1999 NATIONAL HOUSEHOLD SURVEY ON DRUG ABUSE

Statistical Inference

Authors:

James R. Chromy
Teresa R. Davis
Lisa E. Packer

Project Manager:

Tom Virag - Project Director

Contract No. 283-98-9008
RTI Project No. 7190

Substance Abuse and Mental Health Services Administration
Office of Applied Studies
5600 Fishers Lane
Room 16-105
Rockville, MD 20857

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service

July 2001

1999 NATIONAL HOUSEHOLD SURVEY ON DRUG ABUSE

Statistical Inference

Authors:

James R. Chromy
Teresa R. Davis
Lisa E. Packer

Project Manager:

Tom Virag - Project Director

Contract No. 283-98-9008
RTI Project No. 7190

Substance Abuse and Mental Health Services Administration
Office of Applied Studies
5600 Fishers Lane
Room 16-105
Rockville, MD 20857

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Public Health Service

July 2001

Table of Contents

STATISTICAL INFERENCE

1. Introduction

The 1999 National Household Survey on Drug Abuse (NHSDA) is part of a 5-year sample design to provide national estimates and state estimates of drug use through 2003. For the 5-year 50-state design, 8 states were designated as large sample states (California, Florida, Illinois, Michigan, New York, Ohio, Pennsylvania, and Texas), with samples large enough to support direct state estimates. For the remaining 42 states and the District of Columbia, smaller--but adequate--samples were selected to support state estimates using small area estimation (SAE) techniques.

Using the 50 state design, states were first stratified into a total of 900 field interviewer (FI) regions (48 regions in each large sample state and 12 regions in each small sample state). Within FI regions, adjacent census blocks were combined to form the first stage sampling units, called area segments. Eight sample segments per FI region were fielded during the 1999 survey year. These sampled segments were allocated equally into four separate samples, one for each 3 month period during the year, so that the survey is essentially continuous in the field.

The 1999 NHSDA final respondent sample of 66,706 persons was representative of the U.S. general population (the civilian noninstitutional population) aged 12 or older. In addition, state samples were representative of their respective state populations.

2. Sampling Error

The national and direct large-state estimates along with the associated variance components were computed using RTI's multi-procedure package Software for Statistical Analysis of Correlated Data (SUDAAN). The final, nonresponse-adjusted and post-stratified analysis weights, were used in SUDAAN to compute unbiased design-based drug use estimates. The variance estimates were calculated using the SUDAAN option,[1] which is unbiased for linear statistics based on multi-stage clustered sample designs where the first-stage (primary) sampling units are drawn with replacement. As had been done in previous years mainly for quality control purposes, two other variance estimates were computed. The second variance was based only on the stratification and unequal weighting effects, and the third was based on no effects or simple random sampling. The reported variance estimate was then the maximum of these three

[1] This SUDAAN option is DESIGN=WR.

estimates, an approach designed specifically for estimates that can be represented as proportions and to ensure that only conservative estimates of sampling error were published.

A review of the current maximum-of-three rule was initiated as the issue evolved of how to compute these conservative estimates of sampling error consistently across various types of analyses. To ensure that all sampling error estimates would be calculated using the same methodology, the decision was made in November of 2000 to eliminate the use of the maximum-of-three rule. The new procedure would use only the stratified and clustering SUDAAN option for computing sampling errors. This adjustment would be implemented for any additional analyses done as of that date forward using the 1999 NHSDA data.

Estimates of means or proportions, such as drug use prevalence, take the form of nonlinear statistics where the variances are not capable of being expressed in closed form. Variance estimation for nonlinear statistics in SUDAAN is based on a first-order Taylor series approximation of the deviations of estimates from their expected values.

Key nesting variables were created to capture explicit stratification and to identify clustering, because of the nature of stratified-clustering sampling design. For the 1999 NHSDA, each FI region consists of its own stratum. Two replicates per year are defined within each variance stratum. Each variance replicate consist of four segments: one segment for each quarter of data collection.

