E. Nagel, P. Suppes, & A. Tarski (Eds.), Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress. Stanford: Stanford University Press, 1962. pp. 252-261.

The Stanford Institute for Mathematical Studies in the Social Sciences

APPLIED MATHEMATICS AND STATISTICS LABORATORIES
STANFORD UNIVERSITY

Reprint No. 57

Models of Data

PATRICK SUPPES

Reprinted from
Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress
1962

MODELS OF DATA

PATRICK SUPPES

Stanford University, Stanford, California, U.S.A.

1. Introduction

To nearly all the members of this Congress the logical notion of a model of a theory is too familiar to need detailed review here. Roughly speaking, a model of a theory may be defined as a possible realization in which all valid sentences of the theory are satisfied, and a possible realization of the theory is an entity of the appropriate set-theoretical structure. For instance, we may characterize a possible realization of the mathematical theory of groups as an ordered couple whose first member is a non-empty set and whose second member is a binary operation on this set. A possible realization of the theory of groups is a model of the theory if the axioms of the theory are satisfied in the realization, for in this case (as well as in many others) the valid sentences of the theory are defined as those sentences which are logical consequences of the axioms. To provide complete mathematical flexibility I shall speak of theories axiomatized within general set theory by defining an appropriate set-theoretical predicate (e.g., "is a group") rather than of theories axiomatized directly within first-order logic as a formal language. For the purposes of this paper, this difference is not critical. In the set-theoretical case it is convenient sometimes to speak of the appropriate predicate's being satisfied by a possible realization. But whichever sense of formalization is used, essentially the same logical notion of model applies.[1]

It is my opinion that this notion of model is the fundamental one for the empirical sciences as well as mathematics. To assert this is not to deny a place for variant uses of the word "model" by empirical scientists, as, for example, when a physicist talks about a physical model, or a psychologist refers to a quantitative theory of behavior as a mathematical model. On this occasion I do not want to argue for this fundamental character of the logical notion of model, for I tried to make out a detailed case at a colloquium in Utrecht last January, also sponsored by the International Union of History and Philosophy of Science [3]. Perhaps the most persuasive argument which might be singled out for mention here is that the notion of model used in any serious statistical treatment of a theory and its relation to experiment does not differ in any essential way from the logical notion of model.

The focus of the present paper is closely connected to the statistical

---

The writing of this paper was partially supported by the Group Psychology Branch of the Office of Naval Research.

[1]For a detailed discussion of axiomatization of theories within set theory, see Suppes [2, Chap. 12].

Reprinted from LOGIC, METHODOLOGY AND PHILOSOPHY OF SCIENCE: PROCEEDINGS OF THE 1960 INTERNATIONAL CONGRESS edited by E. Nagel, P. Suppes, and A. Tarski,

analysis of the empirical adequacies of theories. What I want to try to show is that exact analysis of the relation between empirical theories and relevant data calls for a hierarchy of models of different logical type. Generally speaking, in pure mathematics the comparison of models involves comparison of two models of the same logical type, as in the assertion of representation theorems. A radically different situation often obtains in the comparison of theory and experiment. Theoretical notions are used in the theory which have no direct observable analogue in the experimental data. In addition, it is common for models of a theory to contain continuous functions or infinite sequences although the confirming data are highly discrete and finitistic in character.

Perhaps I may adequately describe the kind of ideas in which I am interested in the following way. Corresponding to possible realizations of the theory I introduce possible realizations of the data. Models of the data of an experiment are then defined in the customary manner in terms of possible realizations of the data. As should be apparent, from a logical standpoint possible realizations of data are defined in just the same way as possible realizations of the theory being tested by the experiment from which the data come. The precise definition of models of the data for any given experiment requires that there be a theory of the data in the sense of the experimental procedure, as well as in the ordinary sense of the empirical theory of the phenomena being studied.

Before analyzing some of the consequences and problems of this viewpoint, it may be useful to give the ideas more definiteness by considering an example.

