Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

Composite estimation of drug prevalences for sub-state areas

by Manas Chattopadhyay, Partha Lahiri, Michael Larsen
and John Reimnitz

June 1999

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 1999
Vol. 25, No. 1, pp. 81-86
Statistics Canada, Catalogue No. 12-001

Composite estimation of drug prevalences for sub-state areas

Manas Chattopadhyay, Partha Lahiri, Michael Larsen and John Reimnitz [1]

Abstract

The Gallup Organization has been conducting household surveys to study state-wide prevalences of alcohol and drug (*e.g.*, cocaine, marijuana, *etc.*) use. Traditional design-based survey estimates of use and dependence for counties and select demographic groups have unacceptably large standard errors because sample sizes in sub-state groups are too small. Synthetic estimation incorporates demographic information and social indicators in estimates of prevalence through an implicit regression model. Synthetic estimates tend to have smaller variances than design-based estimates, but can be very homogeneous across counties when auxiliary variables are homogeneous. Composite estimates for small areas are weighted averages of design-based survey estimates and synthetic estimates. A second problem generally not encountered at the state level but present for sub-state areas and groups concerns estimating standard errors of estimated prevalences that are close to zero. This difficulty affects not only telephone household survey estimates, but also composite estimates. A hierarchical model is proposed to address this problem. Empirical Bayes composite estimators, which incorporate survey weights, of prevalences and jackknife estimators of their mean squared errors are presented and illustrated.

Key Words: Alcohol abuse; Drug abuse; Empirical Bayes; Jackknife; Mean squared error; Small area estimation; Synthetic estimation.

1. Introduction

The Gallup Organization has been conducting a series of household surveys for different states to study state-wide prevalences of the use of alcohol and drugs (*e.g.*, cocaine, marijuana) among civilian, non-institutionalized adults and adolescents. The common goal of these surveys is to estimate the use and dependence prevalences for alcohol and drugs and, on that basis, to project the treatment needs of dependent users. For planning and resource allocation, states need precise estimates of prevalences for certain subgroups of the target population. For example, it is of interest to estimate prevalences for sub-state planning regions and counties in demographic subpopulations (*e.g.*, older white males).

Traditional design-based procedures to estimate use and dependence for subpopulations have two drawbacks. First, if the traditional design-based survey estimate for a subgroup is positive, but sample size is small, then the corresponding standard error is unacceptably large. Second, since the problem is to estimate the proportion of a rare event, it is possible that the design-based procedure produces an estimate of zero and standard error estimation formulas for a particular subgroup, if applied, would give a false impression of the true underlying variability.

To improve on the traditional design-based estimators, one can use certain supplementary information usually available from administrative records in conjunction with the telephone survey data. This generally is done by using either implicit or explicit models that "borrow strength" or incorporate additional information that relates the various groups, counties, and planning regions to one another. The method proposed here combines information across counties in order to deal with problem of zero estimates in some counties. It is derived from a model that bounds the proportions away from 1, which is reasonable in an application with proportions expected to be very small, and estimates parameters using empirical Bayes methods. The procedure also incorporates the survey sampling weights in estimation.

For a detailed account of small-area estimation methods, see Ghosh and Rao (1994). Other recent references can be found in Farrell, MacGibbon and Tomberlin (1997) and Malec, Sedransk, Moriarity and Leclere (1997). Farrell *et al.*, (1997) propose estimating small-area proportions with empirical Bayes procedures. They model the proportions via a logistic regression that relates expected proportions to respondent variables and includes random effects for the small areas. Malec *et al.*, (1997) use hierarchical Bayes models. They use logistic regression models to relate individual characteristics to probabilities of an outcome and then use a linear regression model to relate coefficients across small areas. Most existing methods, including those of Farrell *et al.*, (1997) and Malec *et al.*, (1997) do not directly use survey sampling weights in estimation.