The sample estimate, $\hat{Y}$, for a population total Y is obtained as the weighted sum of the stratum totals, or

$$\hat{Y} = \sum_{h=1}^{H} \left(\hat{Y}_{h1} + \hat{Y}_{h2} \right)$$

where $\hat{Y}_{h1}$ and $\hat{Y}_{h2}$ are the replicate level contributions to the weighted sample total from stratum h. The variance of the sample total is estimated by

$$var\left(\hat{Y} \right) = \sum_{h=1}^{H} \left(\hat{Y}_{h1} - \hat{Y}_{h2} \right)^2$$

For ratio estimates, such as the NHSDA drug use estimates, for instance P=(Y/X), the linearization variance employed by SUDAAN replaces the replicate totals $\hat{Y}_{hi}$ with linearized

variates $Z_{hi} = (\hat{Y}_{hi} - \hat{P}\hat{X}_{hi})/\hat{X}$ where the replicate totals $\hat{Y}$ of the ratio estimate are a subset of the ratio denominator $\hat{X}$.

3. Confidence Intervals

In some NHSDA publications, sampling error was quantified using 95% confidence intervals. Because the estimates in the NHSDA are frequently small percentages, the confidence intervals based on logit transformations. Logit transformations yield asymmetric interval boundaries that are more balanced with respect to the probability that the true value falls below or above the interval boundaries than is the case for standard symmetric confidence intervals for small proportions.

To illustrate the method, let the proportion P_d represent the true prevalence rate for a particular analysis domain "d". Then the logit transformation of P_d, commonly referred to as the "log odds," is defined as:

$$L = \ln[P_d/(1-P_d)]$$

where "1n" denotes the natural logarithm.

Letting p_d be the estimate of the proportion, the log odds estimate becomes $\hat{L} = \ln[p_d/(1-p_d)]$. Then the lower and upper confidence limits of L are formed as

$$A = \hat{L} - K\left[\frac{\sqrt{var(p_d)}}{p_d(1-p_d)}\right]$$

$$B = \hat{L} + K\left[\frac{\sqrt{var(p_d)}}{p_d(1-p_d)}\right]$$

where var(p_d) is the variance estimate of p_d, the quantity in brackets estimates the standard error of $\hat{L}$, and K is the constant chosen to yield a level of confidence. (e.g., K = 1.96 for 95% confidence limits).

Applying the inverse logit transformation to A and B above yields a confidence interval for p_d as follows:

$$p_{d,lower} = \frac{1}{1 + \exp(-A)}$$

$$p_{d,upper} = \frac{1}{1 + \exp(-B)}$$

where "exp" denotes the inverse log transformation. The lower and upper confidence interval endpoints for percentage estimates are obtained by multiplying the lower and upper endpoints of p_d by 100.

Corresponding to the percentage estimates, the number of drug users, Y_d, can be estimated as

$$\hat{Y}_d = \hat{N}_d \bullet p_d$$

where

$\hat{N}_d$ = estimated population total for domain d

p_d = estimated proportion for domain d.

The confidence interval for $\hat{Y}_d$ is obtained by multiplying the lower and upper limits of the proportion confidence interval by $\hat{N}_{d.}$ This approach is theoretically correct when the domain size estimates $\hat{N}_d$ are among those forced to Census Bureau population projections by our final weight adjustments. In these cases, $\hat{N}_d$ is clearly not subject to sampling error. For domain totals $\hat{Y}_d$ where $\hat{N}_d$ is not fixed, the confidence interval approximation assumes that the sampling variation in $\hat{N}_d$ is negligible relative to the error in p_d.

4. Variance of Prevalence Rates

For a given variance estimate, the associated design effect is the ratio of the design-based variance estimate over the variance that would have been obtained from a simple random sample of the same size. The NHSDA design involves stratification, clustering, and unequal weighting. Clustering and unequal weighting usually increase the design-based variance (design effect greater than 1), but stratification along with effective allocation of the sample can actually decrease the design-based variance relative to what would be obtained using a simple random sample (design effect less than 1). The maximum-of-three rule was developed for the sample designs used prior to 1999 when it was generally believed the combined effects of stratification, clustering, and unequal weighting would always lead to a design effect greater than 1. Since there was concern about declaring unwarranted significant results when interpreting data from

published reports, using the maximum of the three separate variance estimates provided additional protection against making such errors. As a result of this rule, no published standard error estimate ever reflected a design effect of less than 1.

