the line, a frame of reference can be developed by studying the results of Norton (Linquist, 1953) and of Boneau (1960). It is also quite instructive to compare p values for parametric and nonparametric tests of the same data.

It may be worth noting that one of the reasons for the popularity of nonparametric tests is probably the current obsession with questions of statistical significance to the neglect of the often more important questions of design and power. Certainly some minimal degree of reliability is generally a necessary justification for asking others to spend time in assessing the importance of one's data. However, the question of statistical significance is only a first step, and a relatively minor one at that, in the over-all process of evaluating a set of results. To say that a result is statistically significant simply gives reasonable ground for believing that

some nonchance effect was obtained. The meaning of a nonchance effect rests on an assessment of the design of the investigation. Even with judicious design, however, phenomena are seldom pinned down in a single study so that the question of replicability in further work often arises also. The statistical aspects of these these two questions are not without importance but tend to be neglected when too heavy an emphasis is placed on p values. As has been noted, it is the parametric procedures which are the more useful in both respects.

Measurement Scale Considerations

The second and principal part of the article is concerned with the relations between types of measurement scales and statistical tests. For convenience, therefore, it will be assumed that lack of equinormality presents no serious problem. Since the F ratio remains constant with changes in unit or zero point of the measuring scale, we may ignore ratio scales and consider only ordinal and interval scales. These scales are defined following Stevens (1951). Briefly, an ordinal scale is one in which the events measured are, in some empirical sense, ordered in the same way as the arithmetic order of the numbers assigned to them. An interval scale has, in addition, an equality of unit over different parts of the scale. Stevens goes on to characterize scale types in terms of permissible transformations. For an ordinal scale, the permissible transformations are monotone since they leave rank order unchanged. For an interval scale, only the linear transformations are permissible since only these leave relative distance unchanged. Some workers (e.g., Coombs, 1952) have considered various scales which lie between the ordinal and interval scales. However, it will not be necessary to take this further refinement of the scale typology into account here.

As before, we suppose that we have a measuring scale which assigns numbers to events of a certain class. It is assumed that this measuring scale is an ordinal scale but not necessarily an interval scale. In order to fix ideas, consider the following example. Suppose that we are interested in studying attitude toward the church. Subjects are randomly assigned to two groups, one of which, reads Communication A, while the other reads Communication B. The subjects' attitudes towards the church are then measured by asking them to check a seven category pro-con rating scale. Our problem is whether the data give adequate reason to conclude that the two communications had different effects.

To ascertain whether the communications had different effects, some statistical test must be applied. In some cases, to be sure, the effects may be so strong that the test can be made by inspection. In most cases, however, some more objective method is necessary. An obvious procedure would be to assign the numbers 1 to 7, say, to the rating scale categories and apply the F test, at least if the data presented some semblance of equinormality. However, some writers on statistics (e.g., Siegel, 1956; Senders, 1958) would object to this on the ground that the rating scale is only an ordinal scale, the data are therefore not "truly numerical," and hence that the operations of addition and multiplication which are used in computing F cannot meaningfully be applied to the scores. There are three different

questions involved in this objection, and much of the controversy over scales and statistics has arisen from a failure to keep them separate. Accordingly, these three questions will be taken up in turn.

Question 1. Can the F *test be applied to data from an ordinal scale?* It is convenient to consider two cases of this question according as the assumption of equinormality is satisfied or not. Suppose first that equinormality obtains. The caveat against parametric statistics has been stated most explicitly by Siegel (1956) who says:

> The conditions which must be satisfied . . . before any confidence can be placed in any probability statement obtained by the use of the *t* test are at least these: . . . 4. The variables involved must have been measured in *at least* an interval scale . . . (p. 19). (By permission, from *Nonparametric Statistics*, by S. Siegel. Copyright, 1956. McGraw-Hill Book Company, Inc.)

This statement of Siegel's is completely incorrect. This particular question admits of no doubt whatsoever. The *F* (or *t*) test may be applied without qualm. It will then answer the question which it was designed to answer: can we reasonably conclude that the difference between the means of the two groups is real rather than due to chance? The justification for using *F* is purely statistical and quite straightforward; there is no need to waste space on it here. The reader who has doubts on the matter should postpone them to the discussion of the two subsequent questions, or read the elegant and entertaining article by Lord (1953). As Lord points out, the statistical test can hardly be cognizant of the empirical meaning of the numbers with which it deals. Consequently, the validity of a statistical inference cannot depend on the type of measuring scale used.

