Latent Class Analysis of Complex Sample Survey Data: Application to Dietary Data

Blossom H. PATTERSON, C. Mitchell DAYTON, and Barry I. GRAUBARD

High fruit and vegetable intake is associated with decreased cancer risk. However, dietary recall data from national surveys suggest that, on any given day, intake falls below the recommended minima of three daily servings of vegetables and two daily servings of fruit. There is no single widely accepted measure of "usual" intake. One approach is to regard the distribution of intake as a mixture of "regular" (relatively frequent) and "nonregular" (relatively infrequent) consumers, using an indicator of whether an individual consumed the food of interest on the recall day. We use a new approach to summarizing dietary data, latent class analysis (LCA), to estimate "usual" intake of vegetables. The data consist of four 24-hour dietary recalls from the 1985 Continuing Survey of Intakes by Individuals collected from 1,028 women. Traditional LCA based on simple random sampling was extended to complex survey data by introducing sample weights into the latent class estimation algorithm and by accounting for the complex sample design through the use of jackknife standard errors. A two-class model showed that 18% do not regularly consume vegetables, compared to an unweighted estimate of 33%. Simulations showed that ignoring sample weights resulted in biased parameter estimates and that jackknife variances were slightly conservative but provided satisfactory confidence interval coverage. Using a survey-wide estimate of the design effect for variance estimation is not accurate for LCA. The methods proposed in this article are readily implemented for the analysis of complex sample survey data.

KEY WORDS: Categorical data; Cluster sample; Design effect; Dietary propensity scores; Jackknife; Latent class model; Sample weight.

1. INTRODUCTION

Frequent consumption of fruit and vegetables has been linked to reduced cancer incidence. For many cancer sites, persons with low intake of these foods experience about twice the cancer risk as do those with high intake (Block, Patterson, and Subar 1992). Dietary surveillance is used to monitor the intake of foods that are important risk factors for cancer and heart disease. A major goal of dietary surveillance is to estimate the distribution of intake of nutrients and foods in the population. At a policy level, information on dietary intake is important for shaping dietary guidance and for the evaluation of dietary intervention programs, such as the national Five A Day program, that encourages the consumption of five or more servings of fruits and vegetables daily (Subar et al. 1994). In this article we focus on vegetable consumption alone.

Twenty-four–hour dietary recall data from national surveys suggest that, on any given day, consumption falls below the recommended three or more daily servings of vegetables (Patterson, Block, Rosenberger, Pee, and Kahle 1990; Patterson, Harlan, Block, and Kahle 1995; Krebs-Smith, Cook, Subar, Cleveland, and Friday 1995). The mean of two nonconsecutive recall days from the 1994–1996 Continuing Survey of Food Intakes by Individuals (CSFII) showed that 55% of the population age 20 years and older consumed three or more servings of vegetables (U. S. Department of Agriculture 1998). A goal of *Tracking Healthy People 2010* (U. S. Department of Health and Human Services 2000) is increasing consumption to 75%. New dietary assessment methods that include estimation of the regularity of consumption are critical to measure progress toward this goal.

Although there is no clear definition of "usual" dietary intake (Guenther 1997), it can be regarded as intake over some long period. Methods currently used for measuring usual intake have been described by Thompson and Byers (1994). One method, the focus of this study, requires the collection of two or more 24-hour recalls or daily food diaries. Several methods for combining dietary records have appeared in the literature. The 1977–1978 Nationwide Food Consumption Survey (Human Nutrition Information Service 1983) estimated the percentage of individuals using a particular food as the number reporting consuming that food at least once in the 3-day survey period, divided by the group size. Hartman et al. (1990) reported mean daily intake for several food groups based on 12 two-day diaries. Popkin, Siega-Riz, and Haines (1996) summarized information on consumption of various foods from a single 24-hour recall into a dietary score for each respondent.

Analyses of dietary data have focused primarily on nutrient intake (e.g., fat, vitamin A) rather than on the intake of particular foods (e.g., butter, carrots). Nutrients are typically consumed daily in some quantity, with the result that zero intake rarely occurs. In contrast, specific foods are consumed less frequently, and zero counts are expected to occur. In the 1985 CSFII dataset, the intake of each food consumed by a respondent was reported in grams based on portion-sized estimates (U. S. Department of Agriculture 1987). Thus, for a given food, either an amount in grams or a 0 is associated with each respondent for each recall day. The distribution of intake for a given food consists of a continuous component that can be modeled by, for example, a lognormal distribution with a point mass at 0. Alternatively, the nonzero values can be classified into a set of ordered categories. The data also can be treated as

Blossom H. Patterson is Mathematical Statistician, Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892 (E-mail: *bp27z@nih.gov*). C. Mitchell Dayton is Professor, Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park, MD 20742. Barry I. Graubard is Senior Investigator, Biostatistics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD 20892. This work was part of the first author's Ph.D. dissertation in the Department of Measurement, Statistics, and Evaluation, University of Maryland, College Park. The authors thank Susan Krebs-Smith for her insightful comments on nutritional issues, Kevin Dodd for his interpretation of our model in terms of propensity scores, and the referees for their suggestions that greatly improved the manuscript.

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Applications and Case Studies
DOI 10.1198/00

counts or "mentions" of a food or foods in a food group, where a mention is any nonzero quantity. The distribution of intake can then be modeled by, for example, a Poisson distribution with overdispersion at 0 (Smith, Graubard, and Midthune —) or by other mixture models. In all of these cases, the distribution can be regarded as arising from a mixture of "regular," that is, relatively frequent consumers and "nonregular," that is, relatively infrequent consumers.

Dietary data also have been analyzed separating consumers and nonconsumers (Subar et al. 1993; Patterson et al. 1995). Such data can be further simplified by identifying a consumer of a given food (or food in a food group) with a 1 and a nonconsumer with a 0. These two approaches have the advantage of mitigating the measurement error inherent in dietary data based on portion size (Smith 1991; Young and Nestle 1995). Reporting the fact of consumption is simpler and likely to be more accurate than the quantity consumed. In fact, such measurement error may be a reason for dichotomizing the data.

Latent class analysis (LCA) is a method of grouping individuals with respect to some underlying, unobservable variable based on data from polytomous indicators or items. This method can be useful in the analysis of the intake of foods, for example, in estimating the regularity of vegetable consumption. Individuals in a sample can be classified into two or more latent classes based on binary data reflecting their consumption/nonconsumption of vegetables.

National dietary data are typically collected in surveys that have complex sample designs involving multistage sampling with sample weighting. The analysis of such designs has been described for mixture models (Wedel, ter Hofstede, and Steenkamp 1998; Patterson 1998). The focus of this study was to fit a population latent class model (LCM) to data from the 1985 CSFII and to develop appropriate LCA methods for complex sample surveys (see Patterson 1998). In Section 2 the CSFII is summarized. In Section 3 the LCM is introduced, and the jackknife is presented as a method of estimating standard errors for the LCM parameters. The CSFII data are analyzed in Section 4, and a simulation is presented in Section 5. Finally, the method and results are discussed in Section 6.

2. THE CONTINUING SURVEY OF FOOD INTAKES BY INDIVIDUALS

The 1985 CSFII comprised a multistage stratified area probability sample of women age 19–50 living in private households in the 48 conterminous states. The conterminous United States was divided into 60 "relatively homogeneous" strata, and 2 primary sampling units (PSUs) were sampled per stratum. Although the survey was designed to be self-weighting, differential sample weights were computed to reflect various levels of nonresponse at the household and individual levels. (For more details see U. S. Department of Agriculture 1987.)

In an attempt to estimate usual intake, six dietary recalls of foods consumed during the previous 24 hours were collected at about 2-month intervals. The first recall was collected in a face-to-face interview; the next five recalls were done by telephone. The public-use CSFII data tape includes all women who participated in the face-to-face interview and completed at least three phone recall interviews. For women who completed the face-to-face interview and four or five phone recall interviews, three phone recalls were randomly selected. Thus recalls 2–4 do not represent the same recall occasions for all of the women. Those women who were lost because of insufficient numbers of interviews were accounted for in the sample weights.

Table 1. Distribution of Days on Which Respondents a Respondents Reported Eating a Vegetable

Number of Days	Weighted Percent	Cumulative Weighted Percent
0	3.1	3.1
1	9.4	12.5
2	22.4	34.9
3	37.7	72.6
4	27.4	100.0

This dataset has been used as an exemplar to test various methods of analysis because it consists of four independent food records on each respondent (Krebs-Smith et al. 1990; Haines, Hungerford, Popkin, and Guilkey 1992; Nusser, Carriquiry, Dodd, and Fuller 1996). The dataset used in this analysis consists of 1,028 women who had nonzero food intake on all of the recall days. (Four women who had zero food intake on a least one of the recall days were eliminated.) Five of the strata in the dataset had a single PSU. For the purposes of variance estimation, these were paired/combined in such a way that the resulting 56 strata each contained 2 PSUs and 1 stratum contained 3 PSUs.

For each interview, a respondent was assigned a value of 1 if she reported consuming any vegetable on the recall day (i.e., one or more mentions) and a 0 otherwise. This broad group of vegetables includes salad, legumes, and such foods as peas, carrots, corn, and other green and deep-yellow vegetables, but not potatoes; this group of vegetables is of special interest because of its nutrient content.

The weighted relative distribution for the number of recall days on which sampled women reported consuming at least one vegetable is shown in Table 1. On average, respondents reported consuming at least one vegetable on 2.8 days out of 4. Approximately 73% of respondents did not consume a vegetable on at least 1 of the 4 recall days and 12.5% did so on at most 1 of the 4 days.

3. LATENT CLASS MODEL FOR SURVEY DATA

An LCM is used to explain underlying, unobservable categorical relationships, or latent structures, that characterize discrete multivariate data (Lazarsfeld and Henry 1968; Goodman 1974; Dayton and Macready 1976; Haberman 1979). When food intake is dichotomized, LCA is a technique uniquely suited to combining dietary information from several food records or 24-hour recalls to characterize the regularity of vegetable consumption of a population (as here) or population subgroup for a food or food group of interest. Here, regularity of vegetable consumption is the underlying structure of interest.

Methods for LCA that take into account sample design features, such as sample weighting, clustering, and stratification used in complex surveys like the CSFII, have not been described in the literature. However, results from regression

analysis have shown that if data are collected under a complex sampling design and simple random sampling (SRS) is assumed in the analysis, then parameter estimates can be biased and standard errors underestimated (Korn and Graubard 1999, pp. 159–172).

Let $\mathbf{Y}_i = \{y_{ij}\}$ be the vector-valued response for J survey items, $j = 1, \ldots, J$, for the ith respondent drawn from a finite population of size N. The polytomous response options take on discrete values $r = 1, \ldots R_j$ for the jth item. The probabilities, θ_l, for the unobserved LCs, c_l, $l = 1, \ldots, L$, are called *LC proportions*. Item-conditional probabilities, $\alpha_{lj1} \ldots \alpha_{ljR_j}$, represent the probabilities of response r to item j given membership in LC l. Thus for each item j there is an R_j-vector of conditional probabilities. To illustrate the notation, consider data based on four polytomous variables with, say, $R_1 = 2, R_2 = 4, R_3 = 3$, and $R_4 = 2$ denoting the number of discrete values taken on by each item. $\mathbf{Y}_i = (y_{i1}, y_{i2}, y_{i3}, y_{i4})' = (1, 3, 2, 2)'$ might represent the responses for the ith respondent.

The traditional LCM can be defined as

$$\Pr(\mathbf{Y}_i \mid c_l) = \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{ljr}^{\delta_{ijr}} \tag{1}$$

and

$$\Pr(\mathbf{Y}_i) = \sum_{l=1}^{L} \theta_l \Pr(\mathbf{Y}_i \mid c_l), \tag{2}$$

where the Kronecker delta is defined as

$$\delta_{ijr} = \begin{cases} 1 & \text{iff } y_{ij} = r, \qquad r = 1, \ldots, R_j \\ 0 & \text{otherwise.} \end{cases}$$

The usual restrictions for item-conditional probabilities apply (i.e., $\sum_{r=1}^{R_j} \alpha_{ljr} = 1 \ \forall j$) and, in addition, the LC proportions sum to 1 (i.e., $\sum_{l=1}^{L} \theta_l = 1$). Note that, unlike models proposed by Clogg and Goodman (1984, 1985), the model in (1) and (2) does not directly reflect grouping of respondents, for example, males and females.

In the context of the CSFII dataset, the response variables, y_{ij}, are (dichotomous) indicators of consumption of some food of interest on each of 4 recall days. For a two-class model, the LC proportions refer to the proportions in "regular" and "nonregular" vegetable consumption groups. Each item-conditional probability refers to the probability or "propensity" for consuming at least one vegetable on the corresponding recall day, given membership in a specific consumption group.

Assuming SRS with a sample of size n, the log-likelihood is

$$\Lambda = \sum_{i=1}^{n} \ln \sum_{l=1}^{L} \theta_l \Pr(Y_i \mid c_l) = \sum_{i=1}^{n} \ln \left\{ \sum_{l=1}^{L} \theta_l \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{ljr}^{\delta_{ijr}} \right\}. \tag{3}$$

Fundamental to classical LCA is the assumption that the observed variables are independent within LCs. Parameter estimation can be accomplished by means of maximum likelihood methods using conventional iterative algorithms such as Newton–Raphson or the EM algorithm (Dempster, Laird, and Rubin 1977; Heinen 1996).

