PSYCHOLOGICAL BULLETIN
Vol. 53, No. 2, 1956

THE PROBLEM OF INFERENCE FROM CURVES BASED ON GROUP DATA

W. K. ESTES
Indiana University

Papers by Sidman (8), Hayes (4), and Merrill (6) have raised serious questions about the validity of inferences from curves of functional relationship based on averaged data. By means of mathematical arguments and numerical illustrations, these writers have shown convincingly that " . . . given a mean curve, the form of the individual curves is not uniquely specified" (8, p. 268). This demonstration strikes close to home for the learning theorist. In the study of learning, we are interested in describing behavioral changes in individuals, but owing to limited control over behavioral variability must frequently depend upon averages for groups of organisms to determine functional relationships. In many areas we could scarcely remain in business if it were actually true that " . . . the mean curve does not provide the information necessary to make statements concerning the function for the individual" (8, p. 268). Unfortunately it is true. More accurately, it is true *if* we regard the mean curve solely as a source of inductive generalizations. This qualification suggests that possibly the fault lies, not in the averaged curves, but in our customary interpretations of them.

It is noteworthy that learning theory, even quantitative learning theory, has made rather steady progress in spite of the widespread acceptance of a false methodological assumption. Apparently inferences from averaged curves, although not necessarily correct, must in fact often be so. This being the case, researchers in learning are unlikely to give up readily the habit of computing mean curves of functional relationship. My purpose in this note is to show that we need not feel obliged to try. The group curve will remain one of our most useful devices both for summarizing information and for theoretical analysis provided only that it is handled with a modicum of tact and understanding.

The principal point to be made is that the valid treatment of averaged curves depends upon the same principles of statistical inference that have become faimilar to all of us in such cases as the analysis of variance and the chi-square test. Just as any mean score for a group of organisms could have arisen from sampling any of an infinite variety of populations of scores, so also could any given mean curve have arisen from any of an infinite variety of populations of individual curves. Therefore no "inductive" inference from mean curve to individual curve is possible, and the uncritical use of mean curves even for such purposes as determining the effect of an experimental treatment upon rate of learning or rate of extinction is attended by considerable risk. These considerations set rather severe limitations upon the use of mean curves in the study of learning. Nonetheless we can anticipate that, as so regularly turns out to be the case in scientific research, our virtue in accepting these limitations will not go unrewarded. The same type of theoretical inquiry that has led to recognition of the need for caution in handling averaged data may be turned in a constructive direction and lead to more effective exploitation of the one defensible and important theoretical application that remains for the averaged curve—the testing of exact hypotheses about individual functions.

The first step in this direction is to recognize that the effects of averaging are not in any way capricious or unpredictable and need not be regarded as artifacts or distortions. Distortion arises only if unwarranted inferences are drawn from the mean curves. But given any specified assumption about the form of individual functions, we can proceed to deduce the characteristics to be expected of an averaged curve and then to test these predictions against obtained data. As in any problem of statistical inference, it will always be true that other assumptions might yield the same predictions. The task undertaken will be, however, to test, not the infinity of possible hypotheses, but only the one hypothesis under consideration.

In testing quantitative theories against averaged data we may be concerned either (*a*) with the form of a functional relationship or (*b*) with parameter values for the population of organisms sampled. Case *a* is illustrated by the formerly popular pastime of trying to determine "the form of the learning curve" or by the attempts to verify Hull's hypothesis that habit strength is an exponential function of number of reinforcements (**5**). Case *b* is illustrated by attempts to determine whether the slope parameter of the habit growth curve depends upon amount of reinforcement (**11**) or whether the rate and asymptote of maze learning are functions of stimulus variability (**9**).

In studies involving Case *a*, it has been customary to operate on the tacit assumption that the form of a mean curve will reflect faithfully the form of the individual curves. Since this assumption is now recognized to be unwarranted, we can no longer expect averaged data to yield any direct answer to the question, "What is the form of the individual function?" We can, however, replace this question with one which can be answered, namely, "Is the form of the mean empirical curve in accord with the assumption that the individual functions are of a given form, say $y=f(x, a, b, \cdots)$?" (In the remainder of the discussion we shall represent by f the function relating a dependent variable y to an independent variable x and parameters a, b, etc.) It becomes a specific mathematical or statistical research problem to determine for any given function f what testable predictions can be made concerning the mean curve for a group of organisms. Some preliminary considerations that may be helpful in dealing with this type of problem will be discussed below.

In studies involving Case *b* the assumption has frequently been made that if the function obtained for the individual organism is $y=f(x, a, b, \cdots)$, then the function describing the mean curve for a group of organisms should be $y=f(x, \bar{a}, \bar{b}, \cdots)$, i.e., a curve of the same form with parameters equal to the means of the corresponding individual parameters. Since the assumption is not generally true, the treatment of this case will require, first, recognizing the instances in which the assumption holds, and, second, investigating instances in which it does not hold in order to determine what information about parameter values is obtainable from the mean curve.