2. Example from learning theory

I have deliberately chosen an example from learning theory because it is conceptually simple, mathematically non-trivial and thoroughly probabilistic. More particularly, I consider linear response theory as developed by Estes and myself [1]. To simplify the presentation of the theory in an inessential way, let us assume that on every trial the organism in the experimental situation can make exactly one of two responses, A_1 or A_2, and after each response it receives a reinforcement, E_1 or E_2, of one of the two possible responses. A possible experimental outcome in the sense of the theory is an infinite sequence of ordered pairs, where the n^{th} term of the sequence represents the observed response — the first member of the pair — and the actual reinforcement — the second member of the pair — on trial n of the experiment.

A possible realization of the theory is an ordered triple $\mathfrak{X} = < X, P, \theta >$ of the following sort. The set X is the set of all sequences of ordered pairs such that the first member of each pair is an element of some set A and the second member an element of some set B, where A and B each have two elements. The set A represents the two possible responses and the set B the two possible reinforcements. The function P is a probability measure

on the smallest Borel field containing the field of cylinder sets of X; and θ, a real number in the interval $0 < \theta \leqq 1$, is the learning parameter. (Admittedly, for theories whose models have a rather complicated set-theoretical structure the definition of possible realization is at points arbitrary, but this is not an issue which affects in any way the development of ideas central to this paper.)

There are two obvious respects in which a possible realization of the theory cannot be a possible realization of experimental data. The first is that no actual experiment can include an infinite number of discrete trials. The second is that the parameter θ is not directly observable and is not part of the recorded data.

To pursue further relations between theory and experiment, it is necessary to state the axioms of the theory, i.e., to define models of the theory. For this purpose a certain amount of notation is needed. Let $A_{i,n}$ be the event of response A_i on trial n, $E_{j,n}$ the event of reinforcement E_j on trial n, where $i, j = 1,2$, and for x in X let x_n be the equivalence class of all sequences in X which are identical with x through trial n. A possible realization of the linear response theory is then a model of the theory if the following two axioms are satisfied in the realization:

Axiom 1. *If* $P(E_{i,n}A_{i',n}x_{n-1}) > 0$, *then*
$$P(A_{i,n+1} \mid E_{i,n}A_{i',n}x_{n-1}) = (1 - \theta)\, P(A_{i,n} \mid x_{n-1}) + \theta.$$

Axiom 2. *If* $P(E_{j,n}A_{i',n}x_{n-1}) > 0$ *and* $i \neq j$, *then*
$$P(A_{i,n+1} \mid E_{j,n}A_{i',n}x_{n-1}) = (1 - \theta)\, P(A_{i,n} \mid x_{n-1}).$$

The first axiom asserts that when a response is reinforced, the probability of making that response on the next trial is increased by a simple linear transformation. The second axiom asserts that when a different response is reinforced, the probability of making the response is decreased by a second linear transformation. To those who are concerned about the psychological basis of this theory it may be remarked that it is derivable from a much more complicated theory that assumes processes of stimulus sampling and conditioning. The linear response theory is the limiting case of the stimulus sampling theory as the number of stimuli approaches infinity.

For still greater definiteness it will be expedient to consider a particular class of experiments to which the linear response theory has been applied, namely, those experiments with simple contingent reinforcement schedules. On every trial, if an A_1 response is made, the probability of an E_1 reinforcement is π_1, independent of the trial number and other preceding events. If an A_2 response is made, the probability of an E_2 reinforcement is π_2. Thus, in summary for every n,

$$P(E_{1,n} \mid A_{1,n}) = \pi_1 = 1 - P(E_{2,n} \mid A_{1,n}),$$
$$P(E_{2,n} \mid A_{2,n}) = \pi_2 = 1 - P(E_{1,n} \mid A_{2,n}).$$

This characterization of simple contingent reinforcement schedules has been made in the language of the theory, as is necessary in order to compute

theoretical predictions. This is not possible for the finer details of the experiment. Let us suppose the experimenter decides on 600 trials for each subject. A brief description (cf. [4, pp. 81–83]) of the experimental apparatus might run as follows.