In the case of non-SRS, where sample weights, w_i, are available for each respondent (e.g., from a public-use data tape), these weights are usually the product of the reciprocal of the sample inclusion probability, a factor that adjusts for nonresponse, and a factor that reflects poststratification adjustment. These weights may be expansion weights that sum to the total population size or relative weights that are scaled to sum to the sample size. A sample-weighted pseudo–log-likelihood can be defined as

$$\begin{aligned} \Lambda_w &= \sum_{i=1}^{n} w_i \ln \sum_{l=1}^{L} \theta_l \Pr(Y_i \mid c_l) \\ &= \sum_{i=1}^{n} w_i \ln \left\{ \sum_{l=1}^{L} \theta_l \prod_{j=1}^{J} \prod_{r=1}^{R_j} \alpha_{ljr}^{\delta_{ijr}} \right\}. \end{aligned} \tag{4}$$

Maximizing the pseudo–log-likelihood simultaneously with respect to θ_l and α_{ljr} provides design-consistent estimates of the underlying population parameters (Pfeffermann 1993).

The LC model used in this article can be expressed as a log-linear model, using either a Poisson or binomial distribution for the cell counts in the finite population. In the survey setting, the weighted pseudo-likelihood is obtained by replacing the unweighted cell counts with the sample-weighted cell counts in the likelihood, as implied by (4). An alternative method for log-linear analysis of sample weighted contingency tables (Clogg and Eliason 1987; Agresti 1990, p. 199) uses the unweighted cell counts with an offset, consisting of the log of the inverse of the average cell sample weight, in each cell of the contingency table. Under a correctly specified log-linear model for the population, this method will produce consistent estimates of model parameters, that is, estimates that are asymptotically equal to the values that would have been obtained had they been computed using the entire finite population. However, if the log-linear model for the population is misspecified, then the two methods will not agree asymptotically. We prefer the weighted pseudolikelihood method because its estimates will be approximately unbiased for values of the population model parameters, regardless of whether the model was correctly specified.

Although not explicit in the models as written, clustering is taken into account when estimating standard errors. Cluster sampling, such as that in the CSFII, induces correlation among responses and typically results in sampling variances that are larger than would be the case under SRS. Further, standard test statistics (such as the Pearson chi-squared) used in LCA are no longer asymptotically distributed as chi-squared random variables when the data arise from a survey with clustered sampling (Hidiroglou and Rao 1987; Roberts, Rao, and Kumar 1987).

Two methods of calculating standard errors for complex sample survey data are adjustment using a design effect (deff), and the use of a replication method such as the jackknife. The first method was used in LCA by Haertel (1984a, 1984b, 1989). The jackknife is applicable to virtually any type of complex sample design (Wolter 1985) and is known to provide reasonable standard errors for many statistics that are smooth (differentiable) functions of the data (Efron 1982). The applicability of the jackknife to the estimation of LC parameters under SRS has been studied empirically by Bolesta (1998).

In complex sample surveys, sample weights and clustering usually inflate the variance, whereas stratification may result in variance reduction. The deff, the ratio of the variance under the full design to the variance assuming SRS, is usually greater than 1. Kish (1965, pp. 258–259) described this "comprehensive factor" as attempting to summarize the effects of "various complexities in the sample design, especially those of clustering and stratification. . . (and) may include effects of. . . varied sampling fractions." He noted that the design effect can be used to obtain an effective sample size, $n' = \frac{n}{deff}$, to be used in place of n, the actual sample size, in the calculation of standard errors. The size of the deff depends on the variable being estimated and may vary among subsets of the population. If a surveywide estimate of the deff is available and applied to all estimates, then the adjustment may be too large or too small and also may give misleading results for population subgroups (Korn and Graubard 1995). Haertel (1984a, 1984b, 1989) took the sample design into account in the calculation of standard errors in LCA by using an external estimate of the overall deff to estimate an effective sample size, which he then used in the calculation of standard errors.

The jackknife was introduced as a method of bias reduction by Quenouille (1949), and the procedure was subsequently used to estimate the variance of a parameter estimate (Mosteller and Tukey 1968; Miller 1968, 1974). Frankel (1971) made an early application of this technique to complex sample survey data. The method proceeds as follows. Let $\hat{\gamma}$ be the sample-weighted estimate of a population parameter of interest, γ, for a sample of size n. In a complex sample survey with stratification and clustering, the PSUs are randomly grouped within strata, where each random group has approximately the same number of PSUs. Let k_h denote the number of random groups in stratum $h, h = 1, \ldots, H$. A random group of PSUs in stratum h is omitted, and the remaining observations in that stratum are reweighted by a multiplicative factor $k_h/(k_h - 1)$. The usual parameter estimates, called *jackknife estimates*, are derived from the reduced sample. This process is repeated sequentially for the entire sample of PSUs. A variance estimator based on the jackknife is (Wolter 1985)

$$\widehat{\text{var}}^J(\hat{\gamma}) = \sum_{h=1}^{H}\left[\sum_{s=1}^{k_h}\frac{k_h - 1}{k_h}(\hat{\gamma}_{(sh)} - \hat{\gamma})^2\right], \tag{5}$$

where $\hat{\gamma}_{(sh)}$ is the jackknife estimate of γ omitting group s in stratum h. Alternatively, $\hat{\gamma}$ may be replaced by the mean of the jackknife estimates, $\hat{\gamma}^J = \sum_{h=1}^{H}\sum_{s=1}^{k_h}\hat{\gamma}_{(sh)}/\sum_{h=1}^{H}k_h$. The foregoing procedure also can be used without grouping the PSUs, treating each PSU as a group of size 1.

In general, resampling methods are applied to PSUs without attention to the form of subsampling within the PSUs. This convenient feature is justified by the fact that when the first-stage sampling fraction remains low (<10% for practical purposes), the standard error may be accurately estimated from the variation between PSU totals. The contribution from second and later stage variances is reflected in the sampling error estimated from the PSUs (Lee, Forthofer, and Lorimor 1986). In addition, jackknife variance estimation correctly estimates the component of variance due to sample weighting.

We are unaware of any commercial LC software appropriate for analyzing complex sample survey data. Weighted estimates of LC parameters are provided by the computer packages LEM (Vermunt 1997) and Latent Gold (Vermunt and Magidson 2000); however, these programs do not provide correct estimates for the standard errors for surveys with stratification and clustering. For the current study, GAUSS (version 3.5) (Aptech Systems, Inc. 1997) programming code was written to perform the LCA and the jackknife and verified for traditional LCMs before being applied to complex survey data.

4. ANALYSIS OF DATA FROM THE CONTINUING SURVEY OF FOOD INTAKES BY INDIVIDUALS

We fit a two-class LCM to the CSFII data taking sample weights into account. Three-class models were not assessed for these data, because the unrestricted three-class model is not identified for four variables (Lindsay, Clogg, and Grego 1991). As shown in Table 2, $\hat{\theta}$, the proportion in the first latent class (LC1) is estimated to comprise 18% of the population. LC1 can be interpreted as consisting of "nonregular," or infrequent, vegetable eaters, that is, those who do not consume vegetables on a regular (daily) basis. The second latent class (LC2), comprising 82% of the population, can be interpreted as consisting of those individuals who consume at least one vegetable as more or less a regular (daily) practice. In LC1, estimates of the item-conditional probabilities for vegetable consumption on a given recall day, $\hat{\alpha}_{1j}$, were variable, ranging from .28 to .46 for vegetable consumption on the jth day, whereas in LC2 these probabilities, $\hat{\alpha}_{2j}$, were similar and consistently higher, ranging from .73 to .78 (see Table 2). Note that we drop the middle subscript (r) for the item conditional probabilities because the responses have only two levels. The jackknife standard error of the LC proportion, .13, was relatively large, as were jackknife standard errors for the item-conditional probabilities in LC1.

In general, the larger the LC, the more observations it represents and the smaller the variability in the estimates for the item-conditional probabilities for that class. We calculated estimates of standard errors based on SRS using a weighted Fisher information based on the (weighted) pseudolikelihood,

Table 2. Latent Class Analysis of Vegetable Consumption Habits: 1985 Continuing Survey of Food Intakes by Individuals

Weighted data			*Unweighted data*		
Estimate	*Mean of jackknife estimates*	*Jackknife standard error*	*Estimate*	*Mean of jackknife estimates*	*Jackknife standard error*
.178	.179	.128	.331	.332	.137
.456	.456	.200	.604	.604	.078
.391	.390	.227	.510	.510	.094
.275	.276	.113	.396	.397	.082
.392	.392	.148	.464	.464	.074
.781	.781	.021	.800	.801	.019
.764	.764	.030	.818	.818	.034
.766	.766	.069	.810	.811	.065
.729	.730	.040	.787	.787	.046

where the weights were normalized to the sample size. These were about one-half the size of the jackknife standard errors. For the LC proportion, the standard error from the Fisher Information was .07, compared to .13 from the jackknife; this translated into a deff of about 4. Deffs for the conditional probabilities ranged from about 1 to 4 (data not shown).

The Akaike information criterion (AIC) has been used to assess goodness of fit for LCMs (Lin and Dayton 1994), but has not been modified for complex sample survey data. We used a Wald test to test goodness of fit for our two-class model, because this test can be adapted to sample survey data by using a design-based estimate of the variance matrix. The Wald test statistic is the quadratic form $W = d'V^{-1}d$, where d is a 15H1 vector of the differences between the observed and expected cell proportions for 15 of the 16 possible outcome cells and V is the estimated variance matrix for d. The jackknife was used to compute V. We compared $W \times (57-15+1)/(57 \times 15)$, where 57 is the number of PSUs (114) minus the number of strata (57), to an F distribution with 15 and 43 degrees of freedom. (See Korn and Graubard 1999, pp. 91–93, for a discussion of Wald tests.) For the two-class model, the test statistic was .72 ($p = .75$), indicating that the model fits the data satisfactorily. To assess bias in parameter estimates incurred by ignoring sample weights, we fit an unweighted two-class model to the data (see Table 2). For the unweighted data, the estimated proportion falling in LC1 was .33 as opposed to .18 for weighted data. Differences for the conditional probabilities were smaller. Overall, the variances tended to be greater when the weights were used than when they were ignored.

We used Wald tests for the difference between the weighted and unweighted estimates to assess whether the sample weights were informative. Because weighted analyses tend to increase the variance of estimated parameters, these tests are known to have low power. Testing the 8 item-conditional probabilities, the F value for the Wald test with 8 and 57 degrees of freedom was 1.02 ($p = .49$). Testing only the LC proportion, the F value for the Wald test with 1 and 57 degrees of freedom was 2.01 ($p = .16$). The results of this analysis suggest that the weights may not be informative.

The U. S. Department of Agriculture computed a single estimated overall deff of 1.43 for analyzing the 4 days of records for the 1985 CSFII. It was computed as $1+\{cv(wts)^2\}$ (Joseph Goldman, personal communication), where $cv(wts)$ is the coefficient of variation of the sampling weights. This deff takes into account the variability associated with the weights, but not the effects of clustering or stratification.

As we had decided to retain the weights, we were interested in obtaining an estimate of the deff due to clustering and stratification alone, apart from that due to the weights. For the sample of 1,028 women, we generated a vector of 1,028 uniform random numbers, each number associated with an observation. Next, we randomly regrouped the response vectors into clusters retaining the original cluster sizes. We then fit the reordered data to a two-class model and used the resulting variances to estimate the deff for each parameter estimate. The deff was estimated as .97 for the LC proportion and ranged from .98 to 1.17 for the item-conditional probabilities (Table 3). These effects were modest compared to the deffs that incorporate weighting as well as clustering, indicating that most of the increase in variance was due to the sample weights.

Table 3. Latent Class Analysis of Vegetable Consumption Habits: 1985 Continuing Survey of Food Intakes by Individuals Weighted Data, Clusters Broken by Random Reordering

Parameter	Estimate	Mean of jackknife estimates	Jackknife variance estimate, without clusters	Jackknife variance estimate, with clusters	Ratio of variances with:without clusters
θ	.178	.179	.017	.016	.972
α_{11}	.456	.455	.041	.040	.976
α_{12}	.391	.390	.050	.052	1.027
α_{13}	.275	.276	.011	.013	1.157
α_{14}	.392	.391	.020	.022	1.111
α_{21}	.781	.781	0	0	.980
α_{22}	.764	.764	.001	.001	1.095
α_{23}	.766	.766	.005	.005	1.012
α_{24}	.729	.730	.001	.002	1.122

5. SIMULATION

We performed a simulation to investigate the validity of the methods used for taking weights and clustering into account for the CSFII data and to assess the accuracy of the jackknife standard errors. This simulation was based on numbers of strata (i.e., 60) and PSUs (i.e., 2 per stratum) similar to those in the CSFII. The size of the PSUs in the simulation was set at 8, the average PSU size in the CSFII. For simplicity, all PSUs were of equal size and the sample size was set at 960, a multiple of 8 and similar to the CSFII sample size. A population with an underlying two-class structure was simulated. We drew the LC proportions for 30 of the strata were drawn from a beta distribution $\beta(1, 9)$, with mean .1, (i.e., $\theta_1 = .1$), and drew the proportions for the remaining 30 strata from a beta distribution $\beta(3, 7)$, with mean .3 (i.e., $\theta_2 = .3$), so that the proportion in LC1 in the overall simulated population, .2, was close to .18, as estimated in the two-class solution for the CSFII data. We randomly generated values of the LC proportions from these beta distributions, inducing intracluster correlations within PSUs. We selected the beta distribution because it is a flexible two-parameter distribution (scale and location parameters), has values lying in the [0, 1] interval, and is the natural conjugate prior distribution for the binomial distribution. In theory, the intraclass correlation coefficient for a beta distribution with parameters (ν, ω) is $(\nu + \omega + 1)^{-1}$ (Brier 1980). We set the item-conditional probabilities at .2 for LC1 and .7 for LC2 to approximate the CSFII values.

