DESIGNING SURVEYS FOR MEASURING CHANGE IN CATEGORICAL DATA STRUCTURES OVER TIME*

by

Stephen E. Fienberg and Richard R. Picard

Technical Report No. 321

June 20, 1978

Department of Applied Statistics
School of Statistics
University of Minnesota
St. Paul, Minnesota 55108

*This work was supported in part by grant NIE-G-76-0096 from the National Institute of Education, U.S. Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the positions or policies of the National Institute of Education.

ABSTRACT

In many surveys involving the study of characteristics of a population over time, data may not be collected on all individuals at all time points in the survey. A methodology for the analysis of several such categorical data structures with loglinear models is outlined, using the two-wave, two-variable panel study as an illustration. An important outgrowth of this methodology is in evaluation of the efficiency in allocating resources in panel studies. Simulations are performed for the two-wave, two-variable panel study to provide some insight into the issues involved.

Keywords

panel studies

frequency tables with indirect observation

loglinear models

missing data

1. LONGITUDINAL STRUCTURES AND MULTIVARIATE METHODS FOR SAMPLE SURVEY DATA

Statistical methods for the design and analysis of sample survey data have traditionally been geared toward the estimation of aggregate quantities for single points in time, such as population rates or averages, and the changes in these quantities from one point in time to another. The development of optimal or efficient sampling designs for such quantities, balancing costs of various sorts and accuracy of estimation, has often led to surveys which involve the repeated measurement of the same individuals or households over time.

For example, ongoing surveys conducted by the U.S. Bureau of the Census, such as The Current Population Survey (CPS) and the National Crime Survey (NCS), are based on several panels of households that are interviewed multiple times. The panel structure is typically balanced so that for any survey period one panel is being interviewed for the last time, and is then replaced by a new panel for the next survey period. Thus in the NCS each household is interviewed 7 times, and then its panel or rotation group is dropped from the sample. The reasons for the rotation group structure are primarily related to cost efficiency relative to sample recruitment and the collection of background information. Of secondary, although not negligible importance are potential gains in efficiency of estimation for the measurement of change as a result of the correlation structure between repeated survey items (see U.S. Bureau of the Census, 1978). These surveys make no other use of the longitudinal or panel structure of the data collected.

There are other surveys where the longitudinal structure is of major importance (see e.g. the National Longitudinal Survey of Labor Force Experience (Parnes, 1973), and the Wisconsin Youth Panel (Sewell and Hauser, 1975)), and data are collected and analyzed to measure changes over time. In these surveys, individuals with a set of common characteristics such as cohort membership are followed over time, and multivariate methods are typically used in their analysis.

In both types of surveys, detailed analysis has focussed on one-dimensional quantities, especially if they are categorical in nature. There were two reasons for this focus. First was the lack of broad-based multivariate methods for categorical data analysis, a lack that has been remedied during the past decade (e.g. see Bishop, Fienberg, and Holland (1975), Fienberg (1977), Haberman (1974), and Plackett (1974)). Second was the fact that most multivariate methods, including those for measurement data, are appropriate primarily for data arising out of simple random sampling, not for the complex sample structures involving multiple levels of stratification and clustering which occur in practice. Most attempts to get around this second difficulty have focussed on the use of design effects (e.g. see Kish and Frankel (1974)), which are essentially adjustment factors used to produce sample sizes equivalent to those based on simple random samples, or the use of individual sample weights to carry out weighted analyses (see e.g. Kish and Lemeshow (1972) and Koch, Freeman, and Freeman (1975)). The justification for both these approaches for several classes of statistical models is tenuous at best. Design effects, while conceptually appealing, tend to vary from one problem to the next depending on the substantive context. The relevance of sample weights for analyses depends on the statistical models being considered, and their use

typically requires greater justification than authors have given in the past (see Porter (1972) and DuMouchel and Duncan (1977)).

The use of superpopulation models has offered yet another way around the problems of the multivariate analysis of complex sample survey data, but those models often lead to new classes of statistical problems that have not been widely explored. For categorical data situations, the recent work of Brier (1978) on models for the analysis of cluster samples is a major development, but his methods are still not directly applicable to national surveys as complex as those carried out by most major survey organizations.

We thus divide the work to be done in the development of methods for measuring change in survey structures involving categorical data over time into two parts: (i) the development of methods based on simple random sampling, and (ii) the extension of these methods the complex sample surveys. This paper is concerned with the first of these problems. The actual utility of the approach we propose will inevitably depend on its appropriateness when extensions to complex sample surveys are considered.

2. DESIGN OF LONGITUDINAL SURVEYS INVOLVING CATEGORICAL VARIABLES

In this paper we consider a methodology for the analysis of categorical data structures over time based on loglinear models. In particular we consider situations where some sample individuals have complete longitudinal records, while others have less than complete records, due to absence from the sample. For concreteness we describe our approach in the relatively simple situation involving two dichotomous variables measured at two points in time, the so-called two-wave, two-variable panel study. In this situation we can have data on individuals for both time points, for only the first, or for only the second. A sample design in this context involves specifying which combination of the three forms of data is to be used and a mechanism for selecting the individuals for the three "subsamples."

