Psychological Review
1962, Vol. 69, No. 1, 54–61

THEORETICAL NOTES

TESTING THE NULL HYPOTHESIS AND THE STRATEGY AND TACTICS OF INVESTIGATING THEORETICAL MODELS

DAVID A. GRANT[1]

University of Wisconsin

Testing the null hypothesis, H_0, against alternatives, H_1, is well established and has a proper place in scientific research. However, this testing procedure, when it is routinely applied to comparing experimental outcomes with outcomes that are quantitatively predicted from a theoretical model, can have unintended results and bizarre implications. This paper first outlines three situations in which testing H_0 has conventionally been done by psychologists. In terms of the probable intentions or strategy of the experimenter testing H_0 turns out to be an appropriate tactic in the first situation, but it is inadequate in the second situation, and it is self-defeating with curious implications in the last situation. Alternatives to this conventional procedure are then presented along with the considerations which make the alternatives preferable to testing the usual H_0.

THREE APPLICATIONS OF H_0 TESTING

Probably the most common application of the tactic of testing H_0 arises when the independent variable has produced a sample difference or set of differences in the magnitude of the dependent variable. Quantitative predictions of the size of the difference or differences are not available. The experimenter wishes to know whether or not differences of the size obtained could have occurred by virtue of the operation of the innumerable nonexperimental factors conventionally designated as random. He sets up H_0 that the differences are zero; chooses a significance level, α; determines the set of hypotheses alternative to H_0 that he is willing to entertain, H_1; selects an appropriate test statistic, t, F, χ^2, U, T, or the like; and proceeds with the test. Rejection of H_0 permits him to assert, with a precisely defined risk of being wrong, that the obtained differences were not the product of chance variation. Failure of the test to permit rejection of H_0, which, unfortunately, is commonly termed "accepting" H_0, means that the obtained differences or greater ones would occur by chance with a probability greater than α. This situation is straightforward. The experimenter has limited aims. He has asked a simple question, and he has received a simple answer, subject only to those ambiguities which attend all experimental and inductive inference. His tactics are admirably suited to his strategic objective.

Another common but less satisfactory instance of testing H_0 arises when the results of pre-experimental matching or pretesting are to be evaluated. Here the experimenter has measured the dependent variable or some related variable before operation of the independent variable, and he devoutly hopes that the experimental and control groups are alike except for random differences. He is now relieved or chagrined, depending upon whether H_0 is "accepted" or "rejected" as a consequence of his test. Even if H_0 is accepted his relief is tempered by some uneasiness. He knows that he has not proved, and indeed cannot prove, that H_0 is "true." His tactics in testing H_0 seem to be appropriate to the impossible strategic aim of proving the truth of H_0. Certainly, if he had a more reasonable aim he has adopted inappropriate tactics. Utilizing these tactics, the best he can do is to beat a strategic retreat, and if H_0 is accepted he can perhaps point out that he has used a very powerful test and that if there were real differences they were most likely very small. Although psychologists have never to my knowledge done so, he might be able to go one step further and point out that his testing procedure would reject H_0 a given percentage of the time, say, 90%, if the "true" difference had been as little as, say, one-tenth of an *SD*. This sort of statement of the power of a test is a commonplace in acceptance inspection (Grant, 1952, Ch. 13).

With the advent of more detailed mathematical models in psychology (e.g., Bush, Abelson, & Hyman, 1956; Bush & Estes, 1959; Goldberg, 1958; Kemeny, Snell, & Thompson, 1957) a new statistical testing situation is arising more and more frequently. The specificity of the predictions and perhaps the whole philosophy behind model construction pose a different kind of statistical problem than those faced by most psychological investigators in the past. It seems obvious that as the use of models becomes more widespread a greater number of investigators will face the problem of evaluating the correspondence between empirical data points and precise numerical predictions of these points. Unfortunately most of the procedures used to date in testing the adequacy of such theoretical predictions set rather bad examples. Probably the least adequate of these procedures has been that in which an H_0 of exact correspondence between theoretical and empirical points is tested against H_1 covering any discrepancy between predictions and experimental results.

