Statistical Science
1992, Vol. 7, No. 2, 183–207

The Impact of Sociological Methodology on Statistical Methodology

Clifford C. Clogg

Abstract. Developments in sociological methodology and in quantitative sociology have always been closely related to developments in statistical theory, methodology and computation. The same statement applies if "methodology for social research" and "quantitative social research" replace the more specific terms in this statement. Statistical methodology, including especially the battery of methods used to estimate and evaluate statistical models, has had a tremendous effect on social research in the post-war period, particularly in the United States. What is less well appreciated is the influence of sociological methodology, or methodology for social research more generally, on modern statistics. I give a brief sketch of the linkages between methodology in social research and methodology in statistics. The focus is on areas where developments in sociological methodology, or at least the *scientific contexts* of social research, have brought forth new methods of general significance to the practice of statistics, in both theoretical and "applied" areas. These remarks should be taken as the impressions of someone who has tried to straddle the fence between statistics and social research throughout his career, not as a careful history of statistical ideas.

Key words and phrases: Social survey, latent variable, log-linear model, latent structure, latent trait, covariance structure, event history data, causal inference.

The science of statistics is essentially a branch of Applied Mathematics, and may be regarded as mathematics applied to observational data. (p. 1)

Statistical methods are essential to social studies, and it is principally by the aid of such methods that these studies may be raised to the rank of sciences. This particular dependence of social studies upon statistical methods has led to the unfortunate misapprehension that statistics is to be regarded as a branch of economics, whereas in truth methods adequate to the treatment of economic data, in so far as these exist, have mostly been developed in biology and the other sciences. (p. 2)

Sir Ronald A. Fisher, *Statistical Methods for Research Workers* (1970)

Clifford C. Clogg is Distinguished Professor of Sociology and Professor of Statistics at Pennsylvania State University, 202 Pond Laboratory, University Park, Pennsylvania 16802. *An earlier version of this paper was presented at the 1989 annual meeting of the American Association for the Advancement of Science in a session entitled "Sociologists and Statisticians: A Sesquicentennial Partnership."*

1. INTRODUCTION

I begin by quoting from Fisher's influential text on statistical methodology because these particular passages represent points of view that are still quite prevalent today, in spite of the fact that they first appeared in the earlier editions of the text written in the 1920s. I do not think any observer of trends in sociology or social research over the last 40 years could deny the validity of Fisher's view that statistical methods are important for "social studies" (i.e., the social sciences, including economics in Fisher's mind). I hasten to add that among quantitative sociologists there is little disagreement about the role of statistical methodology as a language for scientific social research. [For dissenting views, see Duncan and Stenbeck (1988) and Freeman (1991).] It seems to me, too, that sociology and other branches of social research have indeed been raised to the status of "sciences," although I realize that the extent to which this is true is debated just about as much now as it was in Fisher's day.

Fisher's view of the process by which statistical methods have developed is the controversial point. Fisher believed that statistical methodology, which is

"the mathematics of observational data," arose primarily in response to problems in the natural sciences, biology and genetics in particular. This was a natural position for him to hold simply because his own career was devoted to solving inferential problems that arose in the analysis of data from biology or genetics, including data derived from agricultural experimentation. The role of the social sciences, even of economics, is downplayed. The developments that we now recognize as the foundation of the modern discipline of statistics, Fisher would have us believe, arose in the scientific contexts of biology and other natural sciences and were brought forth by statisticians with close ties to those areas.

In what follows I shall refer to statistical methodology rather than statistics, partly because I believe that what Fisher defined as statistics is what we would today call statistical methodology. By statistical methodology I mean procedures that actually find some use in the analysis of data, of some kind or another, or have direct bearing on procedures that are used for this purpose. A time lag is allowed. Statistical methodology need not be used right away to qualify as such, but if it does not find some use or change the way that inferences are obtained from some real data set, in a decade or so after it first appears, then we are probably not talking about statistical methodology but rather statistical theory, or statistical types of mathematics (not the applied kind), or something else altogether. A great many of the papers or results given in contemporary statistics journals do not actually have much effect on the analysis of data, besides sharpening the wits of some of those who do that job on occasion. Serious statistical methodology almost always has some practical purpose, which is found sooner as opposed to later, and the standard of use or usefulness ought to be taken seriously in our field. (I have tried to apply that standard as an editor, for example.)

Recent histories by Stigler (1986) and Duncan (1984) largely invalidate Fisher's views on the origin of statistical ideas. These scholars demonstrated that we owe a great deal to social statisticians in the 19th century as well as to those who came from the biological camp. Here are a few examples drawn from these histories:

- Quetelet gave us the concept of the "average man," an important first step in separating fixed from random (or individualistic) determinants of behavior. He thus gave the framework now used for nearly all behavioral models in modern social science.
- Lexis provided statistical foundations for the study of survivorship and laid the basis for demographic models of "event histories" that are so ubiquitous in economics and sociology today. John Graunt's invention of the life table much earlier was also very important, and demographers at least think of Graunt as a founder of demography, a social science that ought not be overlooked.
- Fechner and the early psychophysicists evidently laid some of the groundwork for the modern experimental method; that method was not invented out of nothing by Fisher and his co-workers.
- Galton and Karl Pearson, scholars whom Fisher would undoubtedly place in the biological camp, contributed in fundamental ways to the study of social as well as physical inheritance. The statistical imagery as well as the regression-type models that they employed are strikingly similar to those employed in modern studies of status attainment and the intergenerational transmission of inequality.
- Edgeworth, an economist, contributed in fundamental ways to the statistical analysis of time series and to the development of modern systems of social and economic indicators. In addition, Edgeworth produced many analytical tools, such as so-called Edgeworth expansions, that are still widely used in mathematical statistics for approximation work. And Edgeworth, according to Stigler, laid the foundations for regression and correlation as applied to social science data, including the groundwork for what we now commonly call "causal" models (or structural equation models) for observational data.
- Yule clearly had social science data in mind, and used examples of social data, when he developed his theories of association. Yule understood the difference between marginal and partial association, and *collapsibility* of variables in contingency tables, nearly 90 years ago. Simpson's paradox (Simpson, 1951), which says that marginal and partial association can differ even in direction, should be called Yule's paradox, for example. It is interesting to note that Yule's framework for studying association between categorical variables leads naturally to the log-linear model developed during the 1960s and 1970s. And Yule was one of the forerunners of the kind of social science methodology that we now call evaluation research or policy analysis; his analysis of the probable effects of the Poor Laws in England represents a foray into that area, for example, using multiple regression and partial correlation for causal inference.

How Fisher could have ignored or been oblivious to the *social science* roots of modern statistics that were so clearly formed prior to 1915 is an open question that I leave for others to resolve. But throughout most of this century Fisher's view has become the dominant view, in my judgment at least. I believe that most

mathematical statisticians, and even most statistical or quantitative social scientists, have internalized this viewpoint to a considerable extent. The main purpose of this essay is to try to change this perception of the history of ideas in statistics. My main point is not purely academic, although it certainly has implications for the organization of statistics as a field or as departments within the academic setting. The debate about statistics as a mathematical discipline versus statistics as a system of methods for scientific analysis (in social, physical or other sciences) still exists. (The introductory statements from Fisher appear to reflect some aspects of both views, so the debate is hardly new.) How this debate is or has been resolved will shape statistics departments and statistical science for a long time to come.