The key to our approach is how we divide the data into complete and incomplete longitudinal observations. When the only difference between the two types of observations is that for the latter some data is missing by design, then the categorical data model for the parameters of the complete data structure reduce to what Haberman (1977) refers to as a special case of product models for frequency tables involving indirect observation. We outline Haberman's approach as applied to our specific problem in Section 4, using his coordinate-free notation so that the results can be generalized to other situations with relative ease. The reduction of our problem to the product model structure follows because the likelihood function associated with our model factors into two components, one involving the parameters of the complete data structure and the other involving the assignment or sample design parameters. Thus we can examine the two parts

separately from a likelihood perspective.

For our problem we can estimate and draw inferences about parameters using all of the data (both complete and incomplete) because the sample assignment mechanism has either a known probabilistic structure, or an unknown one which we can model separately in our analysis. There is a very close analogy between the type of assignment structure assumed here, and that considered by Rubin (1978) as being "ignorable" for the purposes of deriving inferences from a Bayesian perspective for causal effects.

When the sample assignment mechanism is based on some other type of stochastic mechanism which is a function of unmeasured or unmeasurable variables, our approach is inapplicable. For example in the NCS, households that are highly victimized tend to move (and thus leave the sample) at higher rates than those who are not victimized at all. These "movers" produce incomplete longitudinal data compared with those who do not move, and the incompleteness is beyond the control of the survey designer. Great care is needed in modelling such data since the dropout mechanism is related in unknown ways not only to the key variable of interest (i.e. victimization) but also to other measured sociodemographic variables, as well as other unmeasured variables and unestimable parameters. The current mechanism used to replace household dropouts in the CPS and NCS is a form of matching, i.e., the location stays in the sample and the dropout household is replaced by a new one which moves into that location. The resulting data are not directly ameniable to analysis by the methods described here.

In practice we need to distinguish between sample assignment mechanisms that are known, and those that are unknown but separately capable of being modelled. If the mechanism is known precisely, then we can assess the fit

of our models to the data directly. If the mechanism requires modelling, then we need to have estimates of the parameters associated with the mechanism in order to assess the goodness-of-fit. We explore some of these features in the following sections.

3. THE TWO-WAVE, TWO-VARIABLE PANEL STUDY

Consider a study involving the repeated measurement on a sample of individuals of two dichotomous variables, e.g. victimication status (victim, nonvictim) and employment status (employed, unemployed). We refer to these variables as A_1 and B_1 at time 1 and A_2 and B_2 at time 2. We envision collecting data related to this study in two ways:

(i) a longitudinal sample involving data on A and B at both times 1 and 2;

(ii) supplementary cross-sectional samples, one for time 1 only and one for time 2 only.

Data from (i) form counts in the form of a 2^4 contingency table with entries $\underset{\sim}{x} = \{x_{ijk\ell}\}$ (where i and j correspond to the states of A and B at time 1, and k and ℓ to the states of A and B at time 2). These counts correspond to cell probabilities $\pi=\{\pi_{ijk\ell}\}$. The mechanism for generating the supplementary data of (ii) is as follows. Originally we have N individuals from which we can in principle get complete data. For an individual whose classification is (i,j,k,ℓ) we chose not to collect the data at time 2 with probability $\lambda_{1(i,j)}$, or not to collect the data at time 1 with probability $\lambda_{2(k,\ell)}$. Thus with probability $(1 - \lambda_{1(i,j)} - \lambda_{2(k,\ell)})$ we collect the complete longitudinal data. Note that if $\{\lambda_{1(i,j)}\}$ are to represent or include differential probabilities for individuals dropping out of the survey, then the model specifies that the dropout mechanism can depend only on the values for A_1 and B_1, and not on any other information.

The partially categorized data for time 1 form a 2×2 supplemental margin with counts $\{y_{ij}\}$, and that for time 2 a second 2×2 supplemental margin with counts $\{z_{k\ell}\}$. The result of this allocation mechanism is a 24-cell multinomial with cell probabilities and corresponding counts:

$$(1 - \lambda_{1(i,j)} - \lambda_{2(k,\ell)})\ \pi_{ijk\ell} \leftrightarrow x_{ijk\ell},$$

$$\lambda_{1(i,j)}\pi_{ij++} \leftrightarrow y_{ij}, \tag{1}$$

$$\lambda_{2(k,\ell)}\pi_{++k\ell} \leftrightarrow z_{k\ell},$$

where a "+" indicates the summation over the corresponding subscript.

The model we have just described is a special version of one first considered by Chen (1972), and adapted to two-dimensional tables in Chen and Fienberg (1974). The likelihood function for the 24-cell multinomial is proportional to

$$L \propto \prod_{ijk\ell}\left[(1-\lambda_{1(i,j)}-\lambda_{2(k,\ell)}\pi_{ijk\ell}\right]^{x_{ijk\ell}} \prod_{ij}\left[\lambda_{1(i,j)}\pi_{ij++}\right]^{y_{ij}} \prod_{k\ell}\left[\lambda_{2(k,\ell)}\pi_{++k\ell}\right]^{z_{k\ell}} \tag{2}$$

Finding maximum likelihood estimates of the $\underset{\sim}{\pi}$ and $\underset{\sim}{\lambda}$ is not difficult. The likelihood function factors, i.e. $L = f_1(\underset{\sim}{\pi})f_2(\underset{\sim}{\lambda})$, and thus the log likelihood is separable. This means we can find maximum likelihood estimates of $\underset{\sim}{\pi}$ without regard to $\underset{\sim}{\lambda}$ and vice versa.