The case in which equinormality does not hold remains to be considered. We may still use *F*, of course, and as has been seen in the first part, we would still have about the same significance level in most cases. The *F* test might have less power than a rank order test so that the latter might be preferable in this simple two group experiment. However, insofar as we wish to inquire into the reliability of the difference between the measured behavior of the two groups in our particular experiment, the choice of statistical test would be governed by purely statistical considerations and have nothing to do with scale type.

Question 2. Will statistical results be invariant under change of scale? The problem of invariance of result stems from the work of Stevens (1951) who observes that a statistic computed on data from a given scale will be invariant when the scale is changed according to any given permissible transformation. It is important to be precise about this usage of invariance. It means that if a statistic is computed from a set of scale values and this statistic is then transformed, the identical result will be obtained as when the separate scale values are transformed and the statistic is computed from these transformed scale values.

Now our scale of attitude toward the church is admittedly only an ordinal scale. Consequently, we would expect it to change in the direction of an interval scale in future work. Any such scale change would correspond to a monotone transformation of our original scale since only such transformations are permissible with an ordinal scale. Suppose then that a monotone transformation of the scale has been made subsequent to the experiment on attitude change.

We would then have two sets of data: the responses as measured on the original scale used in the experiment, and the transformed values of these responses as measured on the new, transformed scale. (Presumably, these transformed scale values would be the same as the subjects would have made had the new scale been used in the original experiment, although this will no doubt depend on the experimental basis of the new scale.) The question at issue then becomes whether the same significance results will be obtained from the two sets of data. If rank order tests are used, the same significance results will be found in either case because any permissible transformation leaves rank order unchanged. However, if parametric tests are employed, then different significance statements may be obtained from the two sets of data. It is possible to get a significant F from the original data and not from the transformed data, and vice versa. Worse yet, it is even logically possible that the means of the two groups will lie in reverse order on the two scales.

The state of affairs just described is clearly undesirable. If taken uncritically, it would constitute a strong argument for using only rank order tests on ordinal scale data and restricting the use of F to data obtained from interval scales. It is the purpose of this section to show that this conclusion is unwarranted. The basis of the argument is that the naming of the scales has begged the psychological question.

Consider interval scales first, and imagine that two students, P and Q, in an elementary lab course are assigned to investigate some process. This process might be a ball rolling on a plane, a rat running an alley, or a child doing sums. The students cooperate in the experimental work, making the same observations, except that they use different measuring scales. P decides to measure time intervals. He reasons that it makes sense to speak of one time interval as being twice another, that time intervals therefore form a ratio scale, and hence a fortiori an interval scale. Q decides to measure the speed of the process (feet per second, problems per minute). By the same reasoning as used by P, Q concludes that he has an interval scale also. Both P and Q are aware of current strictures about

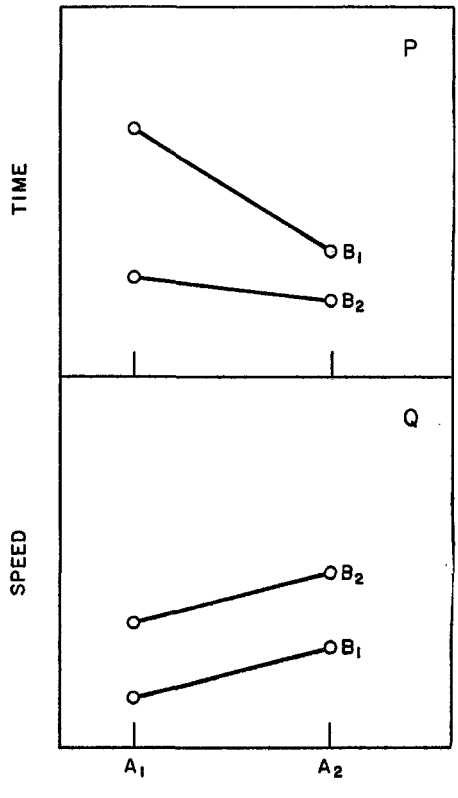

FIG. 1. Temporal aspects of some process obtained from a 2×2 design. (The data are plotted as a function of Variable A with Variable B as a parameter. Subscripts denote the two levels of each variable. Note that Panel P shows an interaction, but that Panel Q does not.)

scales and statistics. However, since each believes (and rightly so) that he has an interval scale, each uses means and applies parametric tests in writing his lab report. Nevertheless, when they compare their reports they find considerable difference in their descriptive statistics and graphs (Figure 1), and in their F ratios as well. Consultation with a statistican shows that these differences are direct consequences of the difference in the measuring scales. Evidently then, possession of an interval scale does not guarantee invariance of interval scale statistics.