The survey design used by Gallup is described in section 2. In section 3, notations used in the paper are introduced. A direct design-based estimator and two synthetic estimators are presented in section 4. In section 5, several composite estimators of prevalences of alcohol and drug use and dependence are given. In this section, certain

1. Manas Chattopadhyay, The Gallup Organization; Partha Lahiri, University of Nebraska/Lincoln; Michael Larsen, Harvard University, Department of Statistics, Science Center, One Oxford Street, Cambridge, MA 02138, U.S.A.; John Reimnitz, The Gallup Organization.

empirical Bayes estimators and jackknife estimators of their mean squared errors (MSE) are proposed. In section 6, estimators presented in sections 4 and 5 are applied to a data set from a particular state. The focus of the analysis in this study is taken to be county level estimates. Sample size planning considerations originally were concerned with larger sub-state planning areas.

2. Survey

For sampling purposes, the state is divided into a few planning regions and samples are collected independently for each planning region using a truncated stratified random digit dialing (RDD) method of Casady and Lepkowski (1993). This design stratifies the Bellcore (BCR) frame into two strata: a high density stratum consisting of 100-banks with one or more listed residential numbers and a low density stratum consisting of all the remaining numbers in the BCR frame. About 52 percent of the numbers in the high-density stratum are estimated to be working residential numbers whereas in the low-density stratum, the corresponding percentage is only about 2 percent. The Casady-Lepkowski procedure exploits the significant difference in the cost of sampling between the two strata by optimally determining the sample size in each stratum. In the truncated version of the procedure, sampling is done only from the high-density stratum.

Sample size in the original study was determined in order to estimate statewide prevalence with a desired degree of accuracy. Sample sizes were allocated to the planning regions using an optimal allocation scheme. Data on drug treatment admissions for the adult population in each county were used to compute the index prevalence (rate of admissions) percent in every planning region. These indices were then used to calculate the optimum sample size for each planning region. As a result of optimal allocation, relatively larger sample sizes were allocated to planning regions with higher index prevalences. The optimal allocation also minimizes the variance of the estimators. Gallup also oversampled the 18-45 age group by planning region, because it is the age group with relatively higher rates of illicit drug use. Due to optimal allocation (which may be disproportional), the age oversampling and the complex design, weighting was needed to compute estimates from the sample data. The necessary weights, commonly known as sampling weights, were computed using current estimates of the population based on census data.

Due to budgetary constraints, it is not possible to increase sample size for all sub-state regions and groups in order to achieve the desired accuracy. To estimate alcohol and drug prevalences, we consider empirical Bayes procedures (see Efron and Morris 1973, Fay and Herriot 1979, Ghosh and Lahiri 1987, among others) to improve on usual design-based estimates of drug prevalences by taking advantage of demographic measurements and social indicator data.

Other variables that possibly are related to use and dependence prevalence by county and that are available from Census include the percent of population that is over 65, under 30, white, male, married, and renters. Local governments can provide data by county on social indicators, such as DUI (Driving Under the Influence) rate, mortality rate, per capita liquor licenses, and drug and alcohol treatment admission rates. The more closely auxiliary variables relate to use and dependence prevalence, the more likely it is that methods that "borrow strength" across areas and groups, such as the empirical Bayes methods presented here, can be employed to meet the desired accuracy levels for sub-state areas.

3. Notations

Let n_i be the sample size allocated to the i^{th} planning region, $i = 1, \ldots, I$ ($n = \sum_{i=1}^{I} n_i$). Samples are drawn independently in each planning region using RDD telephone surveys. After the sample is observed, suppose each region is post-stratified into K demographic groups. These groups are formed by cross-classifying gender (Male, Female) and age (18-24, 25-44, 45-64, 65+), resulting in $K = 2 \times 4 = 8$ groups. Suppose there are J_i counties in the i^{th} planning region ($i = 1, \ldots, I$) and n_{ijk} observations within the k^{th} demographic group in the j^{th} county belonging to the i^{th} planning region ($i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $k = 1, \ldots, K$). Since typically n_{ijk} is small, there is a good chance that some of the k demographic groups are not represented in a particular county. Let S_{ij} be the set of demographic groups in the j^{th} county within the i^{th} stratum ($i = 1, \ldots, I$; $j = 1, \ldots, J_i$) for which individuals have completed surveys.

