Journal of Official Statistics, Vol. 13, No. 4, 1997, pp. 401–416

Ratio Estimation of Hardcore Drug Use

Doug Wright, Joe Gfroerer, and Joan Epstein[1]

Levels of hardcore drug use have been especially difficult to estimate because of the relative rareness of the behavior, the difficulty of locating hardcore drug users, and the tendency to underreport stigmatized behavior. We present a new application of ratio estimation combining sample data from the National Household Survey on Drug Abuse (NHSDA) together with population counts of the number of persons arrested in the past year (from the Uniform Crime Report) and the number of persons in drug treatment programs in the past year (from the National Drug and Alcoholism Treatment Unit Survey). The population counts serve as a benchmark accounting for undercoverage and underreporting of hard-drug users.

Key words: Ratio estimation; benchmarking; hard core drug use.

1. Introduction

The need for accurate estimates of the size of the ''hardcore'' drug using population is substantial. Regardless of how it is specifically defined, this population of heavy drug users is likely to need significant resources for treatment of their drug problem and associated medical and other problems. Hardcore drug users have also been shown to be responsible for a disproportionate amount of crime.

This article describes a method for estimating the prevalence of hardcore drug use based on the National Household Survey on Drug Abuse (NHSDA) in conjunction with outside sources and the methodology of ratio estimation. In ratio estimation, one can often obtain a better estimate of a population total if there is a known population total for a related variable. Then the estimate of the total is $X' = (x/y) * Y$, where x is the variable of interest, y is the related variable, and Y is the known population total for the related variable.

Another way of describing this method is to say that it ''inflates'' (i.e., gives more weight to) the drug prevalence data from the NHSDA for populations with characteristics that are known to be related to hardcore drug use but are also underestimated. In this case we know that the NHSDA undercounts arrestees and drug treatment populations, so we ''ratio adjust'' the NHSDA hardcore drug use estimates upward to externally derived counts of arrestees and treatment clients that are believed to be accurate.

In survey sampling theory, ratio estimation is often associated with the desire to improve the precision of an estimate. The ratio estimate will be better, in the sense that it will have a smaller variance, than the simple expansion estimator $X'' = \sum w_i x_i$ that

[1] Prevalence Branch, Office of Applied Studies, Substance Abuse and Mental Health Services Administration, Health and Human Services, 5600 Fishers Lane, Room 16-105, Rockville, MD 20857, U.S.A.

is commonly used, when certain conditions are met. (See Section 7 on precision of the estimates.)

However, in this application, we are less interested in variance reduction and more interested in bias reduction. Ratio estimates have been used for a number of years to adjust for nonresponse and to adjust to known population counts, often based on a census. This application represents an extension of those earlier uses to one in which we use known population counts to adjust NHSDA sample estimates for underreporting and undercoverage.

2. Background

The National Household Survey on Drug Abuse (NHSDA) is the primary source of statistical information on the use of illegal drugs by the United States population. Conducted periodically by the Federal Government since 1971, the survey collects data by administering questionnaires to a representative sample of persons age twelve and older living in the nation. Since October 1, 1992 the survey has been sponsored by the newly created Substance Abuse and Mental Health Services Administration (SAMHSA). The primary purpose of the survey is to estimate the prevalence of illegal drug use in the United States, and to monitor trends in prevalence over time. Prevalence rates for selected population subgroups and types of drugs are generated from the survey data, providing insight into the population groups most at risk for illicit drug use and the drugs that are most commonly used.

The respondent universe for the 1992 NHSDA is the civilian noninstitutionalized population aged twelve years old and older within the United States, including the residents of noninstitutional group quarters (e.g., shelters, rooming houses, dormitories) and residents of civilian housing on military bases. Persons excluded from the universe include the homeless not found in shelters, residents of institutional group quarters, such as jails and hospitals, and active military personnel.

The 1992 survey employed a multistage area probability sample of 28,832 persons interviewed from January through December 1992. The screening and interview response rates were 95 per cent and 83 per cent, respectively, for an overall response rate of 79 per cent. The sample design incorporates varying selection probabilities which result in oversampling of blacks, Hispanics, and young people, to improve the reliability of estimates for those populations. Also incorporated in the NHSDA sample design were special samples of about 2,500 respondents in each of six large metropolitan areas (New York, Washington, DC, Miami, Chicago, Denver, and Los Angeles).