We begin with the likelihood equations for $\underset{\sim}{\pi}$. These reduce to

$$\begin{aligned} N\hat{\pi}_{ijk\ell} &= x_{ijk\ell} + \hat{\pi}_{ijk\ell}\left(\frac{y_{ij}}{\hat{\pi}_{ij++}} + \frac{z_{k\ell}}{\hat{\pi}_{++k\ell}}\right) \\ &= \mathcal{E}(m_{ijk\ell}\,|\,\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z},\hat{\underset{\sim}{\pi}}) \end{aligned} \tag{3}$$

where $m_{ijk\ell} = N\pi_{ijk\ell}$. If we impose the usual log-linear structure $\underset{\sim}{\pi}$, i.e.

$$\log \pi_{ijk\ell} = u+u_{1(i)}+u_{2(j)}+u_{3(k)}+u_{4(\ell)}+u_{12(ij)}+ \cdots +u_{1234(ijk\ell)} \tag{4}$$

where each u-term sums to zero over each subscript, we can find maximum likelihood estimates for $\underset{\sim}{\pi}$ under various models for the u-terms in a manner

similar to the standard complete-data situation. In the latter, the likelihood equations are found by setting the minimal sufficient statistics equal to their expected values. Here a similar situation results in that we set the corresponding margin totals of $\mathcal{E}(m_{ijk\ell}|\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z},\hat{\underset{\sim}{\pi}})$ in (3) equal to their expected values. We illustrate by an example.

Suppose our model posits 1st-order interactions between A_1 and B_1, B_1 and A_2, B_1 and B_2, and A_2 and B_2, as well as a 2nd-order interaction involving B_1, A_2, and B_2. In the notation of Fienberg (1977) the minimal sufficient statistics in the complete data situation are described by $[A_1B_1]$ $[B_1A_2B_2]$, and the likelihood equations are:

$$\begin{aligned} x_{ij++} &= \hat{m}_{ij++} \quad \forall\, i,j' \\ x_{+jk\ell} &= \hat{m}_{+jk\ell} \quad \forall\, j,k,\ell. \end{aligned} \tag{5}$$

In the partially categorized data situation, the corresponding likelihood equations are

$$\begin{aligned} \sum_{k,\ell} \mathcal{E}(m_{ijk\ell}|\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z},\hat{\underset{\sim}{\pi}}) &= \hat{m}_{ij++} \quad \forall\, i,j \\ \sum_{i} \mathcal{E}(m_{ijk\ell}|\underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z},\hat{\underset{\sim}{\pi}}) &= \hat{m}_{+jk\ell} \quad \forall\, j,k,\ell\ . \end{aligned} \tag{6}$$

In general, the likelihood equations for $\underset{\sim}{\pi}$ in the partially categorized data situation must be solved iteratively. In certain cases closed form estimates do exist but, unlike the standard case, there are no simple rules for detecting when this is so. An iterative procedure to solve for $\hat{\underset{\sim}{\pi}}$ under various models for the u-terms suggested by Chen (1972) is quite similar to the iterative proportional fitting scheme used in the standard case. For the above example, the algorithm would proceed as follows:

Let $\pi_{ijk\ell}^{(0)} = 1/16$. The nth cycle of the iteration consists of the following steps:

$$\begin{aligned}
\pi_{ij++}^{(4n-3)} &= \frac{1}{N} \sum_{k,\ell} \mathcal{E}(m_{ijk\ell} | \underset{\sim}{x}, \underset{\sim}{y}, \underset{\sim}{z}, \underset{\sim}{\pi}^{(4n-4)}), && \forall i,j \\
\pi_{ijk\ell}^{(4n-2)} &= \pi_{ijk\ell}^{(4n-4)} \frac{\pi_{ij++}^{(4n-3)}}{\pi_{ij++}^{(4n-4)}}, && \forall i,j,k,\ell \\
\pi_{+jk\ell}^{(4n-1)} &= \frac{1}{N} \sum_{i} \mathcal{E}(m_{ijk\ell} | \underset{\sim}{x}, \underset{\sim}{y}, \underset{\sim}{z}, \underset{\sim}{\pi}^{(4n-2)}), && \forall j,k,\ell \\
\pi_{ijk\ell}^{(4n)} &= \pi_{ijk\ell}^{(4n-2)} \frac{\pi_{+jk\ell}^{(4n-1)}}{\pi_{+jk\ell}^{(4n-2)}}, && \forall i,j,k,\ell
\end{aligned} \tag{7}$$

This algorithm can be viewed as a simple variant on the EM-algorithm of Dempster, Laird, and Rubin (1977), and their proof of convergence is applicable here.