Let y_{ijkl} be the l^{th} observation (0 or 1) for the k^{th} demographic group in the j^{th} county belonging to the i^{th} planning area ($i = 1, \ldots, I$; $j = 1, \ldots, J_i$; $k \in S_{ij}$; $l = 1, \ldots, n_{ijk}$). Let w_{ijkl} be the corresponding sampling weight available from the survey. The goal is to estimate π_{ij}, the true prevalence of substance use or dependence for the j^{th} county within i^{th} planning area ($i = 1, \ldots, I$; $j = 1, \ldots, J_i$).

4. Direct survey estimator and synthetic estimators

The direct sample survey estimator of π_{ij} is given by

$$\hat{\pi}_{ij}^{D} = \frac{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl} y_{ijkl}}{\sum_{k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}}.$$

The sample size available from a county could be very small (sometimes as small as 3 or 4). Thus, the estimator is highly unreliable. Other direct survey estimators are defined similarly. For example, the direct survey estimator

of π_{ik}, the true prevalence in the k^{th} demographic group in the i^{th} planning region is

$$\hat{\pi}_{ik}^{D} = \frac{\sum_{j:\, k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}\, y_{ijkl}}{\sum_{j:\, k \in S_{ij}} \sum_{l=1}^{n_{ijk}} w_{ijkl}},$$

where the notation $j: k \in S_{ij}$ means that the summation is over counties j in which demographic group k is observed.

Additional problems arise when estimating the proportion of rare events. It is quite likely that all observations in a county may be zero, resulting in a zero estimate for a county. If usual estimates of standard error were applied, an estimated zero standard error of the estimate would give a false impression of the uncertainty of the estimate. Thus, it is very important to improve on the direct survey estimator.

Synthetic estimators borrow strength from related counties through implicit modeling of supplementary data from the U.S. Census Bureau along with the telephone survey data. A synthetic estimator, which has been used in the past to estimate alcohol prevalence at the county level, is given by

$$\hat{\pi}_{ij}^{S1} = \sum_{k=1}^{K} a_{ijk}\, \hat{\pi}_{k}^{D},$$

where $\hat{\pi}_{k}^{D}$ is the statewide direct survey estimator of prevalence of alcohol for the k^{th} demographic group and a_{ijk} is the proportion of individuals belonging to the k^{th} demographic group in the j^{th} county within the i^{th} planning area ($i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K$). The value a_{ijk} is available from current census estimates. For the household survey reported in this paper, the a_{ijk} values were obtained from database vendors like Claritas Data Services of Ithaca, New York. Based on latest available census data, the a_{ijk} values are typically estimated using projection models. In practice, therefore, the a_{ijk} values are not true proportions but are current census estimates of reasonable precision. Outdated or inaccurate a_{ijk} values cause the estimators using them to be biased. If population projections are used to calculate poststratification weighting adjustments in the survey, the direct survey estimator also suffers from this source of bias. It is beyond the scope of this paper to study the impact of alternate population projections. In proposing $\hat{\pi}_{ij}^{S1}$, it is implicitly assumed that the prevalences for alcohol and drug use for the k^{th} group in all the counties are the same (or nearly the same).

A less restrictive synthetic estimator of prevalence of alcohol and drug use is given by

$$\hat{\pi}_{ij}^{S2} = \sum_{k=1}^{K} a_{ijk} \hat{\pi}_{ik}^{D},$$

where $\hat{\pi}_{ik}^{D}$ is a direct survey estimator of π_{ik}, the prevalence for alcohol or drug use for the k^{th} demographic group in the i^{th} planning region. It is implicitly assumed that the prevalences in the k^{th} group are the same (or nearly the same) for all the counties in a planning region. This assumption is more "regional" or less restrictive than the one made in proposing $\hat{\pi}_{ij}^{S1}$. A similar direct survey estimator $\hat{\pi}_{ijk}^{D}$ for the k^{th} demographic group within a specific county j in region i may be defined by restricting the sample to county j only. As compared to $\hat{\pi}_{ijk}^{D}$, the estimator $\hat{\pi}_{ik}^{D}$ will have relatively lower variance although it may have some bias since it does not distinguish the counties. $\hat{\pi}_{ijk}^{D}$, on the other hand, may be based on a very small sample size and hence may be significantly less reliable in terms of its variability.

The above synthetic estimators achieve reductions in variances at the cost of increasing bias. The synthetic estimators distinguish counties only through an indirect variable a_{ijk} obtained from the census, whereas the direct estimator treats each county separately.