> The subject sits at a table of standard height. Mounted vertically in front of the subject is a large opaque panel. Two silent operating keys (A_1 and A_2 responses) are mounted at the base of the panel 20 cm. apart. Three milk-glass panel lights are mounted on the panel. One of these lights, which serves as the signal for the subject to respond, is centered between the keys at the subject's eye level. Each of the other two lights, the reinforcing events E_1 and E_2, is mounted directly above one of the keys. On all trials the signal light is on for 3.5 sec.; the time between successive signal exposures is 10 sec. A reinforcing light comes on 1.5 sec. after the cessation of the signal light and remains on for 2 sec.

It is not surprising that this description of the apparatus is not incorporated in any direct way at all into the theory. The important point is to take the linear response theory and this description as two extremes between which a hierarchy of theories and their models is to be fitted in a detailed analysis.

In the class of experiments we are considering, the experimenter records only the response made and reinforcement given on each trial. This suggests the definition of the possible realizations of the theory that is the first step down from the abstract level of the linear response theory itself. This theory I shall call the *theory of the experiment*, which term must not be taken to refer to what statisticians call the theory of experimental design—a topic to be mentioned later. A possible realization of the theory of the experiment is an ordered couple $\mathscr{Y} = < Y, P >$, where (i) Y is a finite set consisting of all possible finite sequences of length 600 with, as previously, the terms of the sequences being ordered pairs, the first member of each pair being drawn from some pair set A and correspondingly for the second members, and (ii) the function P is a probability measure on the set of all subsets of Y.

A possible realization $\mathscr{Y} = < Y, P >$ of the theory of the experiment is a model of the theory if the probability measure P satisfies the defining condition for a simple contingent reinforcement schedule. Models of the experiment thus defined are entities still far removed from the actual data. The finite sequences that are elements of Y may indeed be used to represent any possible experimental outcome, but in an experiment with, say, 40 subjects, the observed 40 sequences are an insignificant part of the 4^{600} sequences in Y. Consequently, a model closer to the actual situation is needed to represent the actual conditional relative frequencies of reinforcement used.

The appropriate realization for this purpose seems to be an N-tuple Z of elements from Y, where N is the number of subjects in the experiment. An N-tuple rather than a subset of Y is selected for two reasons. The first is that if a subset is selected there is no direct way of indicating that two

distinct subjects had exactly the same sequence of responses and reinforcements — admittedly a highly improbable event. The second and more important reason is that the N-tuple may be used to represent the time sequence in which subjects were run, a point of some concern in considering certain detailed questions of experimental design. It may be noted that in using an N-tuple as a realization of the data rather than a more complicated entity that could be used to express the actual times at which subjects were run in the experiment, we have taken yet another step of abstraction and simplification away from the bewilderingly complex complete experimental phenomena.[2]

The next question is, When is a possible realization of the data a model of the data? The complete answer, as I see it, requires a detailed statistical theory of goodness of fit. Roughly speaking, an N-tuple realization is a model of the data if the conditional relative frequencies of E_1 and E_2 reinforcements fit closely enough the probability measure P of the model of the experiment. To examine in detail statistical tests for this goodness of fit would be inappropriate here, but it will be instructive of the complexities of the issues involved to outline some of the main considerations. The first thing to note is that no single simple goodness of fit test will guarantee that a possible realization Z of the data is an adequate model of the data. The kinds of problems that arise are these: (i) (*Homogeneity*) Are the conditional relative frequencies (C.R.F.) of reinforcements approximately π_i or $1-\pi_i$, as the case may be, for each subject? To answer this we must compare members of the N-tuple Z. (ii) (*Stationarity*) Are the C.R.F. of reinforcements constant over trials? To answer this practically we sum over subjects, i.e., over members of Z, to obtain sufficient data for a test. (iii) (*Order*) Are the C.R.F. of reinforcements independent of preceding reinforcements and responses? To answer this we need to show that the C.R.F. define a zero order process — that serial correlations of all order are zero. Note, of course, that the zero order is with respect to the conditional events E_i given A_j, for $i, j = 1,2$. These three questions are by no means exhaustive; they do reflect central considerations. To indicate their essentially formal character it may be helpful to sketch their formulation in a relatively classical statistical framework. Roughly speaking, the approach is as follows. For each possible realization Z of the data we define a statistic $T(Z)$ for each question. This statistic is a random variable with a probability distribution — preferably a distribution that is (asymptotically) independent of the actual C.R.F. under the null hypothesis that Z is a model of the data. In statistical terminology, we "accept" the null hypothesis if the obtained value of the statistic $T(Z)$ has a probability equal to or greater than some significance level α on the assumption that indeed the null hypothesis is true.