Fisher's view was that statistical methodology of general importance arose in the hard sciences and was developed primarily by those, like himself, who had close ties to them. [This viewpoint is also consistent with my reading of Box's (1978) biography of Fisher. I do not think my statement of Fisher's view on this matter is controversial.] Statistical methods in the social sciences (Fisher calls these "social studies") largely copy those used in the natural sciences, without important amendments or modifications. Fisher's point seems to be that the main statistical methods have developed in response to scientific problems in the natural sciences, and that with time they trickle down to the social sciences. I will take this as Fisher's view, but it requires some modification to reflect the mathematization of statistics as a discipline since the Second World War. A revision that I believe might be consistent with Fisher's view were he alive today, and which I believe is consistent with views widely held among mathematical statisticians at the present time, is that generally important statistical methods in our era have arisen from two sources. The first is the pure mathematical source: good methodology arises from developments in highly mathematized areas of modern statistics, from the theorems, proofs and approximations reported at a bewildering rate, for example, in *The Annals of Statistics*. The second source is the hard-science source: good methodology also arises from solving problems in the natural sciences, biology and the experimental sciences in particular, and it emanates from the theoretical and methodological work of those statisticians with close ties to the natural sciences.

I shall attempt another revision of Fisher's views on the subject of the history of ideas, by giving what I think are convincing examples where social science methodology and/or the *context* of social research have played a major role in shaping modern statistical methodology. Because I know more about sociology or sociological methodology than I know about other social science areas, I necessarily concentrate on the impact of sociological methodology on modern statistical methodology. I believe it would be possible, however, to make many of the same claims, with as much or possibly more supporting evidence, if arguments from the vantage point of psychometrics or of econometrics were treated in an analogous fashion. I simply do not know enough about those areas to include them very much here, although I think it would be very interesting to hear from social statisticians who represent these other areas.

2. STATISTICAL MODELS

What statistical methodology refers to in most areas today is virtually synonymous with statistical modeling. A statistical model can be thought of as an equation, or set of equations, that (a) links "inputs" to "outputs" (factors to responses, exogenous variables to endogenous variables, independent variables to dependent variables, etc.), (b) have both fixed and stochastic components, (c) include either a linear or a nonlinear decomposition between the two types of components, and (d) purport to explain, summarize or predict levels of or variability in the "outputs." I consider two general classes of statistical models that are very important in contemporary social research: the *log-linear model* and the *event-history model.*

The Log-Linear Model

The log-linear model for discrete variables (or discrete dependent variables) has had a major impact on sociological methodology since 1970. Many articles developing or extending the log-linear model for "applications" in social research have appeared regularly in sociology journals such as the *American Journal of Sociology*, the *American Sociological Review* and *Sociological Methodology*, to mention just a few. Many of the key articles from these sociological or social science outlets are referenced in the main *statistical* monographs that summarize the log-linear model, including Bishop, Fienberg and Holland (1975) and Agresti (1990). One of the most popular textbooks in this area (Fienberg, 1980) utilizes special methods actually developed in the social science context that constitute nearly one-half of the work. (These include path analysis, methods for assessing collapsibility of categories, scaling models and methods for dealing with ordinal variables, among others.) The log-linear model has become a standard component of the methodological arsenal of modern statistics, and the specialty in statistics referred to as "categorical data analysis" has become particularly prominent in the last decade or so. Most statisticians today find it necessary to know something about log-linear models, logistic regression, odds ratios, partitioning chi-squared statistics in contingency tables, and so on. Every major software pack-

age for statistical analysis now includes modules or procedures for log-linear analysis. Such models and methods, including logistic regression, probably form the most-used battery of statistical techniques in contemporary applied statistics after the ordinary linear model. Practically every issue of both theoretical and applied journals in statistics features some new development in this general area. In short, the log-linear model is an integral part of both statistics and the methodology of social research. This model has certainly become a tool "essential to social studies," and it is fitting to begin with a discussion of it.

Where did the log-linear model originate? Who was responsible for developing it? What scientific contexts led to the major innovations in the methodology of categorical data analysis? A brief summary of the development of the log-linear model, at least as I understand it, is as follows:

1. Pearson (1900) develops the chi-squared statistic for testing goodness of fit and later (Pearson, 1904) applies it to test independence or homogeneity in two-way contingency tables. [Fisher (1922) corrects Pearson's mistake on degrees of freedom.] The independence *model* plays the role of the baseline or "null" model in the log-linear model for contingency tables.

2. Yule (1900), dissatisfied with the implicit continuity (and normality) assumptions in Pearson's related work on tetrachoric and polychoric correlations for contingency tables, develops the odds ratio, or cross-ratio, as well as several measures of association based on it. The odds ratio is the fundamental parameter in the log-linear model.

3. Fisher, in many papers, develops the seemingly unrelated technique of analysis-of-variance (AOV). The AOV model, including the famous AOV table of sums of squares, plays a special role in the log-linear model for contingency tables; logarithms of expected cell frequencies rather than cell means of continuous variables are decomposed using the same framework. Fisher also made extensive use of logits and cross-product ratios, and in Fisher (1935) gave maximum likelihood procedures for quantal response models.

4. Bartlett (1935) presents the model of no three-factor interaction for the three-way contingency table, sometimes called the model of constant partial association. Iterative procedures would be required, which meant that practical work with the general model as well as serious theoretical work would have to wait for the computer age.

5. Deming and Stephan (1940) present the famous algorithm that bore their name for many years, as a method of raking sampled frequencies in a contingency table from a survey to "known" marginals from a census. The generalization came to be called the iterative-proportional-fitting (IPF) method, and this algorithm, or algorithms closely related to it (see, e.g., Goodman, 1968, 1970; Fienberg, 1970) were widely used in the early stages of the development of the general log-linear model.

6. Birch (1963) reconsiders the main log-linear models for the three-way table, develops maximum likelihood theory (based on properties of exponential families), for "unsaturated" or restricted models. Birch's notation and theorems were utilized subsequently by others; see, for example, Mantel (1966).

7. Goodman (e.g., Goodman, 1964) gave procedures for simultaneous testing of interactions (logarithms of cross-product ratios) in three-way tables, which was suited for at least some inferences that would be made with the saturated or unrestricted model.

8. Goodman (1968, 1969, 1970) Mosteller (1968), Fienberg (1970), Bishop (1969), Bock (1970) and Haberman (1970, 1974a) largely completed the task. Also see Grizzle, Starmer and Koch (1969). These sources provide the complete taxonomy of models, the algorithms, examples, partitioning strategies, software, sampling theory, relation to logit models, extension to incomplete tables, and so on. During the period from 1963 to 1972 or so, the log-linear model as we know it today had come into being.

Since 1970 or so, the methodology of log-linear models has been shaped a great deal by social scientists in general and by sociologists in particular. I think that even a casual inspection of Bishop, Fienberg and Holland (1975) or Agresti (1990), especially the "chapter notes" in the latter, indicate this dramatically. To be sure, logistic regression and other log-linear models became popular in the medical literature and other areas during the same period. But were there social science roots in the log-linear model during the formative stages covered by the above summary of main developments?

The log-linear model originated in response to problems in data analysis encountered in the social sciences and in other areas, including biomedical areas. Pearson, Fisher and Birch are not noted for their contributions to the methodology of social research; however, every other scholar listed in the above account had (or has) close ties to sociology or social research. Yule was concerned with summarizing association between categorical variables that were as often as not social or sociological variables. Deming devised his famous algorithm with Stephan (a sociologist who spent most of his career at Princeton) in response to data adjustment problems in *social* surveys. Goodman, Mosteller, Fienberg and Haberman – the modern pioneers of categorical data analysis – are perhaps the best examples of statisticians with close ties to the social sciences. Fisher's statement clearly requires revision with respect to the development of the log-linear model. The methodol-

ogy of log-linear analysis, in truth, has been developed at least as much in the social science context, by statisticians and sociological methodologists, as in the biological or biostatistical context.