5. Composite estimators of π_{ij} using telephone survey and census data

A compromise between a direct survey estimator and a synthetic estimator is a composite estimator. A number of different composite estimators are proposed here based on the following identity:

$$\pi_{ij} = \sum_{k \in S_{ij}} a_{ijk}\, \pi_{ijk} + \sum_{k \notin S_{ij}} a_{ijk}\, \pi_{ijk},$$

where π_{ijk} is the prevalence for alcohol and drug use and a_{ijk}, as defined above, is the proportion of individuals belonging to the k^{th} demographic group in the j^{th} county within the i^{th} planning area ($i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K$).

A simple composite estimator of π_{ij} is obtained when, for $k \in S_{ij}$, π_{ijk} is estimated by $\hat{\pi}_{ijk}^{D}$, the direct survey estimator of π_{ijk}, and for $k \notin S_{ij}$, π_{ijk} is estimated by $\hat{\pi}_{ik}^{D}$. The estimator is then given by

$$\hat{\pi}_{ij}^{C} = \sum_{k \in S_{ij}} a_{ijk}\, \hat{\pi}_{ijk}^{D} + \sum_{k \notin S_{ij}} a_{ijk} \hat{\pi}_{ik}^{D}.$$

In the above formula, π_{ijk} ($k \in S_{ij}$) is estimated using a small sample and thus there is the possibility for improving on $\hat{\pi}_{ijk}^{D}$ (and, hence, on $\hat{\pi}_{ij}^{C}$) by borrowing strength from relevant resources. To this end, an empirical Bayes estimate of π_{ij} is proposed based on the following model.

Model

1. Given the π_{ijk}'s, the y_{ijkl}'s are uncorrelated with one another with $\mathrm{E}(y_{ijkl} \mid \pi_{ijk}) = \pi_{ijk}$ and $\mathrm{Var}(y_{ijkl} \mid \pi_{ijk}) = \pi_{ijk}(1 - \pi_{ijk})$ for $i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K;\ l = 1, \ldots, n_{ijk}$.

2. The π_{ijk}'s are uncorrelated with $E(\pi_{ijk}) = \mu_{ik}$; $\mathrm{Var}(\pi_{ijk}) = d\mu_{ik}^2 (i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K)$.

If $\pi_{ijk} \sim \mathrm{Uniform}(0, 2\mu_{ik})$, then in statement (2) $d = 1/3$. Thus, unlike the implicit assumption made in the synthetic estimator $\hat{\pi}_{ij}^{S2}$ (*i.e.*, $\pi_{ijk} = \mu_{ik}$), some variability of proportions across counties within a region for a particular demographic group is allowed.

The first assumption of the model implies that given π_{ijk}, the $\hat{\pi}_{ijk}^D$'s are uncorrelated with one another, $E(\hat{\pi}_{ijk}^D | \pi_{ijk}) = \pi_{ijk}$, and $\mathrm{Var}(\hat{\pi}_{ijk}^D | \pi_{ijk}) = c_{ijk}\pi_{ijk}(1 - \pi_{ijk})$, where $c_{ijk} = \sum_{l=1}^{n_{ijk}} w_{ijkl}^2 / (\sum_{l=1}^{n_{ijk}} w_{ijkl})^2$ for $i = 1, \ldots, I;\ j = 1, \ldots, J_i;\ k = 1, \ldots, K$. The linear Bayes estimator of π_{ij}, under the model and squared error loss function, is given by

$$\hat{\pi}_{ij}^B = \sum_{k \in S_{ij}} a_{ijk}(B_{ijk}\hat{\pi}_{ijk}^D + (1 - B_{ijk})\mu_{ik}) + \sum_{k \notin S_{ij}} a_{ijk}\mu_{ik},$$

where $B_{ijk} = d\mu_{ik}^2/(d\mu_{ik}^2 + c_{ijk}(\mu_{ik} - (d + 1)\mu_{ik}^2))$.