---

[2]The exact character of a model $\mathscr{M}$ of the experiment and a model Z of the data is not determined uniquely by the experiment. It would be possible, for instance, to define $\mathscr{M}$ in terms of N-tuples.

For the questions of homogeneity, stationarity, and order stated above, maximum likelihood or chi-square statistics would be appropriate. There is not adequate space to discuss details, but these statistics are standard in the literature. For the purposes of this paper it is not important that some subjectivists like L. J. Savage might be critical of the unfettered use of such classical tests. A more pertinent caveat is that joint satisfaction of three statistical tests (by "satisfaction" I mean acceptance of the null hypothesis with a level of significance $\geqq .05$) corresponding to the three questions does not intuitively seem completely sufficient for a possible realization Z to be a model of the data.[3] No claim for completeness was made in listing these three, but it might also be queried as to what realistic possibility there is of drawing up a finite list of statistical tests which may be regarded as jointly sufficient for Z to be a model of the data. A skeptical non-formalistic experimenter might claim that given any usable set of tests he could produce a conditional reinforcement schedule that would satisfy the tests and yet be intuitively unsatisfactory. For example, suppose the statistical tests for order were constructed to look at no more than fourth-order effects, the skeptical experimenter could then construct a possible realization Z with a non-random fifth-order pattern. Actually the procedure used in well-constructed experiments makes such a dodge rather difficult. The practice is to obtain the C.R.F. from some published table of random numbers whose properties have been thoroughly investigated by a wide battery of statistical tests. From the systematic methodological standpoint it is not important that the experimenter himself perform the tests on Z.

On the other hand, in the experimental literature relevant to this example it is actually the case that greater care needs to be taken to guarantee that a possible realization Z of the data is indeed a model of the data for the experiment at hand. A typical instance is the practice of restricted randomi-

---

[3]For use at this point, a more explicit definition of models of the data would run as follows. Z is an N-fold model of the data for experiment $\mathscr{Y}$ if and only if there is a set Y and a probability measure P on subsets of Y such that $\mathscr{Y} = <Y, P>$ is a model of the theory of the experiment, Z is an N-tuple of elements of Y, and Z satisfies the statistical tests of homogeneity, stationarity, and order. A fully formal definition would spell out the statistical tests in exact mathematical detail. For example, a chi-square test of homogeneity for E_1 reinforcements following A_1 responses would be formulated as follows. Let N_j be the number of A_1 responses (excluding the last trial) for subject j, i.e., as recorded in Z_j — the j^{th} member of the N-tuple Z, and let ν_j be the number of E_1 reinforcements following A_1 responses for subject j. Then

$$\chi^2_H(Z) = \sum_{j=1}^{N} \frac{(\nu_j - N_j\pi_1)^2}{N_j\pi_1} + \frac{(N_j - \nu_j - N_j(1-\pi_1))^2}{N_j(1-\pi_j)}$$

$$= \sum_{j=1}^{N} \frac{(\nu_j - N_j\pi_1)^2}{N_j\pi_1(1-\pi_1)},$$

and this χ^2 has N degrees of freedom. If the value $\chi^2_H(Z)$ has probability greater than .05 the null hypothesis is accepted, i.e., with respect to homogeneity Z is satisfactory.