Next we turn to the maximization of that component of the likelihood function involving the λ-parameters. The likelihood equations are:

$$\begin{aligned}
\hat{\lambda}_{1(i,j)} &= \frac{y_{ij}}{\sum_{k,\ell} \frac{x_{ijk\ell}}{1-\hat{\lambda}_{1(i,j)}-\hat{\lambda}_{2(k,\ell)}}} \\
\hat{\lambda}_{2(k,\ell)} &= \frac{z_{k\ell}}{\sum_{i,j} \frac{x_{ijk\ell}}{1-\hat{\lambda}_{1(i,j)}-\hat{\lambda}_{2(k,\ell)}}}
\end{aligned} \tag{8}$$

As with the $\underset{\sim}{\pi}$ parameters, these equations must also be solved iteratively. An algorithm to do this is given in Chen and Fienberg (1974). If the λ's have a "nice" structure closed-form estimates can exist. In particular if we get supplemental data "at random," i.e.

$$\begin{aligned}
\lambda_{1(i,j)} &\equiv \lambda_1 && \forall i,j, \\
\lambda_{2(k,\ell)} &\equiv \lambda_2 && \forall k,\ell,
\end{aligned} \tag{9}$$

then $\hat{\lambda}_1 = y_{++}/N$ and $\hat{\lambda}_2 = z_{++}/N$.

Finally we can combine the MLE's of the λ's and $\underset{\sim}{\pi}$ as in (1) to produce estimated expected values for the 24-cell multinomial. These are needed if we are to check the goodness-of-fit of the overall model in the usual manner.

One way to compare different sample allocation mechanisms is by the precision with which we estimate the various u-terms in our loglinear model for $\underset{\sim}{\pi}$. To compute the estimated asymptotic covariance matrix of $\hat{u}$-terms for a given loglinear model, we can compute the information matrix $I(\underset{\sim}{u})$, substitute maximum likelihood estimates $\hat{\underset{\sim}{u}}$ for $\underset{\sim}{u}$, and invert, getting $\left[I(\hat{\underset{\sim}{u}})\right]^{-1}$, an estimate of the asymptotic covariance matrix of $\hat{\underset{\sim}{u}}$. Details are described in the Appendix.

There do not appear to be many shortcuts to this procedure. This is unfortunate in that computing $I(\underset{\sim}{u})$ directly, a fair amount of algebra is required - even for the simple 2^4 case. Basically the approach is to express the likelihood in terms of the u parameters directly, using the fact that

$$\pi_{ijk\ell} = \frac{e^{\pm u \; \pm u_{1(1)} \; \pm u_{2(1)} \; \pm \cdots}}{\sum_{i,j,k,\ell} e^{\pm u_{1(1)} + \cdots}}, \tag{10}$$

where the + or - is determined by (i,j,k,ℓ). (For example, if $i = 1$ and $j = 2$, we have $+u_{1(1)} - u_{2(1)} - u_{12(11)} \cdots$). The matrix of second partials can be computed directly and expectations taken, and thus $I(\underset{\sim}{u})$ can be derived. For larger problems this approach seems intractable. Haberman (1977) gives general results for asymptotic properties of $\hat{\underset{\sim}{\pi}}$ under models which include those mentioned here, but they do not appear to lead to any computational simplifications.

A computer program to do the calculations outlined in the Appendix is available from the authors.

4. ESTIMATION USING PRODUCT MODEL FORMULATION

We can recast some of the results of the previous section to utilize results on product models for frequency tables due to Haberman (1977). Even though the likelihood function in expression (2) is not exactly of product model form, when we route it in the factored form

$$L = f_1(\underset{\sim}{\pi})\ f_2(\underset{\sim}{\lambda}) \tag{11}$$

$f_1(\underset{\sim}{\pi})$ does have the product model structure. This means that we can perform the more difficult parts of the likelihood maximization using Haberman's coordinate-free framework. In particular, if we assume the simplest model for the λ-terms ("missing at random"), then we can apply Haberman's results directly. In what follows, we describe the two-wave, two-variable panel study in this fashion. Generalizations to more complex problems (e.g. more than two time points or more than two categories per variable of interest) are straightforward.

For the 2^4 table situation, there exists an underlying (partially unobserved) frequency table $\underset{\sim}{n}_{(48\times 1)}$ which consists of three independent multinomial vectors $\underset{\sim}{n}_1$, $\underset{\sim}{n}_2$, and $\underset{\sim}{n}_{3(16\times 1)}$ with sample sizes N_1, N_2, and N_3, corresponding to the "complete," "time 1 only," and "time 2 only" observations, respectively. There is a probability vector $\underset{\sim}{p}(\underset{\sim}{\pi})$ corresponding to $\underset{\sim}{n}$, the ith entry of which is of product model form:

$$p_i(\underset{\sim}{\pi}) = d_i \prod_{h\in H} \pi_h^{c(h,i)} \qquad i = 1, 2, \ldots, 48. \tag{12}$$

Here H is an index set with 16 elements (corresponding to all possible cells (i,j,k,ℓ)) and the $c(h,i)$ are either 0 or 1. For fixed i, exactly one $c(h,i)$ is 1, indicating which $\pi_{ijk\ell}$ corresponds to the

ith (possibly unobserved) cell. For the simplest model for the λ terms, we may set $d_i \equiv 1$.