Since the Bayes estimator involves the unknown parameter μ_{ik}, it cannot be used in practice. The following empirical Bayes estimator of π_{ij} is obtained when μ_{ik} is replaced by an estimator, say $\hat{\mu}_{ik}$, of μ_{ik}:

$$\hat{\pi}_{ij}^{EB} = \sum_{k \in S_{ij}} a_{ijk}(\hat{B}_{ijk}\hat{\pi}_{ijk}^D + (1 - \hat{B}_{ijk})\hat{\mu}_{ik}) + \sum_{k \notin S_{ij}} a_{ijk}\hat{\mu}_{ik},$$

where $\hat{B}_{ijk} = d\hat{\mu}_{ik}^2 / (d\hat{\mu}_{ik}^2 + c_{ijk}(\hat{\mu}_{ik} - (d + 1)\hat{\mu}_{ik}^2))$. The weight or shrinkage factor $\hat{B}_{ijk}$ is a ratio of the variance of π_{ijk} in the model to the (unconditional) variance of $\hat{\pi}_{ijk}$. The estimator of μ_{ik} is taken to be $\hat{\mu}_{ik} = \hat{\pi}_{ik}^D$.

Mean square errors

The mean squared error (MSE) of the Bayes estimator $\hat{\pi}_{ij}^B$ is defined as $\mathrm{MSE}(\hat{\pi}_{ij}^B) = E(\hat{\pi}_{ij}^B - \pi_{ij})^2$, where (unconditional) expectation is taken with respect to the model. It can be checked that

$$\begin{aligned}
\mathrm{MSE}(\hat{\pi}_{ij}^B) &= \mathrm{Var}(\hat{\pi}_{ij}^B - \pi_{ij}) \\
&= \mathrm{Var}(\hat{\pi}_{ij}^B) + \mathrm{Var}(\pi_{ij}) - 2\mathrm{Cov}(\hat{\pi}_{ij}^B, \pi_{ij}) \\
&= \mathrm{Var}(\pi_{ij}) - \mathrm{Var}(\hat{\pi}_{ij}^B) \\
&= d\left(\sum_{k \in S_{ij}} a_{ijk}^2(1 - B_{ijk})\mu_{ik}^2 + \sum_{k \notin S_{ij}} a_{ijk}^2\mu_{ik}^2\right).
\end{aligned}$$

It is customary to take $\mathrm{MSE}(\hat{\pi}_{ij}^B)$ as the MSE of the empirical Bayes estimator $\hat{\pi}_{ij}^{EB}$. However, $\mathrm{MSE}(\hat{\pi}_{ij}^B)$ will underestimate the MSE of $\hat{\pi}_{ij}^{EB}$ since it does not incorporate the uncertainty due to the estimation of the parameter μ_{ik}. See Prasad and Rao (1990) and Lahiri and Rao (1995) in this context. Using a standard Bayesian argument, it can be shown that

$$\mathrm{MSE}(\hat{\pi}_{ij}^{EB}) = \mathrm{MSE}(\hat{\pi}_{ij}^B) + E(\hat{\pi}_{ij}^{EB} - \hat{\pi}_{ij}^B)^2.$$

It is necessary to estimate $\mathrm{MSE}(\hat{\pi}_{ij}^{EB})$ since it contains the unknown parameter μ_{ik}. The first term $\mathrm{MSE}(\hat{\pi}_{ij}^B)$ can be estimated by

$$\begin{aligned}
\mathrm{mse}_J(\hat{\pi}_{ij}^B) &= \mathrm{mse}(\hat{\pi}_{ij}^B) \\
&\quad - \frac{J_i - 1}{J_i}\sum_{u=1}^{J_i}(\mathrm{mse}_{(-u)}(\hat{\pi}_{ij}^B) - \mathrm{mse}(\hat{\pi}_{ij}^B)),
\end{aligned}$$

where

$$\mathrm{mse}(\hat{\pi}_{ij}^B) = d\left(\sum_{k \in S_{ij}} a_{ijk}^2(1 - \hat{B}_{ijk})\,\hat{\mu}_{ik}^2 + \sum_{k \notin S_{ij}} a_{ijk}^2\,\hat{\mu}_{ik}^2\right)$$

and

$$\mathrm{mse}_{(-u)}(\hat{\pi}_{ij}^B) = d\left(\sum_{k \in S_{ij}} a_{ijk}^2(1 - \hat{B}_{ijk(-u)})\,\hat{\mu}_{ik(-u)}^2 + \sum_{k \notin S_{ij}} a_{ijk}^2\hat{\mu}_{ik(-u)}^2\right),$$