The maximum-of-three rule continued to be applied to 1999 reports published through about November 2000. The new 50-state design provides very effective geographic stratification and 900 degrees of freedom for estimating sampling error for national estimates. An empirical review of the relationships among the three variance estimates and a study of simple variance components, lent support to the credibility of some design effects being less than 1. The stability of the design-based variance estimates was considered much improved under the new design and larger sample. In addition, the suppression rules used in NHSDA reports would help prevent spurious interpretations of data. As a result, the maximum-of-three rule was discontinued and only the design-based variances and standard errors were used in subsequent reports.

5. Suppression of Estimates with Low Precision

Direct survey estimates, noted by asterisks (*), are not reported as they are considered to be unreliable because of unacceptable large sampling errors. The criterion used for suppressing all direct estimates was based on the relative standard error (RSE) of the estimate. The RSE is defined as the ratio of the standard error of the estimate over the estimate itself. For proportion estimates (p) within the range [1<p<1], rates and corresponding estimates numbers of users were suppressed if:

$$\frac{SE(p)/p}{-\ln(p)} > 0.175 \quad \text{when } p \le 0.5 \quad \text{or}$$

$$\frac{SE(p)/(1-p)}{-\ln(1-p)} > 0.175 \quad \text{when } p > 0.5$$

where SE(p) equals the standard error estimate of p. This is an ad hoc rule that requires an effective sample size exceeding 50 when $0.10 \le p \le 0.90$. As (p) approaches 0.00 or 1.00, it requires increasingly larger effective sample sizes. The log transformation of p is used to provide a more balanced treatment of measuring the quality of small, large, and intermediate p values. The switch to (1-p) for p greater than 0.5 provides a symmetric suppression rule across the range

of possible p values. Estimates were also suppressed if they were close to zero or 100% (if $p < .00005$ or if $p \geq .99995$).

For estimates of other totals, and means (not bounded between 0 and 1) estimates were suppressed if $SE(p)/p > 0.5$. Additionally, estimates of mean age were suppressed if the sample size was smaller than 10 respondents.

6. Incidence Estimates

To assist in the evaluation of trends in initiation of drug use, the National Household Survey on Drug Abuse (NHSDA) data was also used to generate estimates of drug use incidence, or initiation (i.e. number of new users during a given year). Incidence rates measure the rapidity with which new drug users arise, and can suggest emerging patterns of drug use.

The measure of incidence is defined as the number of new cases of drug initiation divided by the person time of exposure. For diseases, the incidence rate for a population is defined as the number of new cases of the disease, ***N,*** divided by the person time, ***PT***, of exposure or

$$IR = \frac{N}{PT}.$$

The person time of exposure can be measured for the full period of the study or for a shorter period. The person time of exposure ends at the time of diagnosis (e.g., Greenberg et al, 1996, pp. 16-19). Similar conventions were followed for defining the incidence of first use of a substance.

Beginning in 1999, the NHSDA questionnaire allows for collection of year and month of first use for recent initiates. Month, day, and year of birth are also obtained directly or imputed in the process. In addition, the questionnaire call record provides the date of the interview. By imputing a day of first use within the year and month of first use reported or imputed, we then have the key respondent inputs in terms of exact dates. Exposure time can be determined in terms of days and converted to an annual value.

Having exact dates of birth and first use also allows us to determine person time of exposure during the targeted period, ***t***. Let the target time period for measuring incidence be

specified in terms of dates; e.g., for the period 1998 we would specify

$$t = [t_1, t_2) = [1\ Jan\ 1998, 1\ Jan\ 1999),$$

a period that includes 1 January 1998 and all days up to but not including 1 January 1999. The target age group can also be defined by a half open interval as $a = [a_1, a_2)$. For example, the age group 12 to 17 would be defined by $a = [12, 18)$ for persons at least age 12, but not yet age 18. If person ***i*** was in age group ***a*** during period ***t***, the time and age interval, $L_{t,a,i}$, can then be determined by the intersection

$$L_{t,a,i} = [t_1,t_2) \cap [DOB_i\ MOB_i\ YOB_i + a_1, DOB_i\ MOB_i\ YOB_i + a_2)$$

where we defined the time of birth as in terms of day (DOB_i), month (MOB_i), and year (YOB_i). Either this intersection will be empty ($L_{t,a,i} = \varnothing$) or it was designed by the half-open interval $L_{t,a,i} = [m_{1,i}, m_{2,i})$, where