Most models predict a considerable number of different aspects of the data, and some of these aspects are predicted with greater success than others (Bush & Estes, 1959, Chs. 14, 15, 17, 18). We shall restrict our discussion to the prediction of values along a curve which might be a learning curve. An idealized version of such a typical situation is presented in Figure 1. Here, the dependent variable, Y, is plotted on the vertical axis against the independent variable, X, on the horizontal axis. The theoretical model has led to an expression, $Y' = f(X)$, giving a set of k theoretical predictions, $Y'_1, Y'_2, \cdots, Y'_k$. The experiment has produced k empirical data points, a set of mean values, $\bar{Y}_1, \bar{Y}_2, \cdots, \bar{Y}_k$, corresponding to the values of the independent variable that were investigated, namely, $X_1, X_2, \cdots, X_k$. Individual observations tend to form normal distributions about each of the $\bar{Y}_i$, and these normal distributions tend to have equal σ's for all data points. In further discussion we shall assume that inaccuracies in the manipulation of the independent variable, X, can be ignored. The problem now is to investigate the goodness of fit of the Y'_i to the $\bar{Y}_i$ or the correspondence between the Y'_i and the $\bar{Y}_i$.

The tactics oriented toward accepting H_0 as corroborating the theory involve breaking down the jth individual obser-

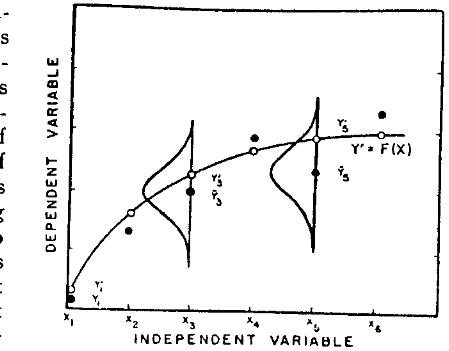

FIG. 1. Idealized situation involving the test of a theoretical function, $Y' = f(X)$. (Theoretical points, Y'_i, are represented by open circles; obtained means, $\bar{Y}_i$, are represented by solid circles.)

[1] The author is indebted to Arnold M. Binder of Indiana University whose arguments inspired him to make explicit some of the issues involved in using conventional analysis of variance procedures in testing the adequacy of a theoretical model. As this paper went through various revisions over a period of time the writer is correspondingly indebted to a number of supporting agencies: the Graduate Research Committee and the College of Letters and Science of the University of Wisconsin, the National Science Foundation, and finally to the Department of Psychology of the University of California, Los Angeles, his host during final preparation of the manuscript.

vation from the general mean of all of the observations, as follows:

$$Y_{ij} - \bar{Y} = (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - Y'_i) + (Y'_i - \bar{Y}) \quad [1]$$

where Y_{ij} is the jth observation in the ith normal distribution, and $\bar{Y}$ is the general mean of all observations.

the total sum of squares may then be partitioned as follows:

$$SS_{\text{Tot}} = SS_{\text{Dev Est}} + SS_{\text{Dev Theory}} + SS_{\text{Theory}} \quad [2]$$

where $SS_{\text{Dev Est}}$ is the sum of squares associated with the variation of individual measures from their means, $SS_{\text{Dev Theory}}$ is the sum of squares associated with the systematic departures of empirical data points from the theoretical points, and SS_{Theory} is the sum of squares associated with departures of the theoretical points from the general mean of the whole experiment.

If we suppose that the linear model for the analysis of variance holds, then:

$$Y_{ij} = \mu + T_i + D_i + e_{ij} \quad [3]$$

where μ is the population mean for all Y_{ij} over the specific values of the independent variable, X_i; T_i is the departure of the "true" theoretical value of Y'_i from μ; D_i is the discrepancy of the "true" value of $\bar{Y}_i$ from the true value of Y'_i; and e_{ij} is a random element from a normal distribution with a mean of zero and variance, σ_e^2, for all i.