A plot of the sample weights from the CSFII suggests that they are approximately lognormal in distribution. We used moments of the empirical distribution of the weights to define a lognormal distribution with a median of .84 and a variance of .616, and generated sample weights for the observations in the simulation from this distribution. The simulation can be viewed as a series of one-stage cluster samples where each cluster consists of b observations and where the LC proportions vary by cluster within each stratum.

Table 4. Estimated True and Jackknife Variances for Simulation

Parameter	Estimated true variance	Jackknife estimate of variance	Ratio of jackknife:true variance
θ	.00202	.00217	1.074
α_{11}	.00518	.00534	1.031
α_{12}	.00507	.0054	1.065
α_{13}	.00512	.00532	1.039
α_{14}	.00499	.00532	1.066
Mean	.00509	.00535	1.050
α_{21}	.00076	.00078	1.026
α_{22}	.00077	.00078	1.013
α_{23}	.00075	.00077	1.027
α_{24}	.00077	.00079	1.026
Mean	.00076	.00078	1.023

Table 5. 95% Confidence Interval Coverage for Simulation, $n = 960$, $\theta_1 \sim \beta(1, 9)$, $\theta_2 \sim \beta(3, 7)$, Lognormal Weights

	Proportion in lower tail	Proportion in upper tail	Coverage
θ	0	.060	.940
α_{11}	.025	.042	.933
α_{12}	.023	.037	.940
α_{13}	.034	.042	.924
α_{14}	.024	.032	.944
Mean	.027	.038	.935
α_{21}	.014	.028	.958
α_{22}	.017	.026	.957
α_{23}	.016	.030	.954
α_{24}	.028	.026	.946
Mean	.019	.028	.954

To investigate the effect of clustering on the jackknifed variance, we generated clustered data from the aforementioned population, and estimated the variance taking the clustering into account. We then calculated the jackknife variance for a sample from the same population constructed using the reordering method described in the previous section in the discussion of the deff. We randomly regrouped observations into clusters of the same size and, using these clusters as PSUs, estimated jackknife variances.

The code for the simulations was written in the matrix language, GAUSS, version 3.5, and the EM algorithm was used to estimate model parameters. The programming criteria used in the simulation were (1) 1,000 replications, (2) convergence criterion of 10^{-6}, and (3) maximum number of 500 iterations allowed to achieve convergence in the LCA algorithm (nonconverging cases were replaced in the simulation).

To assess the validity of the jackknife variances from the simulations, we generated proxy population variances by calculating mean squared errors for the parameter estimates based on 10,000 replications using the same parameter values as in the simulations. The ratio of the jackknife variance estimate to the corresponding proxy variance was taken as a measure of the accuracy of the jackknife estimate. A 95% two-tailed confidence interval (CI) was calculated for the simulation parameter values as

$$CI = \left\{ \hat{\theta} - t_{\frac{\alpha}{2}, df} \sqrt{\widehat{\text{var}}^J}, \hat{\theta} + t_{\frac{\alpha}{2}, df} \sqrt{\widehat{\text{var}}^J} \right\}, \tag{7}$$

where $\hat{\theta}$ can be either the LC proportion or an item-conditional probability, α is the type I error rate, df is the number of (jackknifed) groups minus the number of strata, and $\widehat{\text{var}}^J$ is the jackknife variance estimate.

As expected (Kish and Frankel 1974), for all parameters, the jackknifed variances modestly overestimated the proxy variances (Table 4). Estimates for the item-conditional probabilities were within 7% of the proxy variances for the smaller LC and within 3% for the larger class. The jackknife overestimated the variance of the LC proportion by 7%. As shown in Table 5, coverage was close to the nominal .95 level for all parameters.

6. DISCUSSION

Multiple dietary records of food intake typically have been summarized by means and proportions. LCA is a new method of combining records to group respondents into categories, or classes, that define patterns of food consumption and provide estimates of class size. We fit an unconstrained model because of the possibility that seasonality or other variables might affect vegetable consumption over the course of the survey year. Fitting a two-class model, we found that about 18% of the population of women age 19–50 consumed a diet deficient in vegetables in that they did not make consumption of these foods a regular practice. LCA also provides estimates of the item-conditional probabilities (class-specific propensity scores). There was a suggestion that respondents tended to be more likely to report consuming at least one vegetable on the first survey day than on later recall days. Because vegetable consumption is advocated as part of a good diet, respondents may have been more likely to report eating a vegetable in the face-to-face interview than when queried by telephone. Although the similarity of the item-conditional probabilities for recalls 2–4, especially for LC2, suggested that a model restricting these probabilities to be equal might be appropriate, we rejected this course because it would have been a post hoc analysis. The similarity of the item-conditional probabilities over the 4 recall days for LC2 suggested a stable propensity to consume vegetables. This was not true for LC1.

In this study, we used LCA to estimate the proportion of women age 19–50 that consume vegetables on a regular basis, a different objective than estimating the number of servings per day as in some other types of analysis. LCA requires only indicators of consumption and can lead to data reduction in some datasets. Thus LCA can be readily performed on data that otherwise may require a multiple-step, perhaps lengthy, analysis involving transformations and distributional assumptions. For example, Nusser et al. (1996) proposed a complex multistep procedure for estimating the distribution of nutrient intake. Finally, LCA provides a new way to describe "usual" dietary intake and to estimate the number and size of subgroups that display different food consumption patterns. Such analyses may be useful in developing public health intervention programs.

There has been a long-standing debate in the statistical literature on whether to do weighted or unweighted analysis (i.e., design-based or model-based analysis) of survey data (Brewer and Mellor 1973; Smith 1976, 1984; Hansen, Madow, and Tepping 1983; Fienberg 1989; Kalton 1989; Korn and Graubard 1995a, 1995b). It is well known that using sample weights will result in approximately unbiased or consistent estimates for population parameter values, but may increase the variances of these estimates, whereas an unweighted analysis may result in biased or inconsistent estimates, but smaller variances. We have described a weighted analysis that uses weighted pseudolikelihood estimation, and applied this method to the analysis of the CSFII data. Consistency of weighted estimates is maintained regardless of whether the posited model is correctly specified. In contrast, unweighted estimates depend on the sample weighting of a particular sample design when the sample weights are informative for the analysis of interest (Pfefferman 1993). Issues to be considered when choosing a weighted versus an unweighted analysis are (1) the purpose of the analysis—analytical versus descriptive; (2) the magnitude of the inefficiency that would result from a weighted analysis if the weighting were unnecessary to correct for bias and whether this inefficiency is small relative to the effect being estimated; (3) the expected bias from an unweighted analysis; and (4) whether sufficient information is known about the sample design and whether variables are available to model the sample design in an unweighted analysis (Korn and Graubard 1999, chap. 4).

For LC modeling, sample weighting can affect the estimation of the item-conditional probabilities, the LC proportions, or both when sampling rates differ across subgroups. Consider an unstratified analysis of a population comprising two subgroups (i.e., a single LC model be fitted to the entire population), where both subgroups have the same number of underlying latent classes but are sampled at different rates. If the LC proportions differ between subgroups, then sample-weighted estimates of the LC proportions will differ from unweighted estimates. If the item-conditional probabilities are homogeneous across subgroups, then the weighted and unweighted estimates of the item-conditional probabilities should be approximately the same, whereas the LC proportions could differ. If these probabilities differ between subgroups, then again weighted and unweighted estimates can differ. If the analysis is stratified so that a separate LCM is fitted to each subgroup, then weighting is no longer necessary. However, stratifying among all population subgroups is rarely feasible.

The CSFII data analysis was a descriptive analysis that used LC modeling without covariates. The objective of the analysis was to obtain unbiased estimates of the LC proportions and item-specific probabilities for the target population. Following the recommendations of Korn and Graubard (1999, pp. 180–182) we used a weighted analysis for this descriptive study. For an analytical study, the trade-off between variance and bias must be carefully considered. An analyst choosing to use unweighted analysis because of large inefficiency due to the weighting should adjust for the sample weighting by including in the analytic model those sample design variables used in determining the sample weighting (Korn and Graubard 1999). Regardless of the type of analysis done, model adequacy should be assessed using diagnostic methods, as we have tried to do here.

Analyses of the CSFII data and the simulations done with and without sample weights demonstrated both the possibility of incurring unacceptable bias by ignoring the weights and the potential increase in variance arising from including them unnecessarily. The CSFII design is described as self-weighting, although weights were used to adjust for eligibility within the household and for nonresponse. The self-weighting aspect of the design might lend support to the notion that the weights could be ignored, although this is not obvious for the present analysis. The Wald test comparing the weighted estimates to the unweighted estimates showed a larger, but not significant, effect on the LC proportions than on the item-conditional probabilities. This test has low power, however.

The jackknife is an easily applied method for obtaining empirical variance estimates for an LCM applied to complex sample survey data. Our simulation suggested that the jackknife standard errors slightly overestimate the actual standard errors. Despite this overestimation, these estimates seem sufficient for most practical applications. However, it may be worthwhile to investigate other resampling methods, such as the bootstrap or modifications to the jackknife. Another proposed approach uses linearization variances based on Taylor series approximations of the estimating equations from the sample weighted pseudolikelihood (Wedel et al. 1998). This approach is less flexible in that it requires developing new software (e.g., for the calculation of second derivatives for each term in the model for each model considered.)

Another approach to analyzing clustered sampled data is using hierarchical modeling with random effects to model the correlation at each stage of cluster sampling. The use of random-effects models applied to survey data is an area of current research with no well-established methods, even in the case of linear models (Korn and Graubard 1998; Pfeffermann, Skinner, Holmes, Goldstein, and Rasbash 1998). This approach is difficult to implement because it requires knowledge of all levels of clustering, which is often unavailable on public use files because of confidentiality concerns. The approach that we have taken, (weighted) pseudolikelihood with design-based jackknife variance estimation, is commonly used to analyze survey data with complex sample designs (Skinner, Holt, and Smith 1989; Korn and Graubard 1999, p. 101).

We do not recommend inflating the variance by an overall survey deff, as done by Haertel (1984a, 1984b, 1989). First, we found that the jackknife standard errors, which take the sample design fully into account, were about twice as large as standard errors based on Fisher information for a sample-weighted likelihood; this difference translates into deffs of approximately 4. These very large deffs were due primarily to the effects of sample weights, with only modest effects due to clustering and stratification. These deffs varied by parameter and were larger than the overall deff of 1.43 estimated by the U. S. Department of Agriculture.

National surveys such as the CSFII, the National Health and Nutrition Examination Survey, and the National Health Interview Surveys are major sources of information on dietary

practices in the general population and in demographic subgroups. In the past, LCA has not been applied to these data. With the development of methods to accommodate weighted and clustered data, LCA can be used to describe food consumption patterns in the whole population, as well as in subsets of interest.

The 1994–1996 CSFII collected only two 24-hour recalls. LCA can be applied to surveys such as this by introducing two or more latent variables, such as separate latent indicators of fruit and of vegetable intake, and fitting models with two or more classes. These may be independent or correlated, as discussed by Hagennars (1990). Alternatively, multiple group analyses can be performed, where the groups relate, to say, sex, to race, or to some other classification variable (Dayton 1999).

An area of future research is the development of goodness-of-fit test statistics for LC models for survey data. Although sample weights might be readily incorporated into statistics based on the log-likelihood, the distribution of test statistics such as the AIC must be modified to take into account clustering or stratification.

[Received Xxxxx xxxx. Revised Xxxxx xxxx.]

REFERENCES

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

Aptech Systems, Inc. (1997), *The GAUSS System* (Version 3.5), Maple Valley, WA: author.

Block, G., Patterson, B. H., and Subar, A. (1992), "Fruit, Vegetables, and Cancer Prevention: A Review of the Epidemiological Evidence," *Nutrition and Cancer*, 18, 1029.

Bolesta, M. S. (1998), "Comparison of Standard Errors Within a Latent Class Framework Using Resampling and Newton Techniques," doctoral dissertation, University of Maryland, College Park.

Brewer, K. R. W., and Mellor, R. W. (1973), *The Australian Journal of Statistics*, 15, 145–152.

Brier, S. S. (1980), "Analysis of Contingency Tables Under Cluster Sampling," *Biometrika*, 67, 591–596.

Clogg, C. C., and Eliason, S. R. (1987), "Some Common Problems in Log-Linear Analysis," *Sociological Methods & Research*, 16, 8–44.

Clogg, C. C., and Goodman, L. A. (1984), "Latent Structure Analysis of a Set of Multidimensional Tables," *Journal of the American Statistical Association*, 79, 762–771.

——— (1985), "Simultaneous Latent Structure Analysis in Several Groups," in *Sociological Methodology 1985*, ed. N. B. Tuma, San Francisco: Jossey-Bass.

Dayton, C. M. (1999), *Latent Class Scaling Analysis*, Thousand Oaks, CA: Sage.

Dayton, C. M., and Macready, G. B. (1976), "A Probabilistic Model for Validation of Behavioral Hierarchies," *Psychometrika*, 41, 189–204.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood Estimation From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, 39, 1–22.

Efron, B. (1982), *The Jackknife, the Bootstrap and Other Resampling Methods*, Philadelphia: Society for Industrial and Applied Mathematics.

Fienberg, S. E. (1989), "Modeling Considerations: Discussion From a Modeling Prospective," in *Panel Surveys*, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh. NewYork: Wiley, pp. 566–574.

Frankel, M. R. (1971), *Inference From Survey Samples*, Ann Arbor, MI: Institute for Social Research, University of Michigan.

Goodman, L. A. (1974), "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models," *Biometrika*, 61, 215–231.

