Variance Models Applicable to the NHSDA

James R. Chromy and Lawrence E. Myers, RTI
James R. Chromy, RTI, 3040 Cornwallis Road, PO Box 12194, Research Triangle Park, NC 27709

Key Words: Intracluster correlation, unequal weighting effects, cluster size variation

We propose a variance model which isolates population and design parameters for the 1999 National Household Survey on Drug Abuse. We then use relatively simple methods to obtain estimates of the key parameters. Finally, we validate the model-based estimates of sampling error against empirical estimates using design-based estimation.

The variance of estimates from sample surveys can take three forms: (1) the theoretical formula, (2) design-based estimates, and (3) models for studying design variation. The full theoretical variance function is expressed in terms of population values. It can be expressed conceptually even if the sample design does not permit unbiased estimation. For nonlinear estimates such as ratios, an approximate variance formulation utilizing the first Taylor series approximation may also be used. The estimated variance used for analytic evaluation of the data is generally design-based, but may involve some assumptions about finite population corrections, collapsing of design strata, formation of replicates, or treatment of missing data. Several software packages, utilizing either theTaylor series approximation or replication methods (e.g., Jackknife or balanced repeated replication), are available. The third form of the variance is a model that allows the statistician to consider the impact of adjusting the sample allocation, changing overall sample size, changing the planned sample clustering, or making other design changes. While the full theoretical model would serve this purpose, it is not often possible to estimate the parameters in the full theoretical model. The unbiased variance estimates used for analytic purposes do not isolate the components of variance or attempt to identify design characteristics that might predict any change in the variance. We are left with developing reasonable models for studying design change.

It is not uncommon to see the computed design effect from a survey estimate equated with the expression, $1 + \rho(\bar{m} - 1)$, and assigning the whole increase over the simple random variance to clustering. While simple models are needed for sample design purposes, we believe this model is an oversimplification. The key parameters used in our models include:
- Variance components
- Unequal weighting effects
- Average cluster sizes and measures of cluster size variation.

For simplicity, all of these are estimated on an unweighted basis.

The NHSDA Sample Design

The basic NHSDA 50-state design initiated in 1999 assigns equal sample sizes to most states; eight large states receive allocations about 4 times as large as the remainder to allow for direct state estimates. Younger age groups are also sampled at disproportionately higher rates. The major analytic objective of the 50-state design is to obtain state estimates for three age groups: 12 to 17, 18 to 25, and 26 or older; equal sample sizes for these age groups were required as a basic design constraint. The 26 or older group was further subdivided and sample allocation to the three age subgroups was optimized to obtain more precise estimates for the overall age group. The national sample size resulting from imposing the state sample size requirements also provides adequately precise estimates for other demographically defined populations (race, gender, more detailed age groups, etc.) at the national level.

The household roster is entered into a hand-held computer during the household screening process. The computer applies specified sampling rates at 5 age levels: 12 to17, 18 to 25, 26 to 34, 35 to 49, and 50 or older. The selection probabilities at the segment, dwelling unit, and person stage of selection are coordinated to achieve approximately equal sampling rates within each age group within each state. The selection algorithm allows selection of 0, 1, or 2 persons per dwelling unit. Because the sample is limited to no more than 2 persons per dwelling unit, the design goal of equal weights within state and age group cannot be met even theoretically.

The main features of the basic design for a typical small sample state and a typical large sample state are summarized in Table 1. The NHSDA design also requires that each quarter's sample is a proper subsample of the annual sample, so that any seasonal effects are properly represented in the annual estimates. This also makes it possible to study seasonal variation. Only data from quarters 2, 3, and 4 were utilized for estimating the parameters presented in this paper.

Table 1. NHSDA 1999-2003 Design

Design feature	Typical small sample state	Typical large sample state
FI regions (strata)	12	48
Area segments	96	384
Listed addresses	2700	10800
Respondents		
12-17	300	1200
18-25	300	1200
26 or older	300	1200

The final analytic weights are design-based with further adjustments for nonresponse and for calibration using quarterly Census estimates by age, race, and gender. We attempted to represent as many as possible of the design and estimation features of the 1999 NHSDA in the variance model.

Variance Components

We estimated variance components for state, FI region, area segment, and person. Since this is a completely nested design, we applied the method of moments techniques in SAS PROC NESTED. All four components were treated as random. This allowed us to identify the proportion of variance associated with stratification; i.e., with state and FI region. These stratification components are then excluded from the modeled variance to reflect the gains attributable to stratification.

Variance components were computed for nine substance use and treatment variables by race (Total, Hispanic, black, and other), and by age (12-17, 18-25, 26-34, 35-49, and 50+). We also estimated variance components and other population and design parameters for age groups defined as 26 or older and 35 or older. Examples of estimated variance components for nine measures are shown for the total population in Table 2.