Classification of Functions

Relative to these problems, the

mathematical functions that we will have occasion to deal with can be classified into three types, each calling for somewhat different treatment. Let us consider briefly the problems that will arise in dealing with each of these types and illustrate some of the procedures that will prove useful in dealing with them.

Class A. *Functions unmodified by averaging.* In these cases the mean curve for the group has the form of the individual function and the parameters of the mean curve are simply the means of the corresponding individual parameters. The chief problem here is that of defining the class of functions so that we will recognize instances of it. The essential characteristics of the class will be apparent from consideration of a few examples:

1. $y=a+bx$
2. $y=a+bx+cx^2$
3. $y=a \log x$
4. $y=a \sin x+b \cos x$
5. $y=a/x$.

A numerical illustration involving one of these examples will show in a concrete way how the averaging process works out for this type of function. Suppose that we have two organisms whose behavior in a learning situation is described by the function $y=a \log x$, where a is a constant which varies in value from one organism to another, but remains fixed in value throughout learning for any one organism. Let y_1 and y_2 be response measures for the two organisms, and let the value of a be 1 for the first organism and 2 for the second. Then the course of learning for the two organisms will be described by the equations

$$y_1=\log x$$

and

$$y_2=2 \log x, \text{ respectively.}$$

Now we compute the "empirical" response measures for each organism for the first four values of the independent variable x as indicated in Table 1. Then by averaging the two response measures at each value of x, we obtain the mean "empirical" curve represented by the values in the column headed $\bar{y}$. It is clear, however, that the column of mean values also represents the values of the function $\bar{y}=1.5 \log x$. Therefore the function describing the mean curve is of the same form as the individual functions, and the parameter of the function describing the mean curve is the mean of the individual parameters.

TABLE 1

EFFECT OF AVERAGING A SIMPLE LOGARITHMIC FUNCTION

x	$\log x$	y_1	y_2	$\bar{y}$	$1.5 \log x$
1	.00	.00	.00	.00	.00
2	.30	.30	.60	.45	.45
3	.48	.48	.96	.72	.72
4	.60	.60	1.20	.90	.90

All functions belonging to this class work out similarly.[1] Stated in the simplest terms, what they all have in common is that each parameter in the function appears either alone or as a coefficient multiplying a quantity which depends only on the independent variable x. In averaging, any quantity of the latter sort factors out at each value of x and appears in the mean curve, multiplying the mean value of the parameter.

Class B. *Functions for which averaging complicates the interpretation of parameters but leaves form unchanged.* Examples of functions falling in this class[2] are

[1] See Mathematical Note 1.

1. $y = \log bx$

2. $y = \dfrac{1}{a} + \dfrac{b}{ax}$.

In the first example, we can rewrite the function in the form

$$y = \log b + \log x;$$

then it is apparent that the mean curve for a group of organisms which differ with respect to parameter b will be logarithmic in form, for the same reasons discussed in the preceding section, but will have the mean value of log b rather than log $\bar{b}$ as the intercept constant. Thus from a mean empirical curve, we can obtain an estimate of the geometric mean of the parameter b for the organisms sampled, but no estimate of the arithmetic mean of b.

In the second example, the mean curve of y vs. $1/x$ will be linear, but the parameters of the mean curve will be the mean values of $1/a$ and b/a for the organisms sampled, so no estimate of $\bar{a}$ or $\bar{b}$ can be obtained from the averaged data.

The testing of hypotheses involving functions in this class raises no difficulties if we are interested only in the form of the function; if we wish to estimate mean parameter values or to test hypotheses involving changes in parameter values as a function of experimental treatments, then care must be taken to allow for the effects of averaging.

Class C. *Functions modified in form by averaging.* A function will fall in this class[3] if it contains any terms involving the independent variable x which will not factor out when we sum values of y over a group of organisms for a constant value of x. The most familiar example of a function belonging to this class is the "growth" curve

$$y = a + be^{-cx}$$

encountered in some guise or other in many learning theories, and given detailed discussion in Sidman's paper (8).

In some cases, a function belonging to this class can be moved into Class B or even Class A by means of an appropriate transformation. Take, for example, the exponential function given above. If the value of the parameter a is known for all individuals, it can be subtracted from the response measure y, leaving us with the simpler equation

$$y' = y - a = be^{-cx}.$$

The latter can be made more tractable by the logarithmic transformation

$$\log y' = \log b - cx$$

which when averaged yields

$$E(\log y') = E(\log b) - \bar{c}x,$$

where $E()$ represents the mean, or expected, value of the term in parentheses. If, then, we take logarithms (base e) of the dependent variable y' and plot the transformed variable as a function of x, both the curve for any individual and the averaged curve for a group will be linear; from the mean curve we can obtain estimates of the mean value of the parameter c and of the geometric mean of the parameter b. By means of this strategem the problem of testing the hypothesis that an exponential function holds for individual organisms has been reduced to the very simple problem of determining whether the mean curve plotted from the transformed data departs significantly from linearity. Similarly, other hypotheses that might be tested against the group data are greatly simplified. Suppose, for example, that a theoretical curve of extinction took the form of this exponential

[2] See Mathematical Note 2.