Guenther, P. M. (1997), "Development of an Approach for Estimating Usual Nutrient Intake Distributions at the Population Level," *Journal of Nutrition*, 127, 1106–1112.

Haberman, S. J. (1979), *Analysis of Quantitative Data*, Vol. 2, New York: Academic Press.

Haertel, E. (1984a), "An Application of Latent Class Models to Assessment Data," *Applied Psychological Measurement*, 8, 333–346.

——— (1984b), "Detection of a Skill Dichotomy Using Standardized Achievement Test Items," *Journal of Educational Measurement*, 21, 59–72.

——— (1989), "Using Restricted Latent Class Models to Map the Skill Structure of Achievement Items," *Journal of Educational Measurement*, 26, 301–321.

Hagennars, J. A. (1990), *Categorical Longitudinal Data*, Newbury Park, CA: Sage.

Haines, P. S., Hungerford, D. W., Popkin, B. M., and Guilkey, D. K. (1992), "Eating Patterns and Energy and Nutrient Intakes of U. S. Women," *Journal of the American Dietetic Association*, 92, 698–704, 707.

Hansen, M. H., Madow, W. N., and Tepping, B. J. (1983), "An Evaluation of Model Dependent and Probability-Sampling Inferences in Sample Surveys," *Journal of the American Statistical Society*, 78, 776–793.

Hartman, A. M., Brown, C. C., Palmgren, P. P., Verkasalo, M., Myer, D., and Virtamo, J. (1990), "Variability in Nutrient and Food Intakes Among Older Middle-Aged Men: Implications for Design of Epidemiologic and Validation Studies Using Food Recording," *American Journal of Epidemiology*, 132, 999–1012.

Heinen, T. (1996), *Latent Class and Discrete Trait Models*, Advanced Quantitative Techniques in the Social Sciences Series, Vol. 6, Thousand Oaks, CA: Sage.

Hidiroglou, M. A., and Rao, J. N. K. (1996), "Chi-Squared Tests With Categorical Data From Complex Surveys: Part I," *Journal of Official Statistics*, 3, 117–132.

Human Nutrition Information Service (1983), "Food Intakes: Individuals in 48 States, Year 1977–78, Nationwide Food Consumption Survey 1977–78," Report I-1, U. S. Department of Agriculture.

Kalton G. (1989), "Modeling Considerations: Discussion From a Survey Sampling Perspective," in *Panel Survey*, eds. D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, New York: Wiley, pp. 575–585.

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

Kish, L., and Frankel, M. P. (1974), "Inference From Complex Samples," *Journal of the Royal Statistical Society*, Ser. B, 36, 1–37.

Korn, E. L., and Graubard, B. I. (1995a), "Analysis of Large Health Surveys: Accounting for the Sampling Design," *Journal of the Royal Statistical Society*, Ser. A, 158, 263–295.

——— (1995b), "Examples of Differing Weighted and Unweighted Estimators From a Sample Survey," *The American Statistician*, 49, 291–295.

——— (1998), Discussion of "Weighting for Unequal Selection Probabilities in Multilevel Models," by D. Pfeffermann et al., *Journal of the Royal Statistical Society*, Ser. B, 60, 23–40.

——— (1999), *Analysis of Health Surveys,* New York: Wiley.

Krebs-Smith, S. M., Cook, A., Subar, A. F., Cleveland, L, and Friday, J. (1995), "U. S. Adults' Fruit and Vegetable Intakes, 1989 to 1991: A Revised Baseline for the *Healthy People 2000* Objective," *American Journal of Public Health*, 85, 1623–1629.

Krebs-Smith, S. M., and Kantor, L. S. (2001), "Choose a Variety of Fruits and Vegetables Daily: Understanding the Complexities," *Journal of Nutrition,* 131, 487–501S.

Lazarsfeld, P. F., and Henry, N. W. (1968), *Latent Structure Analysis*, Boston: Houghton Mifflin.

Lee, E. S., Forthofer, R. N., and Lorimor, R. J. (1986), "Analysis of Complex Sample Survey Data: Problems and Strategies," *Sociological Methods & Research*, 15, 69–100.

Lin, T. H., and Dayton, C. M. (1997), Model Selection Information Criteria for Non-Nested Latent Class Models," *Journal of Educational and Behavioral Statistics*, 22, 249–264.

Lindsay, B., Clogg, C. C., and Grego, J. (1991), "Semiparametric Estimation in the Rasch Model and Related Models, Including a Simple Latent Class Model for Item Analysis," *Journal of the American Statistical Association*, 86, 96–107.

Miller, R. G. (1968), "Jackknifing Variances," *Annals of Mathematical Statistics*, 39, 567–582.

——— (1974), "The Jackknife—A Review," *Biometrika*, 61, 1–14.

Mosteller, F., and Tukey, J. W. (1968), "Data Analysis, Including Statistics," in *The Handbook of Social Psychology*, Vol. 2, eds. G. Lindsay and E. Aronson, Reading, MA: Addison-Wesley, pp. 80–203.

Nusser, S. M., Carriquiry, A. L., Dodd, K. W., and Fuller, W. A. (1996), "A Semiparametric Transformation Approach to Estimating Usual Daily Intake Distributions," *Journal of the American Statistical Association*, 91, 1440–1449.

Patterson, B. H. (1998), "Latent Class Analysis of Sample Survey Data," doctoral dissertation, University of Maryland, College Park.

Patterson, B. H., Block, G., Rosenberger, W. F., Pee, D., and Kahle, L. L. (1990), "Fruit and Vegetables in the American Diet: Data From the NHANES II Survey," *American Journal of Public Health,* 80, 1443–1449.

Patterson, B. H., Harlan, L. C., Block, G., and Kahle, L. (1995), "Food Choices of Whites, Blacks and Hispanics: Data From the 1987 National Health Interview Survey," *Nutrition and Cancer*, 23, 105–119.
Pfeffermann, D. (1993), "The Role of Sampling Weights When Modeling Survey Data," *International Statistical Review*, 61, 317–337.
Pfeffermann, D., Skinner, C. J., Holmes, D. J., Goldstein, H., and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multilevel Models," *Journal of the Royal Statistical Society*, Ser. B, 60, 23–40.
Popkin, B. M., Siega-Riz, A. M., and Haines, P. S. (1996), "A Comparison of Dietary Trends Among Racial and Socioeconomic Groups in the United States," *New England Journal of Medicine*, 335, 716–720.
Quenouille, M. H. (1949), "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society*, Ser. B, 11, 68–84.
Roberts, G., Rao, J. N. K., and Kumar, S. (1987), "Logistic Regression Analysis of Sample Survey Data," *Biometrika*, 74, 1–12.
Skinner, C. J., Holt, D., and Smith, T. M. F. (eds.) (1989), *Analysis of Complex Surveys*, New York: Wiley.
Smith, A. F. (1991), "Cognitive Processes in Long-Term Dietary Recall," National Center for Health Statistics. Vital Health Statistics.
Smith, P. J., Graubard, B. I., and Midthune, D. N. ( ), "Mixed Poisson Model for Clustered Count Data With Covariates and Application to Dietary Intake Data," unpublished manuscript.
Smith, T. M. F. (1976), "The Foundations of Survey Sampling: A Review" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 139, 183–204.
——— (1984), "Present Position and Potential Developments: Some Personal Views—Sample Surveys" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 147, 208–221.
Subar, A. F., Frey, C. M., Harlan, L. C., and Kahle, L. (1993), "Differences in Reported Food Frequency by Season of Questionnaire Administration: The 1987 National Health Interview Survey," *Epidemiology*, 5, 226–233.
Subar, A. F., Heimendinger, J., Patterson, B. H., Krebs-Smith, S. M., Pivonka, E., and Kessler, R. (1994), "Fruit and Vegetable Intake in the United States: The Baseline Survey of the Five A Day for Better Health Program," *American Journal of Health Promotion*, 9, 352–360.
Thompson, F. E., and Byers, T. (1994), "Dietary Assessment Manual," *The Journal of Nutrition*, 24, 2245S–2317S.
U. S. Department of Agriculture, Agricultural Research Services (1998), *Food and Nutrient Intakes by Individuals in the United States, by Age and Sex, 1994–1996*, Nationwide Food Surveys Report 96-2, Beltsville, MD: U. S. Department of Agriculture.
U. S. Department of Agriculture, Human Nutrition Research Service (1987), *CSFII: Nationwide Food Consumption Survey. Continuing Survey of Food Intakes by Individuals, Women 19–50 and Their Children 1–5 Years, 4 Days, 1985*, CSFII Report 85-4, Washington DC: U. S. Government Printing Office.
U. S. Department of Health and Human Services (2000), *Tracking Healthy People 2010*, Washington DC: U. S. Government Printing Office.
Vermunt, J. K. (1997), "The LEM User Manual," WORC paper, Tilburg University, The Netherlands.
Vermunt, J. K., and Magidson, J. (2000), *Latent Gold: User's Guide*. Belmont, MA: Statistical Innovations, Inc.
Wedel, M., ter Hofstede, F., and Steenkamp, J. E. M. (1998), "Mixture Model Analysis of Complex Surveys," *Journal of Classification*, 15, 225–244.
Wolter, K. M. (1985), *Introduction to Variance Estimation*, New York: Springer-Verlag.
Young, L. R., and Nestle, M. (1995), "Portion Sizes in Dietary Assessment: Issues and Policy Implications," *Nutrition Reviews*, 53, 149–158.

Comment

Alicia L. CARRIQUIRY and Sarah M. NUSSER

It is well known that collecting and analyzing dietary intake data can be challenging (e.g., Beaton et al. 1979; Basiotis, Welsh, Cronin, Kelsay, and Mertz 1987; Dwyer and Coleman 1997). Yet despite the difficulties inherent in accurately measuring food intakes and in drawing useful inferences from those measurements, the U. S. government relies on complex dietary intake surveys to guide nutrition and health policy, monitor the performance of food assistance programs, and design interventions such as national food fortification programs. In this light, the work of Patterson, Dayton, and Grauband is welcome in that it seeks to capitalize on the rich data available for dietary assessment.

In this discussion we focus on the subject matter interpretation and the statistical aspects of the variable used to indicate dietary intake as well as the latent class model used to make inferences using this variable. In the next section we discuss the importance of informative dietary intake measures with respect to policy development. In Section 2 we focus on the model itself. Finally, in Section 3 we provide some conclusions and thoughts.

1. VARIABLES OF INTEREST TO POLICY MAKERS

In nationwide surveys such as the Continuing Survey of Food Intakes by Individuals (CSFII), respondents are asked to report on the amounts of food consumed during the previous 24 hours. The amounts of the various foods consumed are expressed in such units as glasses, cups, grams, slices, tablespoons, and so forth. Even though the interviewer arrives at a respondent's home armed with two- and three-dimensional models that are meant to help the respondent to accurately quantify the amount of each food consumed, it is still well known that correctly gauging portion sizes can be difficult (Hartman et al. 1994; Haraldsdottir, Tjonneland, and Overvad 1994; Dwyer and Coleman 1997). When interviews are conducted over the phone, measurements are likely to be even more inaccurate. In this sense, the authors correctly argue that the measurement error in dietary intake data can be significant. They believe that the presence of this measurement error, compounded by the fact that respondents tend to underreport

Alicia L. Carriquiry is Associate Provost Professor, and Sarah M. Nusser is Associate Professor, Department of Statistics, Iowa State University, IA, 50011 (E-mail: *alicia@iastate.edu* and *nusser@iastate.edu*).

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Applications and Case Studies
DOI 10.1198/00

the amount of food eaten, sometimes by omitting foods altogether, justifies discretizing the amounts of foods consumed to a binary 0–1 indicator of whether the individual reported consumption of at least one fruit or vegetable during each survey day.

One could argue that the presence of measurement error does not imply complete lack of information, and that discretizing continuous variables results in loss of information that can be useful in drawing inferences about food consumption patterns in the population. Indeed, the binary variable indicating consumption of vegetables is not immune to under reporting or over reporting of the amounts of food consumed, because it has been shown that respondents tend to omit "sinful" foods such as candy and alcohol and to report consumption of fruits and vegetables that did not take place (e.g., Hebert et al. 1997; Kristal, Feng, Coates, Oberman, and George 1997). These reporting biases may depend on individual characteristics such as body mass index (Tarasuk and Beaton 1991; Hebert et al. 1997; Kristal et al. 1997; Johnson, Soultanakis, and Matthews 1998).

A second motivation to use a dichotomous indicator of consumption is the authors' perception that usual intake is not really defined in the dietary assessment community. However, the concept of *usual intake* of a food or a nutrient is typically defined as the long-run average intake of the food or nutrient. Formally, if Y_{ij} denotes the observed intake of a food or a nutrient by individual i on day j of the survey, then usual intake is defined as

$$y_i = E\{Y_{ij} \mid i\},$$

the conditional (on individual) expectation of daily intake. This definition is widely used by researchers (including Guenther et al. 1997 cited by the authors) and policy makers, and was proposed by the National Research Council (1986) in its report (on nutrient adequacy 1986). More recently, the same definition has appeared in various reports by the Institutes of Medicine (1998, 2000, 2001) that discuss (among other topics) the use of nationwide food consumption surveys to assess the intakes of individuals. In some applications, the usual intake of a food is expressed in terms of grams of the food consumed, whereas in some others, intake is expressed as the number of portions of the food consumed.

The approach taken by Patterson et al. assumes that two classes of individuals are present in the population: those that consume vegetables on a "regular" basis and those that do not. This term is not defined operationally, making it difficult to understand how to interpret results from the analyses. In particular, it is not completely clear how regularity of consumption relates to a concise concept of usual intake, and indeed seems to be more descriptive of consumption patterns. We return to this point later.

