Analysis of change in discrete variables

Sorensen, Aage B.

Aage B. Sørensen

Analysis of Change in Discrete Variables*

1. Introduction

Longitudinal data are analyzed using a variety of techniques and methods in the various social and behavioral sciences. Over-time data comes in many forms – as panel data, time series, and event-histories[1]. Different disciplines have tended to focus on one particular type of over-time data. Econometricians have concentrated on time series, demographers on a particular form of event-histories, sociologists on panel data, psychometricians on change scores. Further, the different disciplines have specialized in particular methodological problems – econometricians in problems of estimation, especially in problems of time dependent errors, psychometricians in the reliability of change scores, and, in classical panel analysis, sociologists have concentrated on developing measures of causal influence. The result is that longitudinal methodology is a confusing affair. Some problems have solutions, others equally important do not, and it is often difficult to see the relevance of a technique for a problem if this technique has been developed in another discipline with a different research tradition.

There is then a need both for codification and for remedying some of the uneven development of existing longitudinal methodology. One set of problems for which longitudinal methodology seems in particular need of attention is composed of those encountered when analyzing change in discrete or categorical variables.

* The research reported here was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Department of Health, Education and Welfare pursuant to the provisions of the Economic Opportunity Act of 1964.

1 Event-history data are longitudinal data where the exact timing of events is known. They are thus continuous time records of events like job shifts, residence shifts, etc., when the units of analysis are individuals. Life-history data are event-history data. For methodological purposes, the important feature of life-history data is the information on the timing of events, not the coverage of people's lives. Further, in some instances, information on the timing of events may be obtained with other designs than the life-history study design. Hence the term event-history is preferred.

Although such variables are often employed by the softer social sciences, there does not exist a readily available set of techniques and methods for the analysis of change.

Until recently there was in fact very little that could be done with categorical variables other than computing percentages. This situation has changed dramatically in recent years. Powerful techniques for log-linear and multiple classification analyses have increasingly become available[2]. These techniques may of course also be applied to change data, treating time as any other independent variable, and analyzing the data using the approach applied to cross-sectional data. These techniques of course also may be applied to over-time data in the same manner as they are applied to cross-sectional data. The over-time variation is then treated in the same manner as the cross-sectional variation among individuals (or other units of analysis).

Using log-linear and multiple classification analysis with over-time data on categorical variables corresponds to the use of regression techniques with algebraic (usually linear) models with cross-sectional and over-time data on continuous variables. In both instances the interest is in estimating the relationship among variables and in both cases the over-time variation is treated in the same manner as the cross-sectional variation in variables. Much social science research has the estimation and interpretation of relationships as the primary objective. These *ad hoc* techniques are then appropriate and the methodology, despite very different estimation techniques, is conceptually similar whether log-linear or multiple regression techniques are applied.

The application of *ad hoc* statistical techniques makes assumptions about the form of the relationship among variables, usually that they are linearly related. Such assumptions are rarely tested and even more rarely justified in terms of substantive consideration of the process under study. Nevertheless, the assumption made may be empirically and conceptually inadequate. This may lead to misleading inferences and limit our ability to understand fully the processes being analyzed. On cross-sectional data there is, however, very little that can be done since the unfolding of the processes that generate observed relationships among variables cannot be observed. On over-time data it is possible to study directly the change processes that generate observed outcomes. However, when over-time variation is treated as cross-sectional variation, this opportunity for obtaining a better understanding of how observed outcomes are generated is missed. Direct study of change is needed. This paper will advocate such an approach to the study of change in discrete variables.

To study change it is necessary to identify the components of change. The first part of this paper will identify these components. The second part of the paper will then briefly outline some strategies for the causal analysis of these components.

[2] A comprehensive survey is provided by Bishop, Y. M. M., et al., Discrete Multivariate Analysis: Theory and Practice, Cambridge 1975.

2. Conceptualizing Change

The concern in longitudinal methodology is with description and analysis of variables that are functions of time. To identify the tasks involved it is necessary to have a representation of the change process that identifies the quantities that should be estimated in empirical analysis. In other words, a conceptualization of the change process should be given in a mathematical representation. The classic approach to the mathematical analysis of change is the one represented by calculus. It applies to variables that are continuous, i. e. variables that can be represented by real numbers. Though the concern in this paper is with discrete variables, the continuous variable treatment serves as a model, and will be briefly outlined.