[3] See Mathematical Note 3.

function, with y being a response measure, x number of trials, and the asymptote a equal to zero, and that we were interested in the question whether some difference in the experimental treatments given two groups of organisms influenced rate of extinction; by means of the suggested transformation, this problem would reduce to that of testing for a difference in slope between two regression lines. A variety of transformations which may be useful in situations of this sort have been discussed by Mueller (7).

Even when functions in Class C cannot be moved into one of the more docile classes by any available transformation, or when for some reason transformation of the data is undesirable (as might be the case if a contemplated transformation produced heterogeneity of variances along the curve), we are not necessarily helpless. The extent to which functional form is modified by averaging will generally depend upon the dispersion of parameter values in the group of organisms sampled; thus in some cases it may be possible by studying individual curves to estimate the dispersion of parameter values in the group and determine whether the form of the mean curve can be expected to conform closely to the form of the individual functions; see, e.g., (3). Further, even in the case of the most refractory functions, it will usually be possible by appropriate mathematical analysis to derive the main characteristics that should be predicted for an averaged curve; an analysis of this sort for a "growth" function has been described in a recent paper (2).

The Role of Experimental Error

The analysis given here might be objected to on the grounds that we have considered only the effects of averaging upon data obtained from idealized organisms which behave strictly in accordance with theoretical functions. Response measures obtained from real organisms may, on the other hand, be influenced by various sources of experimental error as well as by the variables taken account of in a given theory. The objection is pertinent, but not fatal. The answer is that in testing a theoretical prediction one must make some explicit assumption about the role of experimental error in the test situation. And as in any statistical test, the validity of the conclusions will be conditional upon the degree to which such assumptions are satisfied. In some instances, it may be reasonable to assume that the contribution of experimental error is negligible; then the analyses given above will apply without modification. Frequently it will be more reasonable to operate under the assumption, routinely made in working with analysis-of-variance models, that error combines additively with treatment effects to determine the observed response measures. In this case, if we wish to test the hypothesis that a function $y=f(x, a, b, \cdots)$ holds for individuals, we will assume that the observed response measure Y for any individual is equal to the sum of y and a random variable e which represents the contribution of experimental error, i.e.,

$$Y=y+e=f(x, a, b, \cdots)+e.$$

Now if the error variable e is independent of x, and if the function f falls in our Class A, averaging of individual curves will yield a mean curve described by the function

$$\bar{Y}=\bar{y}+\bar{e}=f(x, \bar{a}, \bar{b}, \cdots)+\bar{e}.$$

If the mean value of e is zero, which will, for example, be the case whenever the distribution of errors is normal, then the form of the mean curve will be unaffected by the error term; if the mean is not zero, then the mean function will be modified only by the addition of a constant and the plotted mean curve will be changed only by a vertical displacement. In some cases the error variable may interact with experimental variables. If the nature of the interaction can be stated explicitly, then its effects upon the averaging process can be determined by appropriate analysis. In situations where error variables and experimental variables interact in complex or unknown ways, exact tests of quantitative hypotheses will generally be impossible.

Summary

These comments are not meant to provide an exhaustive treatment of the problem of averaging. The one point I have tried to bring out clearly is that the valid interpretation of group curves[4] depends on the principles common to all problems of statistical inference. Although the form of a group mean curve does not determine the forms of the individual curves, it does provide a means of testing exact hypotheses about them. In each particular case, the procedure must be to state explicitly the hypothesis under test, and then to derive the properties that should hold for the averaged curve if the hypothesis is correct. If the predictions thus derived are in accord with data, the hypothesis remains tenable; if they are not, then the hypothesis can be rejected at some specified level of confidence. Utilized within this framework, the averaged curve can be expected to remain one of the most valuable techniques for the analysis of behavioral data, and in fact to increase progressively in value as mathematical and statistical research continue to enlarge our repertory of special devices for the handling of particular problems.

[4] Throughout this discussion we have spoken in terms of mean curves obtained from groups of organisms. Similar problems arise, and similar considerations apply, however, in the case of a curve whose points represent means of repeated measures on the same organism. Parameter values associated with an individual organism may vary either systematically or randomly during the course of an experiment. In either case, we may think of each possible combination of parameter values as determining a hypothetical curve, this population of curves being sampled at each value of the independent variable. Whether the obtained mean curve should be expected to have the same form as the hypothetical individual curves will depend on the nature of the mathematical function describing the latter and on the role of experimental error, just as in the case of a group curve.