The entire vector $\underset{\sim}{n}$ is not observed, but instead we see $\underset{\sim}{n}^*_{24\times 1}$, a somewhat collapsed version of $\underset{\sim}{n}$. The first 16 components of $\underset{\sim}{n}^*$ are the same as the first 16 of $\underset{\sim}{n}$ -- the "complete observations." The next 4 components of $\underset{\sim}{n}^*$ correspond to the next 16 components of $\underset{\sim}{n}$, collapsed over the time 2 responses. The final 4 components of $\underset{\sim}{n}^*$ correspond to the last 16 components of $\underset{\sim}{n}$, collapsed over the time 1 responses. Let $\underset{\sim}{p}^*(\underset{\sim}{\pi})_{24\times 1}$ be $(p_1(\underset{\sim}{\pi}), \ldots, p_{48}(\underset{\sim}{\pi}))$ collapsed in the same manner as $\underset{\sim}{n}$ is to give $\underset{\sim}{n}^*$. For

$$m_h(\underset{\sim}{\pi}|\underset{\sim}{n}^*) \qquad\qquad h = 1, \ldots, 16 ,$$

the conditional expected values of $y_h = \sum_{i=1}^{48} c(h,i)n_i$, the likelihood equations are

$$m_h(\underset{\sim}{\pi}|\underset{\sim}{n}^*) = \sum_{j=1}^{24} n_j^* \left\{ \sum_{i\in J_j} c(h,i) \frac{p_i(\underset{\sim}{\pi})}{p_j^*(\underset{\sim}{\pi})} \right\} \qquad h = 1, \ldots, 16 \qquad (13)$$

where

$$J_j = \{\text{cells in } \underset{\sim}{n} \mid \text{those cells are collapsed to give } n_j^*, \text{ the jth entry of } \underset{\sim}{n}^*\}. \qquad (14)$$

In the notation of Section 3, we can express the right hand side of expression (13) as

$$\begin{aligned} \mathcal{E}(m_{ijk\ell} \mid \underset{\sim}{x},\underset{\sim}{y},\underset{\sim}{z},\hat{\underset{\sim}{\pi}}) &= x_{ijk\ell} \cdot 1 + y_{ij} \frac{\pi_{ijk\ell}}{\pi_{ij++}} + z_{k\ell} \frac{\pi_{ijk\ell}}{\pi_{++k\ell}} \\ &= x_{ijk\ell} + \pi_{ijk\ell} \left(\frac{y_{ij}}{\pi_{ij++}} + \frac{z_{k\ell}}{\pi_{++k\ell}} \right) . \end{aligned} \qquad (15)$$

For $N_k / \sum_\ell N_\ell \to \tau_k$ as $N = \sum N_\ell \to \infty$, the asymptotic properties of $\hat{\underset{\sim}{\pi}}$ for the simplest $\underset{\sim}{\lambda}$-model depend on its asymptotic mean and covariance:

$$m(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)_{16x1} = \{m_h(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)\} = \{\lim_{N\to\infty} m_h(\underset{\sim}{\pi}|\underset{\sim}{n}^*)\} \tag{16}$$

and

$$C(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)_{16x16} = \lim_{N\to\infty} C(\underset{\sim}{\pi}|\underset{\sim}{n}^*) , \tag{17}$$

where

$$\underset{\sim}{\mu}^* = \{\tau_k \, p_j^*(\underset{\sim}{\pi}) \mid j \varepsilon A_k, \; k = 1,2,3\} , \tag{18}$$

$$A_k = \{j \mid J_j \text{ is a subset of the kth multinomial}\} , \tag{19}$$

and

$$\left[C(\underset{\sim}{\pi}|n^*)\right]_{hh'} = \sum_{j=1}^{24} n_j^* \left(\left[\sum_{i\varepsilon J_j} c(h,i)c(h',i) \frac{p_i(\underset{\sim}{\pi})}{p_j^*(\underset{\sim}{\pi})} \right] - \left[\sum_{i\varepsilon J_j} c(h,i) \frac{p_i(\underset{\sim}{\pi})}{p_j^*(\underset{\sim}{\pi})} \right] \left[\sum_{i\varepsilon J_j} c(h,i) \frac{p_i(\underset{\sim}{\pi})}{p_j^*(\underset{\sim}{\pi})} \right] \right) . \tag{20}$$

To set up the asymptotic structure we require the following notation:

$$B(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)_{16x16} = \text{diag}\{m_h(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)\} ,$$

$$\Pi^{-}(\underset{\sim}{\pi})_{16x16} = \text{diag}\{\pi_h^{-1}\} ,$$

$$E(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)_{16x16} = \Pi(\underset{\sim}{\pi}) \left[B(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*) - C(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)\right] \Pi(\underset{\sim}{\pi}) ,$$

Ξ is an affine subspace of the space R^H of functions from H to R in which $\underset{\sim}{\pi}$ is assumed to lie,

$$\Omega = \{z - w \mid z,w \varepsilon \Xi\} ,$$

$$\Omega'(\underset{\sim}{\pi}) = \{x \varepsilon \Omega \mid \pi_h = 0 \Rightarrow x_h = 0\} ,$$

$P(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)$ the projection on $\Omega'(\underset{\sim}{\pi})$ with respect to $E(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*)$,

$$\Phi(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*) = P(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*) \, (E(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*))^{-} (P(\underset{\sim}{\pi}|\underset{\sim}{\mu}^*))^{A} ,$$

where $(\,)^{-}$ denotes a generalized inverse, and $(\,)^{A}$ denotes an adjoint. Then, subject to appropriate regularity conditions, Haberman (1977) shows that

$$\sqrt{N}\,(\hat{\underset{\sim}{\pi}} - \underset{\sim}{\pi}) \xrightarrow{d} \mathcal{N}(\underset{\sim}{0},\ \Phi(\underset{\sim}{\pi}|\underset{\sim}{\mu}^{*})) \quad . \tag{21}$$