Surely one of the main factors that has driven social research since the Second World War is the social survey. Of course, the federal government relied increasingly on sample surveys throughout this period (not all of which could be called *social* surveys), and this was not unrelated to the growth of surveys in empirical social research. Survey methods and sampling techniques as developed in relation to census data are perhaps equally important. The growth of educational testing as an industry (ability or achievement tests are at least similar to social surveys) during this era is also an important context to consider. Survey measures of public opinion, such as election polls and marketing surveys, differ in purpose but not in execution. The social survey became the main source of data in sociology, at least in the parts of it that became statistical and/or quantitative in orientation. (Samples drawn from census data should be included as "social surveys.")

Social surveys typically provide categorical measurements (nominal, discrete-ordinal, discrete-quantitative) of the key "dependent" variables of most interest in social research. Categorical rather than continuous measures are the norm rather than the exception in the social survey. Social surveys usually provide multiple measurements of the key variables, such as sets of items measuring political ideology or poverty, and so special methods for combining multiple measurements were placed high on the agenda. The typical social survey, including both attitudinal surveys and the more "objective" surveys carried out by or for the federal government, now collects hundreds of categorical measurements on large "representative" samples (with n's anywhere from 1000 to 200,000). One of the major tasks faced in post-war social research was to develop methodology for analyzing multivariate categorical data of this sort.

The context of the social survey is very different from the scientific areas that Fisher knew best. Indeed, the entire context of social research was and is different from the context of field plots and agricultural experimentation. Controlled experiments isolating just a few "treatment" variables are the norm in the latter areas. Random assignment of treatments, to randomly chosen subjects, has seldom been carried out in social research, at least not in areas of major concern to the disciplines involved. Instead of continuous response variables (in experiments), surveys give *specified* response variables that are most often categorical. (Fisher, of course, often encountered categorical variables in genetics, but categorical variables are ubiquitous in social surveys, not quite so ubiquitous in biology or genetics.) Instead of a few specified factors of special interest, surveys give scores of *specified* predictor variables (true covariates) that researchers treat as factors. Instead of a clear-cut causal inference and an obvious partitioning of sources of variability, the survey gives ambiguous causal inferences; many believe that the ambiguity can be minimized by using panel data, other forms of longitudinal data collected in the survey format, or even retrospective information collected in a conventional cross-sectional survey. Such data give a less obvious partitioning of variability into sources. [The difference between a social survey and the classical experiment is covered in many sources. The one I like best is Kish (1987).]

Sociologists in general, and many of the statisticians who worked with sociologists or social researchers in this century, were uneasy about normality or continuity assumptions that were implicit with the standard methods. Yule's theories on association (odds ratios) were considered seriously, and in my judgment they won out over Pearson's tradition of correlation analysis and normal-theory models. The dominant sociological methodologist prior to the 1960s, Paul Lazarsfeld, developed algebraic decompositions for systems of dichotomous variables that were viewed, at the time, as alternatives to the normal-theory regression approach (Lazarsfeld, 1961). Leo Goodman can be credited with the generalization (Goodman, 1972), which was based on the log-linear model. Prior to the log-linear model, the main method of dealing with categorical variables was the method of measures of association as put forth in a series of papers by Goodman and Kruskal (see Goodman and Kruskal, 1954, 1979). The Goodman-Kruskal measures departed from Pearson's correlation approach also, but they dealt mostly with two-variable relationships and were difficult to generalize to truly multivariate settings. See Haberman (1982) for procedures that tie together (asymmetric) measures of association, including the Goodman-Kruskal tau, and logistic regression (or log-linear models).

In the early stages, the log-linear model was developed primarily by statisticians with close ties to the social sciences (Yule, Goodman, Mosteller, Haberman, Fienberg). To be sure, there were important linkages with biological or biomedical areas, so Fisher's view would hold at least partly with respect to this statistical innovation. [See Imrey, Koch, and Stokes (1981) for a rather different history of some aspects of the log-linear model.] It is probably fair to say that sociological methodologists who were not statisticians, as most readers of this journal would define the latter, had little to do with the main technical developments in the area. But since 1975 or so, several sociological methodologists have made major contributions to the methodology. For example, Duncan (1979) proposed some models for the analysis of cross-classified ordinal

variables which were acknowledged and generalized in Goodman's (1979) fundamental paper on the subject. The importance of association models for ordinal data, especially as contrasted with the method of correspondence analysis, has been substantial, both in social science areas and in statistics (see Goodman, 1991). Special models for mobility tables (or for other tables where there is a one-to-one correspondence between row categories and column categories), graph-theoretic models for path analysis, and special models for the analysis of panel data were also put forth in social research. (The reader can refer to contributions in *Sociological Methodology* since the mid-1970s to corroborate these claims.) These contributions are noted in several of the most popular monographs on the log-linear model (Fienberg, 1980; Agresti, 1984, 1990; Andersen, 1980), so it is not necessary to elaborate.

Perhaps one of the best examples of the social science role in the development of this area of methodology is the *quasi-log-linear model.* The familiar quasi-independence model for a two-way contingency table is a special case. The quasi-log-linear model also uses an AOV decomposition of log-frequencies, but the decomposition is posited to hold only for a subset of the cells in the table. The subset excluded might consist of *structural zeroes* or cells that have frequencies that are particularly large (or particularly small) for various reasons. I have given a short history of this model in Clogg (1986b). The analysis of incomplete contingency tables arose in earlier work by Pearson, who gave incorrect results. Prior to 1960, valid general methods for analyzing such contingency tables were simply unavailable. In a series of papers culminating in the 1968 Fisher Memorial Lecture (Goodman, 1968), valid methods along with algorithms that are very similar to the algorithms now used for log-linear analysis were developed. The main empirical context for this branch of the log-linear method was the analysis of social mobility, with the standard (sociological) occupational mobility table serving as the primary example to which these models were applied. Another very important special case is the *quasi-symmetry model* generally associated with Caussinus (1965).

This brief and perhaps idiosyncratic account of the log-linear model should prove three points:

1. A major component of modern statistical methodology, the log-linear model, has roots in the context of social research, in the earlier as well as the later years of this century.
2. Statisticians with close ties to the social sciences have played a major role, perhaps the most important role, in its development.
3. The flexible methodology that is now associated with the log-linear model owes a great deal to the scientific or inferential problems encountered in social research, especially in the stimulation created by the need to analyze social surveys, which are ubiquitous in social science.

The Event-History Model

The term "event-history data" was coined by Nancy Tuma, a former editor of *Sociological Methodology* (see, e.g., Tuma, Hannan and Groeneveld, 1979). A general class of models for these data can be called the event-history model, and the most complete account of it is Tuma and Hannan (1984), one of the most ambitious methodological monographs in sociology since Coleman (1964). In reviewing this work and the status of the model in social research a few years ago, I emphasized the degree to which this general model borrowed from the methodology of survival analysis in statistics and biostatistics (Clogg, 1986a). Poisson, Weibull, Gompertz, and other parametric models of "time dependence," along with partially-parametric models associated with Cox regression (proportional hazards), are certainly used extensively in the analysis of event histories. In this respect, it is undoubtedly true that the context of biology (or biostatistics) and the work of statisticians in these areas has had great effect on what we now call the event-history model. Fisher's view would seem to be appropriate with respect to this general model in some sense at least, but in my judgment this is an incomplete picture.