It seems natural to represent change as the difference in values of the variable of interest obtained over some time interval. Denote the time dependent variable y (t). The difference $y(t_2)-y(t_1)$ observed over the interval t_2-t_1 would be the quantity of interest. Presumably this difference is brought about by some causal variables, possibly including time, that act on y (t) in a certain way. In descriptive analysis the objective is to specify the resulting time variation in y (t). In causal analysis we go further and attempt to specify the various causal forces acting on y (t), and estimate their influence. In other words, for causal analysis it is necessary to specify: (1) the mechanisms that bring about change, and (2) to assess the causal influences transmitted by these mechanisms.

2.1. *Specifying the Mechanism of Change*

The specification of the change mechanisms depends first on the timing of change. If y(t) changes continuously in time so that y(t) is continuously differentiable with respect to time over the interval of interest, relating $y(t_2) - y(t_1)$ to $t_2 - t_1$ presents the problem that as y(t) changes, so does t. The classic solution is to focus on the change in y(t) obtained in an infinitesimal small interval of time[3]. This conceptual abstraction makes it possible to relate change to the value of time (and other variables) rather than to intervals of time. Hence we focus on the quantity dy(t)/dt, i. e. the instantaneous rate of change in y(t). The specification of the dependency of y(t) on time and other variables may then be carried out in a differential equation:

$$\frac{dy(t)}{dt} = f(\underline{x}(t), \alpha, t) \qquad (1)$$

[3] See for further discussion Coleman, J. S., The Mathematical Study of Change, in: Blalock, Hubert M., and Blalock, Ann B. (eds.), Methodology in Social Research, New York 1968.

where the vector $\underline{x}(t)$ represents causal variables, possibly including time and y(t) itself, and the vector α represents a set of parameters.

The specification of f in the differential equation should represent assumptions about how change is produced. Some simple examples will illustrate the strategy.

The simplest process is obtained assuming that y(t) changes by a constant amount in each small interval of time, or:

$$\frac{dy(t)}{dt} = k \qquad (2)$$

A slightly more complicated expression, that is a useful representation of many processes, assumes that change in y(t) is dependent on y(t):

$$\frac{dy\,(t)}{dt} = k + by(t) \qquad (3)$$

If the quantity b represents a feedback either positive or negative. In many growth processes this feedback will be negative, and (3) describes a process where y(t) changes rapidly in the start of the process, but decreases as y(t) increases and eventually reaches zero at the equilibrium level of y(t), where dy(t)/dt is zero. Though stable processes will have this property, there may be considerable interest in processes with positive feedback where the variables of interest will take an explosive course. One example is arms races leading to wars that have been modeled by Richardson in a simultaneous differential equation model with basic properties like (3) though mathematically more complicated[4].

Since dy(t)/dt is a conceptual abstraction, differential equations cannot be used directly with empirical data. In order to estimate parameters and test the models it is necessary to solve the equations using methods of integration. For example, the solution to equation (2) is:

$$y(t) = y(0) + kt \qquad (4)$$

where y(0) is the value of y(t) obtained at the start of the process, at time 0. The solution to (3) is:

$$y(t) = \frac{k}{b}\,(e^{bt} - 1) + y(0)\,e^{bt} \qquad (5)$$

Expressions such as (4) and (5) may be used with empirical observations on y(t) and y(0) either for a set of individuals (or whatever is the unit of analysis) or obtained through repeated observations on the same individual. These formulations are necessary to test the models and estimate parameters, but the conception of the change process is given by the differential equation, from which the parameters derive their interpretation.

It is important to note that the solution (5) to (3) only holds if the parameters k and b are assumed constant over time and identical for all individuals. Failure of

4 Richardson, L. G., Arms and Insecurity, Pittsburg 1960.

these assumptions of stationarity and homogeneity will result in models that do not describe the observed course of processes adequately. Failure of the assumption means that characteristics of individuals and/or time periods cause variation in the components of change. Such variation should be modeled. The specification of the sources of variation provide the desired information on the causes of change, as shown below.

