PROBABILITY
AND
MATHEMATICAL STATISTICS
Vol. 15 (1995), pp. 21–27

STOCHASTIC MODELS AND THEIR APPLICATIONS TO SOCIAL PHENOMENA

BY

WILLIAM KRUSKAL AND JERZY NEYMAN

Presented at the joint session of the Institute of Mathematical Statistics, American Statistical Association and American Sociological Society, Detroit, September 1956

The imposing sweep of this session's title — *Critical Problems in New Quantitative Techniques*, the high reputations of the discussants and the other speakers, together with sponsorship by three scholarly associations, combine to fill us with humility. Would then we were able to present to you in a neatly organized form a map of the battle lines, quantitative techniques sector, in the struggle between social scientists and obdurate social nature.

But this is clearly impossible, even if time and ability permitted. In the first place, the struggle is extensive and complex, as is inevitable and even desirable in a vigorous fight. In the second place, lines of communication are spotty: there are many media of communication and, to some degree, each patrol speaks its own special patois. (Among certain units, it is even considered a mark of distinction and scholarship to possess such a special patois.) It is simply hopeless for a statistician, who has after all other fish to fry as well, to become and to keep well informed of all quantitative developments in the social sciences. Even to keep informed about one area, say that of latent structure, is a difficult task, somewhat facilitated perhaps if one has friendly spies in such bastions as the Rand Corporation and the Columbia Bureau for Applied Social Research.

Still it behoves statisticians to be as aware as possible of quantitative developments in all the sciences, and to react to these developments with critical sympathy.

Our remarks will be divided into two parts — about ten minutes on each. The first part will discuss a mode of description for scientific models that helps, we think, to organize and interrelate the very numerous models proposed and used as interpretations of nature. The second part will consider briefly the

question of the identification of models; that is, the question of telling two models apart. This question will, we suspect, become increasingly important in the social sciences and may act as a considerable stimulus to progress.

Let us turn first, then, to the description of models, particularly models for social phenomena, as to the degree of which they are, on the one hand, *interpolatory devices*, and as to the degree to which they are, on the other hand, genuine *explanatory theories*.

By *theory* or *model building* we mean the formulation of a set of reasonable assumptions regarding the underlying mechanism of the phenomena studied. From these assumptions, conclusions – often expressed as formulas – are deduced describing those aspects of the phenomena that are observable. Agreement between the deduced conclusions and appropriate observations then becomes an important criterion for judging the adequacy of the theory. How good agreement should be, and how this goodness may be efficiently judged, is an area in which substantive scientist and statistician should, in principle at least, cooperate fruitfully. An example of a theory of the kind we have in mind is Newtonian mechanics. This of course is a nonstochastic example; a stochastic example would be any of the theories of learning discussed by Mosteller, Estes, and others, although this statement is a relative one, since we doubt that any social scientist would claim that there exist in the social sciences any theories comparable to Newtonian mechanics.

An interpolatory device, by contrast, consists of the selection of a relatively *ad hoc* family of functions, not deduced from underlying assumptions, and indexed by a set of parameters. Given a specific set of observations, one of these functions is then chosen, by some process called *fitting the parameters*, so that a close agreement is reached between the function and the observations. Theories, as well as interpolatory devices, usually also have free parameters; the essential distinction is that in the case of theories the parameters flow directly from basic assumptions that are mechanistic (in a wide sense), while, in the case of interpolatory devices, the parameters appear merely as a convenient index to the members of the *ad hoc* family of functions.

As one example, a nonstochastic one, of an interpolatory device, we suggest the method used by Kepler to find the elliptical paths of the planets. In contrast, Newton's *theory* of gravitation, starting from far more basic considerations, *implies* the elliptical character of the planetary paths. This classical relationship contributes considerably to the intellectual and aesthetic satisfaction of Newton's theory. It also suggests its great utility as a predictive device for related phenomena. As stochastic examples of interpolatory devices, parallel to present-day learning theory roughly as Kepler's work is parallel to Newtonian theory, we point to the many investigations in which so-called learning curves are "fitted" by a member of some parametric family of functions, chosen only because its members have the right qualitative features and are computationally malleable.

Although the distinction between true theory and interpolation is sometimes quite sharp in specific cases, the two modes of analysis really represent somewhat extreme points of a continuous spectrum. Theory construction must always contain some interpolatory elements in its selection of assumptions. For example, Newton's assumption that the force of gravity is inversely proportional to the *square* of distance (rather than to its cube, say, or to some other function of it) is of an interpolatory nature.