For ordinal scales, we would expect to obtain invariance of result by using ordinal scale statistics such as the median (Stevens, 1951). Let us suppose that some future investigator finds that attitude toward the church is multidimensional in nature and has, in fact, obtained interval scales for each of the dimensions. In some of his work he chanced to use our original ordinal scale so that he was able to find the relation between this ordinal scale and the multidimensional representation of the attitude. His results are shown in Figure 2. Our ordinal scale is represented by the curved line in the plane of the two dimensions. Thus, a greater distance from the origin as measured along the line stands for a higher value on our ordinal scale. Points A and B on the curve represent the medians of Groups A and B in our experiment, and it is seen that Group A is more pro-church than Group B on our ordinal scale. The median scores for these two groups on the two dimensions are obtained simply by projecting Points A and B onto the two dimensions. All is well on Dimension 2 since there Group A is greater than Group B. On Dimension 1, however, a reversal is found: Group A is less than Group B,

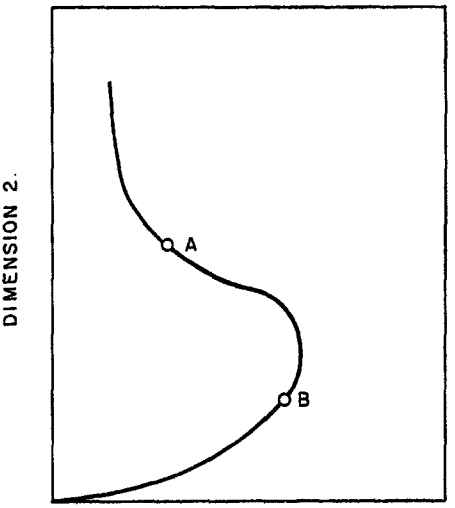

FIG. 2. The curved line represents the ordinal scale of attitude toward the church plotted in the two-dimensional space underlying the attitude. (Points A and B denote the medians of two experimental groups. The graph is hypothetical, of course.)

contrary to our ordinal scale results. Evidently then, possession of an ordinal scale does not guarantee invariance of ordinal scale statistics.

A rather more drastic loss of invariance would occur if the ordinal scale were measuring the resultant effect of two or more underlying processes. This could happen, for instance, in the study of approach-avoidance conflict, or ambivalent behavior, as might be the case with attitude toward the church. In such situations, two people could give identical responses on the one-dimensional scale and yet be quite different as regards the two underlying processes. For instance, the same resultant could occur with two equal opposing tendencies of any given strength. Representing such data in the space formed by the underlying dimensions would yield a smear of points over an entire

region rather than a simple curve as in Figure 2.

Although it may be reasonable to think that simple sensory phenomena are one-dimensional, it would seem that a considerable number of psychological variables must be conceived of as multidimensional in nature as, for instance, with "IQ" and other personality variables. Accordingly, as the two cited examples show, there is no logical guarantee that the use of ordinal scale statistics will yield invariant results under scale changes.

It is simple to construct analogous examples for nominal scales. However, their only relevance would be to show that a reduction of all results to categorical data does not avoid the difficulty with invariance.

It will be objected, of course, that the argument of the examples has violated the initial assumption that only "permissible" transformations would be used in changing the measuring scales. Thus, speed and time are not linearly related, but rather the one is a reciprocal transformation of the other. Similarly, Dimension 1 of Figure 2 is no monotone transformation of the original ordinal scale. This objection is correct, to be sure, but it simply shows that the problem of invariance of result with which one is actually faced in science has no particular connection with the invariance of "permissible" statistics. The examples which have been cited show that knowing the scale type, as determined by the commonly accepted criteria, does not imply that future scales measuring the same phenomena will be "permissible" transformations of the original scale. Hence the use of "permissible" statistics, although guaranteeing invariance of result over the class of "permissible" transformations, says little about invariance of result over the class of scale changes which must actually be considered by the investigator in his work.