Table 2. Variance Components for Selected Measures for All Persons 12 or Older

Variable	1999 Unwtd. Mean	Variance component as a percent of total variance			
		σ^2_{state}	$\sigma^2_{FI\,region}$	$\sigma^2_{segment}$	σ^2_{person}
Past year, dependent on alcohol	0.037	0.0011	0.0041	0.0166	0.9782
Past month alcohol use	0.472	0.0116	0.0133	0.0563	0.9187
Past month cigarette use	0.259	0.0088	0.0050	0.0372	0.9489
Past month cocaine use	0.007	0.0002	0.0028	0.0036	0.9934
Past year received treatment for illicit drug use	0.008	0.0004	0.0000	0.0108	0.9888
Past year received treatment for alcohol use	0.010	0.0012	0.0008	0.0019	0.9960
Past month illicit drugs except marijuana	0.029	0.0004	0.0002	0.0097	0.9896
Dependent on illicit drugs	0.016	0.0007	0.0046	0.0086	0.9860
Past month any illicit drug	0.068	0.0026	0.0041	0.0263	0.9669

Table 3   Unequal Weighting Effects for the National Estimates (Analytic Weights, Quarters 2, 3, and 4)

Age group	Race/Ethnicity			
	Total	Hispanic	Black	Other
12 or older	4.638	5.324	4.902	4.495
12 to 17	1.560	1.654	1.515	1.548
18 to 25	1.786	1.774	1.777	1.780
26 to 34	1.600	1.719	1.538	1.585
35 to 49	1.696	1.714	1.686	1.690
50 or older	1.940	1.780	1.979	1.929

Unequal Weighting Effects

The unequal weighting effect was treated in two ways: (1) the general variation in weights caused by failure of the sample to maintain equal probability selections and the effects of calibration (nonresponse adjustments and control to known population totals), and (2) intentionally induced unequal weighting due to planned oversampling by age or state.

Unequal weighting effects were computed for a domain as

$$UWE_d = \frac{n_d \sum_{i \in d} W_i^2}{(\sum_{i \in d} W_i)^2}$$

Alternately, one can express the unequal weighting effect in terms of the coefficient of variation of the

weights for the domain of interest:
$$UWE_d = 1 + CV_W^2.$$
Unequal weighting effects based on the final analytic weights are shown in Table 3. Note that the unequal weighting effects for all persons 12 or older are quite high due to the differential sampling rates applied at the age group level. Much more reasonable unequal weighting effects hold within age groups.

By design, states are sampled at different rates. We therefore may want to make use of models which exclude the between-state component of unequal weighting. To obtain average within-state unequal weighting effects for a domain d, we computed the state specific values and computed a weighted average where the weights are proportional to the state domain sample sizes.

Step 1: Compute $UWE_{d,s} = \dfrac{n_{d,s} \sum\limits_{i \in (d,s)} W_i^2}{(\sum\limits_{i \in (d,s)} W_i)^2}$

for each state s.

Step 2: Compute the weighted average
$$\overline{UWE}_d = \dfrac{\sum\limits_s n_{d,s} UWE_{d,s}}{\sum\limits_s n_{d,s}}.$$

Table 4 shows the average within-state unequal weighting effects. Table 4 estimates of the unequal weighting are more appropriate for modeling the variance in a typical state since they exclude the effect of disproportionate allocation across states. In order to achieve annual target sample sizes, the 1999 quarterly

allocations were adjusted; as a result, some additional unequal weighting was introduced. Table 5 illustrates that the unequal weighting effect can be reduced further by examining only one quarter.

Cluster Sizes and Cluster Size Variation

We feel that cluster size variation in conjunction with intracluster correlation helps explain the total effects of clustering. A heuristic model can be developed based on the average of the simple clustering effect, $1 + \rho(m_i - 1)$, taken over varying cluster sizes and then taking the super-population expectation of m_i^2:

$$\sum \frac{m_i}{\overline{nm}} \{1 + \rho(m_i - 1)\} = 1 + \rho(\sum \frac{m_i^2}{\overline{nm}} - 1)$$
$$\doteq 1 + \rho(\overline{m}^* - 1)$$

where $\overline{m}^* = \overline{m}(1 + CV_m^2)$. Table 6 shows average cluster sizes and their coefficients of variation for the subpopulations studied.

Note that while the average cluster size for some domains is less than 1, which would imply a reduction in variance with a positive intraclass correlation, the value of m^* is always greater than 1 for these domains. This is intuitively appealing, since we should always expect clustering to increase the variance if the intracluster correlation coefficient is positve. Note also that the smallest domains, for which we have little control on cluster size, also have the highest cluster size coefficient of variation.