There are many aspects of event-history models, as these are presently defined in sociology and economics, that owe much to the contexts of social and economic research. The special features of the general model owe a great deal to statisticians with close ties to the social sciences. In biostatistics, the goal is usually to model the time until a single nonrepeatable event (such as death) occurs. In contrast, when social researchers speak of event histories they usually mean multiple types of events (e.g., several labor force states, not just dead or alive) followed through time, with one or more of the events repeatable. Event-history data gives the type of event experienced along with the time that it occurred over the course of the observation period. In contrast, in biostatistics until very recently the focus, at least in applied or methodological work, was on the analysis of the time until a single (nonrepeatable) event takes place. (Of course, we can define the *first* occurrence of an event as a nonrepeatable event, the second occurrence as the first occurrence after the first event, and so on, but this type of reduction is fairly restrictive.)

In modern event-history analysis in the social sciences, inferences are sought about the types of transitions, the dependence of the transition rates on time and on covariates, and about the influence of prior events or durations in previously occupied states. The

general event-history model in sociology or economics (Tuma and Hannan, 1984; Heckman and Walker, 1987) is related to the parametric or partially parametric survival models in biostatistics in the sense that the latter are special cases of the former. Social statisticians such as Tuma, Heckman, Hoem, Burt Singer and James Coleman have played a major role in developing this general model, at least as a branch of statistical methodology. Coleman's (1964) treatise on mathematical sociology, where stochastic models of various kinds were put forth, was an important impetus to this area in the social sciences, for example. Special computer programs associated with Tuma (RATE) and Heckman (CTM) are much more general than would have been necessary for most analyses in biomedical areas, and the portfolio of special models in the latter are especially imaginative. In my judgment, the event-history model would not have developed as it has without the social science context (the growing importance of event-history data) or without the contributions of social statisticians and econometricians.

3. LATENT VARIABLES

It is impossible to appreciate modern quantitative sociology without coming to grips with the concept of the latent variable. In its most elementary form, a latent variable is simply a variable that cannot be measured directly. Stated this way, latent variables abound in statistics: we do not "see" parameters or even distributions. But the latent-variable concept has a more specific meaning in the social sciences. Instead of observing the variable we would like to have (Y^*), we instead observe or measure a contaminated form of the variable, say $Y = Y^* + \epsilon$. Usually we try to observe or measure multiple Y's as reflections of one or more Y^*'s. Measurement error of a particular kind—*random* measurement error—can be tolerated for the dependent variables in linear models; ϵ becomes just another factor that increases the error variance. This creates some problems (loss of precision, loss of power, etc.), but we think we know how to solve those problems (e.g., take a larger sample). But measurement errors in predictor variables, even if they are random, have perverse consequences, and these are noted in statistical, psychometric and econometric literature (and lately in biostatistical literature as well). Measurement error, and hence the idea of a latent variable, has been part of statistics for decades. See Madansky (1959) for an early survey including coverage of standard psychometric methods of correcting correlations for attenuation. See Fuller's (1987) treatise for a more complete survey, many new results, and a definitive treatment of measurement error as a statistical problem. (These sources mostly deal with continuous measurements and measurement-error models suited for them. Other sources ought to be consulted for the analysis of measurement error in discrete measurements; some of this is covered briefly below.)

In technical or mathematical statistics, measurement-error models have been developed largely without recourse to the latent-variable concept. For example, in Fuller's (1987) treatise, which I admire greatly, the term occurs just once, on page 2: "the unobserved variable ... is called a latent variable in some areas of application." The common social science terms—latent-structure model, latent-class model and latent factor—appear just once each (on pp. 60, 272 and 273, respectively). In some other quarters of statistics, however, latent-variable concepts figure more prominently. Dillon and Goldstein (1984), for example, present the tools of multivariate analysis largely from a social science point of view, including summaries of *latent-variable models* that are widely used in the social sciences. Dillon and Goldstein cover factor analysis, including classical methods that are based on methods of rotation and modern methods that rely on other types of restrictions that are more natural to impose in social research, latent structure analysis, and covariance structure models—all social science products—along with the more traditional methods such as multivariate AOV, principal components, discriminant analysis, etc. The Dillon-Goldstein text has been received well as a modern text on statistical methodology [see the review by Schervish (1987)].

Just as there are discrete and continuous observable variables, latent variables can be discrete or continuous. Models for discrete latent variables in sociology and other social science areas are commonly called *latent class models* (LCM's). Models for continuous latent variables in psychology, sociology, educational testing and other areas have various names, such as *factor models*, *latent trait models* or *covariance structure models*. The general term for all such models, in the social sciences at least, is *latent structure analysis*. That term was coined by the eminent sociological methodologist, Paul Lazarsfeld (see Lazarsfeld and Henry, 1968). Models for both kinds of latent variables were developed mostly in the social sciences, and they have achieved considerable stature as general-purpose statistical models. Fuller's book demonstrates that, although the terminology of latent variables is used sparingly. Schervish (1987) makes the point exceedingly well when he contrasts a technical contribution (Anderson, 1984) and the more applied contribution (Dillon and Goldstein, 1984) developed to a considerable extent with social research in mind.

The primary reason for the prominence of latent-variable models in the social sciences is that we do not know how to measure the "key variables" of theoretical or substantive interest very well. The point is made best in the area of educational testing: no one would

take seriously the results of a test with just one item or problem, and in this area we usually think we can measure better by including more rather than fewer items or problems on an ability test. In the social sciences we typically obtain *multiple measurements* (or multiple indicators) of the variables that we would really like to study. The existence of multiple measurements of both Y's and X's, I have argued, is one of the main distinctive features of statistical analysis in social research (see Clogg and Dajani, 1991). Multiple measurements and the uncertainty associated with them are taken into account in practice mostly by formulating models for latent variables. It is usually the case that a "true" measurement or benchmark is unavailable, whereas in biomedical work we can sometimes calibrate fallible measurements against infallible ones. (An example might be choice of tests for screening blood sera for HIV+ where the "true" diagnosis might be known for a small sample.) I consider two important classes of such models next.

The Latent Class Model

The latent-class model arises in the following way. Suppose that we have a set of categorical measures (say, $Y_1, \ldots, Y_k$) of some "true" variable (say, Y^*). The true variable might be "attitude" toward the death penalty; the available measures might be dichotomous (yes or no) items such as "Should capital punishment be used for persons guilty of (crime x) under (given circumstances)?" If we suppose that the true variable is itself discrete, with two or more levels or "latent classes," we obtain the LCM. (The latent classes might be thought of as unordered or nominal, ordered from low to high, or even as discrete-quantitative categories.) The basic idea is to infer the distribution of the latent variable and its relation to the observed variables from the observed information (the joint distribution of the observed variables). [The restriction to a discrete Y^* even in cases where it is more natural to think of a continuous Y^* is not as restrictive as it might first appear. See Lindsay, Clogg and Grego (1991).] In modern statistics, the same model is called a finite mixture model. The latent classes in the LCM represent the unobservable components or groups in the mixture, and when the observed variables are categorical, it is most natural to think of the observed joint distribution of the Y_i's as a mixture of multinomials. This is an alternative way to define the LCM. It will be appreciated that some restrictions have to be imposed in order to identify the parameters, and it is customary to assume that the observed Y's are independent conditional on the level of Y^*, which is the so-called axiom of local independence (Lazarsfeld and Henry, 1968). Other assumptions are possible.

LCM's (or finite mixture models) are currently quite popular in statistics, both theoretical and applied. The treatise by Titterington, Smith, and Makov (1985), as well as several other recent monographs, can be consulted for verification of this claim. LCM's are also quite popular in social research, particularly in sociology; see McCutcheon (1987). Where did this model originate? Who was responsible for developing it? What scientific contexts provided the basis for its development?