We do not wish to deprecate interpolatory procedures, for these are essential, especially in the earlier exploratory stages of an investigation. In fact, the traditional bread-and-butter topics of statistics — for example, analysis of variance — are, at least as commonly applied, of an interpolatory nature. But we urge the importance of not remaining satisfied with the essentially descriptive successes of interpolatory procedures, and of going ahead to the more satisfying and — in the long run — more useful construction of explanatory models. Needless to say, explanatory models should be based firmly on the wide empirical knowledge of sophisticated scientists in order to avoid that armchair vapidity of which we have been warned many times.

In summary, social phenomena result from the varying preferences and actions of the individuals forming society. An important mode of understanding such phenomena is the creation of an explanatory model or theory — often stochastic — for the behavior of individuals, such that deductions from the theoretical assumptions are in concordance with properly conducted statistical investigations of the phenomena in question.

The continuum we have described is by no means precise, nor is it meant to be so. Still it is amusing, and possibly productive of fruitful argument, to make up a fanciful scale and to try to place on it various models. Let us draw a horizontal line to represent such a scale and place at its left end a zero point corresponding, perhaps, as a thorough-going interpolatory device, to traditional least-squares fitting of a line to a swarm of observed points that cry out — to children of our culture — to be fitted by a straight line.

As a reasonable example of a *non*interpolatory theory let us locate at "10" on the right the sort of theory discussed by Estes, Mosteller, and Bush for learning in simple *T*-maze situations.

Latent structure analysis, as it now stands, we would place at about "5", partly because of the multiplicity of free parameters its formulas entail, and partly because of *ad hoc* elements in the specification of the so-called trace lines. Perhaps the father of latent structure analysis will take exception to our position for it.

Factor analysis, in most of its forms of which we are aware, we would put a bit lower on the scale, perhaps at "4". Classical economics is hard to assign, because it is largely nonstochastic, dealing with averages over large human populations. It does, nonetheless, deal with definite mechanisms, frequently only qualitatively defined. Perhaps "9" would be a fair position. In discussing

this scale with friends, wide differences of opinion have appeared, in part of a semantic nature, depending on what one means by the "*xyz* theory." For example, a theory such as that of latent structure could be regarded as of a purely curve-fitting type or as including extensive sociological structure depending upon the meaning one gives to the term "latent structure theory."

We find this game of recreational interest, since it entails so many personal evaluations. However, perhaps it should best be continued under more recreational circumstances.

I turn now to the second part of our talk, to the question of the identification of a model. As we shall see, this question is really only of a serious nature when the model in question is a true theory, resting on conceptions and assumptions about empirical individuals, however much these conceptions and assumptions may be swathed in abstract terminology.

Let us describe the point graphically. We choose first a set of observable quantities that our theory might explain. Suppose that every possible joint probability distribution of these observable quantities is represented by a point on the blackboard. So each point corresponds to a different, and possibly quite complex, statement describing fully the distribution of the observables.

Now consider a specific theory with its basic concepts and assumptions. If it is not trivial, the theory will lead to only *some* of the points; that is, only some of the joint distributions, usually relatively few, will be consistent with the theoretical assumptions. For convenience, represent these distributions consistent with the theory as the points enclosed by this curve. If the theory is wholly specific, there will be only one point within the curve. In so far as the theory has free parameters, the curve will enclose a multiplicity of points, but almost always a small fraction of the totality of points.

Let us consider an example that may have intrinsic interest. Back in the twenties, some prominent English statisticians began an intensive study of accident statistics. The observables might, for example, be the number of nondisabling accidents suffered by London bus drivers in a two-year period. Some drivers had been involved in no accidents, others in one, others in two, and so on. Thus to each driver studied would be attached the number of his accidents, and this was regarded as an observation on an underlying distribution of interest. If the drivers were chosen at random, and if they exercised negligible effect on each other's accident frequencies, then the observations might be regarded as a random sample from a specific, but unknown, univariate probability distribution over the integers 0, 1, 2, etc. There are infinitely many such distributions; they correspond to the totality of points on our diagram.

The English statisticians now wanted to introduce theoretical structural assumptions so that they could come to sensible grips with this situation. The simplest assumption would be that for each driver there is a *constant* small risk, each second of the working day, that an accident may occur, *and* that this

timewise constant risk is the same for each driver. From this assumption it is readily derived that the observations follow some Poisson distribution, thus reducing the many possible distributions on the nonnegative integers to a relative few, the Poisson distributions. But this conclusion was quickly found to be hopeless at variance with the observed facts. The numbers simply did not look anything like a Poisson distribution; they were much more scattered.

What to do? One argument ran somewhat as follows: people are different, and bus drivers are people. Suppose that to each bus driver there corresponds his *own* risk, constant over time *but* now variable from person to person. As distributions of risks over persons, the family of Pearson Type III functions might be, and was, chosen on the grounds of manipulative simplicity and reasonable qualitative shape. Note the interplay between explanatory theory and interpolatory convenience. Applying this variable risk theory, the family of resulting distributions for the observables is that family called the *negative binomial* distributions. This family includes as limiting cases the Poisson distributions, but it is much more extensive: it is a two- not a one-parameter family. Thus the variable risk theory supposes that the observations are governed by a kind of *mixture* of Poisson distributions. In this the theory bears a resemblance to Lazarsfeld's latent structure theory.