For a fixed set of X_i the Ts and Ds may be defined so that $\sum T_i = \sum D_i = 0$. Under H_0 each $D_i = 0$. Under H_1 some $D_i \neq 0$, and the variance of the D_i, $\sigma_D^2 \neq 0$. This last variance may be termed the true variance of the discrepancies from the theory over the particular set of X_i that was investigated.

The foregoing is a conventional analysis of variance model, and the F ratio of the $MS_{\text{Dev Theory}}$ divided by the $MS_{\text{Dev Est}}$ provides an excellent and powerful test of H_0 against H_1. The number of degrees of freedom for $SS_{\text{Dev Est}}$ will be $k(n - 1)$ where n is the number of observations per data point, and the degrees of freedom for $SS_{\text{Dev Theory}}$ will be $k - n_T$, where n_T is the number of degrees of freedom lost in the process of fitting the model to the data. If this F is significant, we reject H_0, concluding that the discrepancies between the $\bar{Y}_i$ and the Y'_i are too great to be accounted for by the observed random variation in the experiment. In this conclusion we accept the 5% or 1% risk implied by our choice of α.

Logical difficulties arise when F fails of significance. H_0 remains tenable but is not proved to be correct. A tenable H_0 provides some support for the theory but only in the negative sense of failing to provide evidence that the theory is faulty. To assert that accepting the H_0 proves that the model provides a satisfactory fit to the data is an inaccurate and misleading statement. We may mean that we are satisfied, but others, especially proponents of other theories, will tend to regard our test as too lenient.

Failure to reject H_0, instead of producing closure, leaves certain annoying ambiguities, but the tactics of testing this particular H_0 imply a strategy that suffers from more serious defects that are readily apparent when the whole conception of testing a theory is carefully considered. To begin with, in view of our present psychological knowledge and the degree of refinement of available theoretical models it seems certain that even the best and most useful theories are not perfect. This means, in terms of the analysis of variance model, there will be some nonzero D_i's. H_0, then, is never really "true." Its "acceptance," rather than "proving" the theory, merely indicates that in this instance the D_i's were too small to be demonstrated by the sensitivity of the experiment in question. The tactics of accepting H_0 as proof and rejecting H_0 as disproof of a theory lead to the anomalous results that a small-scale, insensitive experiment will most often be interpreted as favoring a theory, whereas a large-scale, sensitive experiment will usually yield results opposed to the theory!

Curiously enough, even rejection of H_0 by means of a very stringent experimental test may be quite misleading as far as casting light on the adequacy of the theory is concerned. If the D_i's are very small indeed the theoretical model may be a great improvement over anything else that is available and satisfactory for many purposes even though an extremely sensitive experiment were to reveal the nonzero D_i's. If our task, as scientists, were to test and accept or reject theories as they came off some assembly line the tactics of testing H_0 could be made in a satisfactory manner simply by requiring that the test be "sufficiently" stringent. In fact, our task and our intentions are usually different from testing products; what we are really up to resembles *quality control* rather than *acceptance inspection*, and statistical procedures suitable for the latter are rarely optimal for the former (Grant, 1952, Chs. 1, 13).

Hypothesis Testing versus Statistical Estimation

An analogy will make clear the relation between testing tactics and the intention of the tester. Suppose that I wish to test a parachute; how should I go about it? How I should test depends upon my general intentions. If I want to sell the parachute and am testing it only to be able to claim that it has been tested, and I do not care what happens to the purchaser, then I should give the parachute a most lenient, nonanalytic test. If, however, I am testing the parachute to be sure of it for my own use, then I should subject it to a very stringent, nonanalytic test. But if I am in the competitive business of manufacturing and selling parachutes, then I should subject it to a searching, analytic test, designed to tell me as much as possible about the locus and cause of any failure in order that I may improve my product and gain a larger share of the parachute market. My contention is that the last situation is the one that is most analogous to that facing the theoretical scientist. He is not accepting or rejecting a finished theory; he is in the long-term business of constructing better versions of the theory. Progress depends upon improvement or providing superior alternatives, and improvement will ordinarily depend upon knowing just how good the model is and exactly where it seems to need alteration. The large D_i's designate the next point of attack in the continuing project of refining the existing model. Therefore attention should be focused upon the various discrepancies between prediction and outcome instead of on the over-all adequacy of the model.