A brief and probably idiosyncratic summary of the development of the LCM is as follows:

1. Lazarsfeld (1950) invents or discovers the LCM as a means to summarize multiple measurements that arose in social-psychological studies, using survey data obtained from military personnel, conducted during the Second World War. The basic terminology of LCM's was already available in these sources, although the statistical foundations of the method (mixture of multinomials) was not made clear.

2. T. W. Anderson, Albert Madansky, Neil Henry and others produce some crude methods of estimation, the most important of which was the so-called determinantal method. This statistical methodology, including maximum likelihood procedures for certain models, was summarized in Lazarsfeld and Henry (1968), the fundamental treatise on the LCM. [One of the most important recent monographs on the subject is Formann (1984); Formann is a psychometrician. I hasten to add that special cases were considered in genetics; see references in Haberman (1979).]

3. The LCM is generalized, and other latent structure models codified, in Lazarsfeld and Henry (1968). The LCM is presented as a statistical model that operationalizes some of the concepts in theoretical sociology associated with Robert Merton, one of the leading social theorists in the post-war period.

4. Goodman (1974a, b) gives a general algorithm for maximum likelihood estimation, for both restricted and unrestricted LCM's, relates the LCM to methods for studying turnover in panel studies, gives methods for studying identifiability, and relates LCM's to log-linear models. The algorithm Goodman proposed in 1974 was equivalent to the EM algorithm presented 3 years later by Dempster, Laird and Rubin (1977).

5. Haberman (1974, 1977) presents the LCM more formally, studying the likelihood equations and the likelihood surface in detail. Many generalizations appear subsequently [see, e.g., Clogg and Goodman (1984) for multiple-group versions of the basic model].

6. Clogg (1977) produces the MLLSA (maximum likelihood latent structure analysis) computer program based on Goodman (1974b). (The importance of software should not be overlooked.) Much applied work and several later programs were based on MLLSA, the EM algorithm or both. [McCutcheon (1987) is an introduction to MLLSA, for example.]

Since the mid-1970s there has been a rebirth of interest in the LCM. In social science areas, most of this work follows the Lazarsfeld–Goodman tradition. A separate tradition has developed in statistics. For example, in Titterington, Smith and Makov (1985), surely a fundamental work on the methodology of finite mixtures, a work that I admire greatly, there are *no* references to Lazarsfeld, only passing references to Goodman's work on the subject, and no references to Haberman's theoretical work on LCM's. At the same time, there is more than passing reference to the LCM tradition in social science. For example, on page 26 it is stated that a "closely related application of finite mixture distributions occurs in *latent structure* analysis, for which there is a large literature, particularly in publications devoted to applications of statistics in the social sciences." Note that LCM's are referred to as *applications* of finite mixture distributions or as "applications of statistics." In point of fact, the LCM arose in the social sciences. The statistical theory and the algorithms necessary for serious empirical work, for the discrete-data case at least, were developed primarily by statisticians with close ties to the social sciences, Goodman and Haberman in particular. I believe that the fundamental role of both the social science context (data involving multiple measurements) and the contributions of social statisticians who worked seriously at solving social science problems (Goodman and Haberman in particular) is easy to miss in current accounts of finite-mixture methods in statistics.

In my judgment, the methodology of LCM's would not have developed as it has without the social science roots indicated above. They almost certainly would not have come to be so prominent in the current portfolio of "methods adequate to the treatment of [sociological] data" without the pioneering work of Lazarsfeld, Goodman and Haberman. While it is true that there was parallel work on similar models in genetics (see Haberman, 1979, for references), to my knowledge this work simply did not lead to a general understanding of the basic model. In contrast, the LCM was already established as a practical and general statistical method in the social sciences by the mid-1970s.

The LCM is another example of a set of statistical methods that have general significance but which arose primarily in the social sciences, with the assistance of statisticians working in the social sciences. The usefulness of this model and algorithms developed to a large extent for its analysis is recognized in much current statistical work. [In addition to sources already cited, see Tanner (1991).]

Models for Continuous Latent Variables

Models for continuous latent variables have a long history in social research; see Duncan (1984) for a selective history of the subject. Such models are probably more familiar to statisticians partly because many current texts or monographs on multivariate analysis feature these models or their relatives. Principal components methods are closely related to these models, and so are many of the models for repeated measures. It is important to distinguish between models where continuous latent variables are assumed to produce categorical measurements (correct/incorrect scoring of ability tests are assumed to be related to a continuous latent variable like ability, for example) and models where continuous latent variables are assumed to underlie continuous measurements. The former are called *latent trait* models; the latter are often called *factor models* or *covariance structure* models. Both kinds of models are mostly social science products. The models were formulated by social scientists, including psychologists, and their statistical aspects have been studied largely by statisticians with close ties to the social sciences. It is difficult to imagine how statistical methods for either type of model could be traced to any large extent to either biological science or to pure mathematical statistics. (In some sense, of course, factor models can be viewed in terms of Pearson's early work on correlation theory, but the connection is vague at best.) Andersen (1980) can be consulted for the latent trait model. Joreskog and Sorbom (1978) or Bollen (1989) can be consulted for the factor model. Journals such as *Psychometrika, Journal of Educational Statistics, Sociological Methods and Research* and the *Journal of Educational and Psychological Measurement* regularly feature developments in these areas.

The latent trait model has a relatively short history. It evidently started in the 1950s in the work of several statisticians with close ties to the social sciences, including Rasch, Birnbaum and Novick. One of the main statistical problems in this area is to make inferences about the latent trait without assuming that it is normally distributed. *Conditional likelihood* (CL) methods for some of these models were developed by Rasch, Erling Andersen and others well before the 1970s. Much of the impetus for CL methods of inference actually arose from the analysis of the latent trait model. [Erling Andersen's fundamental work on the subject is cited, for example, in Cox (1972); note that Cox originally conceived of his approach for dealing with proportional hazards models as CL. Andersen (1980) is still one of the best places to learn about CL methods.] Although sociological methodologists (or statisticians working on sociological problems) have not contributed a great deal to the development of latent trait models as *statistical methods*, it is clear that they have contributed to the subject in important ways. For example, the interesting observation that CL solutions for Rasch's version of the latent trait model can be obtained from a special log-linear model

for the cross-classification of item responses is due to Tue Tjur, Noel Cressie and Paul Holland (as was independently suggested by Otis Dudley Duncan), all statisticians or methodologists with close ties to the social sciences. See the background material and references in Lindsay, Clogg and Grego (1991).

4. MORE EXAMPLES

There are a good many other convincing examples that can be used to demonstrate our point. We conclude with just a few.

Methodology for Missing Data

The recent treatise by Little and Rubin (1987) covers this important new class of statistical methods. Judging from recent literature there can be no doubt that statisticians are now taking seriously all of the issues involved with missing data. The main context for such methods is primarily that of the social survey, one of the great achievements of modern social science as well as of modern government. In typical situations involving survey data, anywhere from 5% to 20% of the data are missing on the key variables chosen for analysis. Missingness is due to "don't know" responses in attitude items, item nonresponse, missing cases (sampling units lost), and attrition of sample units in panel surveys. Missing data are often simply ignored in other areas, partly because missing information is comparatively rare in experimental work with nonhuman subjects. Even a casual inspection of Little and Rubin (1987) or Rubin (1987) will indicate that this area developed the most in the context of social statistics, and it was developed primarily by survey researchers or statisticians with close ties to the social sciences. Heckman's (1976) method of correcting for sample-selection bias has been central in econometric analysis, for example; Heckman gives a technique for handling nonresponse on dependent variables when the nonresponse mechanism is "nonignorable," to use the Little-Rubin terminology. Statisticians contributing to the analysis of missing data must reckon with sample-selection adjustment as that subject developed in econometrics.