Haberman suggests solving the likelihood equations using a Newton-Raphson procedure, with a typical iteration of the form:

$$\underset{\sim}{\pi}^{(\nu+1)} = \underset{\sim}{\pi}^{(\nu)} + \Phi(\underset{\sim}{\pi}^{(\nu)}|\,\underset{\sim}{n}^{*})\ \Pi^{-}(\underset{\sim}{\pi}^{(\nu)})\ m(\underset{\sim}{\pi}^{(\nu)}|\,\underset{\sim}{n}^{*}) \quad . \tag{22}$$

Although more difficult to implement computationally than the functional iterative algorithm described in Section 3, this method can result in a savings of computer time in that it converges quadratically. A bonus is that an estimate of the large sample covariance matrix is obtained. If sample sizes are small, however, convergence is not assured.

This approach can be easily adapted to the estimation of $\underset{\sim}{\pi}$ for rotating panel structure such as that used in the surveys described in Section 1.

5. EXAMINING ALTERNATIVE SAMPLE ALLOCATION SCHEMES

A major use of the modelling results of the preceding sections is in the evaluation of alternative sample allocation schemes. For example, in the two-wave, two-variable panel study, we would like to be able to chose from among alternative allocations between longitudinal and cross-sectional data. In this section we illustrate by means of a series of examples how a choice between such alternative might be made. A more detailed study of such a choice utilizing Monte Carlo methods is beyond the scope of this paper, and we will report on it at a later time.

For the two-wave, two-variable panel study we intend initially to compare allocation schemes which involve interviews with the same number of individuals at both time points. Thus N (the sum of the total number of individuals interviewed) will vary from one allocation scheme to the next. In Table I we list three allocations to be examined here. In each case 300 individuals are to be intereviewed at each time point.

Table I goes about here

In Table II we list 6 comparisons among the three allocation schemes of Table I, based on 6 different choices of parameter values for the model:

$$\begin{aligned}\log \pi_{ijk\ell} = u &+ u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{4(\ell)} \\ &+ u_{12(ij)} + u_{13(ik)} + u_{14(i\ell)} \\ &+ u_{23(jk)} + u_{24(j\ell)} + u_{34(k\ell)}.\end{aligned}$$

For each comparison we generated a single random sample for each allocation scheme, computed the estimated u-terms and their estimated asymptotic variance matrix, assuming the true model. The $\hat{u}$-terms and the estimates of

their asymptotic variances are listed in Table II.

Table II goes about here

From the limited examination of these three allocation schemes in Table II, two features are clear.

(i) The variance estimates of $\hat{u}$-terms measuring the cross-time links (e.g. u_{13}), are monotonic functions of the number of longitudinal observations.

(ii) The variance estimates of main effects (e.g. u_1), and of same-time interactions (e.g. u_{12}) seem to vary relatively little from one scheme to the next, and from one parameter set to the next.

A cursory look at the estimated correlation matrices (not listed here) exhibit the following qualities:

(i) remarkable similarities among the matrices in both magnitude of entries and sign patterns,

(ii) off diagonal elements are primarily in the (-.4,.2) range,

(iii) in each matrix, the largest (abs. value) off-diagonal entries are almost exclusively negative,

(iv) the same time interactions tend to have small correlation with each other and larger, negative correlation with the cross time links,

(v) the correlations between $\hat{u}_{13}$ and $\hat{u}_{14}$, $\hat{u}_{13}$ and $\hat{u}_{23}$, and $\hat{u}_{14}$ and $\hat{u}_{24}$ are, on average, larger in absolute value than those between $\hat{u}_{13}$ and $\hat{u}_{24}$, and $\hat{u}_{14}$ and $\hat{u}_{23}$.

Several outstanding questions remain. For example,

(i) How much information is contained in the cross-sectional data regarding parameters involving cross-time links?

(ii) Are the asymptotic variances of the $\hat{u}$-terms for cross-time links approximately inversely proportional to the number of complete observations?

These and other questions will be explored in a Monte Carlo study.

APPENDIX: CALCULATION OF THE INFORMATION MATRIX FOR $\underset{\sim}{\pi}$

From expression (2), we have that

$$\log L = \sum_{i,j,k,\ell} x_{ijk\ell} \log \pi_{ijk\ell} + \sum_{i,j} y_{ij} \log \pi_{ij++} + \sum_{k\ell} z_{k\ell} \log \pi_{++k\ell} + \begin{matrix}\text{terms not} \\ \text{involving} \\ \underset{\sim}{\pi}\end{matrix}$$

To compute the information matrix of $u = (u_{1(1)}, u_{2(1)}, \ldots)$ we take partial derivatives with respect to each of the three terms above, or

$$\frac{\partial^2 \log L}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} = \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum x_{ijk\ell} \log \pi_{ijk\ell} + \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum y_{ij} \log \pi_{ij++} + \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum z_{k\ell} \log \pi_{++k\ell},$$