zation. To illustrate, if $P(E_{1,n} \mid A_{1,n}) = .6$, then some experimenters would arrange that in every block of 10 A_1 responses exactly 6 are followed by E_1 reinforcements, a result that should have a probability of approximately zero for a large number of trials.[4]

The most important objection of the skeptical experimenter to the importance of models of the data has not yet been examined. The objection is that the precise analysis of these models includes only a small portion of the many problems of experimental design. For example, by most canons of experimental design the assignment of A_1 to the left (or to the right) for every subject would be a mistake. More generally, the use of an experimental room in which there was considerably more light on the left side of subjects than on the right would be considered mistaken. There is a difference, however, in these two examples. The assignment of A_1 to the left or right for each subject is information that can easily be incorporated into models of the data — and requirements of randomization can be stated. Detailed information about the distribution of physical parameters characterizing the experimental environment is not a simple matter to incorporate in models of data and is usually not reported in the literature; roughly speaking, some general *ceteris paribus* conditions are assumed to hold.

The characterization of models of data is not really determined, however, by relevant information about experimental design which can easily be formalized. In one sense there is scarcely any limit to information of this kind; it can range from phases of the moon to I.Q. data on subjects.

The central idea, corresponding well, I think, to a rough but generally clear distinction made by experimenters and statisticians, is to restrict models of the data to those aspects of the experiment which have a parametric analogue in the theory. A model of the data is designed to incorporate all the information about the experiment which can be used in statistical tests of the adequacy of the theory. The point I want to make is not as simple or as easily made precise as I could wish. Table 1 is meant to indicate a possible hierarchy of theories, models, and problems that arise at each level to harass the scientist. At the lowest level I have placed *ceteris paribus* conditions. Here is placed every intuitive consideration of experimental design that involves no formal statistics. Control of loud noises, bad odors, wrong times of day or season go here. At the next level formal problems of experimental design enter, but of the sort that far exceed the limits of the particular theory being tested. Randomization of A_1 as the left or right response is a problem for this level, as is random assignment of subjects to different experimental groups. All the considerations that enter

[4]To emphasize that conceptually there is nothing special about this particular example chosen from learning theory, it is pertinent to remark that much more elaborate analyses of sources of experimental error are customary in complicated physical experiments. In the literature of learning theory it is as yet uncommon to report the kind of statistical tests described above which play a role analogous to the physicists' summary of experimental errors.

TABLE 1

HIERARCHY OF THEORIES, MODELS, AND PROBLEMS

Theory of	Typical Problems
Linear response models	Estimation of θ, goodness of fit to models of data
Models of experiment	Number of trials, choice of experimental parameters
Models of data	Homogeneity, stationarity, fit of experimental parameters
Experimental design	Left-right randomization, assignment of subjects
Ceteris paribus conditions	Noises, lighting, odors, phases of the moon

at this level can be formalized, and their relation to models of the data, which are at the next level, can be made explicit — in contrast to the seemingly endless number of unstated *ceteris paribus* conditions.

At the next level, models of the experiment enter. They bear the relation to models of the data already outlined. Finally at the top of the hierarchy are the linear response models, relatively far removed from the concrete experimental experience. It is to be noted that linear response models are related directly to models of the data, without explicit consideration of models of the experiment. Also worth emphasizing once again is that the criteria for deciding if a possible realization of the data is a model of the data in no way depend upon its relation to a linear response model. These criteria are to determine if the experiment was well run, not to decide if the linear response theory has merit.

The dependence is actually the other way round. Given a model of the data we ask if there is a linear response model to which it bears a satisfactory goodness of fit relation. The rationale of a maximum likelihood estimate of θ is easily stated in this context: given the experimental parameters π_1 and π_2 we seek that linear response model, i.e., the linear response model with learning parameter $\hat{\theta}$, which will maximize the probability of the observed data, as given in the model of the data.