In view of our long-term strategy of improving our theories, our statistical tactics can be greatly improved by shifting emphasis away from over-all hypothesis testing in the direction of statistical estimation. This always holds true when we are concerned with the actual size of one or more differences rather than simply in the existence of differences. For example, in the second instance of hypothesis testing cited at the beginning of this paper, where the investigator tests a pre-experimental difference, he would do better to obtain 95% or 99% confidence interval for the pre-experimental difference. If the interval is small and includes zero, he (and any other moderately sophisticated person) knows immediately that he is on fairly safe ground; but if the interval is large, even though it includes zero, it is immediately apparent that the situation is more uncertain. In both instances H_0 would have been accepted.

Testing a Revised H_0

Before turning to estimation procedures that are useful in examining the correspondence between experimental outcomes and predictions from a mathematical model, I shall digress briefly to outline a statistical testing method which can legitimately be used in appraising the fit of a model to data as shown in Figure 1. Basically the statistical argument in the proper test is reoriented so that rejection of H_0 constitutes evidence favoring the theory. The new H_0 is that the correlation between the predicted values, Y'_i, and the obtained values, $\bar{Y}_i$, is zero, after all correlation due to the fitting process has been eliminated. The alternative, H_1, against which H_0 is tested is that there is a correlation greater than zero between theoretical and empirical points. The four simple steps required to obtain the necessary F test are as follows:

1. Calculate $t_i = Y'_i - \hat{Y}$ for all i. $\Sigma t_i = 0$.[2]
2. Calculate $SS_{Correspondence} = n(\Sigma t_i \hat{Y}_i)^2/\Sigma t_i^2$, where n is the number of observations upon which each $\hat{Y}_i$ is based. Negative values of $(\Sigma t_i \hat{Y}_i)$ are treated as zero.
3. Obtain $MS_{Correspondence} = SS_{Correspondence}/n_T$, where n_T, the number of degrees of freedom involved in fitting the theoretical points to the empirical data, will ordinarily be the number of linearly independent fitting constants in the mathematical expression of the model.
4. Divide $MS_{Correspondence}$ by $MS_{Dev\ Est}$ to give $F_{Correspondence}$ which has n_T degrees of freedom for its numerator and $k(n-1)$ degrees of freedom for its denominator, k being the number of $\hat{Y}_i$. The test is one-tailed in the sense that negative values of $\Sigma t_i \hat{Y}_i$ are treated as zero values, so that the probability values of the F distribution must be halved, an unusual procedure with F tests in analysis of variance.

Following the above procedure, rejection of H_0 now means that there is more than random positive covariation between predicted and obtained values of the dependent variable.

This test is admirable in that it puts the burden of proof on the investigator, because a small-scale, insensitive experiment is unlikely to produce evidence favoring the model. Furthermore, if the model has any merit, the more sensitive the experiment, the more likely it is that a significant F, favoring the theory, will be obtained. Actually, the test is extremely sensitive to virtue in the theory, and therefore in the case of a moderately successful model and a moderately sensitive experiment both this F and the one testing the significance of systematic deviations from the model $(F = MS_{Dev\ Theory}/MS_{Dev\ Est})$ will tend to be significant. This outcome is no anomaly; it merely indicates that the model predicts some but not all of the systematic variation in the data. In short, progress is being made, but improvement is possible. The fact that simultaneous significance of both Fs, indicating general success and specific failures of a model, should be a commonplace points up the necessity of turning to methods of statistical estimation for a more adequate examination of the workings of a theoretical model.