Mathematical Notes

1. A more formal criterion for class inclusion is desirable for some purposes, and may be formulated as follows.[5] Let us consider a function $y=f(x, a, b, \cdots)$. At any given value of x, we may regard y as a function of the parameters a, b, etc., and expand the function in a Taylor's series around the mean values of the parameters (6, 10), obtaining the relation

$$y=f(x, \bar{a}, \bar{b}, \cdots)+(\Delta a)f_a+(\Delta b)f_b + \cdots + \frac{(\Delta a)^2}{2}f_a{}^2+\cdots$$

where $\bar{a}+\Delta a$ is the value of the a parameter for a given organism; $f_a{}^i$

[5] A criterion proposed by Bakan (1), which involves expanding the function in a Maclaurin series around the point $x=0$, is not entirely satisfactory. For one thing it is frequently inapplicable. Take, for example, the functions $y=a \log x$ or $y=x^a$; in neither case are the derivatives all continuous at $x=0$, so in neither case will the series generally represent the function. The criterion suggested in the present paper will hold for all functions which can be expanded by Taylor's theorem, a class which includes all the elementary functions and, in fact, all explicit functions that the psychologist is apt to have dealings with.

represents the ith derivative of y with respect to a, evaluated at $a=\bar{a}$; and so on. When the function is averaged over a group of individuals, we obtain

$$\bar{y}=f(x, \bar{a}, \bar{b}, \cdots)+\tfrac{1}{2}\sigma_a^2 f_a^2 + \tfrac{1}{2}\sigma_b^2 f_b^2 + \cdots.$$

Our criterion for inclusion of a function in Class A may now be stated: if in the Taylor's series development, all second and higher order partial derivatives of the function with respect to parameters are zero, then the function is unmodified by averaging. Applying the criterion to $y=a \log x$, we have $f_a=\log x$; $f_a^2=0$; and therefore $\bar{y}=\bar{a}\log x$, in agreement with the conclusion reached above by a more informal route.

2. A sufficient criterion for inclusion of a function $y=f(x, a, b, \cdots)$ in Class B is that it does not satisfy the criterion of Class A when expanded around $\bar{a}$, $\bar{b}$, etc., but does satisfy that criterion when rewritten $y=f(x, u, v, \cdots)$ and expanded around $\bar{u}$, $\bar{v}$, etc. (u, v, etc. being functions of the parameters a, b, $\cdots$). In the first example under Class B above, this criterion is satisfied if we let $\log b=u$; in the second example, it is satisfied if we let $1/a=u$ and $b/a=v$.

3. If a function falls in Class C, then in the Taylor's series developments described above, some of the second or higher order derivatives will depend on x regardless of how u, v, etc. are chosen, and thus the criteria for Class A or Class B cannot be satisfied.

It will be noted that these formal criteria provide more rigorous definitions of the various classes than can be given in nonmathematical terms. However, it should be emphasized that the conclusions about inference from averaged curves that we have reached in this paper do not depend on abstruse mathematical analyses. In many practical situations, questions concerning the effects of averaging can be handled by simple numerical methods of the type illustrated in an earlier section.

REFERENCES

1. Bakan, D. A generalization of Sidman's results on group and individual functions and a criterion. *Psychol. Bull.*, 1954, **51**, 63–64.
2. Estes, W. K., & Burke, C. J. A theory of stimulus variability in learning. *Psychol. Rev.*, 1953, **60**, 276–286.
3. Estes, W. K., & Straughan, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *J. exp. Psychol.*, 1954, **47**, 225–234.
4. Hayes, K. J. The backward curve: a method for the study of learning. *Psychol. Rev.*, 1953, **60**, 269–275.
5. Hull, C. L. *Principles of behavior.* New York: Appleton-Century-Crofts, 1943.
6. Merrill, Margaret. The relationship of individual growth to average growth. *Hum. Biol.*, 1931, **3**, 37–70.
7. Mueller, C. G. Numerical transformations in the analysis of experimental data. *Psychol. Bull.*, 1949, **46**, 198–223.
8. Sidman, M. A note on functional relations obtained from group data. *Psychol. Bull.*, 1952, **49**, 263–269.
9. Wolfle, D. L. The relative efficiency of constant and varied stimulation. III. The objective extent of stimulus variation. *J. comp. Psychol.*, 1936, **22**, 375–381.
10. Woods, F. S. *Advanced calculus.* New York: Ginn, 1934.
11. Zeaman, D. Response latency as a function of the amount of reinforcement. *J. exp. Psychol.*, 1949, **39**, 466–483.

Received February 7, 1955.