with

$$\hat{\mu}_{ik(-u)} = \sum_{j \neq u}^{J_i}\sum_{l=1}^{n_{ijk}} w_{ijkl}y_{ijkl} \Big/ \sum_{j \neq u}^{J_i}\sum_{l=1}^{n_{ijk}} w_{ijkl},$$

and

$$\hat{B}_{ijk(-u)} = d\hat{\mu}_{ik(-u)}^2 \big/ (d\hat{\mu}_{ik(-u)}^2 + c_{ijk}(\hat{\mu}_{ik(-u)} - (d + 1)\hat{\mu}_{ik(-u)}^2)).$$

See Jiang, Lahiri, and Wan (1998) for comment on these estimators. The second term $E(\hat{\pi}_{ij}^{EB} - \hat{\pi}_{ij}^B)^2$ can be estimated with the following jackknife estimator:

$$E_J(\hat{\pi}_{ij}^{EB} - \hat{\pi}_{ij}^B)^2 = \frac{J_i - 1}{J_i}\sum_{u=1}^{J_i}(\hat{\pi}_{ij(-u)}^{EB} - \hat{\pi}_{ij}^{EB})^2,$$

where

$$\begin{aligned}
\hat{\pi}_{ij(-u)}^{EB} &= \sum_{k \in S_{ij}} a_{ijk}\,(\hat{B}_{ijk(-u)}\hat{\pi}_{ijk}^D + (1 - \hat{B}_{ijk(-u)})\hat{\mu}_{ik(-u)}) \\
&\quad + \sum_{k \notin S_{ij}} a_{ijk}\hat{\mu}_{ik(-u)}.
\end{aligned}$$

Thus $\mathrm{MSE}(\hat{\pi}_{ij}^{EB})$ is estimated by

$$\mathrm{mse}(\hat{\pi}_{ij}^{EB}) = \mathrm{mse}_J(\hat{\pi}_{ij}^B) + E_J(\hat{\pi}_{ij}^{EB} - \hat{\pi}_{ij}^B)^2.$$

Jackknife methods are reviewed in the recent text by Shao and Tu (1995).

6. An example

In this study, the primary objective is to provide information about treatment need. Anyone who meets the criteria for lifetime dependence or abuse as defined by the National Technical Center's DSM-III-R criteria, is considered a member of the group of respondents who may have needed treatment during the last year. Several indicator variables were created in the dataset to identify respondents with a diagnosis for substance dependence or abuse for alcohol or drugs. For the purpose of numerical calculations, these indicator variables with 0 and 1 as possible values were treated as response variables (y_{ijkl}).

In order to save space, results are presented for the outcome variable Alcohol Dependence only. Results on other response variables can be obtained from the authors. In order to preserve confidentiality, results for only 40 counties, identified as counties 1 through 40, are reported. Table 1 contains five different estimates of prevalence for alcohol dependence. In general, the direct estimates are highly variable and are often zero. The first synthetic estimator (S1) is the most stable, producing no zero estimates and estimates with little variability. The second synthetic estimator (S2) is similar to S1, but not as restrictive. The first synthetic estimates are very homogeneous, while the second synthetic estimates are homogeneous within the four planning areas. The estimates produced by the composite estimator are more variable than the other estimates. The empirical Bayes estimator produces estimates very similar to those of S2. In the model leading

Table 1
Five estimators of alcohol dependence prevalence expressed as percents for forty counties. Estimated standard errors for direct (Est.se) and square root of estimated mean square error for empirical Bayes ($\sqrt{\text{Est.mse}}$) estimates in parentheses also as percents