The use of differential equations to mirror change processes depends on the continuous differentiability of y(t) with respect to time. If change does not take place continuously, but only after certain intervals of time, a different formulation is necessary. Change may then be modeled in a difference equation treating time as a discrete (integer) variable:

$$\Delta y = f(\underline{x}(n), \alpha, n) \qquad (6)$$

where n is used to represent time, often trials or other discretely occurring events. A different equation may be estimated directly, since the quantity Δy usually is observable. This is sometimes seen as an advantage, and difference equations are for example often used in economics because observations are obtained at fixed intervals of time (e. g. at yearly intervals). On the other hand difference equations will still need to be solved in order to study the over-time behavior of the process and test the models, and the standard methods of calculus are not available for this purpose. Further the conception of change, not the timing of observations, should govern the formulation of a model of change. This will usually dictate the continuous time formulation in a differential equation model.

2.2. *Specifying the Causes of Change*

The examples brought above were examples of models expressing the mechanism of change in time, but not the dependency on other variables. One useful way of introducing causal variables is to express the parameters of the models as functions of a set of independent variables. In equation (3), the quantity k may for example be written as a linear function of a set of exogenous variables, i. e. $k = c_o + c_1x_1 + \ldots + c_nx_n$. This will result in:

$$\frac{dy(t)}{dt} = c_o + by(t) + c_1x_1 + c_2x_2 \ldots + c_nx_n \qquad (7)$$

The solution to (7) is parallel to (5) with the linear eypansion of k. It is important to note that if $b > 0$ and as $t \to \infty$ the solution to (7) will reduce to:

$$y(e) = -\frac{c_o}{b} - \frac{c_1}{b} x_1 \ldots - \frac{c_n}{b} x_n \qquad (8)$$

The equilibrium formulation of (7) is thus the simple linear model for a variable so often used on cross-sectional data. Note that the derivation from (7) shows that the quantities $-\frac{c_i}{b}$ that are the observed coefficients to the independent variables depend on the feedback term b. In other words, starting out with the model of change, equation (3) results in a formulation of the relationship among variables that may be observed in a cross-section in terms of the fundamental quantities that govern change. Only over-time data can identify these quantities, and only modeling change directly will specify the components of change. Over-time analysis that treats over-time variation as cross-sectional variation will not provide this information, as it will amount to using models such as (7) with time as an independent variable; an inappropriate conception if (3) governs the change process. (For further implications of this and other results of modelling change directly see Sørensen[5].)

Writing parameters in simple change models as functions of causal variables is only a meaningful way of modeling the causes of change if it can be assumed that the independent variables are unaffected by y(t), i. e. that there is no interdependence among y(t) and the x_i variables. If this cannot be assumed more complicated simultaneous differential equation models are needed to mirror the change process. These complications will not be discussed here.

The specification of the variation in quantity k of equation (3) in terms of the x_i variables also should make the model more empirically adequate since the heterogeneity in k is taken into account. Further modification may allow for time dependency, though the resulting models are quite complicated[6].

Causal analysis of change processes then demands first a specification of the mechanism of change in a differential of difference equation. The causal variables may be introduced directly in the defining equation. In many situations it is, however, simpler to see the causal variables as acting on the parameters that govern change. This is the approach that will be suggested for the analysis of discrete variables, to be discussed in the remainder of the paper.

3. Conceptualizing Change in Discrete Variables

Analysis of change in continuous and in discrete variables differ in one all important respect. Change cannot meaningfully be represented as differences in the values of variables when the variable is discrete. Hence differential or difference equations cannot be used to represent the change process, and calculus cannot be applied directly to the variables.

[5] Sørensen, Aage B., Causal Analysis of Cross-Sectional and Overtime Data: With Special Reference to the Occupational Achievement Process, in: Weselovski, W. (ed.), Social Mobility in a Comparative Perspective, forthcoming.

[6] See Coleman, Introduction to Mathematical Sociology, New York 1964, for an example.

The problem is sometimes solved by treating discrete variables as though they have a stronger metric. It is, for example, common in sociology to treat the standard measures of occupational prestige as though they possess interval level metric, though they are ordinal measures. Similarly dichotomous variables are often treated in the same manner as continuous variables in regression analysis. This solution is, however, often conceptually unsatisfactory, and the obtained estimates have undesirable statistical properties. An alternative solution is to study change in discrete variables by „proxy"; by mapping the categories of the discrete variables onto a probability distribution. The probabilities provide the desired metric, and can be studied as change in probability distributions over the state space given by the categories of the discrete variable. Probability theory (of course often using a great deal of calculus) becomes the relevant mathematical language for the study of change, and the resulting models will be stochastic process models.