This point is no doubt pretty obvious, and it should not be thought that those who have taken up the scale-type ideas are unaware of the problem. Stevens, at least, seems to appreciate the difficulty when, in the concluding section of his 1951 article, he distinguishes between psychological dimensions and indicants. The former may be considered as intervening variables whereas the latter are effects or correlates of these variables. However, it is evident that an indicant may be an interval scale in the customary sense and yet bear a complicated relation to the underlying psychological dimensions. In such cases, no procedure of descriptive or inferential statistics can guarantee invariance over the class of scale changes which may become necessary.

It should also be realized that only a partial list of practical problems of invariance has been considered. Effects on invariance of improvements in experimental technique would also have to be taken into account since such improvements would be expected to purify or change the dependent variable as well as decrease variability. There is, in addition, a problem of invariance over subject population. Most researches are based on some handy sample of subjects and leave more or less doubt about the generality of the results. Although this becomes in large part an extrastatistical problem (Wilk & Kempthorne, 1955), it is one which assumes added importance in view of Cronbach's (1957) emphasis on the interaction of experimental and subject variables. In the face of these assorted difficulties, it is not easy to see what utility the scale typology

has for the practical problems of the investigator.

The preceding remarks have been intended to put into broader perspective that sort of invariance which is involved in the use of permissible statistics. They do not, however, solve the immediate problem of whether to use rank order tests or F in case only permissible transformations need be considered. Although invariance under permissible scale transformations may be of relatively minor importance, there is no point in taking unnecessary risks without the possibility of compensation.

On this basis, one would perhaps expect to find the greatest use of rank order tests in the initial stages of inquiry since it is then that measuring scales will be poorest. However, it is in these initial stages that the possibly relevant variables are not well-known so that the stronger experimental designs, and hence parametric procedures, are most needed. Thus, it may well be most efficient to use parametric tests, balancing any risk due to possible permissible scale changes against the greater power and versatility of such tests. In the later stages of investigation, we would be generally more sure of the scales and the use of rank order procedures would waste information which the scales by then embody.

At the same time, it should be realized that even with a relatively crude scale such as the rating scale of attitude toward the church, the possible permissible transformations which are relevant to the present discussion are somewhat restricted. Since the F ratio is invariant under change of zero and unit, it is no restriction to assume that any transformed scale also runs from 1 to 7. This imposes a considerable limitation on the permissible scale transformations which must be considered. In addition, whatever psychological worth the original rating scale possesses will limit still further the transformations which will occur in practice.

Although rank order tests do possess some logical advantage over parametric tests when only permissible transformations are considered, this advantage is, in the writer's opinion, very slight in practice and does not begin to balance the greater versatility of parametric procedures. The problem is, however, an empirical one and it would seem that some historical analysis is needed to provide an objective frame of reference. To quote an after-lunch remark of K. MacCorquodale, "Measurement theory should be descriptive, not proscriptive, nor prescriptive." Such an inquiry could not fail to be fascinating because of the light it would throw on the actual progress of measurement in psychology. One investigation of this sort would probably be more useful than all the speculation which has been written on the topic of measurement.

Question 3. Will the use of parametric as opposed to nonparametric statistics affect inferences about underlying psychological processes? In a narrow sense, Question 3 is irrelevant to this article since the inferences in question are substantive, relating to psychological meaning, rather than formal, relating to data reliability. Nevertheless, it is appropriate to discuss the matter briefly in order to make explicit some of the considerations involved because they are often confused with problems arising under the two previous questions. With no pretense of covering all aspects of this question, the following two examples will at least touch some of the problems.

The first example concerns the two students, P and Q, mentioned above, who had used time and speed as dependent variables. We suppose that their experiment was based on a 2×2 design and yielded means as plotted in Figure 1. This graph portrays main effects of both variables which are seen to be similar in nature in both panels. However, our principal concern is with the interaction which may be visualized as measuring the degree of nonparallelism of the two lines in either panel. Panel P shows an interaction. The reciprocals of these same data, plotted in Panel Q, show no interaction. It is thus evident in the example, and true in general, that interaction effects will depend strongly on the measuring scales used.