The loss of information in a dichotomized consumption variable comes from ignoring the amounts of a food consumed, which may be problematic when intake amounts are associated with health outcomes such as cancer. An individual who consumes a small portion of vegetables on most survey days is likely to be classified as a "regular" consumer by the approach proposed by Patterson et al. but so is another individual who consumes the recommended amounts of fruits and vegetables each day. The breadth of this class may make it difficult to draw meaningful inferences about the impact of diet on such diseases.

2. THE LATENT CLASS MODEL FOR DIETARY ANALYSIS

Mixture models, and in particular latent class (LC) models, can be useful to represent observations hypothesized as coming from different subgroups in the population. Sometimes the investigator has a strong scientific argument to decide a priori on a number L of classes. In those cases, the analyst would typically fit an L class model to the data, and then would test whether models with more than L classes might fit the data better (in some predetermined sense). If the data are consistent with the investigator's hypothesis, then it is to be expected that the L class model will fit the data at least as well as models with a larger number of classes.

Patterson et al. do a very nice job of presenting the methodology for fitting a two-class LC model to data collected in a complex survey. However, they do not offer convincing evidence for the choice of the two-class model. Indeed, the choice of model appears to be due to data limitations rather than scientific or statistical arguments. The modified Wald test for goodness of fit presented in the article does not indicate whether the two-class model fits the data at least as well as a model with a different number of classes.

The authors correctly point out that the unrestricted model is not identified for $L > 2$. In fact, identifiability problems arise even if we place "reasonable" restrictions on the model. Consider, for example, the case where we assume that dietary intake information, collected in person during the first survey day is more accurate than that collected via phone interviews and on later sample days. The validity of this assumption has been demonstrated. In fact, a more realistic model, from a nutrition standpoint, would be one where the first day is considered to be different from the rest of the days. Using the author's notation, we would then define

$$\begin{aligned}\delta_{ij} &= 1 \text{ iff } y_{ij} = 1\\ &= 0 \text{ otherwise}\end{aligned}$$

(for the case where $R_j = 1$ used in this article) and

$$x_{ij} = \sum_{j=2}^{J} \delta_{ij}.$$

The new variable x would then take on the values $\{0, 1, 2, \ldots, J-1\}$, depending on how many days after the first the respondent reported eating at least one fruit or vegetable. The likelihood function in expression (4) in Patterson et al. would then be rewritten as

$$\Lambda_w = \sum_{i=1}^{n} w_i \ln \sum_{l=1}^{L} \theta_l \alpha_{l1}^{\delta_{ij}} \prod_{j=2}^{J} \alpha_{l2}^{x_{ij}},$$

where now α_{l1} and α_{l2} represent the frequency of consumption of fruits or vegetables in the first survey day or in later survey days, given latent class l.

Notice that in the two-class LC model for this restricted case, the number of parameters to be estimated is smaller than in the unrestricted model. Here, we must estimate four item probabilities and one mixing parameter. However, the number of degrees of freedom available for estimation is also smaller, because now we have only 7 $(8-1)$ degrees of freedom to work with. The three-class model, with the same number of degrees of freedom, has eight unknown parameters to be estimated; therefore, any model with more than two classes is unidentified even after restrictions.

This is typical for binary data. The dimensionality of the table with which we can work decreases as we collapse the model; thus we would be limited to fitting a two-class model to these dichotomized data even under strong model restrictions. This perhaps could be seen as another reason to avoid discretizing continuous data into binary categories. With intake data expressed as either number of servings or even amounts consumed, it would have been possible to fit a richer class of models and to test whether the two-class model LC is reasonable for these data.

3. AN ALTERNATIVE APPROACH

Are there different "classes" of individuals in the population when it comes to consumption of fruits and vegetables? Perhaps, but determining just how many will require a different type of analysis. Nusser et al. (1997) proposed an alternative approach to assessing intakes of foods from complex survey data. In their work, the underlying assumption is that there exists a *distribution of probabilities of consumption* in the population. In fact, Nusser et al. (1997) take the analysis one step further and estimate the distribution of usual food intake (for a given food) by combining the distribution of probability of consumption with the distribution of amounts consumed. Although it is true that the procedure is considerably more complex, it is also true that the outcome from the analysis provides a richer assessment of the intake of a food (both in terms of frequency and of amounts) in the population of interest. The Nusser et al. (1997) procedure accounts for differential weights resulting from the design of the survey or from nonresponse, but assumes that the probability of consuming a food is independent of the amount consumed. This assumption may not hold for foods such as milk and soft drinks, however.

We thank the authors for bringing to the fore the importance of correctly analyzing and interpreting dietary intake data, and for presenting a new modeling approach for food consumption patterns.

ADDITIONAL REFERENCES

Basiotis, P. P., Welsh, S. O., Cronin, F. J., Kelsay, J. L., and Mertz, W. (1987), "Number of Days of Food Intake Records Required to Estimate Individual and Group Nutrient Intakes With Defined Confidence," *J Nutr* 117, 1638–1641.

Beaton, G. H., Milner, J., Corey, P., McGuire, V., Cousins, M., Stewart, E., de Ramos, M., Hewitt, D., Grambsch, P. V., Kassim, N., and Little, J. A. (1979), "Sources of Variance in 24-Hour Dietary Recall Data: Implications for Nutrition Study Design and Interpretation," *Am J Clin Nutr* 32, 2546–2559.

Dwyer, J. T., and Coleman, K. A. (1997), "Insights into Dietary Recall From a Longitudinal Study: Accuracy Over Four Decades," *Am J Clin Nutr* 65, 1153S–1158S.

Guenther, P. M., Kott, P. S., and Carriquiry, A. L. (1997), "Development of an Approach for Estimating Usual Nutrient Intake Distributions at the Population Level," *J Nutr* 127, 1106–1112.

Haraldsdottir, J., Tjonneland, A., and Overvad, K. (1994), "Validity of Individual Portion Size Estimates in a Food Frequency Questionnaire," *Int J Epidemiol* 23, 787–796.

Hartman, A. M., Block, G., Chan, W., Williams, J., McAdams, M., Banks, W. L. Jr., and Robbins, A. (1996), "Reproducibility of a Self-Administered Diet History Questionnaire Administered Three Times Over Three Different Seasons," *Nutr Cancer* 25, 305–315.

Hebert, J. R., Ma, Y., Clemow, L., Ockene, I. S., Saperia, G., Stanek, E. J., Merriam, P. A., and Okene, J. K. (1997), "Gender Differences in Social Desirability and Social Approval Bias in Dietary Self-Report," *Am J Epidemiol* 146, 1046–1055.

Institutes of Medicine (1998), *Dietary Reference Intakes for Thiamin, Riboflavin, Niacin, Vitamin B_6, Folate, Vitamin B_{12}, Panthotenic Acid, Biotin, and Cholin*, Washington, DC: National Academy Press.

——— (2000), *Dietary Reference Intakes for Vitamin C, Vitamin E, Selenium, and Carotenoids*, Washington, DC: National Academy Press.

——— (2001), *Dietary Reference Intakes: Applications in Dietary Assessment*, Washington, DC: National Academy Press.

Johnson, R. K., Soultanakis, R. P., and Matthews, D. E. (1998), "Literacy and Body Fatness are Associated With Underreporting of Energy Intake in U. S. Low-Income Women Using the Multiple-Pass 24-Hour Recall: A Doubly-labeled Water Study," *J Am Diet Assoc* 98, 1136–1140.

Kristal, R. A., Feng, Z., Coates, R. J., Oberman, A., and George, V. (1997), "Associations of Race/Ethnicity, Education, and Dietary Intervention With the Validity and reliability of a Food Frequency Questionnaire: The Women's Health Trial Feasibility Study in Minority Populations," *Am J Epidemiol* 146, 856–869.

National Research Council (1986), *Nutrient Adequacy: Assessment Using Food Consumption Surveys,* Washington, DC: National Academy Press.

Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1997), "Estimating Usual dietary intake distributions: Adjusting for measurement error and nonnormality in 24-hour food intake data," in *Survey Measurement and Process Quality*, eds. Lyber, Biemer, Collins, de Leeuw, Dippo, Schwartz, and Trewin, New York: Wiley.

Tarasuk, V., and Beaton, G. H. (1991), "The Nature and Individuality of Within- Subject Variation in Energy Intake," *Am J Clin Nutr* 54, 464–470.

Comment

Michael R. ELLIOTT and Mary D. SAMMEL

Latent class analysis (LCA) has been an increasingly popular useful data reduction and analysis tool, used not only in the analysis of categorical data, as Patterson, Dayton, and Graubard (PDG) do here, but also to classify count or continuous longitudinal data or mixed continuous and categorical data as well (see e.g., Roeder, Lynch, and Nagin 1999; Muthen and Shedden 1999). Despite dozens of journal articles appearing each year using LCA in areas as diverse as epidemiology, psychology, and economics, PDG are correct that virtually no attempts have been made to use LCAs in a manner that takes into account complex sample design schemes. This makes their contribution particularly timely and useful.

We discuss several features of LCA that we believe will enhance the understanding of their use in this context. These features include the choice of model (i.e., data reduction) of the observed data, the probabilities of individual class membership and the effect of weighting on these probabilities, the view of the effect of weights as an interaction between the LCA structure and the probability of inclusion, and some additional model-checking procedures.

In the context of nutrition research, PDG's model addresses a somewhat different perspective, that of summarizing events with low prevalence and categorizing subjects with little or no consumption. LC and trait models are very useful for summarizing multiple outcomes and has the potential to increase the power to identify effects (Sammel and Ryan 1996; Holmes et al. 1987). PDG mention that LC methods are useful tools for data reduction; however, they do not use this attribute to the fullest extent. Before any modeling, the data for each interview are reduced to a single Bernoulli outcome, any vegetable consumption (yes/no). LCA models have already been developed to handle any number and type of outcome (Bernoulli, count, measured), and we encourage PDG to extend their method. The reduction method that they use carries with it the strong assumption that any versus none is all that is relevant, whereas additional data concerning low consumption would enhance the model's ability to discriminate among subjects of different consumption classes. A simple extension would be to model the number of vegetables reported at each time point as a Poisson count. Alternatively, an additional level could be incorporated into the model for subject responses to each individual item of the recall at each time point. This would add the potential for evaluating more than two classes and would increase the power of the analysis by increasing the effective sample size unless the correlation among the items is extreme [(Legler, Lefkopoulou, and Ryan 1995).] Extending the model in this direction would allow PDG to evaluate the contribution of each item and search for clustering of items. In addition, given estimates for the item parameters (α) and jackknife estimates of the covariance matrix, hypotheses about these parameters can be tested. For example, formal testing of the hypothesis "sporadic vegetable eaters (those in class 1) are more likely to report consumption in a face-to-face interview" could be formulated with a contrast statement and an F test.

In their model, PDG focus on estimation of population parameters for the proposed LC model. Although they are not concerned with a predictive model for individual class membership, determining the probability of class membership for each subject conditional on the observed data is useful for several reasons. First, as a model-checking procedure, we would prefer LC models that sharply delineate individuals into the L distinct classes over those that assign a probability of approximately $1/L$ to all subjects. Using the notation of PDG, it is straightforward to show that the probability of subject i being a member of class l is given by

$$\Pr(c_l \mid y_i) = p_{li} / \sum_l p_{li}, \quad p_{li} = \theta_l \prod_{j=1}^{J} \prod_{r=1}^{R} \alpha_{ljr}^{\delta_{ijr}}, \tag{1}$$

where δ_{ijr} is the Kronecker delta equal to 1 iff $y_{ij} = r$. Comparing the histogram of these posterior probabilities computed from (1) using the unweighted and weighted maximum likelihood estimates from PDG's Table 2, Figure 1 shows that the weighted posterior probabilities (a) delineate the posterior class probabilities more sharply than the unweighted estimates (b). In particular, if we assign those with $\Pr(c_l = 1 \mid y_i) > .5$ to class 1, the "sporadic vegetable eaters," we see a nice alignment with the observed data: The 12% of subjects that reported eating vegetables on a maximum of 1 day are assigned to class 1.

Moreover, these posterior probabilities can themselves be used as outcomes for further analyses that relate the latent class assignments to observed covariates. Table 1 considers the log-odds of being classified as a "sporadic" vegetable eater relative to age, region of residence, income, and race, where subjects are classed as sporadic or consistent vegetable eaters depending on whether their posterior probability of being in one class or the other is $> .5$. By considering the unweighted and weighted classification schemes, we obtain insight into the informative nature of the weighting. Both models are similar with respect to region (no association) and income (those with income $< 130\%$ of the poverty level are much more likely to be sporadic vegetable eaters than those with higher incomes), whereas the weighted classification shows a somewhat stronger positive association between

Michael R. Elliott and Mary D. Sammel are Assistant Professors, Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 (E-mail: *melliott@cceb.upenn.edu* and *msammel@cceb.upenn.edu*).