Practical Estimation Methods for Investigation of Models

As is true of statistical tests, each method of statistical estimation has its advantages and limitations. In the investigation of the adequacy of theoretical curves in psychology there are reasons to believe that the simpler estimation methods have practical advantages over some of the more elegant procedures. To give a fairly complete view of the situation, methods of point and interval estimation of σ_D^2 and of the individual D_i will be described, and a brief evaluation of each method will be given.

Estimating σ_D^2. The variance of the discrepancies between the Y'_i and the $\hat{Y}_i$ condenses into a single number the adequacy of fit of the theoretical model. As such it is an excellent index for the evaluation of the model. The smaller the variance, σ_D^2, the better the model, and vice versa. As an estimate of the size of the discrepancies one might expect in future similar applications of the model, σ_D^2 is far more informative than any F test. Furthermore σ_D^2 is readily estimated in the case of homogeneity of the error variance, σ_e^2. The expected values of the relevant mean squares are as follows:

[2] In the unusual event where the general mean of the observations, $\bar{Y}$, is not used as a fitting constant for $Y' = f(X)$, the t_i must be computed as deviations from the mean of all the Y'_i, $\bar{Y}'$. The test will then be insensitive to discrepancies between $\bar{Y}'$ and $\bar{Y}$, and the interpretation will be somewhat equivocal. A separate test of H_0 that $\bar{Y}_{population}$ equals $\bar{Y}'$, is feasible, but here the experimenter is forced into the illicit posture of seeking to embrace H_0.

$$\text{Exp}(MS_{Dev\ Theory}) = \sigma_e^2 + n\sigma_D^2 \quad [4]$$

$$\text{Exp}(MS_{Dev\ Est}) = \sigma_e^2 \quad [5]$$

A maximum likelihood estimate of the variance of the discrepancies, $\hat{\sigma}_D^2$ is then:

$$\hat{\sigma}_D^2 = (MS_{Dev\ Theory} - MS_{Dev\ Est})/n \quad [6]$$

The accuracy of this estimator depends upon the number of degrees of freedom associated with $SS_{Dev\ Theory}$ and $SS_{Dev\ Est}$. The latter rarely poses any practical problem, but the former, in view of the predilection of psychologists for minimizing the number of data points, is quite critical. This is readily apparent when interval estimation of σ_D^2 is attempted.

Bross (1950) gives a convenient method for accurate approximation of the fiducial interval for σ_D^2, and in this case the fiducial and confidence intervals are essentially equal. The method will be outlined below for the 5% interval.

1. Obtain $\hat{\sigma}_D^2$ from Equation 6, above. (If the estimate is negative or zero, meaningful limits cannot be obtained.)
2. Find:

$$L = \frac{\dfrac{F}{F_{025(k-n_T,\ k[n-1])}} - 1}{\dfrac{F \cdot F_{025(k-n_T,\ \infty)}}{F_{025(k-n_T,\ k[n-1])}} - 1}$$

where:

$$F = MS_{Dev\ Theory}/MS_{Dev\ Est}$$

$F_{025(k-n_T,\ k[n-1])}$ is the entry in the 2.5% F table (Pearson & Hartley, 1954) for $n_1 = k - n_T$ and $n_2 = k[n-1]$; and $F_{025(k-n_T,\ \infty)}$ is the entry for $n_1 = k - n_T$ and $n_2 = \infty$.

3. Find:

$$\hat{L} = \frac{F \cdot F_{025(k[n-1],\ k-n_T)} - 1}{\dfrac{F \cdot F_{025(k[n-1],\ k-n_T)}}{F_{025(\infty,\ k-n_T)}} - 1}$$

where $F_{025(k[n-1],\ k-n_T)}$ is the entry in the 2.5% F table for $n_1 = k[n-1]$, and $n_2 = k - n_T$; $F_{025(\infty,\ k-n_T)}$ is the entry for $n_1 = \infty$, and $n_2 = k - n_T$.