Modern (Complex) Sampling

We live in an era of sample surveys, most of which are social surveys in the broad sense of the term. We also live in an era of *complex* sampling. Almost every major survey today contains both stratification and clustering, the latter of which invalidates the "iid" assumption so central to textbook statistics (and, I might add, to our main journals in statistics). How to sample efficiently, taking into account bias, cost and precision, is something that is mostly dealt with by survey statisticians or "samplers." A pioneer in this area is Leslie Kish, whose monograph (Kish, 1965) continues to be one of the main source books. Many fine statisticians and many excellent monographs can be added to Kish, including Cochran (1977) and Hansen, Hurwitz and Madow (1953). Modern survey sampling arose primarily in the context of social statistics, often in the setting of census operations in the U.S., Canada, Sweden and other nations. Kish was and is a member of the sociology department and the Institute for Social Research at the University of Michigan. The unique flavor of modern sampling methods can be appreciated in current works such as Groves (1989). There are surely other scientific contexts besides social surveys and census operations that have had major impacts on modern statistical methodology, such as capture-recapture sampling in wildlife and fish management, but there can be no doubt about the importance of social science (or social statistics) roots here.

Econometrics

No discussion of modern statistical methods or of statistical methods for social research can afford to ignore the tremendous role of econometrics. Maddala (1983) or Amemiya (1985) can be consulted to appreciate the flavor of this field. By chapter two of the former monograph we have left the modeling apparatus of standard statistics, and yet the tools developed by econometricians are indispensable to modern social science work. So-called discrete-choice models are similar to the methods for categorical data in statistics, but the link to rational choice and utility theory is brought out in econometrics. Quite a few of the major innovations in time series analysis, in simultaneous equations models, and in other areas so central to modern empirical work in the social sciences were produced by econometricians or by statisticians with close ties to econometrics. It is often easier to find "appropriate" statistical tools for analysis of social data in econometric software packages than it is to find them in general purpose statistical software packages. (Of course, perhaps as much as one third of what we find in most software packages for statistical analysis is in fact econometric in nature.) While no one today would ever think of statistics as a branch of economics (perhaps some in Fisher's day did indeed have this misapprehension), the impact of econometrics on statistical methodology is enormous, and I wish I knew the subject better in order to bolster my argument.

Causal Inference

To Fisher, Yates and others who pioneered the modern experimental method, causal inference was valid only when "treatments" could be *randomly* assigned to (randomly chosen) subjects. The treatment effects that are given to us by AOV methods are surely causal effects (of treatments). This conception of causal analy-

sis underlies the conception of causation that is most often utilized in statistics. It serves as the ideal model of causal inference, and one of the main points of the Rubin model of causal inference (see, e.g., Rubin, 1978) is that the experimental model must be used as a guide even in cases where it is difficult or impossible to run classical experiments.

In the social sciences, which rely mainly on so-called observational studies including surveys (see Cochran, 1983; Kish, 1987), the classical Fisher-Yates model cannot be applied. Observational studies, or surveys, are characterized by either the absence of assignment of treatments or by the absence of random assignment of treatments. (There are notable exceptions to this generalization.) Causal inference is more complex for the data typically available in social science work. Practically all of the empirical work conducted in the social sciences that is sponsored by the federal government deals with causal inferences of some type or another, although often the purpose is disguised somewhat (e.g., by using the term, "structural model").

In social science work, there are essentially two brands of causal inference available. The first might be called the regression strategy, or perhaps the econometric or psychometric strategy. This approach involves specifying a "correct" model by including all relevant "causal" variables in a single-equation or a multiple-equation model. An early account of the strategy can be found in Blalock (1962). Much of the literature cited earlier in connection with latent continuous variables is also relevant here. The objective is to include the "right" variables so that the error term is not correlated with the variables whose causal effects are to be estimated, perhaps including variables that ostensibly correct for the possible bias in selecting the sample used for estimating the relationship. By including all relevant covariates, the researcher tacitly assumes that causal inferences can be estimated as if an experiment had been carried out. (Zero correlation between the error term and the treatment levels is guaranteed by the classical experimental method. In situations where it is not natural to think of error terms, analogous statements pertaining to the conditional distribution of Y given "the right" X's apply.)

Statistical methods of this sort have been developed mainly in the social sciences, including econometrics. Where else would such heroic methods be required or even attempted? Of course, the early contributions of Sewell Wright, a population biologist, along with many contributions from econometrics, make the modern methodology of causal inference what it is. Good references to the models and the logic of this brand of causal inference are Duncan (1975) and Bollen (1989) who refer to a vast literature in psychometrics, econometrics and sociological methodology. The pioneers in this area were mostly statisticians with close ties to the social sciences (Wright is the main exception). This is not the place to debate the relative merits of the method of causal inference that derives from covariance-structure analysis. It is important to distinguish between the technology, which is very good, and the language or logic of making causal inferences, which can be criticized. It is difficult to test the key assumptions necessary for the *causal inferences* without making other assumptions that cannot be tested. My point is just that the statistical methodology suited for this popular form of causal inference was woven primarily from social science cloth.

The second main brand of causal inference, which is now subsumed under the general Rubin model, is associated with Cochran, Donald Campbell, Rubin and Holland. (Note that each has or had close ties to the social sciences, including evaluation research.) This branch of causal inference is more closely related to the Fisher-Yates framework, but it relies on matching methods and the modeling of propensity scores to a great extent. This area has generated considerable interest in statistical science of late, and it is not clear to me how it will eventually be tied to the more traditional regression-based strategy of econometrics or sociological methodology. See Heckman and Hotz (1989) for one approach that tries to do just that. But a point that can surely be made is that the context for developing these statistical methods for causal analysis has been the social science context as much as anything else. Careful surveys of the Rubin model (Holland, 1986) justify that assessment. Once again it is difficult to imagine how statistical methods for causal inference would have developed as they have without the impetus provided by the context of social research.

5. CONCLUSION

The reader will have realized well before now that this impressionistic account of the relationship between the methodology for social research and statistical methodology is one-sided. But I hope that it is not too one-sided to invalidate the basic message, which I summarize in three general observations. First, statistical methodology for sociological or social science work, as judged by current practice anyway, has not developed as a simple process of borrowing from statistical methodology developed for biological work, nor has it trickled down from the delta-epsilon rigor of mathematical statistics. Social science methods in general, and sociological methodology in particular, have a distinct flavor that cannot be appreciated very well by looking hard at analysis-of-variance methods for experimental data or, for that matter, by thinking hard about the mathematically neat problems that we see so often in our most technical journals. Second, there

are many compelling examples that can be put forth where statistical methodology created in response to the real (and difficult) problems of social research has come to have truly general significance in both mathematical and applied statistics. The ubiquity of log-linear models, event-history models, latent variable models, causal inference procedures, modern econometric models and methods, complex sampling, and other methodologies in contemporary statistical literature is testimony to this. These methods have been driven by the needs of social research, and they have been developed to a considerable extent by statisticians with close ties to the social sciences. Finally, when we celebrate the history of statistics, as we recently did in sesquicentennial activities of the American Statistical Association, we would do well to realize that statistics as a field has always been closely tied to social statistics and social science research. A serious history of the growth of statistics in this century would, I believe, demonstrate conclusively that the field owes a great deal to the social sciences and to the diverse statistical problems and statistical methodologies associated with them.