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Applications and Case Studies
DOI 10.1198/00

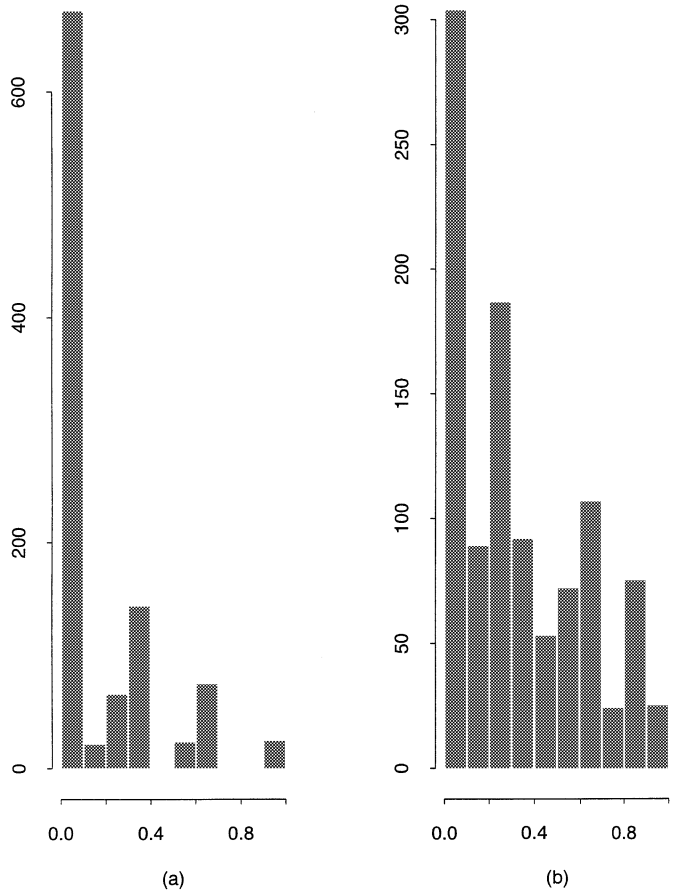

Figure 1. Histogram of Posterior Probability of Class Membership $Pr(c_l = 1 \mid y_i)$ ("sporadic vegetable eating class"), using Weighted (a) and Unweighted (b) Estimates.

youth and sporadic vegetable eating and a much stronger positive association between race and sporadic vegetable eating, with African-Americans being more likely than Caucasians to be classified as sporadic vegetable eaters. A polytomous logistic regression relating the number of days on which vegetables were consumed to the predictors in Table 1 would be an alternative analysis. This is more complex to interpret, however, particularly if the proportional odds assumption is violated.

Table 1. Log-Odds Ratios for Odds of Being in the "Sporadic" Vegetable Eater Class, using Weighted and Unweighted Maximum Likelihood Estimates of θ_l and α_{ljr} to Determine Posterior Probability of Classification Given by (1)

Covariate	Weighted	Unweighted
Age (years)	−.036 (.015)	−.028 (.010)
Region (vs. Northeast)		
Midwest	.24 (.32)	−.07 (.20)
South	.21 (.34)	−.21 (.23)
West	−.42 (.46)	−.09 (.23)
Income (vs. 300+% of poverty level)		
130%–300%	1.07 (.30)	1.00 (.21)
<130%	−.04 (.26)	.26 (.20)
African-American (vs.)	.77 (.35)	.14 (.29)

NOTE: Standard errors (estimated by jackknifing) are in parentheses.

This result leads to the next point, which PDG also allude to in their discussion—namely, that the effect of the weighting can be thought of as an interaction between sampling selection probability and LC structure. Hence noninformative weighting is equivalent to no interaction between the LC structure and the probability of inclusion, and using the weights in such a circumstance is equivalent to estimating an interaction from a sample when none is present in the population and using this estimated interaction together with known population values to determine marginal main effect estimates; the resulting estimators will remain consistent (assuming that the interaction estimate itself is consistent) but will be less efficient. Conversely, ignoring informative weighting is equivalent to estimating a nonzero population interaction from a sample and determining marginal main effects using incorrect population estimates; the resulting estimates will be biased to the extent that the interaction is large and the population estimates misspecified. As PDG note, for the weighting to be noninformative in LCA, the LC structure must be unrelated to selection inclusion with respect to *both* the class probabilities and the conditional probabilities of the observed variables; if the former holds but not the latter, then weighted estimation of both will be inconsistent. Evidence of such informative weighting can be seen in a somewhat crude fashion by stratifying the data into "low," "medium," and "high" probability-of-inclusion strata (as defined by the weights), and then performing an LCA analysis within each stratum. (We report a weighted analysis, but the effect of the weighting is greatly reduced within each weight stratum.) Table 2 shows that sporadic and consistent vegetable eating classes exist in each weight stratum, but that the "sporadic" class is larger and less distinct in the medium and high probability of inclusion strata than in the low probability of inclusion stratum. Consequently, the weighted analysis leads to a smaller and better-defined "sporadic" class than the unweighted analysis.

Thinking of the weights in this manner also suggests an alternative to the "all-or-nothing" approach of analyzing fully weighted or unweighted data. For example, to estimate a population mean $\overline{Y} = N^{-1} \sum_i y_i$ in a population of size N based on a sample s of size n, one could could stratify the data into H design-based strata where the proportion of the population P_h within each stratum is (assumed) known, and assume that the location parameter $\mu_h = E(y_{hi} \mid \mu_h)$ has a prior distribution

Table 2. LCA Analysis, Stratified by Probability of Inclusion

	Probability of inclusion		
Parameter	Low	Medium	High
θ	.10	.34	.28
α_{11}	0	.62	.64
α_{12}	.26	.34	.53
α_{13}	.25	.40	.41
α_{14}	.30	.47	.35
α_{21}	.75	.76	.82
α_{22}	.69	.88	.82
α_{23}	.76	.77	.79
α_{24}	.68	.79	.82
Estimated % of population in stratum	.33	.33	.34

with common mean μ and variance τ^2 within each stratum (Holt and Smith 1979). When $\tau^2 = \infty$, each μ_h is treated as a fixed and independent quantity with no sharing of information across the design strata, and the posterior mean of the population mean $E(\overline{Y} \mid y \in s)$ is given by the fully weighted mean estimator $\bar{y}_w = \sum_{i \in s} w_i y_i / \sum_i w_i$, where w_i is the weight associated with the ith element in the sample. Similarly, when $\tau^2 = 0$, $\mu_h \equiv \mu$ for all h, and $E(\overline{Y} \mid y)$ is given by the unweighted estimator $\bar{y} = n^{-1} \sum_i y_i$ that pools the data across the strata. However, for $0 < \tau^2 < \infty$, $E(\overline{Y} \mid y)$ is given by estimators that can reduce mean squared error by modulating a bias–variance trade-off between unbiasedness (fully weighted) and minimum variance (unweighted). By adding structure to the location prior mean and variance, this bias-variance trade-off can be tuned to the design and data structure in question (Elliott and Little 2000).

Extending this approach, one could stratify the data into H design-based strata where the proportion of the population P_h within each stratum is assumed known and consider a hierarchical model for the latent class assignment and conditional probabilities of the multinomial observed data. Assuming that the Y_{ij} are dichotomous, one could consider a model of the form

$$\begin{aligned} Y_{hij} \mid c_l, \alpha_{hlj} &\overset{ind}{\sim} BERNOULLI(\alpha_{hlj}), \\ \alpha_{hlj} \mid c_l &\overset{ind}{\sim} BETA(a_{lj}, b_{lj}), \\ c_l \mid \theta_{hl} &\overset{ind}{\sim} MULTINOMIAL(1, \theta_{hl}, L), \end{aligned} \tag{2}$$

and

$$\theta_{hl} \overset{ind}{\sim} DIRICHLET(d_1, \ldots, d_L)$$

with weak or noninformative hyperpriors for the beta and Dirichlet prior parameters. Alternatively, one could apply logit transformations to the probability parameters $\zeta = (\theta, \alpha)$ and use normal priors. The population values of ζ could then be estimated from the population score equation, by solving for ζ in (3),

$$\sum_{h=1}^{H} P_h \sum_{i=1}^{n_h} \frac{\partial \Lambda_i(\zeta)}{\partial \zeta} = \sum_{h=1}^{H} P_h \sum_{i=1}^{n_h} \frac{\partial \Lambda_i(\zeta_h)}{\partial \zeta}. \tag{3}$$

Here Λ_i is the log-likelihood contribution of the ith subject from (3) of PDG and the ζ_h in the right side of (3) are the posterior (mean) estimates of ζ_h from the model given by (2). Such an analysis would permit pooling or shrinkage of the estimators across the strata, allowing compromise between a fully weighted analysis (in which the α_{hlj} and θ_{hl} would be estimated separately in each stratum) and an unweighted analysis (in which it is assumed that $\alpha_{hlj} \equiv \alpha_{lj}$ and $\theta_{hl} \equiv \theta_l$).

Because two classes are the maximum that can be identified from four elements of Y_i, PDG focus on a confirmatory rather than an exploratory LCA. Nonetheless, additional model checking can be performed to determine whether two classes are sufficient to maintain the conditional independence assumption among the elements of Y_i. Although PDG do not take a Bayesian approach, an ad hoc "posterior predictive check" similar to that described by Garrett and Zeger (1999) can be obtained by generating y_{ij} from α_{lj} independently for the jth day, where the ith subject is assigned to class c_l with probability $\Pr(c_l \mid y_i)$ given by (1). Replicated values of the (weighted) 16 cells counts can then be compared against the observed cell counts to detect discrepancies between the data and the model assumption of latent/conditional independence. Uncertainty in the estimation of $\Pr(c_l \mid y_i)$ can be taken into account by drawing θ and α_{jl} from beta distributions whose parameters are estimated by method of moments, using the means and standard errors given by the weighted data columns of PDG's Table 2. The resulting "p values" range between .22 and .90, with the exception of cell 0110 (vegetable use on reporting days 2 and 3 only), where the 95% of the replicated cell counts lay between 22 and 50, as compared with the observed cell count of 22. Hence the conditional independence assumption appears largely reasonable with the possible exception of the vegetable use on reporting days 2 and 3 only. (For a similar, yet less structured test of this assumption, see Bandeen-Roche, Miglioretti, Zeger, and Rathouz 1997).

ADDITIONAL REFERENCES

Bandeen-Roche, K., Miglioretti, D. L., Zeger, S. L., and Rathouz, P. J. (1997), "Latent Variable Regression for Multiple Discrete Outcomes," *Journal of the American Statistical Association*, 92, 1375–1386.

Elliott, M. R., and Little, R. J. A. (2000), "Model-Based Approaches to Weight Trimming," *Journal of Official Statistics*, 16, 191–210.

Holmes, L. B., Harvey, E. A., Kleiner, B. C., Leppig, K. A., Cann, C. I., Muñoz, A., and Polk, B. F. (1987), "Predictive Value of Minor Anomalies: II. Use in Cohort Studies to Identify Teratogens," *Teratology*, 36, 291–297.

Garret, E. S., and Zeger, S. L. (2000), "Latent Class Model Diagnosis," *Biometrics*, 56, 1055–1067.

Holt, D., and Smith, T. M. F. (1979), "Poststratification," *Journal of the Royal Statistical Society*, Ser. A, 142, 33–46.

Muthen, B., and Shedden, K. (1999), "Finite Mixture Modeling With Mixture Outcomes Using the EM Algorithm," *Biometrics*, 55, 463–469.

Roeder, K., Lynch, K. G., and Nagin, D. S. (1999), "Modeling Uncertainty in Latent Class Membership: A Case Study in Criminology," *Journal of the American Statistical Association*, 94, 766–776.

Comment

Marilyn M. SEASTROM

In their article, Patterson, Graubard, and Dayton explore new territory on three separate fronts: nutritional measurement, the application of a new technique to measurement in the field of nutrition, and the application of methods for analyzing complex sample survey data to LCs models. I would like to focus my comments on the last of these contributions. Much of the data disseminated by federal statistical agencies are derived from complex sample surveys. Although descriptive secondary analysis of these data requires the use of weights and variance estimation programs equipped to handle the complex sample designs, all too often the design features of survey data are ignored and the data are instead treated as though based on a simple random sample. In this work, the authors demonstrate the potential problems resulting from such oversights.

Patterson, Graubard, and Dayton are to be commended for the detailed and thorough job they did exploring, operationalizing, and confirming the extension of complex sample survey techniques for variance estimation to LCA. They establish the utility of LCA to the measurement of the underlying classes of regular and infrequent vegetable eaters. They carefully explicate both the substantive application and the theoretical aspects of their model, stopping along the way to describe other possible alternatives and to explain their choice at each step. The authors use data from the 1985 CSFII because it allows them to analyze data from 1,028 women with food recall records for 4 separate days. The CFSII is based on a multistage stratified probability sample of U.S. women age 19–50. Although the survey was originally designed to be self-weighting, weight adjustments were incorporated to reflect nonresponse at the household and individual levels.

Having selected LCA as their preferred approach, the authors acknowledge that methods that take into account such sample design features as weighting, clustering, and stratification in LCA have not been described in the literature. Given the documented effects of ignoring complex sample survey design elements with other statistical techniques, this poses a problem. In particular, the authors note that sample weights and clustering usually inflate the variance, and stratification may reduce the variance when the complex sample design is not taken into account. Thus estimates of variance computed ignoring the sample design tend to be underestimated; similarly, parameter estimates can be biased. The authors demonstrate many of these outcomes in the course of their analysis.

The authors note that there are two approaches to calculating standard errors for complex sample survey data: making adjustments using design effects and the using of approximation techniques to estimate variances. Although the first of these two approaches was previously used by Haertel to calculate standard errors in LCA (Haertel 1984a, 1984b, 1989, the authors use their data to demonstrate the fact that size of the deff changes across variables or subgroups of the population. Thus using deffs may yield adjustments that are too large or too small and yield incorrect results.

By comparison, approximation techniques use the data in specific analyses and thus tend to be more accurate than the deff approach. Because the jackknife was previously used to estimate LC parameters under simple random sampling (Bolesta 1988), the authors choose the jackknife version of the sample replication techniques for their analysis. However, to apply these techniques, the authors first had to develop the necessary computer programs and test them against existing software under assumptions of SRS.