4. The upper and lower limits are then $\hat{L}\hat{\sigma}_D^2$ and $L\hat{\sigma}_D^2$, respectively. With less than 15–20 data points these limits will be found to be uncomfortably wide, a fact to bear in mind when designing an experimental test of a theoretical model. For example, in Figure 1, with 6 data points and two degrees of freedom for curve fitting, the limits might plausibly be 0–40, whereas with 14 data points the limits might be 0–12.

Aside from the considerable variability in the estimate of σ_D^2 which can be reduced by increasing the number of data points, there are two other important limitations to the use of estimates of the variance of the discrepancies in evaluating a model. First of all, the population value of σ_D^2 is completely dependent upon the particular values of the independent variable, X, which are chosen for the test of the model. Choice of two different sets of Xs could well lead to two entirely different values of σ_D^2, and both of these values could be perfectly accurate. Secondly, although σ_D^2 gives an over-all index of the adequacy of the model being tested, it condenses so much information into one measure that it does not permit pinpointing the especially large D_i's so that they can be given proper attention in considering revision of the model.

Estimating the D_i. The individual D_i may be estimated as points, or intervals may be established for the D_i, collectively or individually. As before, each method has its good points and its limitations.

Point estimation of the individual D_i's consists simply in comparing the individual data points, the $\bar{Y}_i$, with the fitted curve. It is a crude method, but it has served well in the past and represents the beginning of wisdom. For example, in Figure 1, the model builder might well note that the first three data points lie below the curve and ask himself if there is some special reason for this. He would also note that the greatest discrepancy occurs at $\bar{Y}_5$, where the neighboring discrepancies are in the other direction. The weakness of this simple method lies in the absence of a criterion which will assist the investigator in deciding which discrepancies should be singled out for further attention and which may be disregarded because they are within the range of expected random variation. This defect is remedied by the interval estimation techniques.

Probably the ideal method of interval estimation is that in which intervals are established for the whole curve in one operation by finding the 95% confidence band. The method takes the theoretical curve as a point of departure, and the result is a pair of curves above and below the theoretical curve, which will tend in the case of random variation to contain between them 95% of the data points. Points lying outside the band are immediately suspect; they are the most promising candidates for attention in the next version of the model. There are two practical difficulties with this method. First, homogeneity of the error variance, σ_e^2, over all the X_i is required. And secondly, because errors in estimation of each fitting parameter must be taken into account, for all but the simplest curves (Cornell, 1956, pp. 184–186) the bands may be difficult[3] to obtain. Although the method is elegant, in practice it will rarely represent sufficient improvement over the final method, given below, to justify its use.

The last method seems to me to be the most useful and most robust and most flexible method. It can be widely applied, and the relative ease of application, coupled with its ability to discriminate between significant and random discrepancies make it superior to the other estimation methods. It also possesses the homely virtue of being readily understood. In contrast to the preceding method, this one takes as its point of departure the empirical means, and consists, simply, in computing the 95% confidence limits for each of the $\bar{Y}_i$. If there is homogeneity of variance, the error variance of each mean is taken simply as $\hat{\sigma}_e^2/n$; in cases of suspected heterogeneity, each mean must have its own estimate of error variance. This will, of course, be the variance of the distribution of Y_{ij} for each i, divided by n. When these limits have been obtained, attention is directed to instances where the theoretical curve lies outside the limits. In some cases, the investigator might choose to establish the 80% or 90% limits in order to direct his attention to less drastic departures of the experimental results from the model. Choice of an optimum level for the limits is hard to establish on a general a priori basis, but it is likely that limits narrower than the traditional 95% will be found more useful than the broader limits. Simple as this method is, it is hard to improve upon in actual practice. Instead of giving an almost meaningless over-all acceptance or rejection of a model, it directs attention to specific defects, its functioning improves as the precision of the experimental test is improved, and the investigator can set the confidence coefficient so as to increase its sensitivity to defect at a cost of a fairly well-specified percentage of false positives or wild goose chases. A final and often crucial advantage is that the confidence intervals, based as they are upon the experimental means, can be obtained in cases where the form of the theoretical function does not permit satisfactory estimation of its parameters, and the analysis of variance and confidence bands methods cannot properly be applied.