Change in continuous variables could also be studied by focussing on change in probability distributions and applying stochastic process models with continuous state space. However, the mathematical complications are considerable, and the mathematical problems often become serious with discrete state models. On the other hand, the use of stochastic process models is the only way of modeling change in discrete variables, and this, rather than a fundamental choice between a stochastic versus a deterministic conception of a process, seems to be the usual reason for the use of stochastic process models with discrete variables and deterministic models with continuous variables.

As with continuous variables, the timing of change determines whether the defining equation is a differential or a difference equation, and as describes above, these equations have to be solved in order to estimate parameters and test the models. However, solutions to stochastic process models, except the very simplest, are usually quite complicated and in fact often impossible to obtain (see for example the epidemiological models presented by Bailey[7]). On the other hand, because stochastic process models permit a microscopic analysis of the process of change, even very simple models may provide a wealth of information for the analysis of the various components of change.

Suppose now that the variable of interest is a dichotomous variable giving rise to a two-state system. Label the two 1 and 2 respectively. A unit of analysis, say an individual, is at a point in time, t, characterized by the probability $p_1(t)$ of being in state 1, and $p_2(t)$ of being in state 2, where $p_1(t) = 1-p_2(t)$. The objective is to formulate the mechanism for change in $p_1(t)$ and by implication, $p_2(t)$.

If change occurs continuously, a continuous time stochastic model is desired and should be defined in a differential equation model. Change in $p_1(t)$ will reflect movement in the state space. Movement may either take place in one direction only – as when the two states refer to life and death – or, there may be movement in both directions – as when the states refer to a positive and a negative attitude. If movements in both direction take place, change will be governed by the probability

[7] Bailey, N. T. J., The Mathematical Theory of Epidemics, New York 1957.

of a move from state 1 to state 2 in an interval of time, and the probability of a move from 2 to 1 in the same interval of time. Denote q_{12} dt the probability of moving from 1 to 2 in dt, and q_{21}dt the probability of moving from 2 to 1. Assume further that these quantities are constant over time. Then the probability of an individual being in state 1 will change in dt according to:

$$\frac{dp_1(t)}{dt} = -q_{12}p_1(t) + q_{21}p_2(t) \qquad (9)$$

where $q_{12}p_1(t)$ is the rate of movement from 1 to 2 times the probability of being in state 1, and $q_{21}p_2$ (t) similarly the rate of movement out of state 2 to state 1 times the probability of being in state 2. The expression is easily generalized to cover a larger number of states:

$$\frac{dp_i(t)}{dt} = -\sum_{i \neq j} q_{ij}p_i(t) + \sum_{i \neq j} q_{ji}p_j(t) \qquad (10)$$

where the two parts of the right hand side govern respectively the outflow and the inflow from and to state i. For a k state system there will be k such equations. Assuming the q_{ij}'s constant, these equations can be solved to give the expression needed for empirical analysis. It becomes, in matrix notation:

$$\underline{P}(t) = \underline{P}(0)e^{Qt} \qquad (11)$$

where $\underline{P}(t)$ is the vector of probabilities at time t, $\underline{P}(0)$ the probability distribution at time 0, and e^{Qt} the matrix analog to e^a with Q a matrix of q_{ij}'s. This is the discrete state, continuous time Markov Model. Its application to social processes has been extensively discussed by Coleman[8].

The discrete time analog to (10) is obtained from quantities r_{ij} that are transition probabilities for moving from state i to state j on a trial. The typical equation for the change in the probability of being in state i on a trial will be:

$$\Delta p_i = -\sum r_{ij}p_i(n) + \sum r_{ij}p_j(n) \qquad (12)$$

The solution to the set of different equations is, in matrix notation:

$$\underline{P}(n) = \underline{P}(0)R^n \qquad (13)$$

analog to (11). Although discrete time processes are most often met in experimental situations, the discrete time Markov model is often applied to continuous time processes. Its advantage is mathematical simplicity. The distinction is often unimportant for prediction. However, for analysis the continuous time model often seems the most appropriate framework. One reason is that change can be further decomposed with continuous time models.