Assessing an interaction does not always cause trouble, of course. Had the lines in Panel P, say, crossed each other, it would not be likely that any change of scale would yield uncrossed lines. In many cases also, the scale used is sufficient for the purposes at hand and future scale changes need not be considered. Nevertheless, it is clear that a measure of caution will often be needed in making inferences from interaction to psychological process. If the investigator envisages the possibility of future changes in the scale, he should also realize that a present inference based on significant interaction may lose credibility in the light of the rescaled data.

It is certainly true that the interpretation of interactions has sometimes led to error. It may also be noted that the usual factorial design analysis is sometimes incongruent with the phenomena. In a 2×2 design it might happen, for example, that three of the four cell means are equal. The usual analysis is not optimally sensitive to this one real difference since it is distributed over three degrees of freedom. In such cases, there will often be other parametric tests involving specific comparisons (Snedecor, 1956) or multiple comparisons (Ducan, 1955) which are more appropriate. Occasionally also, an analysis of variance based on a multiplicative model (Williams, 1952) will be useful (Jones & Marcus, 1961). A judicious choice of test may be of great help in dissecting the results. However, the test only answers set questions concerning the reliability of the results; only the research worker can say which questions are appropriate and meaningful.

Inferences based on nonparametric tests of interaction would presumably be less sensitive to certain types of scale changes. However, caution would still be needed in the interpretation as has been seen in Question 2. The problem is largely academic, however, since few nonparametric tests of interaction exist.[3] It might be suggested that the question of interaction cannot arise when only the ordinal properties of the data are considered since the interaction involves a comparison of differences and such a comparison is illegitimate with ordinal data. To the extent that this suggestion is correct, a parametric test can be used to the same purposes equally well if not better; to the extent that it is not correct, nonparametric tests will waste information.

One final comment on the first example deserves emphasis. Since both time and speed are interval scales, it cannot be argued that the

[3] There is a nomenclatural difficulty here. Strictly speaking, nonparametric tests should be called more-or-less distribution free tests. For example, the Mood-Brown generalized median test (Mood, 1950) is distribution free, but is based on a parametric model of the same sort as in the analysis of variance. As noted in the introduction, the usual terminology is used in this article.

difficulty in interpretation arises because we had only ordinal scales.

The second example, suggested by J. Kaswan, is shown in Figure 3. The graph, which is hypothetical, plots amount of aggressiveness as a function of amount of stress. A glance at the graph leads immediately to the inference that some sort of threshold effect is present. Under increasing stress, the organism remains quiescent until the stress passes a certain threshold value, whereupon the organism leaps into full scale aggressive behavior.

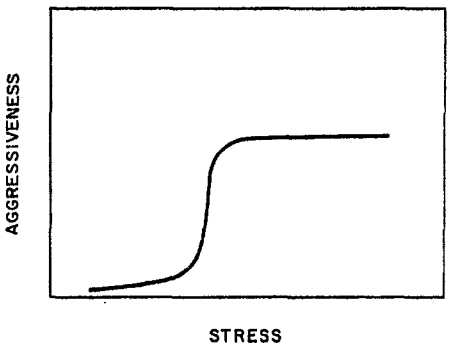

FIG. 3. Aggressiveness plotted as a function of stress. (The curve is hypothetical. Note the hypothetical threshold effect.)

Confidence in this interpretation is shaken when we stop to consider that the scales for stress and aggression may not be very good. Perhaps, when future work has given us improved scales, these same data would yield a quite different function such as a straight line.

One extreme position regarding the threshold effect would be to say that the scales give rank order information and no more. The threshold inference, or any inference based on characteristics of the curve shape other than the uniform upward trend, would then be completely disallowed. At the other extreme, there would be complete faith in the scales and all inferences based on curve shape, including the threshold effect, would be made without fear that they would be undermined by future changes in the scales. In practice, one would probably adopt a position between these two extremes, believing, with Mosteller (1958), that our scales generally have some degree of numerical information worked into them, and realizing that to consider only the rank order character of the data would be to ignore the information that gives the strongest hold on the behavior.

From this ill-defined middle ground, inferences such as the threshold effect would be entertained as guides to future work. Such inferences, however, are made at the judgment of the investigator. Statistical techniques may be helpful in evaluating the reliability of various features of the data, but only the investigator can endow them with psychological meaning.

SUMMARY

This article has compared parametric and nonparametric statistics under two general headings: practical statistical problems, and measurement theoretical considerations. The scope of the article is restricted to situations in which the dependent variable is numerical, thus excluding strictly categorical data.