[3] A sufficient estimate of the error variance of each parameter must be available and independent of the estimates of all other parameters or else the covariances of all parametric estimates must be found and the theoretical function must have continuous first partial derivatives with respect to the parameters in order that the confidence bands may be found in the asymptotic case (Rao, 1952, pp. 207–208). Where an asymptote is involved in the fitting of the theoretical function, satisfactory independent estimators can rarely be obtained.

SUMMARY AND CONCLUSIONS

In this paper I have attempted to show that the traditional procedure of testing a null hypothesis (H_0) of a zero difference or set of zero differences is quite appropriate to the experimenter's intentions or scientific strategy when he is unable to predict differences of a specified size. When theory or other circumstances permit the prediction of differences of specified size, using these predictions as the values in H_0 is tactically inappropriate, frustrating and self-defeating. This is particularly true when a theoretical curve has been predicted, and H_0 is framed in terms of zero discrepancies from the curve. If rejection of H_0 is interpreted as evidence against the theory, and "acceptance" of H_0 is interpreted as evidence favoring the theory, we find that the larger and more sensitive the experiment is, the more likely it will lead to results opposed to the theory; whereas the smaller and less sensitive the experiment, the more likely the results will favor the theory. Aside from this anomaly, which can be corrected by recasting H_0 in terms of a zero covariance between theoretical prediction and experimental outcome, hypothesis testing as a statistical tactic in this case implies an acceptance-inspection strategy. Acceptance-inspection properly involves examination of finished products with a view to accepting them if they are good enough and rejecting them if they are shoddy enough. The theoretician is not a purchaser but rather he is a producer of goods in a competitive market so that his examination of his theory should be from the standpoint of quality control. His idealized intentions are to detect and correct defects, if possible, so that he can produce a more adequate, more general theoretical model. Because his ideal strategy is not to prove or disprove a theory but rather to seek a better theory, his appropriate statistical tactics should be those involving estimation rather than hypothesis testing.

Examination of alternative techniques, available for point or interval estimation of discrepancies between theoretical predictions and experimental outcomes or the over-all variance of these discrepancies suggests strongly that estimation of the confidence intervals for the means found along a theoretical curve is the most practical and most widely applicable general procedure. Other writers have recently emphasized the values of various estimation as opposed to hypothesis testing techniques (e.g., Bolles & Messick, 1958; Gaito, 1958; Savage, 1957) and it is hoped that considerations pointed out by them and points raised in this paper will be helpful to investigators who are in the process of examining theoretical models which lead to specific numerical predictions of experimental outcomes.

REFERENCES

BOLLES, R., & MESSICK, S. Statistical utility in experimental inference. *Psychol. Rep.*, 1958, **4**, 223–227.

BROSS, I. Fiducial intervals for variance components. *Biometrics*, 1950, **6**, 136–144.

BUSH, R. R., ABELSON, R. P., & HYMAN, R. *Mathematics for psychologists: Examples and problems.* New York: Social Science Research Council, 1956.

BUSH, R. R., & ESTES, W. K. (Eds.) *Studies in mathematical learning theory.* Stanford, Calif.: Stanford Univer. Press, 1959.

CORNELL, F. G. *The essentials of educational statistics.* New York: Wiley, 1956.

GAITO, J. The Bolles-Messick coefficient of utility. *Psychol. Rep.*, 1958, **4**, 595–598.

GOLDBERG, S. *Introduction to difference equations.* New York: Wiley, 1958.

GRANT, E. L. *Statistical quality control.* New York: McGraw-Hill, 1952.

KEMENY, J. G., SNELL, J. L., & THOMPSON, G. L. *Introduction to finite mathematics.* Englewood Cliffs, N. J.: Prentice-Hall, 1957.

PEARSON, E. S., & HARTLEY, H. O. (Eds.) *Biometrika tables for statisticians.* Cambridge, England: Cambridge Univer. Press, 1954.

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.

(Received November 10, 1960)