The quantities q_{ij} of the continuous time model give the rate of movement from state i to state j. It is often appropriate to conceive of this rate as resulting from the

[8] Coleman, J. S., Introduction.

occurrence of events randomly in time, and the outcome of events. Thus, in occupational mobility processes, a shift of occupation is the result of a job shift with a certain outcome, i. e. a possible shift of occupation. The occurrence of events and the outcome of events may be analyzed separately. Formally this means that the quantities q_{ij} may be decomposed as:

$$q_{ij} = \begin{cases} \lambda m_{ij} & i \neq j \\ \lambda(m_{ii} - 1) & i = j \end{cases} \qquad (14)$$

where λ governs the occurrence of events, and the m_{ij}'s are the probabilities of moving from i to j given that event occurs.

With this decomposition, equation (11) can be written:

$$\underline{P}(t) = \underline{P}(0)\, e^{\lambda(M-I)t} \qquad (15)$$

as the matrix Q of (11) = M—I, where I is the identity matrix. This formulation has been extensively discussed by Singer and Spilerman[9].

In the simple continuous time Markov Chain, the occurence of events is governed by a Poisson process. This means that the probability $p_o(t)$ of no event occurring by time t will change according to the differential equation:

$$\frac{dp_o(t)}{dt} = -\lambda p_o(t) \qquad (16)$$

The space for the Poisson process is a count of the number of events. The probability distribution corresponding to this state space is the Poisson distribution:

$$p_i(t) = e^{-\lambda t} \frac{(\lambda t)^i}{i!} \qquad (17)$$

where $p_i(t)$ is the probability that i event has occurred by time t. The mean of the distribution is λt, a property that may be used to estimate λ.

Of considerable interest for analysis is the distribution. In a Poisson process this distribution will be exponential, with probability density:

$$f(s) = \lambda e^{-\lambda s} \qquad (18)$$

where s stands for the time interval between events. The mean of s is $1/\lambda$, a property that again can be used in analysis of the occurrence of events.

The continuous time Markov Chain and the associated Poisson process for the occurrence of events are very simple. In fact the Poisson process is the analog to the simplest model for change in a continuous variable given as equation (2) with a constant increment in y(t) in each interval of time, and the Markov Chain is the analog to equation (3) where change is also assumed to depend on the current state of the system (in equation (3) on the value of y(t)). These simple stochastic models may

9 Singer, B., and Spilerman, S., Social Mobility Models for Heterogeneous Populations, in: Costner, H. L. (ed.), Sociological Methodology 1973–74, San Francisco 1974.

appear quite unrealistic models for change in discrete variables. They do, however, mirror the basic components of change in discrete variables, the distinctions between the occurence of events and the outcome of events particularly important for analysis of change. Their appropriateness and one's willingness to live with their simplicity to some extent depends on the objective of the analysis of change, as the next section will describe.

4. Objectives for the Analysis of Change

Models such as those described in the preceding section are introduced because of a desire to model the behavior of a process. This desire may reflect an interest in predicting the future course of a process, or an interest in formulating a theory of the process, or to provide a framework for a causal analysis of the components of change. Ultimately these three objectives may merge, but before the ultimate is achieved, different criteria for the usefulness of the models may be applied depending on which objective is emphasized.

If the objective is to predict, or if the objective is to formulate a theory, the primary emphasis is on the modelling task. The analysis of empirical data on change is carried out primarily to test the predictions from the model and validate its assumptions, not because of an interest in observed patterns of change and their empirical causes.

As a theory of a process the simple Markov model is quite uninteresting, and it has been repeatedly shown that the process does not predict well many social processes. The model's failure may have numerous causes, and an extensive literature exists on how to modify the simple model in order to improve its empirical or theoretical adequacy. Much of the literature on empirical adequacy addresses two problems: one is the problem of non-stationarity – that is, the fact that parameters change over time, the other is the problem of population heterogeneity – that is, parameters vary among individuals or whatever are the units of analysis to which the model is applied. Both non-stationarity and population heterogeneity will result in failure of the model to predict observed processes. Numerous solutions have been suggested in the literature that will improve the fit of the simple Markov model. They will not be reviewed here.