	Direct		Estimator Synthetic 1	Synthetic 2	Composite	Empirical Bayes			Number of Groups Observed in County
County	$\hat{\pi}_{ij}^{D}$	(Est. se)	$\hat{\pi}_{ij}^{S1}$	$\hat{\pi}_{ij}^{S2}$	$\hat{\pi}_{ij}^{C}$	$\hat{\pi}_{ij}^{EB}$	($\sqrt{\text{Est.mse}}$)	Sample Size	
1	1.7	(2.4)	3.4	1.6	0.9	1.6	(0.33)	30	8
2	4.4	(2.0)	3.8	1.8	7.2	2.1	(0.35)	111	8
3	0.0	(0.0)	3.6	3.3	0.0	3.0	(0.85)	36	8
4	0.0	(0.0)	3.3	5.6	1.6	5.3	(1.79)	6	5
5	9.4	(4.8)	3.3	5.6	14.1	6.9	(1.78)	37	8
6	1.6	(1.1)	3.4	3.0	1.7	2.7	(0.67)	136	8
7	9.3	(5.8)	3.4	3.1	9.9	3.1	(0.81)	25	6
8	0.0	(0.0)	3.6	3.2	0.4	3.1	(0.84)	20	7
9	0.0	(0.0)	3.4	5.8	5.6	5.8	(1.93)	3	3
10	1.5	(1.3)	3.4	2.1	0.7	1.9	(0.54)	81	8
11	0.0	(0.0)	3.3	1.6	0.0	1.5	(0.33)	58	8
12	7.0	(6.8)	3.5	1.7	5.0	1.8	(0.35)	14	6
13	5.7	(3.8)	3.3	5.5	12.9	6.4	(1.75)	37	8
14	0.0	(0.0)	3.5	1.7	0.8	1.6	(0.33)	12	4
15	2.4	(1.4)	3.3	5.6	2.0	4.4	(1.56)	120	8
16	4.1	(3.5)	3.3	3.0	2.5	3.0	(0.77)	32	7
17	2.8	(2.4)	3.8	1.8	1.3	1.8	(0.37)	48	8
18	3.9	(1.1)	3.4	3.0	3.2	3.2	(0.60)	316	8
19	0.0	(0.0)	3.4	5.7	3.7	5.7	(1.95)	19	5
20	3.1	(3.9)	3.6	3.2	14.9	3.2	(0.82)	20	6
21	2.7	(1.6)	3.3	5.6	4.1	5.8	(1.50)	102	8
22	4.2	(1.8)	3.3	2.1	1.8	2.2	(0.42)	124	8
23	9.7	(2.7)	4.3	8.0	11.8	8.8	(2.11)	121	8
24	0.0	(0.0)	3.3	2.0	0.2	1.9	(0.54)	22	6
25	7.8	(4.7)	3.3	1.6	2.8	1.8	(0.33)	32	6
26	0.0	(0.0)	3.5	1.7	0.0	1.6	(0.37)	28	7
27	2.2	(1.8)	3.2	5.6	1.6	4.9	(1.74)	63	8
28	10.5	(13.7)	3.4	1.6	14.2	1.7	(0.35)	5	5
29	0.0	(0.0)	3.5	3.1	1.8	3.0	(0.81)	12	5
30	0.0	(0.0)	3.2	1.5	0.0	1.5	(0.33)	11	6
31	4.6	(3.2)	3.5	5.9	17.0	5.8	(1.87)	44	8
32	8.4	(3.8)	3.7	3.4	8.4	4.1	(0.84)	52	8
33	2.5	(1.3)	3.4	2.2	2.5	2.1	(0.50)	144	8
34	2.9	(2.4)	3.6	1.7	1.3	1.7	(0.35)	49	7
35	0.0	(0.0)	3.3	3.0	0.0	2.8	(0.77)	22	8
36	0.0	(0.0)	3.4	3.1	0.3	2.9	(0.82)	17	6
37	4.2	(4.0)	3.0	2.0	3.4	2.1	(0.54)	26	6
38	0.0	(0.0)	3.4	5.8	3.7	5.7	(1.97)	16	6
39	0.0	(0.0)	3.5	3.1	0.6	3.0	(0.81)	10	6
40	5.3	(1.9)	3.4	3.1	2.9	3.5	(0.69)	144	8

Table 2
Summary of five estimators of alcohol dependence prevalence for all counties. Results expressed as percents

Estimator	minimum	1st quartile	median	3rd quartile	maximum	mean	standard deviation
Direct	0.0	0.0	2.2	4.3	10.5	2.8	3.2
Synthetic 1	3.0	3.3	3.4	3.5	4.3	3.5	0.2
Synthetic 2	1.5	1.8	3.0	4.4	8.0	3.2	1.7
Composite	0.0	0.4	1.7	4.6	17.5	3.7	4.8
Empirical Bayes	1.5	1.8	2.8	4.2	8.8	3.2	4.8

to the empirical Bayes estimator, d was chosen to be one third.