It is necessary at this point to break off rather sharply discussion of this example from learning theory, but there is one central point that has not been sufficiently mentioned. The analysis of the relation between theory and experiment must proceed at every level of the hierarchy shown in Table 1. Difficulties encountered at all but the top level reflect weaknesses in the experiment, not in the fundamental learning theory. It is unfortunate that it is not possible to give here citations from the experimental literature of badly conceived or poorly executed experiments that are taken to invalidate the theory they presume to test, but in fact do not.

3. The theory of models in the empirical sciences

I began by saying that I wanted to try to show that exact analysis of the relation between empirical theories and relevant data calls for a hierarchy of models of different logical type. The examination of the example from learning theory was meant to exhibit some aspects of this hierarchy. I would like to conclude with some more general remarks that are partially suggested by this example.

One point of concern on my part has been to show that in moving from the level of theory to the level of experiment we do not need to abandon formal methods of analysis. From a conceptual standpoint the distinction between pure and applied mathematics is spurious — both deal with set-theoretical entities, and the same is true of theory and experiment.

It is a fundamental contribution of modern mathematical statistics to have recognized the explicit need of a model in analyzing the significance of experimental data. It is a paradox of scientific method that the branches of empirical science that have the least substantial theoretical developments often have the most sophisticated methods of evaluating evidence. In such highly empirical branches of science a large hierarchy of models is not necessary, for the theory being tested is not a theory with a genuine logical structure but a collection of heuristic ideas. The only models needed are something like the models of the experiment and models of the data discussed in connection with the example from learning theory.

Present statistical methodology is less adequate when a genuine theory is at stake. The hierarchy of models outlined in our example corresponds in a very rough way to statisticians' concepts of a sample space, a population, and a sample. It is my own opinion that the explicit and exact use of the logical concept of model will turn out to be a highly useful device in clarifying the theory of experimental design, which many statisticians still think of as an "art" rather than a "science." Limitations of space have prevented working out the formal relations between the theory of experimental design and the theory of models of the data, as I conceive it.

However, my ambitions for the theory of models in the empirical sciences are not entirely such practical ones. One of the besetting sins of philosophers of science is to overly simplify the structure of science. Philosophers who write about the representation of scientific theories as logical calculi then go on to say that a theory is given empirical meaning by providing interpretations or coordinating definitions for some of the primitive or defined terms of the calculus. What I have attempted to argue is that a whole hierarchy of models stands between the model of the basic theory and the complete experimental experience. Moreover, for each level of the hierarchy there is a theory in its own right. Theory at one level is given empirical meaning by making formal connections with theory at a lower level. Statistical or logical investigation of the relations between theories at these different levels can proceed in a purely formal, set-theoretical manner. The

more explicit the analysis the less place there is for non-formal considerations. Once the empirical data are put in canonical form (at the level of models of data in Table 1), every question of systematic evaluation that arises is a formal one. It is important to notice that the questions to be answered are formal but not mathematical — not mathematical in the sense that their answers do not in general follow from the axioms of set theory (or some other standard framework for mathematics). It is precisely the fundamental problem of scientific method to state the principles of scientific methodology that are to be used to answer these questions — questions of measurement, of goodness of fit, of parameter estimation, of identifiability, and the like. The principles needed are entirely formal in character in the sense that they have as their subject matter set-theoretical models and their comparison. Indeed, the line of argument I have tried to follow in this paper leads to the conclusion that the only systematic results possible in the theory of scientific methodology are purely formal, but a general defense of this conclusion cannot be made here.

References

[1] Estes, W. K., and P. Suppes. Foundations of linear models. Chap. 8 in *Studies in Mathematical Learning Theory*, R. R. Bush and W. K. Estes, eds., Stanford, Calif., Stanford Univ. Press, 1959, 432 pp.

[2] Suppes, P. *Introduction to logic*, Princeton, Van Nostrand, 1957, 312 pp.

[3] Suppes, P. A comparison of the meaning and uses of models in mathematics and the empirical sciences, *Synthese*, Vol. 12 (1960), 287–301.

[4] Suppes, P., and R. C. Atkinson. *Markov Learning Models for Multiperson Interactions*, Stanford, Calif., Stanford Univ. Press, 1960, xii+296 pp.