The discussion in the preceding section is intended to provide a point of departure for empirical analysis of the causes of change. Such analysis will focus on the sources of variation in the parameters that govern change, using continuous and discrete independent variables to account for this variation in the parameters that govern change in a manner analogous to the specification of equation (3) in equation (7). The utility of the simple models then lies in their identification of the components of change. Non-stationarity and heterogeneity become of interest not because they are sources of failure of the models, but because they are the phenomena we would like to account for by causal variables. They are the objects of analysis, rather than something to get rid of.

5. Panel Versus Event-History Data

The representation of the Markov model presented in equation (15) suggests that analysis of change in discrete variables may focus on the variation in what governs the occurrence of events, and on variation in the m_{ij}'s that govern the outcome of events. However, separate analysis of the two components of change is only possible if the data provide the necessary information. Most data on change in discrete variables in sociology are obtained from panels. Panels are usually only observations at two or three points in time on a group of respondents. Such data can be used to estimate transition probabilities and from these transition rates may be computed.

However, since only a few observations are made on the process, information on the components of change will be very fragmentary. The resulting difficulties have recently been extensively analyzed by Singer and Spilerman[10]. With a large sample some analysis may be performed of variation in transition rates among sub-groups, but individual level analysis is impossible.

Event-history data are still rare, but far superior to panel data for causal analysis of change. With continuous observations on a group of respondents, waiting times between events may be directly observed in order to study variation in . Counts of the outcome of events may be used to obtain information on the m_{ij}'s. Event-history data thus provide much richer possibilities for analysis than do panel data, particularly for analysis of the rate at which events occur. The suggestions that follow for such analysis assumes that life-history data are used.

[10] Singer, Social Mobility; Singer, B., and Spilerman, S., Representation of Social Process by Markov Models, in: American Journal of Sociology, 82 (1976), pp. 1–54.

6. Analysis of the Occurrence of Events

Event-histories of the kind I am assuming will provide information on the timing of certain events and their outcomes. The histories may pertain to individuals and the events may be acts carried out by them such as a change of job or of residence. Or, the event-histories may pertain to societies, and events may be wars or elections (if elections can occur in any time interval). The purpose of the analysis would be to study the causes of variation in the occurrence of events.

With the Poisson process as the framework there are two ways of carrying out such analysis. One is to rely on the Poisson distribution and use counts of events to estimate the rate at which they occur, the other is to rely on information on waiting times between events.

If counts of events are relied on, the rationale is that the probability distributions over the state space given by the count has a mean that is λt. Since t is known, a count of the number of events that have occurred to a person or a group of persons will provide the desired estimate. More precisely, we may, for example, carry out a count for each respondent over a period of time to give separate estimates of λ, say λ_j, for each person. These λ_j's can then be used as dependent variables in a causal analysis by relating their variation to characteristics of the respondents or their situation.

Relying on counts of events is, however, often an inefficient use of the information available in life history data and may in fact provide misleading inferences. The basic assumption of the Poisson process is that events occur with a constant probability in each interval of time. Counts will have to take place over a time period, and with infrequent events this period may be quite long. It is likely that the causal variables relevant to the occurrence of events change over this period. This information is ignored when relying on counts. In other words, intra-individual variation cannot be studied when counts of events are used to study rates. Furthermore, the over-time variation in rates means that the counts do not estimate means in Poisson distributions, so what is studied is not well defined.

An example that illustrates this point occurred in an analysis of job shifts that I did some years ago. One reasonable hypothesis about the occurence of job shifts is that they are more likely to occur the larger the discrepancy between a person's occupational resources (education, ability, etc.) and the returns obtained in the job in the form of status and earnings. Such a hypothesis cannot be tested using counts of events to estimate the rate of shifts, since the returns a person obtains from jobs will change over time as a result of the very job shifts that are analyzed. A different approach is needed, and it is offered by relying on waiting times.

The rationale used for waiting times is that if the occurrence of events is Poisson, waiting times will be exponentially distributed with mean $1/\lambda$. The assumptions of course are the same as for the Poisson distribution. However, waiting times need not be summed over time as in the case of counts of events. Rather each waiting time may be treated as a unit of analysis. This means that if there are N individuals in the sample and k events for each individual, there will be N · k units of observation available for analysis. Each configuration of values of the independent variables may be seen as defining a different Poisson process with its associated exponential distribution, and the procedure of treating waiting times as units will provide a set of means for these processes. The procedure thus provides meaningful quantities also with within-individual variation.