The variable risk theory turned out in fact to be nicely consistent with the observations. So all seemed well. In fact, the concept of accident-proneness (corresponding to a high risk level for an individual) became popular and personnel managers tried hard not to hire accident-prone people. We understand that some psychoanalysts have constructed far-reaching conclusions on the basis of the accident-proneness concept. But now the question of identification of model arose. It was pointed out that another underlying concept and another set of assumptions of a wholly different kind would lead to precisely the *same* set of distributions for the observables.

This second approach was first suggested, as far as we know, by Pólya and Eggenberger. It is simple and assumes that all the individuals are governed by the *same* law. Suppose that they all start out with no accidents and with the same constant risk of an accident λ. Any particular individual goes along with this constant risk until his *first* accident occurs. Then he proceeds at a new level of risk, $\lambda+\alpha$, where α is positive. When his second accident occurs, his risk level jumps to $\lambda+2\alpha$. And so on..., after k accidents and before $k+1$ accidents the risk level is $\lambda+k\alpha$. This model is called that of *positive linear contagion*; it means that the risk of accident increases with number of accidents. The important point is that the family of distributions for the observables under the positive linear contagion model is precisely the same as that under the variable risk or mixed Poisson model. That is, both theories give rise to exactly the same two-parameter family of negative binomial distributions for the observables. Hence even if we knew perfectly the distribution of the observables, we would not be able to discriminate between the two theories, relative, of course,

to the given data; thus nonidentifiability. And remember that the two theories correspond to two very different conceptual mechanisms.

Of course, there may well be more than two theories leading to the same family of distributions for the observables.

Further, more complex things may happen. For example, the families of distributions for the observables under two theories may not be the same but one may be included in the other, or there may be genuine overlapping. Or it may be that the two families are disjoint, so that identification is in principle possible, but so close together that resolution would require a wholly impractical sample size; this might be called *practical nonidentifiability.*

What then? The presence of nonidentifiability forces us to try to differentiate between the theories by considering other and wider classes of observables. And this is in the direct interest of scientific advance. In the accident case, for example, current work is exploring such observables as the time intervals between accidents and the numbers of accidents suffered by the same person in two time periods or in two environments. Of course, considerations of elegance and simplicity — Occam's razor — may also play a role, a consideration some writers adduce when comparing the Ptolemaic and Copernican theories of planetary motion.

Let us comment briefly, in closing, on the theory called a *latent structure.* This theory supposes a mixture of multivariate populations, each with independent marginals. Often other conditions are added. Perhaps the simplest interesting case is that in which a questionnaire has four dichotomous questions, so that there are $2^4 = 16$ patterns of answers, or 15 "degrees of freedom" for specification of the distribution of the observables with respect to a given population. In appropriate circumstances, the latent structure theory says that the population is really made up of two groups, in proportions ν and $1-\nu$, such that among the individuals of group 1 the four questions are answered independently, and among the individuals of group 2 the four questions are also answered independently *but* with different probabilities. Thus the general 15-dimensional "space of distributions" is reduced to a 9-dimensional space corresponding to the following 9 parameters: ν (the mixing parameter), four probabilities of "yes" for group 1, and four probabilities of "yes" for group 2. This is the so-called *latent dichotomy* with 4 questions discussed by Lazarsfeld and his colleagues. The two groups are considered to be sociologically important.

We raise the following question, without attempting to answer it: Might it not be possible to begin with some *different* reasonable conceptual model and show that it implies the same or an overlapping family of distributions for the observables? That is, may another model exist, as in the case of the accident statistics example, that will lead to the same family of probability distributions as does the latent structure model, or at least to a family that overlaps to a large extent? We suspect that the answer may be "yes."

It might be noted here that work on medical problems with the help of models much like those of latent structure was initiated by Muench in 1936 and has been extended by Neyman and Chiang. The major differences between these medical models and those of Lasarsfeld and his colleagues lie in the natures of the restrictions on conditional probabilities of positive responses. Chiang in a paper in Human Biology (1951) has discussed the question of "practical nonidentifiability" of models in the context of tuberculosis diagnosis. He suggests a number of different models and explores the sample sizes necessary to discriminate between them. They turn out to be, in his situation, generally quite large.

In summary, it seems important to us that the model builder and tester consider the possible existence of alternate conceptual structures equally well explaining his observables. For then, in attempting to discriminate, he will be naturally led to wider observational areas that may prove to be of considerable importance.

Received on 24.5.1994