Table 1 also displays the estimated standard errors of the direct estimates and the square root of the estimated mean squared errors (see section 4) of the empirical Bayes estimates. The standard errors of the direct estimates, which are calculated as

$$\sqrt{\hat{\pi}_{ij}^{D}(1-\hat{\pi}_{ij}^{D})/n_{ij}},$$

are often (incorrectly) estimated to be zero and are quite variable. The square roots of the estimated MSE of the empirical Bayes estimates are relatively stable and always below 0.025.

Table 2 summarizes alcohol dependence estimates in the previous table for all counties in the state. The means of the synthetic and composite estimates are higher than the mean of the direct estimates, because there are fewer zero estimates and the means in the summary tables are unweighted.

7. Conclusion

We have proposed simple empirical Bayes estimators to estimate county level prevalences. Empirical Bayes estimators are found to be very effective when sample sizes for the counties are small and when prevalences are extremely small. We have introduced a measure of uncertainty of the proposed empirical Bayes estimator based on the jackknife method. The proposed measure incorporates additional sources of variability due to estimation of various model parameters. In our model, presented in this paper, we have implicitly assumed that the selection probabilities are unrelated to y_{ijkl}. In the household study reported in this paper, the selection probabilities were unequal and depended on several factors like number of telephone lines and number of adult household members in the household. None of these variables were related to y_{ijkl}. The sample allocation to different regions, however, was done based on the number of "treatment admissions" in each region. Hence, the selection probabilities might be indirectly related to y_{ijkl}. In this paper, we have not addressed the issue of sample selection bias, which can be handled appropriately by following procedures discussed in Pfeffermann (1993).

In this paper, we have not considered the use of auxiliary variables in the model to relate small areas to one another and to facilitate improved estimation. The use of available auxiliary data from the U.S. Census and other administrative records may be a sensible use of resources that can be used to improve planning for treatment of drug and alcohol abuse and dependence. We plan to do further work in this area with an actual example in a future paper.

Acknowledgements

We wish to thank an anonymous referee who had many useful suggestions on improving our paper and the Gallup Organization for partial support. Additionally, Partha Lahiri wishes to acknowledge partial support from U.S. National Science Foundation Grant SBR-9705574.

References

Casady, R.J., and Lepkowski, J.M. (1993). Stratified telephone survey designs. *Survey Methodology*, 19, 103-113.

Efron, B., and Morris, C. (1973). Stein's estimation rule and its competitors - An empirical Bayes approach. *Journal of the American Statistical Association*, 68, 117-130.

Farrell, P.J., MacGibbon, B. and Tomberlin, T.J. (1997). Empirical Bayes estimators of small area proportions in multistage designs. *Statistica Sinica*, 7, 1065-1083.

Fay, R., and Herriot, R. (1979). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M., and Lahiri, P. (1987). Robust empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 82, 1153-1162.

Ghosh, M., and Rao, J.N.K. (1994). Small area estimation: An appraisal. *Statistical Science*, 9, 55-93.

Jiang, J., Lahiri, P. and Wan, S. (1998). Jackknifing Mean Squared Error of Empirical Best Predictor. Unpublished manuscript.

Lahiri, P., and Rao, J.N.K. (1995). Robust estimation of mean squared error of small area estimators. *Journal of the American Statistical Association*, 90, 758-766.

Malec, D., Sedransk, J., Moriarity, C.L. and Leclere, F.B. (1997). Small area inference for binary variables in the National Health Interview Survey. *Journal of the American Statistical Association*, 92, 815-826.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. *International Statistical Review*, 61, 317-337.

Prasad, N.G.N., and Rao, J.N.K. (1990). The estimation of mean squared errors of small area estimators. *Journal of the American Statistical Association*, 85, 163-171.

Shao, J., and Tu, D. (1995). *The Jackknife and the Bootstrap*. New York: Springer-Verlag.