In the analysis of job shift each duration of a job was treated as an observation of the dependent variable, and this variable was then analyzed for its dependency on variables characterizing individuals and their jobs. The aforementioned hypothesis was substantiated. Straightforward OLS regression was used. This was probably not the best choice of estimation technique. A maximum likelihood procedure has been developed by Tuma that has more desirable statistical properties and also permits the use of independent variables, such as age, that vary continuously over the period of observation[11].

The proposed procedure is then to use observations on intervals of time between events to estimate expressions of the form:

$$\lambda = b_o + \sum_i b_i x_i \qquad (19)$$

and to use estimates of the b_i coefficients to make inferences about the causes of variation in the occurrence of events. The linear specification may seem a convenient choice. There is, however, one important reason for choosing a different specification. What is analyzed are rates, and they are non-negative quantities. Hence, for example:

$$\lambda = \exp\,(b_o + \sum_i b_i x_i) \qquad (20)$$

may be a better choice.

The use of waiting times gives rise to a rather intriguing problem. It will usually be the case that observations are terminated at an arbitrary point in time in relation to the process. This means that the last waiting time until an event will be interrupted by, for example, the interview. The problem is what to do with this interval. It can be shown that if all other intervals of time are exponentially distributed the truncated interval will be gamma distributed with a mean that is twice that of other intervals. Intuitively the reason for this surprising result is that longer intervals of time have a greater chance of capturing the interruption than shorter intervals. The problem does affect estimation but several solutions are available[12]. It is, incidentally, not a solution to discard the truncated intervals, as serious bias may result.

[11] Tuma, N. B., Rewards, Resources and the Rate of Mobility: a Non-Stationary Multivariate Stochastic Model, in: American Sociological Review, 39 (1976), pp. 338–360.

[12] Sørensen, Aage B., Estimating Rates from Retrospective Data, in: Heise, D. (ed.). Sociological Methodology, 1977, San Francisco 1977.

The conditional probabilities of moving from state to state on the discrete variable given that an event occurs, the m_{ij}'s, may also be subjected to causal analysis. They can be estimated from event history data by counting the number of moves from each state of origin to each state of departure on each event. Thus, in an analysis of occupational mobility using event history data each job shift will result on a move from occupational category i to category j, where i may equal j. The unit of analysis is the shift. If there are N respondents and k shifts, there will be Nx(k–1) shifts available for analysis of the variation in the m_{ij}'s. The timing of shifts, and events in general, is of course irrelevant – it is analyzed using the approach described above.

The m_{ij}'s may be analyzed using an approach proposed by Spilerman[13]. For each row and cell in the m_{ij} matrix a variable y_{ij} is defined so that $y_{ij} = 1$ if there is an entry in the ij'th cell and $y_{ij} = 0$ otherwise. For those outcomes originating in the i'th row a regression analysis with y_{ij} as the dependent variable is performed, i. e. the expression:

$$y_{ij} = a_o + \Sigma a_i x_i \qquad (21)$$

is estimated. There will be k^2 such equations with k states or categories of the discrete variable being analyzed.

Spilerman proposed the procedure for the analysis of transition probabilities in a discrete time Markov model, not for analysis of the m_{ij}'s. However, the discrete time transition probabilities estimated, for example, from panel data confounds the rate at which events occur with the outcome of events when they are estimated from a continuous time process. Though the abundance of panel data makes it tempting to treat event-history data with techniques developed for panel data the result is an inefficient use of the information contained in life histories. Direct analysis of the m_{ij}'s that govern the outcome of events is preferable.

Equation (21) is a linear probability model and the use of ordinary least squares is inefficient and the linear form is probably a misspecification. Log-linear analysis of the m_{ij}'s should be preferable.

An interesting parallel between the continuous variable and the discrete variable case should be noted. In the survey of models for change in continuous variables it was pointed out that the equilibrium state of the model for change with feedback is the simple linear model used in regression analysis of cross-sectional data to the relationship among variables. A similar result may be obtained for the discrete variable case, at least in the two state situation.