Some Other Neglected Topics

The most common criticism received from colleagues on earlier drafts of this paper was that way too many topics *that make my case stronger* had been omitted or given short shrift. I am not always pleased by criticisms of what I have to say, but this sort of criticism is a real treat. Network analysis in sociology, mathematical and statistical demography, evaluation research, methodology for the study of sample selection bias, econometric methods for panel data, methods for pooling cross-sections and multi-level or hierarchical regression models for contextual analysis are just some of the topics that others thought I should feature in order to make my case a stronger one. I do not know enough about most of these areas to give much more than annotated bibliographies, and I apologize to those critics who might feel that my sample selection mechanism has led to a biased inference about the "population" of ideas that this paper is about. The relative neglect of econometrics in this review is very serious; that subject has had and continues to have major effects on virtually all areas of modern statistical methodology.

ACKNOWLEDGMENTS

The author is indebted to Kenneth A. Bollen, Thomas A. DiPrete, Otis Dudley Duncan, Leo A. Goodman, Glenn Firebaugh, Kenneth C. Land, Michael P. Massagli, Alan Sica, Stephen M. Stigler, Christopher Winship, the editor and two referees for helpful comments.

REFERENCES

AGRESTI, A. (1984). *Analysis of Ordinal Categorical Data*. Wiley, New York.

AGRESTI, A. (1990). *Categorical Data Analysis*. Wiley, New York.

AMEMIYA, T. (1985). *Advanced Econometrics*. Harvard Univ. Press.

ANDERSEN, E. B. (1980). *Discrete Statistical Models and Social Science Applications*. North-Holland, Amsterdam.

ANDERSON, T. W. (1984). *Introduction to Multivariate Statistical Analysis*, 2nd ed. Wiley, New York.

BARTLETT, M. S. (1935). Contingency table interactions. *J. Roy. Statist. Soc. Ser. B* **2** 248–252.

BIRCH, M. W. (1963). Maximum likelihood in three-way contingency tables. *J. Roy. Statist. Soc. Ser. B* **26** 220–233.

BISHOP, Y. M. M. (1969). Full contingency tables, logits, and split contingency tables. *Biometrics* **27** 119–128.

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press.

BLALOCK, H. M. (1962). Four-variable causal models and partial correlations. *American Journal of Sociology* **68** 182–194.

BOCK, R. D. (1970). Estimating multinomial response relations. In *Contributions to Statistics and Probability* (R. C. Bose, ed.) 453–479. Univ. North Carolina Press, Chapel Hill.

BOLLEN, K. A. (1989). *Structural Equations with Latent Variables*. Wiley, New York.

BOX, J. F. (1978). *R. A. Fisher: Life of a Scientist*. Wiley, New York.

CAUSSINUS, H. (1965). Contribution a l'analyse statistique des tableau de correlation. *Ann. Fac. Sci. Toulouse Math. (5)* **29** 77–182.

CLOGG, C. C. (1977). *Unrestricted and Restricted Maximum Likelihood Latent Structure Analysis: A Manual for Users*. Working Paper 1977-09, Population Issues Research Cent., Pennsylvania State Univ.

CLOGG, C. C. (1986a). Invoked by RATE. *American Journal of Sociology* **96** 696–706.

CLOGG, C. C. (1986b). Quasi-independence. *Encyclopedia of Statistical Sciences* **7** 460–464.

CLOGG, C. C. and DAJANI, A. (1991). Sources of uncertainty in modeling social statistics: an inventory. *Journal of Official Statistics* **7** 7–24.

CLOGG, C. C. and GOODMAN, L. A. (1984). Latent structure analysis of a set of multidimensional contingency tables. *J. Amer. Statist. Assoc.* **79** 672–771.

COCHRAN, W. G. (1977). *Sampling Techniques*, 3rd ed. Wiley, New York.

COCHRAN, W. G. (1983). *Planning and Analysis of Observational Studies*. Wiley, New York.

COLEMAN, J. S. (1964). *Introduction to Mathematical Sociology*. Free Press, New York.

COX, D. R. (1972). Regression models and life tables (with discussion). *J. Roy. Statist. Soc. Ser. B* **34** 187–220.

DEMING, W. E. and STEPHAN, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Ann. Math. Statist.* **11** 427–444.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Roy. Statist. Soc. Ser. B* **39** 1–38.

DILLON, W. R. and GOLDSTEIN, M. (1984). *Multivariate Analysis: Methods and Applications* Wiley, New York.

DUNCAN, O. D. (1975). *Introduction to Structural Equation Models*. Academic, New York.

DUNCAN, O. D. (1979). How destination depends on origin in the occupational mobility table. *American Journal of Sociology* **84** 793–803.

DUNCAN, O. D. (1984). *Notes on Social Measurement, Historical and Critical.* Russell Sage Foundation, New York.

DUNCAN, O. D. and STENBECK, M. (1988). Panels and cohorts: design and model in the study of voting turnout. In *Sociological Methodology 1988* (C. C. Clogg, ed.) 1-36. American Sociological Association, Washington, D.C.

FIENBERG, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41** 907-917.

FIENBERG, S. E. (1980). *The Analysis of Cross-Classified Categorical Data*, 2nd ed. MIT Press.

FISHER, R. A. (1922). On the interpretation of chi-square from contingency tables, and the calculation of P. *J. Roy. Stat. Soc.* **85** 87-94.

FISHER, R. A. (1935). Appendix to article by C. Bliss. *Annals of Applied Biology* **22** 164-165.

FISHER, R. A. (1970). *Statistical Methods for Research Workers*, 14th ed. Hafner, New York.

FORMANN, A. (1984). *Die Latent-Class-Analyse.* Beltz Verlag, Weinheim/Basel.

FREEMAN, D. A. (1991). Statistical models and shoe leather. In *Sociological Methodology 1991* (P. V. Marsden, ed.) 291-313. Basil Blackwell, Oxford.

FULLER, W. A. (1987). *Measurement Error Models.* Wiley, New York.

GOODMAN, L. A. (1964). Interactions in multidimensional contingency tables. *Ann. Math. Statist.* **35** 716-725.

GOODMAN, L. A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables with or without missing entries. *J. Amer. Statist. Assoc.* **63** 1091-1131.

GOODMAN, L. A. (1969). On partitioning chi-square and detecting partial association in three-way contingency tables. *J. Roy. Stat. Soc. Ser. B* **31** 486-498.

GOODMAN, L. A. (1970). The multivariate analysis of qualitative data: interactions among multiple classifications. *J. Amer. Statist. Assoc.* **65** 226-256.

GOODMAN, L. A. (1972). A general model for the analysis of surveys. *American Journal of Sociology* **77** 1035-1086.

GOODMAN, L. A. (1974a). The analysis of systems of qualitative variables when some of the variables are unobservable. Part I: A modified latent structure approach. *American Journal of Sociology* **79** 1179-1259.

GOODMAN, L. A. (1974b). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215-231.

GOODMAN, L. A. (1979). Simple models for the analysis of cross-classifications having ordered categories. *J. Amer. Statist. Assoc.* **74** 537-552.

GOODMAN, L. A. (1991). Measures, models, and graphical displays in the analysis of cross-classified data (with discussion). *J. Amer. Statist. Assoc.* **86** 1085-1138.

GOODMAN, L. A. and KRUSKAL, W. H. (1954). Measures of association for cross-classifications. *J. Amer. Statist. Assoc.* **49** 732-764.

GOODMAN, L. A. and KRUSKAL, W. H. (1979). *Measures of Association for Cross-Classifications.* Springer, New York.