The authors report that a two-class model fits the data satisfactorily. The authors use a simulation to investigate the validity of the methods that they used to account for weighting and clustering and to assess the accuracy of the jackknife standard errors. Suffice it to say that the underlying assumptions and the programming criteria that are used in the simulation are reasonable. Consistent with expectations, the jackknifed variances modestly overestimate the proxy variances. Despite this, they find that coverage was close to the nominal .95 level for all parameters.

Given the success of this work, the logical next step is to consider possible applications to other disciplines. Recall that the LCA model developed in the article allows an analyst to take a set of related measures and identify underlying LCs. In their example, the authors used four separate 24-hour food recall records to create a vegetable/no vegetable consumption variable with scores on each of the 4 days. The model allowed the authors to determine the proportions of the respondents who were regular vegetable eaters and infrequent vegetable eaters. The model also yielded probabilities of vegetable consumption in each class on each of the 4 days.

Another possible application that occurs in a number of fields of social science might be to identify LCs for respondents who are at risk or not at risk for a particular outcome, say, a negative education or health outcome. In the case of the education example, a number of dichotomized variables are frequently used to study children at-risk of an adverse educational experience (West, Denton, and Germino-Hausken 2000). These variables include single-parent household, welfare recipient, mother with less than a high school diploma, and primary language other than English. In this case, instead of having repeated measures of the same phenomena, there are

Marilyn M. Seastrom is Chief Statistician, National Center for Education Statistics, Washington, DC 20006 (E-mail: *marilyn.mcmillan@ed.gov*)

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Applications and Case Studies
DOI 10.1198/00

four separate measures that are each assumed to be related to the possibility of being at risk for a negative outcome. These measures could be used in an LCA to identify the class of children who are at risk and the class of children not at risk and to determine the probability that members in each class have each of the four characteristics that were used to identify the LCs.

In the case of health, the phenomena could be any one of a number of illnesses (e.g., hypertension, heart attack, stroke); and the variables would be the related physical traits or behaviors associated with the illness. Take the example of hypertension in adults. The risk factors might include obesity, sedentary lifestyle, high salt intake, and excessive alcohol consumption. The LCA would then identify the class of adults who are at risk of hypertension and the class who are not, and it would yield estimates of the probabilities that members of each class would have each of the four characteristics.

As a next step, in any of these examples the respondent's class could be assigned to each respondent to create a new variable that could be used in analysis. In the education example, the analyst could then examine both the outcomes and additional characteristics associated with membership in each class.

This technique has many possible applications in a number of fields. Hopefully, the work of Patterson, Graubard, and Dayton will encourage others to adopt these techniques in LCA and to take care to incorporate the design element of complex sample surveys in other statistical analyses as well.

ADDITIONAL REFERENCE

West, J., Denton, K., and Germino-Hausken, E. (2000), *America's Kindergartners: Findings from the Early Childhood Longitudinal Study, Kindergarten Class of 1998*, Washington, DC: U. S. Department of Education, National Center for Education Statistics.

Comment

Jeroen K. VERMUNT

Patterson, Graubard, and Dayton (PGD) have shown how to take into account complex sampling designs in LC modeling. Sampling weights are dealt with by pseudo–maximum likelihood (PML) estimation, a method also used by Wedel, ter Hofstede, and Steenkamp (1998) for mixture modeling and implemented in some LC software packages, including Latent Gold (Vermunt and Magidson 2000). Because standard asymptotic theory is no longer valid, PGD propose estimating standard errors by means of a simple but computationally intensive jackknife procedure that simultaneously corrects for stratification, clustering, and weighting.

In this comment I focus on the question of whether to use sampling weights in LC modeling. I advocate the linearization variance estimator, present a maximum likelihood (ML) estimator, propose a random-effects LC model, and give an alternative analysis of the dietary data that takes into account the data's longitudinal nature.

WEIGHTING: YES OR NO?

I am not convinced that in the presented application the weighted solution is better than the unweighted solution. To clarify this point, it is important to make a distinction between the two types of parameters in the LC model, the LC proportions θ_l and the item-conditional probabilities α_{ljr}. It is clear that the unweighted estimates of θ_l will be biased if characteristics correlated with the sampling weights are also correlated with class membership. However, it is important to note that the results obtained with a standard LC analysis are valid only if the population is homogenous with respect to the α_{ljr}. If this assumption holds, then there is no need to use sampling weights for estimation of the α_{ljr}, and if it does not hold, then using sampling weights does not solve the problem. Heterogeneity in α_{ljr} should be dealt with by introducing the relevant grouping variables in a multiple-group LC analysis.

Taking into account the much larger standard errors in the weighted analysis, I prefer the unweighted $\hat{\alpha}_{ljr}$. Possible biases in the unweighted $\hat{\theta}_l$ can be corrected by reestimating the LC probabilities by, say, PML, fixing the α_{ljr} at their unweighted ML estimates. This two-step estimator yields an estimated LC proportion of .35, which is quite close to the unweighted estimate of .33. Such a small upward correction of the number of low consumers is what could be expected from the fact that weighting increases the observed proportion of nonconsumers. A weighted analysis with the PML method, however, yields a downward correction of the proportion of low consumers ($\hat{\theta}_1 = .18$).

LINEARIZATION ESTIMATOR

Wedel et al. (1998) proposed using a linearization or robust variance estimator in mixture modeling with complex samples. This method was described in detail by Skinner, Holt, and Smith (1989). PGD state that this approach is less flexible in that it requires developing new software. I do not agree with this statement, because the method is easily implemented in any LC software that already computes first and second derivatives of the pseudolikelihood function. It should be noted that in contrast to PGD's jackknife method, the additional computation time is negligible.

Jeroen K. Vermunt is Senior University Lecturer, Department of Methodology and Statistics, Faculty of Social and Behavioral Sciences, Tilburg University, The Netherlands (E-mail: *J.K.Vermunt@KuB.NL*)

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Applications and Case Studies
DOI 10.1198/00

The standard errors that I obtained with the linearization estimator are very close to the jackknife standard errors. Actually, they are slightly smaller, which indicates that they are not only easier and faster to obtain, but also somewhat better given that PGD's simulation study showed that the jackknife slightly overestimates the standard errors.

MAXIMUM LIKELIHOOD ESTIMATION OF LATENT CLASS MODELS WITH SAMPLING WEIGHTS

Clogg and Eliason (1987) and Magidson (1987) proposed a ML estimator for log-linear models with sampling weights under Poisson sampling. Let k denote a particular response pattern, and let δ_{ik} be 1 if case i has response pattern k and 0 otherwise. The unweighted frequency in cell k, n_k, equals $\sum_i \delta_{ik}$, and the weighted frequency, $n_k^{(w)}$, is obtained by $\sum_i \delta_{ik} w_i$. The inverse of the cell-specific sampling weight, z_k, equals $n_k/n_k^{(w)}$. The log-linear model used in a weighted analysis has the form

$$m_k = \exp(\mathbf{x}_k\beta)\, z_k.$$

The term $\exp(\mathbf{x}_k\beta)$ defines an expected cell entry in the population, whereas the corresponding expected cell entry in the "biased population," m_k, is obtained by multiplying it by z_k.

Under Poisson sampling, ML estimation of the unknown β parameters involves maximizing $\log L = \sum_k [n_k \ln(m_k) - m_k]$. This function correctly reflects the data-generating process as far as the unequal selection (or nonresponse) probabilities are concerned. Note that the PML method maximizes $\log PL = \sum_k [n_k^{(w)}(\mathbf{x}_k\beta) - \exp(\mathbf{x}_k\beta)]$, which is clearly not the same.

The foregoing method can be easily generalized to LC models if we write the LC model as a log-linear model for an incomplete table. Using l as the index for the latent classes, the model for m_k is now

$$m_k = \left[\sum_l \exp(\mathbf{x}_{lk}\beta)\right] z_k,$$

where the linear term $\mathbf{x}_{lk}\beta$ defines the LC model (see Haberman 1979). The Newton (Haberman 1988) and LEM (Vermunt 1997) programs for log-linear modeling with incomplete tables can be used to implement this method.

Application of this ML method to the dietary data yields results similar to PGD's PML results. But an advantage is that standard goodness of-fit measures can be used to assess model fit. The likelihood ratio statistic L^2 equals 18.32 (df = 6 and $p = .01$), indicating that the two-class model does not fit the data.

RANDOM-EFFECTS LATENT MODELS

A standard method for dealing with clustering effects is random-effects modeling. In the application, a cluster is a PSU within a stratum, say PSU h in stratum s, denoted by sh. Let us assume that the LC proportions are coefficients that vary between PSUs. A simple random-effects two-class model is obtained by assuming that $\ln(\theta_{1(sh)}/\theta_{2(sh)}) \sim N(\mu, \sigma^2)$. The contribution of cluster sh to the log-likelihood function equals

$$\ln L_{sh} = \ln \int \left\{ \prod_{\text{all } i \text{ in cluster } sh} \left(\sum_{l=1}^{L} \theta_{l(sh)} P(\mathbf{Y}_i|c_l) \right) \right\} \times f(\theta_{(sh)}|\mu, \sigma^2)\, d\theta_{(sh)}.$$

The integral can be solved by, for instance, Gauss–Hermite quadrature.

Application of this random-effects LC model to the (unweighted) dietary data revealed no evidence for variation of the LC proportions between clusters. This is in agreement with PGD's results.

MEASUREMENT ERROR OR CHANGE?

As indicated by PGD, the four dietary recalls were obtained at six time points; that is, recalls 2–4 do not represent the same recall occasions for all of the women. To be able to take the longitudinal nature of the data into account, I reanalyzed the (unweighted) data using six occasions instead of four, where each woman has two missing values. It should be noted that as long as the missing data can be assumed to be missing at random, it does not cause special problems within a ML framework.

First, I estimated standard LC models with different numbers of classes. The two-class model turned out to be the best in terms of fit ($L^2 = 52.07,\ \text{df} = 50,\ p = .39$). Equating all time-specific intake probabilities for the high-consumption class and the ones of the first three time points for the low-consumption class did not cause the fit to deteriorate ($L^2 = 55.82,\ \text{df} = 57,\ p = .52$). The estimated intake probability was .80 for the high- and stable-consumption class. The low-consumption class had .57 at the first three time points, dropped to .38 and .20, and increased to .46 at the last time point.

PGD do not pay attention to the fact that there is not only measurement error in the reported intake, but also changes in the intake over time. The LC model, however, cannot make a distinction between measurement error and change, however; a better-suited model for this purpose is a hidden or latent Markov model. A simple hidden Markov with two latent states and time-invariant measurement errors fits almost as good as the two-class LC model ($L^2 = 54.37,\ \text{df} = 50,\ p = .31$), but tells a more interesting story about the same dataset. The high-intake class has an intake probability of .83 at each time point and the low-intake class of .36. Note that these measurement errors (.17 and .36) are smaller than those in the standard LC model. Between occasions 1 and 3 are similar numbers of moves from high to low intake as from low to high, between time points 3 and 5 are much more moves from high to low, and between time points 5 and 6 are much more moves from low to high. This indicates that besides measurement error, there is a seasonal effect in the consumption of vegetables; the proportion of low consumers depends on the time of year.

ADDITIONAL REFERENCES

Haberman, S. J. (1988), "A Stabilized Newton–Raphson Algorithm for Log-Linear Models for Frequency Tables Derived by Indirect Observations," *Sociological Methodology*, 18, 193–211.

Magidson, J. (1987), "Weighted Log-Linear Modeling," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 171–174.

Rejoinder

Blossom H. PATTERSON, C. Mitchell DAYTON, and Barry I. GRAUBARD

We wish to thank the discussants for their thoughtful and provocative comments. In our article we have presented a new approach to modeling dietary consumption patterns, as well as methodology permitting the application of LCA to sample survey data. We had two goals. First, extending a long-standing interest in vegetable consumption, we were interested in using LCA to find some overall (if crude) measure of the proportion of the population falling into a "regular" vegetable consumption class and the proportion falling into a less regular, or infrequent, consumption class. This type of information could be useful in formulating public health programs. A national dataset comprising 4 (independent) days of dietary intake data sampled from women age 19–50 from the CSFII offered this opportunity. Second, because these data were not from a simple random sample, we had to develop methodology to apply sample weights to the data and to estimate standard errors that take into account the complex sample design. In our "first cut" at achieving these goals, we sought to fit a simple, straightforward LC model. We recognize that there are many different and more complex approaches than our simple model, and we encourage others to pursue them. A major contribution, both in our eyes and in the eyes of the discussants, was to develop LCA methods that can be used to analyze sample survey data.

The discussions cover a broad range of topics. We begin our rejoinder by returning to our motivating problem, characterizing "usual" vegetable consumption in the United States. Central to this problem is the need to measure intake over some time period. We explain our approach to this problem, which involves a new definition of usual consumption that uses LC modeling with the consequent data reduction to binary observations. Some of the discussants questioned our data reduction and suggested alternative methods of analysis, both frequentist and Bayesian. We review and comment on some of these. The question of how (and whether) to use sample weights was of special interest to our discussants, as was the estimation of standard errors. The question of model fit arose in several of the discussions. We make the point that no adequate measure of goodness of fit for latent class models has yet been developed for sample survey data; such measures need to be developed. We address these three topics (weights, variance estimation, and goodness of fit) under the broad heading of accounting for the sample design. Finally, some of the discussants proposed new uses of our techniques, and we briefly review these.

1. CHARACTERIZING DIETARY INTAKE

As pointed out by Carriquiry and Nusser, the U. S. government relies on dietary intake data from national surveys for the development of nutritional and health policies. For example, in presenting a revised baseline for the *Healthy People 2000* objectives, Krebs-Smith et al. (1995) showed that 8.2% of the population age 20 years and older consumed less than a single serving of a vegetable per day based on 3 days of dietary intake data for 3 consecutive years. Because of the inverse association between vegetable consumption and several cancers, we were interested in using national survey data to estimate the proportion of the population that does and that does not consume vegetables on a "regular" basis, where regular can be regarded as a way of defining "usual." The National Cancer Institute is currently investigating other approaches to estimating regularity.