These can be evaluated directly by the use of expression (10), yielding

$$\frac{\partial^2 \log L}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} = \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum_{i,j,k,\ell} x_{ijk\ell} \log (e^{\pm u_{1(1)} \pm \cdots}) - \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum_{i,j,k,\ell} x_{ijk\ell} \log \Big(\sum_{\substack{\text{all} \\ \text{cells}}} e^{\pm u_{1(1)} \pm \cdots} \Big)$$

$$+ \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum_{i,j} y_{ij} \log \Big(\sum_{\substack{k,\ell \\ \text{fixed } ij \\ (=\pi_{ij++})}} e^{\pm u_{1(1)} \pm \cdots} \Big) - \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum_{i,j} y_{ij} \log \Big(\sum_{\substack{\text{all} \\ \text{cells}}} e^{\pm u_{1(1)} \pm \cdots} \Big)$$

$$+ \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum_{k,\ell} z_{k\ell} \log \Big(\sum_{\substack{i,j \\ \text{fixed} \\ k,\ell}} e^{\pm u_{1(1)} \pm \cdots} \Big) - \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \sum_{k,\ell} z_{k\ell} \log \Big(\sum_{\substack{\text{all} \\ \text{cells}}} e^{\pm u_{1(1)} \pm \cdots} \Big)$$

$$= \frac{\partial^2}{\partial \underset{\sim}{u}\, \partial \underset{\sim}{u}} \left\{ \left(\sum_{i,j} y_{ij} \log \Big[\sum_{\substack{k,\ell \\ \text{fixed} \\ i,j}} e^{\pm u_{1(1)} \pm \cdots} \Big] \right) + \left(\sum_{k,\ell} z_{k\ell} \log \Big[\sum_{\substack{i,j \\ \text{fixed} \\ k,\ell}} e^{\pm u_{1(1)} \pm \cdots} \Big] \right) \right.$$

$$\left. - N \log \sum_{\substack{\text{all} \\ \text{cells}}} e^{\pm u_{1(1)} \pm \cdots} \right\} \qquad \text{(A1)}$$

The information matrix is just the negative expectation of the above. We can see from (A1) how this compares to the information matrix if all N observations were completely categorized. The third term,

$$- N \log \sum_{\substack{\text{all} \\ \text{cells}}} e^{\pm u_1(1)^{\pm} \cdots} \qquad \text{(A2)}$$

corresponds to the information when all observations are completely categorized, while the first two terms represent the penalty, or loss in information, due to the partial categorizations that exist.

REFERENCES

BISHOP, Y.M.M, FIENBERG, S.E., and HOLLAND, P.W. (1975). Discrete Multivariate Analysis: Theory and Practice. M.I.T. Press, Cambridge, Mass.

BRIER, S.S. (1978). Categorical data models for complex structures. Unpublished Ph.D. dissertation, School of Statistics, University of Minnesota.

CHEN, T. (1972). Mixed-up frequencies and missing data in contingency tables. Unpublished Ph.D. dissertation, Department of Statistics, University of Chicago.

CHEN, T. and FIENBERG, S.E. (1974). Two-dimensional contingency tables with both completely and partially cross-classified data. Biometrics 30 629-642.

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

DUMOUCHEL, WM.H. and DUNCAN, G.J. (1977). Using sample survey weights to compare various linear regression models. Department of Statistics , University of Michigan, Technical Report No. 72.

FIENBERG, S.E. (1977). The Analysis of Cross-classified Categorical Data. M.I.T. Press, Cambridge, Mass.

HABERMAN, S.J. (1974). The Analysis of Frequency Data. University of Chicago Press, Chicago, Ill.

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

KISH, L. and FRANKEL, M.R. (1974). Inference from complex samples (with discussion). J. Roy. Statist. Soc. B 36 1-37.

KOCH, G.G. and LEMESHOW, S. (1972). An application of multivariate analysis to complex sample survey data, J. Amer. Statist. Assoc. 67 750-782.

KOCH, G.G., FREEMAN, D.H.,Jr., and FREEMAN, J.L. (1975). Strategies in the multivariate analysis of data from complex surveys. Int. Statist. Rev. 43 59-78.

PARNES, H.S. (1975). Sources and uses of panels of microdata -- The National Longitudinal Surveys: new vistas for labor market research. Amer. Econ. Rev. 65 244-249.

PLACKETT, R.L. (1974). The Analysis of Categorical Data. Hafner Press, New York.

PORTER, R.D. (1972). On the use of survey sample weights in the linear model. Ann. Econ. Soc. Meas. 1 141-158.

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

SEWELL, W. and HAUSER, R. (1975). Education, Occupation, and Earnings. Academic Press, New York.

U.S BUREAU OF THE CENSUS (1978). The Current Population Survey: Design and Methodology. Technical Paper 40.