$$m_{1,i} = Max\{t_1, (DOB_i\ MOB_i\ YOB_i + a_1)\}$$

and

$$m_{2,i} = Min\{t_2, (DOB_i\ MOB_i\ YOB_i + a_2)\}.$$

The date of first use, $t_{fu,d,i}$, is also expressed as an exact date. An incident of first drug ***d*** use by person ***i*** in age group ***a*** occurs in time ***t*** if $t_{fu,d,i} \in [m_{1,i}, m_{2,i})$. The indicator function $I_i(d, a, t)$ used to count incidents of first use is set to 1 when $t_{fu,d,i} \in [m_{1,i}, m_{2,i})$, and to 0 otherwise. The person time exposure measured in years and denoted by $e_i(d,a,t)$ for a person ***i*** of age group ***a*** depends on the date of first use. If the date of first use precedes the target period ($t_{fu,d,i} < m_{1,i}$), then $e_i(d,a,t) = 0$. If the date of first use occurs after the target period or if person ***i*** has never used drug ***d***, then

$$e_i(d,a,t) = \frac{m_{2,i} - m_{1,i}}{365}.$$

If the date for first use occurs during the target period $L_{t,a,i}$, then

$$e_i(d,a,t) = \frac{t_{fu,d,i} - m_{1,i}}{365}.$$

Note that both $I_i(d,a,t)$ and $e_i(d,a,t)$ are set to zero if the target period $L_{t,a,i}$ is empty; i.e., person ***i*** is not in age group ***a*** during time ***t.*** The incidence rate is then estimated as a weighted ratio estimate:

$$IR(d,a,t) = \frac{\sum_i w_i I_i(d,a,t)}{\sum_i w_i e_i(d,,a,t)}$$

where the w_i are the analytic weights.

In previous years, before the exact date data were available for computing incidence rates, a person was considered to be of age ***a*** during the entire time interval ***t*** , if his/her ***a***th birthday occurred during time interval ***t*** (generally, a single year). If the person initiated use during the year, the person time exposure was approximated as one-half year for all such persons rather than computing it exactly for each person.

Because of the new methodology, the 1999 NHSDA incidence estimates are not strictly comparable to prior year estimates. However, because they are based on retrospective reports by survey respondents (as was the case for earlier estimates), they may be subject to some of the same kinds of biases.

Bias resulting from differential mortality occurs because some persons who were alive and exposed to the risk of first drug use in the historical periods shown in the tables died before the 1999 NHSDA was conducted and is probably very small. Incidence estimates are also affected by memory errors, including recall decay (tendency to forget events occurring long ago) and forward telescoping (tendency to report that an event occurred more recently than it actually did). These memory errors would both tend to result in estimates for earlier years (i.e., 1960s and 1970s) that are downwardly biased (because of recall decay) and estimates for later years that are upwardly biased (because of telescoping). There is also likely to be some underreporting bias because of social acceptability of drug use behaviors and respondents' fears of disclosure. This is likely to have the greatest impact on recent estimates, which reflect more recent use and reporting by younger respondents. Finally, for drug use that is frequently initiated at age 10 or younger, estimates based on retrospective reports one year later underestimate total incidence because 11-year-old children are not sampled by the NHSDA. Prior analyses showed that alcohol and cigarette (any use) incidence estimates could be significantly affected by this. Therefore, there are no 1998 estimates made for these drugs.

REFERENCES

Bowman, K. R., Penne, M. A., Chromy, J. R., & Odom, D. M. (2000). 1999 National Household Survey on Drug Abuse: sample design report. Research Triangle Park, NC: Research Triangle Institute.

Chromy, J.R. (1981). Variance estimators for a sequential sample selection procedure. In D. Krewski, R. Platek, & J.N.K. Rao (Eds.) *Current topics in survey sampling*, (pp. 329-347).

Hansen, M. H., Hurwitz, W. N., & Madow, G. W. (1953). *Sample survey methods and theory*. New York, NY: Wiley.

Office of Applied Studies. (2000b). *Summary of findings from the 1999 National Household Survey on Drug Abuse*. (National Household Survey on Drug Abuse Series: H-12, DHHS Publication No. SMA 00-3466. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Shah, B. V., Barnwell, B. G., & Bieler, G. S. (1997). *SUDAAN user's manual: Version 7.5.* Research Triangle Park, NC: Research Triangle Institute.

Woodruff, R. S. (1971). A simple method for approximating the variance of a complicated estimate. *Journal of the American Statistical Association*, 66, 411-414.