As Carriquiry and Nusser note, the definition of "usual" intake as the "long-run average intake of a food" is widely used, although there are other methods in the literature, some of which we cite in our article. However, there is no consensus on how to define "long run". Furthermore, average intake may not be a measure of the regularity of intake. We took a new approach to this problem, considering "regularity" of vegetable consumption to be an unobservable or latent variable. This definition is conceptual but can be operationalized via latent class modeling. We fitted a two-class model to the data. In this context, the item-conditional probabilities, measures of the probability of consuming a vegetables on each recall day given membership in a specific class, are dietary propensity scores (Sue Krebs-Smith and Kevin Dodd, personal communication). These were remarkably consistent for the class of "regular" vegetable consumers but appeared to vary for the infrequent consumers.

Concerns were raised about our dichotomization of the data, which consisted of the number of grams of each individual food consumed by each respondent. We agree that dichotomization of the data does not necessarily reduce measurement error. Carriquiry and Nusser contend that the amount of food consumed on most of the survey days, which is lost in dichotomization, may be crucial in making inferences about the impact of diet on cancer. In our method, an individual consuming small amounts of food on most of the intake days would be classified differently than an individual consuming a large amount on a single recall day, yet the average intake for both individuals could be the same. Whether frequency of consumption of vegetables or the quantity consumed is critical in disease prevention is an open question. Kant, Schatzkin, Graubard, and Schairer (2000) developed a recommended foods score (RFS) that summarizes food frequency questionnaire replies for 23 items, using the report of consumption but not the quantity consumed. They found that dietary diversity as reflected in the RFS was inversely related to cancer and other diseases as well as to all-cause mortality. Our method could be used to examine the relationship between

Journal of the American Statistical Association
September 2002, Vol. 97, No. 459, Applications and Case Studies
DOI 10.1198/00

disease and diet. In a study similar to that described here, but with a larger sample size, respondents would be followed for morbidity or mortality, as is being done with respondents from the second National Health and Nutrition Study. A LC model could be fitted to the data, each subject assigned to a LC, and the eventual outcomes compared to these assignments. Our method could be extended to look at amounts consumed.

Elliott and Sammel suggest that a count of vegetable servings on each occasion of measurement represents a better modeling opportunity than the simpler 0–1 representation that we used. In general, we agree that this is both desirable and possible, using, for example, a Poisson representation for the counts. In the present case our judgment was that the reliability of the measurements better supported the simpler coding, but it would be interesting to compare these models.

2. MODEL CHOICE

2.1 Latent Class and Alternative Models

As discussed earlier, our choice of a two-class model was based on our interest in the proportion of the population that does and that does not consume vegetables on a regular basis. Our data constrained us to a two-class model, as noted by Carriquiry and Nusser. But this was not a problem, because the two-class model was the model of interest to us. Carriquiry and Nusser suggest an approach that would distinguish day 1 from the other recall days. In fact, a LC constrained to equate recalls for days 2–4 for each class would accomplish this objective. The similarity seen in the item-conditional probabilities for class 2 (but not class 1) suggests such a constrained model, especially for class 2. However, because this was a *post hoc* finding, we did not pursue this particular model.

Elliott and Sammel suggest extending our method to take into account all the various vegetables reported by all respondents on all 4 recall days, to create the potential for evaluating more than two classes. The resulting cross-tabulation would likely result in a very sparse table, with the accompanying problems of numerical instability and lack of convergence. An alternative method, grouping vegetables by their characteristics (e.g., deep yellow, dark-green leafy), may be a feasible extension of our model. Yet another approach would be to apply definitions of servings to grams reported by each subject, and to use mixture analysis on these variables. Measurement error in reporting portion size is a problem with this approach, as we noted in our article, and it adds a level of difficulty to the analysis.

We agree with Carriquiry and Nusser that it would be possible to fit a richer class of models using the continuous data. They summarize a method of obtaining the distribution of consumption of foods. However, their method makes the assumption that the probability of consuming a food in independent of the amount consumed, an assumption likely to be untrue for vegetables, and also requires strong distributional assumptions. Similarly, the random-effects model and the hidden Markov model suggested by Vermunt require heavy model assumptions. An assessment of the robustness of these methods to model specification is recommended.

Vermunt suggests alternative analyses that take advantage of the fact that the data were collected on six occasions. Here, as in many large surveys, the data cannot all be collected at a single time point, even when the time of interest might be as long as a season. The six data points do not represent the same time intervals during the year, and each occasion actually represents a period of several overlapping weeks (e.g., observation three for one respondent many be collected in the same month as observation four for a different respondent). Further, the two missing time points may represent occasions deleted randomly by the U.S. Department of Agriculture for subjects with five or six responses or actually may be missing data, so that the mechanism of missingness differs between respondents and is not known to the analyst. For this reason, we chose not to analyze data for all six occasions and cannot agree with Vermunt's interpretation that in these data, "consumption of vegetables. . . depends on the time of year." Pairwise z tests using jackknifed standard errors (our Table 2) among the four conditional probabilities within each LC result in no absolute z value greater than 1.13, suggesting, on a post hoc basis, that homogeneity is not an unreasonable assumption for the rates of vegetable consumption.

Elliott and Sammel propose a post hoc Bayesian approach, using Bayes's theorem to calculate a predicted LC membership for each sample member. Given these classifications, odds ratios can be computed for outside variables such as age and region. Because of concerns about the validity of the two-stage procedure, we did not report analyses of this type. However, a recently completed simulation study (Kuo 2001) suggests that, at least for simple random samples, the two-stage procedure for logistic covariate models performs quite well in estimating the parameters for the covariate function for cases with well-defined latent structures (i.e., cases where the conditional probabilities for the two classes are distinctly different). The vegetable data seem to satisfy this requirement.

From a theoretical perspective, the best strategy would be to use the outside variables as covariates directly within the latent class model (Dayton and Macready 1988). In brief, for a two-class model, the latent class proportion for class 1, say, is modeled by a function of the form

$$\pi^X_{1|Z} = g(Z, \beta),$$

where Z is in general a vector-valued covariate, β is a vector of parameters, and $g(\square)$ is a monotone function with a $0, 1$ range over the domain of Z. For example, a logistic covariate model with J covariates could be defined as

$$\pi^X_{1|Z} = 1/[1 + e^{-\beta_0 - \sum_{j=1}^{J} \beta_j Z_j}],$$

where $\pi^X_{1|Z}$ is the proportion of cases in the first latent class (X) conditional on the covariate vector, Z.

Conditional probabilities for the manifest variables and the parameters of the covariate function are estimated simultaneously. Programs such as Latent Gold (Vermunt and Magidson 2000) and LEM (Vermunt 1997) provide estimates for logistic covariate models with case weights but do not take into account clustering. In the context of a complex survey design, one is faced with assessing the contribution of a covariate. The jackknife is recommended as an easily applicable and valid method for generating standard errors of the regression coefficients of the covariates for complex samples.

2.2 Goodness of Fit

Model fit is mentioned in several of the discussions. As we state in our article, we are unaware of any good measure of fit that is appropriate for LCA of complex sample survey data. We also note that for data from a simple random sample, the likelihood ratio statistic cannot be used for comparison of models with differing numbers of classes. Vermunt fits a weighted model based on a Poisson sampling model and concludes that a two-class model "does not fit the data." Although his analysis takes into account sampling weights, it ignores the role of stratification and clustering of the sample selection in survey design. Our analysis, based on a Wald test that took into account both stratification and clustering, suggested satisfactory fit for a two-class model.

3. ACCOUNTING FOR THE SAMPLE DESIGN

3.1 Sample Weights

The CSFII has a complex sample design involving stratified multistage cluster sampling with sample weighting for nonresponse and postratification adjustment. Vermunt argues that weighting should be used for estimation of the LC proportions but not for estimation of the item-conditional probabilities. He prefers a two-stage approach, in which the unweighted data are used to estimate the conditional probabilities and these are then held constant during a weighted analysis that estimates the LC proportions. He argues that if the sample weights are informative for estimating the items conditional probabilities, then these probabilities are not homogenous across subgroups of the population, and the LC model is misspecified for the population. We address these issues in Section 6 of our article and also address the bias and efficiency trade-off between weighted and unweighted analyses. Vermunt also proposes using a method described by Clogg and Eliason (1987); however, this method does not adjust for clustering in the data and also assumes that the model is correctly specified. It is important to note that in general, it is not possible to know whether a model is "correctly" specified, and even if this were possible, the "correct" model would likely be unduly complex and difficult to interpret. When the posited LC model is misspecified, Vermunt's two-stage approach does not estimate the "census" model, that is, the model that would have been obtained if the entire population had been sampled. In contrast, the weighted pseudolikelihood approach that we use does estimate the census model. This approach has the advantage that if the model is misspecified, estimates from different probability sample designs on average will be approximately the same. Vermunt's suggestion of dealing with heterogeneity of the item-conditional probabilities by identifying homogeneous groups and then using multiple-group LC analysis seems impractical and difficult to carry out.

As shown in Figure 1, the impact of weighting is to lower the magnitude of the conditional probabilities, although The effect is much greater in the low-consumption class than in the high-consumption class.

Elliott and Sammel report a more elaborate analysis based on stratification of the sample by the magnitude of the weights themselves. This appears to show that the estimated item-conditional probabilities differ between the low weight stratum and the medium and high weight stratum. They did not test whether these differences are statistically significant. Based on our null results for a Wald test comparing weighted to unweighted estimates, we doubt that theirs would be statistically significant. As pointed out in our article, the Wald test for informativeness of the weights has low power. Based on our results and those of Elliott and Sammel, we are inclined to believe that the weighted estimate is the more reasonable estimate for the population.

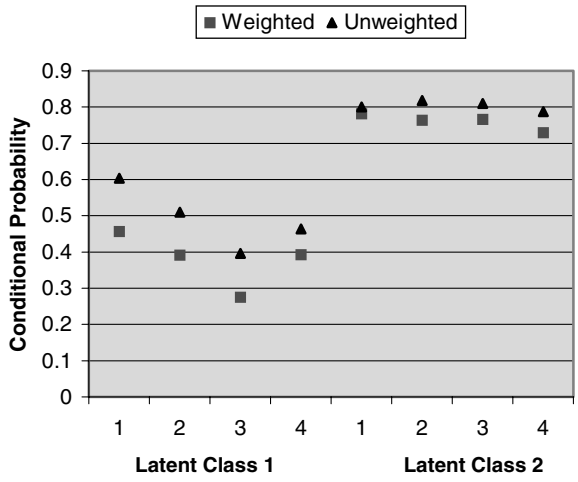

Figure 1. Impact of Sampling Weights on Conditional Probabilities.

Elliott and Sammel also propose an interesting alternative to the "all-or-nothing" approach to weighting. They divide the data into design strata and use an estimator that is a combination of weighted and unweighted estimates. The weighted estimate has more influence on the overall estimates if there is evidence of substantial variability across the strata in the parameter of interest and the unweighted estimate has less influence if there is little evidence of variability. Because, as this approach requires a prior distribution over the strata-specific parameters, its robustness to the distribution of the assumed prior should be investigated before using it. Elliott and Sammel also propose an extension to this model, a hierarchical model that requires both hyperpriors and priors. Such a model would require substantial robustness testing.

3.2 Variance Estimation

Vermunt recommends using linearization variance estimation rather than the jackknife variance estimation that we used. We agree that linearization variances will be faster to compute and can be programmed for various LCAs. We chose to use a jackknife method because of its ease of use; that is, it does not requiring extensive programming. In addition, jackknife variance estimation, through the use of jackknife replicate weights (Rust and Rao 1996), is more flexible than linearization in that it is able to account for variation inherent in commonly used adjustments to the sample weights, such as nonresponse adjustments and postratification. Similar types of replicate weights can be formed from other variance replication methods, such as balanced half-sample replication. National surveys such as the third National Health and Nutrition Survey

(Ezzati, Massey, Waksberg, Chu, and Maurer 1995) are now routinely providing replicate weights for variance estimation.

4. ALTERNATIVE APPLICATIONS

Seastrom suggests other applications of our LC model. Especially useful is her idea of modeling LCs that reflect level of risk for an adverse health, behavioral, or social outcome. It is reasonable to hypothesize that a population may have risk patterns that can be classified into discrete unobservable categories. Also, by identifying these LCs and their relative sizes in the population, intervention programs can be constructed that could be directed at the highest risk classes of nontrivial size. Because complex surveys are used extensively in behavioral and social research, our results for using design-based analyses are potentially of great value for carrying out such analyses with survey data.

ADDITIONAL REFERENCES

Dayton, C. M., and Macready, G. B. (1988), "Concomitant-Variable Latent Class Models.," *Journal of the American Statistical Association*, 83, 173–178.

Ezzati, T. M., Massey, J. T., Waksberg, J., Chu, A., and Maurer, K. R. (1992), "Sample Design: Third National Health and Nutrition Examination Survey," *Vital and Health Statistics* 2(113).

Kant, A. K., Schatzkin, A., Graubard, B. I., and Schairer, C. (2000), "A Prospective Study of Diet Quality and Mortality in Women," *Journal of the American Medical Association*, 283, 2109–2115.

Rust, K. F., and Rao, J. N. K. (1996), "Variance Estimation for Complex Surveys Using replication techniques," *Statistical Methods in Medical Research,* 5, 283–310.