TABLE I

Three Allocation Schemes Used to Examine Two-wave, Two-variable Panel Study

Allocation Scheme	No. Complete Observations	No. Observations in Each Supplemental Margin	N
1.	100	200	500
2.	150	150	450
3.	200	100	400

TABLE II

Six Comparisons Among the Allocation Schemes of Table I

(a) u-term	value of u-term	$\hat{u}$ 1	$\hat{u}$ 2	$\hat{u}$ 3	$\widehat{var}(\hat{u})$ (×100) 1	$\widehat{var}(\hat{u})$ (×100) 2	$\widehat{var}(\hat{u})$ (×100) 3
1	.2	.23	.17	.26	.53	.45	.56
2	.2	.12	.01	.03	.54	.47	.53
3	.2	.23	.22	.28	.48	.44	.45
4	.2	.22	.21	.29	.50	.43	.49
12	.2	.26	.22	.39	.45	.40	.47
13	.1	.11	.06	.13	1.27	.86	.71
14	.1	-.04	.13	-.21	1.32	.85	.79
23	.1	.09	.16	.04	1.28	.83	.69
24	.1	.18	.09	.21	1.25	.84	.70
34	.2	.13	.22	.17	.44	.44	.45

(b) u-term	value of u-term	$\hat{u}$ 1	$\hat{u}$ 2	$\hat{u}$ 3	$\widehat{var}(\hat{u})$ (×100) 1	$\widehat{var}(\hat{u})$ (×100) 2	$\widehat{var}(\hat{u})$ (×100) 3
1	.2	.23	.13	.20	.61	.63	.52
2	.2	.15	.24	.30	.63	.60	.53
3	.2	.15	.30	.11	.52	.61	.58
4	.2	.17	.20	.27	.54	.55	.57
12	.4	.57	.29	.29	.56	.55	.49
13	.1	.10	.29	.09	1.67	1.08	.78
14	.1	.04	.05	.09	1.72	1.07	.81
23	.1	.06	.16	.16	1.67	1.15	.81
24	.1	.17	.17	.14	1.66	1.04	.85
34	.3	.23	.30	.39	.44	.59	.49

TABLE II (Continued)

(c)	u-term	value of u-term	$\hat{u}$ 1	$\hat{u}$ 2	$\hat{u}$ 3	$\widehat{var}(\hat{u})$ (× 100) 1	$\widehat{var}(\hat{u})$ (× 100) 2	$\widehat{var}(\hat{u})$ (× 100) 3
	1	.2	.16	.15	.30	.64	.84	.58
	2	.2	.28	.18	.14	.62	.87	.59
	3	.2	.04	.31	.21	.83	.69	.63
	4	.2	.23	.27	.21	.75	.71	.63
	12	.5	.51	.66	.47	.59	.60	.51
	13	.1	.19	.03	.08	2.15	1.82	1.06
	14	.1	.05	.00	.11	2.24	1.86	1.05
	23	.1	.14	.03	.09	2.25	1.89	1.00
	24	.1	.09	.22	.05	2.29	1.86	1.00
	34	.6	.60	.63	.55	.62	.64	.53

(d)	u-term	value of u-term	$\hat{u}$ 1	$\hat{u}$ 2	$\hat{u}$ 3	$\widehat{var}(\hat{u})$ (× 100) 1	$\widehat{var}(\hat{u})$ (× 100) 2	$\widehat{var}(\hat{u})$ (× 100) 3
	1	.2	.18	.09	.21	.51	.50	.48
	2	.2	.09	.19	.16	.61	.50	.48
	3	.2	.26	.29	.20	.45	.45	.46
	4	.2	.16	.22	.28	.57	.45	.47
	12	.2	.19	.24	.15	.58	.42	.45
	13	.1	.19	.16	.16	1.27	.84	.63
	14	.1	.13	.03	.19	1.37	.84	.67
	23	.1	-.07	.08	.06	1.43	.89	.63
	24	.2	.37	.16	.19	1.29	.84	.65
	34	.1	.11	.12	.07	.62	.45	.47

TABLE II (Continued)

(e)	u-term	value of u-term	$\hat{u}$ 1	$\hat{u}$ 2	$\hat{u}$ 3	$\widehat{var}(\hat{u})$ (× 100) 1	$\widehat{var}(\hat{u})$ (× 100) 2	$\widehat{var}(\hat{u})$ (× 100) 3
	1	.2	.15	.23	.28	.59	.64	.56
	2	.2	.33	.06	.15	.62	.67	.53
	3	.2	.25	.29	.19	.55	.49	.44
	4	.2	.10	.28	.11	.68	.56	.53
	12	.4	.31	.40	.43	.62	.53	.54
	13	.1	.16	.12	.03	1.38	1.06	.80
	14	.1	.19	.20	.31	1.42	1.07	.78
	23	.1	.03	.08	.06	1.58	1.03	.72
	24	.3	.31	.25	.20	1.48	1.02	.75
	34	.1	.03	.10	.04	.60	.56	.47

(f)	u-term	value of u-term	$\hat{u}$ 1	$\hat{u}$ 2	$\hat{u}$ 3	$\widehat{var}(\hat{u})$ (× 100) 1	$\widehat{var}(\hat{u})$ (× 100) 2	$\widehat{var}(\hat{u})$ (× 100) 3
	1	.2	.14	.12	.08	.97	.71	.73
	2	.2	.02	.42	.25	1.48	1.67	.89
	3	.2	.28	.30	.23	.53	.56	.49
	4	.2	.42	.00	.27	.99	1.57	.69
	12	.5	.64	.57	.61	1.26	1.70	.84
	13	.1	-.14	.23	.03	2.13	1.20	.94
	14	.1	.11	.08	.15	2.63	2.14	1.12
	23	.1	.23	-.16	.21	2.51	2.32	1.18
	24	.6	.60	.92	.52	2.68	2.34	1.18
	34	.1	.15	.19	.08	1.11	1.35	.71