Regarding practical problems, it was noted that the difference between parametric and rank order tests was not great insofar as significance level and power were concerned. However, only the versatility of parametric statistics meets the everyday needs of psychological research. It was concluded that parametric procedures are the standard tools of psychological statistics although nonparametric procedures are useful minor techniques.

Under the heading of measurement

theoretical considerations, three questions were distinguished. The well-known fact that an interval scale is not prerequisite to making a statistical inference based on a parametric test was first pointed out. The second question took up the important problem of invariance. It was noted that the practical problems of invariance or generality of result far transcend measurement scale typology. In addition, the cited example of time and speed showed that interval scales of a given phenomenon are not unique. The discussion of the third question noted that the problem of psychological meaning is not basically a statistical matter. It was thus concluded that the type of measuring scale used had little relevance to the question of whether to use parametric or nonparametric tests.

REFERENCES

ANDERSON, N. H. Education for research in psychology. *Amer. Psychologist*, 1959, **14**, 695–696.

BONEAU, C. A. The effects of violations of assumptions underlying the *t* test. *Psychol. Bull.*, 1960, **57**, 49–64.

COCHRAN, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics*, 1947, **3**, 22–38.

COCHRAN, W. G. Some methods for strengthening the common χ^2 tests. *Biometrics*, 1954, **10**, 417–451.

COOMBS, C. H. A theory of psychological scaling. *Bull. Engrg. Res. Inst. U. Mich.*, 1952, No. 34.

COTTON, J. W. A re-examination of the repeated measurements problem. Paper read at American Statistical Association, Chicago, December 1959.

CRONBACH, L. J. The two disciplines of scientific psychology. *Amer. Psychologist*, 1957, **11**, 671–684.

DIXON, W. J., & MASSEY, F. J., JR. *Introduction to statistical analysis.* (2nd ed.) New York: McGraw-Hill, 1957.

DUNCAN, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, **11**, 1–41.

FEDERER, W. T. *Experimental design.* New York: Macmillan, 1955.

FELDT, L. S. A comparison of the precision of three experimental designs employing a concomitant variable. *Psychometrika*, 1958, **23**, 335–354.

GAITO, J. Nonparametric methods in psychological research. *Psychol. Rep.*, 1959, **5**, 115–125.

GRANT, D. A. Analysis-of-variance tests in the analysis and comparison of curves. *Psychol. Bull.*, 1956, **53**, 141–154.

GREENHOUSE, S. W., & GEISSER, S. On methods in the analysis of profile data. *Psychometrika*, 1959, **24**, 95–112.

JONES, F. N., & MARCUS, M. J. The subject effect in judgments of subjective magnitude. *J. exp. Psychol.*, 1961, **61**, 40–44.

LINDQUIST, E. F. *Design and analysis of experiments.* Boston: Houghton Mifflin, 1953.

LORD, F. M. On the statistical treatment of football numbers. *Amer. Psychologist*, 1953, **8**, 750–751.

MOOD, A. M. *Introduction to the theory of statistics.* New York: McGraw-Hill, 1950.

MOSTELLER, F. The mystery of the missing corpus. *Psychometrika*, 1958, **23**, 279–290.

RAO, C. R. *Advanced statistical methods in biometric research.* New York: Wiley, 1952.

SAVAGE, I. R. Nonparametric statistics. *J. Amer. Statist. Ass.*, 1957, **52**, 331–344.

SAWREY, W. L. A distinction between exact and approximate nonparametric methods. *Psychometrika*, 1958, **23**, 171–178.

SENDERS, V. L. *Measurement and statistics.* New York: Oxford, 1958.

SIEGEL, S. *Nonparametric statistics.* New York: McGraw-Hill, 1956.

SNEDECOR, G. W. *Statistical methods.* (5th ed.) Ames: Iowa State Coll. Press, 1956.

STEVENS, S. S. Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley, 1951.

TATE, M. W., & CLELLAND, R. C. *Nonparametric and shortcut statistics.* Danville, Ill.: Interstate, 1957.

WILK, M. B., & KEMPTHORNE, O. Fixed, mixed, and random models. *J. Amer. Statist. Ass.*, 1955, **50**, 1144–1167.

WILLIAMS, E. J. The interpretation of interactions in factorial experiments. *Biometrika*, 1952, **39**, 65–81.

(Received April 8, 1960)