GRIZZLE, J. E., STARMER, C. F. and KOCH, G. G. (1969). Analysis of categorical data by linear models. *Biometrics* **25** 489-504.

GROVES, R. M. (1989). *Survey Errors and Survey Costs.* Wiley, New York.

HABERMAN, S. J. (1970). The general log-linear model. Ph.D. dissertation, Univ. Chicago.

HABERMAN, S. J. (1974a). *The Analysis of Frequency Data.* Univ. Chicago Press.

HABERMAN, S. J. (1974b). Log-linear models for frequency tables derived by indirect observation: Maximum-likelihood equations. *Ann. Statist.* **2** 911-924.

HABERMAN, S. J. (1977). Product models for frequency tables involving indirect observation. *Ann. Statist.* **5** 1124-1147.

HABERMAN, S. J. (1979). *The Analysis of Qualitative Data, Vol. II. New Developments.* Academic, New York.

HABERMAN, S. J. (1982). Analysis of dispersion of multinomial responses. *J. Amer. Statist. Assoc.* **77** 568-580.

HANSEN, M. H., HURWITZ, W. N. and MADOW, W. G. (1953). *Sample Survey Methods and Theory, Vol. I.* Wiley, New York.

HECKMAN, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models. *Annals of Economic and Social Measurement.* **5** 475-492.

HECKMAN, J. J. and HOTZ, V. J. (1989). Choosing among alternative nonexperimental methods for estimating the impact of social programs: the case of manpower training programs. *J. Amer. Statist. Assoc.* **84** 862-874.

HECKMAN, J. J. and WALKER, J. R. (1987). Using goodness of fit and other criteria to choose among competing duration models: a case study of Hutterite data. In *Sociological Methodology 1987* (C. C. Clogg, ed.) 247-308. American Sociological Association, Washington, D.C.

HOEM, J. (1972). Inhomogeneous semi-Markov processes, select actuarial tables and duration dependence in demography. In *Population Dynamics* (T. Greville, ed.) 251-296. Academic, New York.

HOLLAND, P. W. (1986). Statistics and causal inference. *J. Amer. Statist. Assoc.* **81** 967-968.

IMREY, P. B., KOCH, G. G. and STOKES, M. E. (1981). Categorical data analysis: Some reflections on the log linear model and logistic regression. Part I: Historical and methodological overview. *Internat. Statist. Rev.* **49** 265-283.

JORESKOG, K. G. and SORBOM, D. (1978). *Advances in Factor Analysis and Structural Equation Models.* Abt Books, Cambridge, Mass.

KISH, L. (1965). *Survey Sampling.* Wiley, New York.

KISH, L. (1987). *Statistical Design for Research.* Wiley, New York.

LAZARSFELD, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In *Studies in Social Psychology in World War II, Vol. IV: Measurement and Prediction* (S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star and J. A. Clausen, eds.) 362-412. Princeton Univ. Press.

LAZARSFELD, P. F. (1961). The algebra of dichotomous systems. In *Studies in Item Analysis and Prediction* (H. Solomon, ed.) 111-157. Stanford Univ. Press.

LAZARSFELD, P. F. and HENRY, N. W. (1968). *Latent Structure Analysis.* Houghton Mifflin, Boston.

LINDSAY, B., CLOGG, C. C. and GREGO, J. (1991). Semi-parametric estimation in the Rasch model and related exponential response models, including a simple latent class model for item analysis. *J. Amer. Statist. Assoc.* **86** 96-107.

LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data.* Wiley, New York.

MADANSKY, A. (1959). The fitting of straight lines when both variables are subject to error. *J. Amer. Statist. Assoc.* **54** 173-205.

MADDALA, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge Univ. Press.

MANTEL, N. (1966). Models for complex contingency tables and polychotomous dosage response curves. *Biometrics* **22** 83-95.

MCCUTCHEON, A. (1987). *Latent Class Analysis.* Sage, Newbury Hills, Calif.

MOSTELLER, F. (1968). Association and estimation in contingency tables. *J. Amer. Statist. Assoc.* **63** 1-28.

PEARSON, K. (1900). On a criterion that given a system of devia-

tions from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine, Series 5* **50** 157–175.

PEARSON, K. (1904). Mathematical contributions to the theory of evolution XIII: On the theory of contingency and its relation to association and normal correlation. *Draper's Company Research Memoirs. Biometric Series, No. 1.* [Reprinted in *Karl Pearson's Early Papers* (E. S. Pearson, ed.). Cambridge Univ. Press.]

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York.

SCHERVISH, M. J. (1987). A review of multivariate analysis (with discussion). *Statist. Sci.* **2** 396–433.

SIMPSON, E. H. (1951). The interpretation of interaction in contingency tables. *J. Roy. Statist. Soc. Ser. B* **13** 238–241.

STIGLER, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900.* Harvard Univ. Press.

TANNER, M. A. (1991). *Tools for Statistical Inference: Observed Data and Data Augmentation Methods. Lecture Notes in Statist.* **67**. Springer, New York.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

TUMA, N. B. and HANNAN, M. T. (1984). *Social Dynamics: Models and Methods.* Academic, Orlando, Fla.

TUMA, N. B., HANNAN, M. T. and GROENEVELD, L. P. (1979). Dynamic analysis of event histories. *American Journal of Sociology* **84** 820–854.

YULE, G. U. (1900). On the association of attributes in statistics. *Philos. Trans. Roy. Soc. London Ser. A* **194** 257–319.

Comment

David J. Bartholomew

This is an interesting and timely reminder of the important role which the social sciences have played, and continue to play, in the development of statistical methodology. I agree with so much of what the author says that it would be all too easy to make this contribution a repetition of the main points or a catalogue of additional supporting examples. Instead I wish to move the discussion in two other closely related directions – first by putting the emphasis on the inhibiting effect of the natural science influence on the development of statistical methodology and secondly by identifying current social science interests which place new demands on methodology.

I think the author is right in turning the spotlight on R. A. Fisher or, more exactly perhaps, the Fisherian tradition. Had it not been for Fisher's immense prestige the needs of social science might have continued to set an agenda for theoreticians as foreshadowed in the pioneering work of Quetelet and others. The core of modern statistical theory, centred on continuous variables, normal distributions, independence and additive models with the analysis of variance as its centrepiece has become the canon around which statistical education is built. The generalized linear model stands today as a fitting culmination of that tradition. The assumptions and the formulation of the models used are those required by the natural science problems which motivated Fisher and his followers and which still nourish much contemporary research. It is thus entirely understandable, though regretable, that the growth of multivariate analysis should, on the theoretical side, have been developed almost entirely around the multivariate normal distribution.

Perhaps the most striking example of this thesis is the laggardly way in which methods for categorical variables have become part of the statistician's portfolio. After the early excursions of Yule little note seems to have been taken of the fact that categorical variables are extremely common and, in a sense, more fundamental than their continuous counterparts. Until quite recently, measurement of association in two-way contingency tables was about as far as the education of most statisticians went. When they have been confronted with categorical data in practice they have had, for want of anything better, to force it into the Procrustean bed made for continuous variables by upgrading the level of measurement in more or less arbitrary ways. This is still very evident in the analysis of covariance structures where methods are developed for continuous variables and then adapted to categorical variables by the introduction of polychoric coefficients and such like. It is only now becoming apparent that there is a common structure underlying many such multivariate techniques which has been hidden by their diverse origins and notational idiosyncracies. To some

David J. Bartholomew is Professor of Statistics, Department of Statistical and Mathematical Sciences, London School of Economics and Political Sciences, Houghton Street, London WC2A 2AE, United Kingdom.