The Markov Chain will result in an equilibrium distribution if certain restrictions on the transition rates are fulfilled (corresponding to the condition $b < 0$ for equa-

[13] Spilerman, S., The Analysis of Mobility Processes by the Introduction of Independent Variables in a Markov Chain, in: American Sociological Review, 37 (1974), pp. 277–294.

tion (7) to reach an equilibrium state). The equilibrium distribution will reflect the m_{ij}'s as the rate at which events occur will determine only the speed with which equilibrium is reached. In the two state case the equilibrium distribution will be the vector $p(\infty)$ with elements $p_1(\infty)$ and $p_2(\infty)$. In terms of the m_{ij}'s these two quantities can be written as:

$$p_1(\infty) = \frac{m_{21}}{m_{12} + m_{21}}$$

and (22)

$$p_2(\infty) = \frac{m_{12}}{m_{12} + m_{21}}$$

Now, let the m_{ij}'s be log-linear functions of independent variables, that is:

$$m_{12} = \exp\,(b_o + \sum_i b_i x_i)$$
$$m_{21} = \exp\,(c_o + \sum_i c_i x_i) \qquad (23)$$

It follows, inserting (23) into (22) and taking the ratio of $p_1(\infty)$ and $p_2(\infty)$

$$\frac{p_1(\infty)}{p_2(\infty)} = \exp\,[b_o - c_o) + \sum_i (b_i - c_i)x_i] \qquad (24)$$

or

$$\log \frac{p_1(\infty)}{p_2(\infty)} = (b_o - c_o) + \sum_i (b_i - c_i)x_i \qquad (25)$$

Equation (25) is the usual form of the logit model, and if the x_i variables are dummy variables then it is just a special case of Goodman's (1972) log-linear model for odds-ratios. Hence the log-linear model for odds-ratios may be seen as the equilibrium formulation of the Markov Chain model for change in discrete variables with an exponential decomposition of the m_{ij}'s in terms of independent variables. The proof for the two-state case has previously been given by Tuma, Hannan and Groenveld, who, however, rely on the transition rates, the q_{ij}'s. of equation (11), rather than the m_{ij}'s[14]. If the m_{ij}'s are written as linear functions of independent variables, it can be shown – slightly modifying an approach suggested by Coleman[15] – that the linear probability model results.

As in the case of continuous variables the *ad hoc* statistical models that may be used to establish the relationships among discrete variables can be seen as equilibrium states of the simplest models for change. It follows conversely, that if the m_{ij}'s are being subject to log-linear analysis, it is assumed that the m_{ij}'s are in equi-

[14] Tuma, N. B., et al., Dynamic Analysis of Social Experiments, paper presented at the 1977 meetings of the American Sociological Association.
[15] Coleman, Introduction.

librium. This assumption is, however, usually more realistic than assuming that the state distribution is in equilibrium. For example, in analysis of occupational mobility the occupational distribution of a cohort will usually change with the age of the cohort as individuals form their occupational careers. Stable m_{ij}'s are consistent with such an outcome. These quantities govern the outcome of moves when they occur and may be assumed to reflect the occupational structure and be quite stable, while the rate of movement changes with age.

8. Conclusion

This paper has advocated an approach to the analysis of change in discrete or categorical variables where stochastic process models are used to identify the components of change and causal analysis of the sources of variation in these components is then carried out. The continuous time Markov Chain has been suggested as the appropriate framework for such analysis. With event-history data this framework can be utilized to analyze the rate at which events occur and outcomes of events as functions of variables assumed to be relevant to change processes.

As mentioned above, the Markov Chain is usually not able to predict the course of observed social processes very adequately. It may seem that choosing this model as a framework is an unfortunate choice. However, the failure of the model is often due to failure of the assumptions of stationarity and homogeneity. The analysis proposed here are directed at identifying and accounting for variation in parameters over time and among individuals, and thus remedy these problems with the Markov model. The choice of this model is in fact not any more unrealistic than choosing the simple model for change with feedback as the framework for causal analysis of change in continuous variables, and this model has the linear equation, used so often in causal analysis, as its equilibrium state. It has been shown that the Markov model similarly has well known statistical models for analysis of relations among discrete variables as equilibrium formulations.

The alternative to the approach here is to use the *ad hoc* statistical techniques on change data and treat the over-time vatiation in the same manner as the cross-sectional variation. This approach has merit, but if the appropriate data on change are available – event-history data – these techniques do not make efficient use of the available information on change. Event-history data permit the direct analysis of change, and a framework that identifies the components of change is needed to take advantage of this opportunity.