Table 4   Average Within-State Unequal Weighting Effects (Analytic Weights, Quarters 2, 3, and 4)

Age group	Race/Ethnicity			
	Total	Hispanic	Black	Other
12 or older	3.379	4.394	3.716	3.139
12 to 17	1.313	1.433	1.333	1.260
18 to 25	1.341	1.570	1.350	1.291
26 to 34	1.312	1.451	1.297	1.253
35 to 49	1.250	1.300	1.242	1.219
50 or older	1.236	1.256	1.237	1.212

Table 5   Average Within-State Unequal Weighting Effects (Analytic Weights, Quarter 3 Only)

Age group	Race/Ethnicity			
	Total	Hispanic	Black	Other
12 or older	3.331	4.321	3.826	2.990
12 to 17	1.269	1.465	1.274	1.177
18 to 25	1.302	1.509	1.269	1.206
26 to 34	1.276	1.287	1.247	1.185
35 to 49	1.201	1.158	1.177	1.153
50 or older	1.200	1.186	1.207	1.139

Table 6. Average Cluster Sizes and Cluster Size Coefficients of Variation

Age	Race	$\bar{m}$	$c.v.(\bar{m})$	$\bar{m}*$
Total	Total	9.780	0.650	13.915
Total	Hispanic	1.254	2.452	8.791
Total	Black	1.220	2.513	8.924
Total	White	7.306	0.804	12.029
12-17	Total	3.725	0.893	6.699
12-17	Hispanic	0.521	2.850	4.756
12-17	Black	0.494	2.985	4.900
12-17	White	2.709	1.068	5.800
18-25	Total	3.241	1.082	7.034
18-25	Hispanic	0.439	2.845	3.991
18-25	Black	0.427	3.000	4.269
18-25	White	2.376	1.300	6.390
26-34	Total	1.152	1.219	2.864
26-34	Hispanic	0.177	3.502	2.350
26-34	Black	0.135	3.570	1.858
26-34	White	0.840	1.405	2.497
35-49	Total	0.896	1.191	2.167
35-49	Hispanic	0.079	4.163	1.446
35-49	Black	0.098	3.994	1.667
35-49	White	0.719	1.366	2.060
50 +	Total	0.765	1.343	2.144
50 +	Hispanic	0.037	5.863	1.310
50 +	Black	0.065	4.664	1.483
50 +	White	0.663	1.487	2.127

Variance Component Models

We then developed variance models relating to various estimates and domains commonly reported in the NHSDA.

Model 1: For one of the 5 age groups (12 to 17, 18 to 25, 26 to 34, 35 to 49, or 50 or older) designated by the subscript a,

$$Var(\hat{p}_a) = \frac{p_a(1-p_a)}{n_a} UWE_a *$$

$$\{\sigma^2_{segment}\, \bar{m}_a (1 + CV^2_m) + \sigma^2_{person}\}$$

Model 2: The variance model for all persons 12 and older (5 age groups combined)

$$Var(\hat{p}_{12+}) = \frac{1}{n_{12+}} \left\{ \sum_{a=1}^{5} W_a^2 \frac{p_a(1-p_a)}{n_a/n_{12+}} UWE_a \right\} *$$

$$\{\sigma^2_{segment}\, \bar{m}_{12+} (1 + CV^2_{m_{12+}}) + \sigma^2_{person}\}$$

Model 3: For all persons 26 or older:

$$Var(\hat{p}_{26+}) = \frac{1}{n_{26+}} \left\{ \sum_{a=3}^{5} W_a^2 \frac{p_a(1-p_a)}{n_a/n_{26+}} UWE_a \right\} *$$

$$\{\sigma^2_{segment}\, \bar{m}_{26+} (1 + CV^2_{m_{26+}}) + \sigma^2_{person}\}$$

Model 4: For all persons 35 and older:

$$Var(\hat{p}_{35+}) = \frac{1}{n_{35+}} \left\{ \sum_{a=4}^{5} W_a^2 \frac{p_a(1-p_a)}{n_a/n_{35+}} UWE_a \right\} *$$

$$\{\sigma^2_{segment}\, \bar{m}_{35+} (1 + CV^2_{m_{35+}}) + \sigma^2_{person}\}$$

Model Validation

Design-based estimates of variance were computed for nine selected variables (Table 2) whose means estimate a prevalence rate for the population and various domains defined by age and race groups. The relation of the modeled relative standard errors to design-based estimates are shown in Figure 1. Only variance estimates that would not be suppressed were included in the Figure 1 plot. This mainly excluded cases where the estimate was zero or where the coefficient of variation from the design-based estimate exceeded 50 percent.

Conclusions

The modeled relative standard errors provide a realistic approximation to those obtained from design-based estimates. Because they express the variance in terms of design parameters, they are useful for evaluating the impact of alternate designs configurations. The simple (unweighted) approach to variance component estimation appears to provide useful results in spite of ignoring the weights.

The impact of unequal weighting is treated as a multiplicative factor. While the unequal weighting effect can be controlled to some extent by the survey design, the impact of nonresponse and weight adjustment for nonresponse and for calibration against external data can only be controlled in a general way. The unequal weighting effects are not easily subject to any optimization strategy.

The model treatment of variable cluster sizes, particularly for small domains, should be useful in developing variance models for a wide variety of applications.

Figure 1. Projected and Design-Based Relative Standard Errors