The household interview takes about an hour to complete and incorporates procedures designed to maximize honest reporting of illicit drug use, including the use of self-administered answer sheets. Data are collected on the recency and frequency of use of various licit and illicit drugs, demographic characteristics, problems associated with drug use, and drug abuse treatment experience. Further information on the methodology of the NHSDA is provided elsewhere (SAMHSA 1993a, 1994).

Comparison of NHSDA estimates with a variety of sociodemographic variables from independent sources (e.g., the Current Population Survey) typically has shown good agreement. However, estimating the number of hardcore drug users has historically

been a difficult problem. Household interview surveys such as the NHSDA were not designed for this type of estimation and are believed to be inadequate tools for measuring hardcore drug use because of the low prevalence of the behavior and difficulties in accessing this population. Underreporting (survey participants who do not report their drug use) and undercoverage (inability to roster hardcore drug users) also affect this estimation.

In comparing the results of NHSDA estimates to those from various administrative records systems (e.g., drug treatment program data, parole, probation, or arrest data from the FBI, etc.), the apparent underreporting of these types of characteristics by the sample respondents has been significant. Research has shown that underreporting of drug use increases as the reference period approaches the present and as the perceived social disapproval increases (Harrell et al. 1986; Turner et al. 1992). This suggests that hardcore drug use is underestimated more than casual drug use. The underestimates could also be the result of undercoverage of the populations with these characteristics.

Various methods have been used to estimate hardcore drug prevalence including capture-recapture techniques, truncated Poisson, and modeling methods generally (Brodsky 1985; Woodward et al. 1985; Hser et al. 1992). These methods have been based primarily on data from administrative records such as treatment admission data, essentially ignoring household survey data. Other methods have supplemented household survey data with other sources of data (e.g., arrest data) to construct composite estimates of hardcore drug use (Wish 1990–91; Rhodes 1993).

There also has been significant research on various data collection methods that encourage honest responses to sensitive questions. Such methods include randomized response, item count, nominative, and computer assisted self interview techniques (Biemer et al. 1991; Miller 1985; Warner 1965; Zdep et al. 1979; Caspar et al. 1996).

Randomized response involves a randomizing device, such as a pair of dice, and two questions, one of which is sensitive (''Have you used heroin in the past year?'') and one of which is innocuous (e.g., ''Were you born in September?''). The respondent uses the randomizing device to determine which question to answer, and the interviewer records the answer (yes or no), but not the question. This method can be used to estimate the proportion having the sensitive characteristic.

In the item count technique, (a random subsample of) respondents are asked to read a list which includes the sensitive item, such as heroin, and to indicate how many of the items in the list they have used in the past year. Other respondents are asked to read a list which does not include the sensitive item and to indicate how many of the items they have used in the past year. From these two samples it is possible to estimate the number of users of the sensitive item without identifying an individual user.

The nominative technique first asks the respondent to indicate how many of his or her close friends have the sensitive characteristic (e.g., use of heroin in the past year). The second question is asked about each of the close friends, ''How many of this person's other close friends (besides yourself) know that he (or she) has used heroin?'' With these two questions the count of the number of past year heroin users can be corrected for duplication.

Audio Computer Assisted Self Interviewing (ACASI) allows the respondent to listen to questions through a headset, or to read the questions on a computer screen, or both. The

computer assistance and the audio supplement can give respondents a greater sense of privacy than is possible with the interviewer-assisted self interview approach. ACASI techniques have been shown to produce higher prevalences of some sensitive behaviors (Duffer et al. 1996; Turner et al. 1996).

Although many studies have generally established the validity of self-report data and the NHSDA methodology has been shown to produce more valid results than some other reporting methods (e.g., by telephone), comparisons of NHSDA data with data from surveys conducted in classrooms suggest that underreporting of drug use by youths in their homes may be substantial (Turner et al. 1992; Gfroerer 1992).

In the following discussion we want to focus on the ratio estimate's ability to correct for bias (in particular, the undercounting of hard drug users in the NHSDA) given a true population value of a related variable. To make the discussion more concrete, we will apply the estimation procedure to four separate, but overlapping, measures of hardcore drug use for 1992: the number of past year users of heroin, weekly users in the past year of cocaine, past year users who are dependent on some illicit drug, and past year intravenous drug users.

3. Basic Methodology

The information that we wish to make use of is the count of the number of persons in treatment centers for drug abuse during the past year (1992) from the National Drug and Alcoholism Treatment Unit Survey (NDATUS) (SAMHSA 1993) and the known count of the number of arrests (for any crime other than minor traffic violations) during the past year (1991) from the FBI Uniform Crime Reporting (Maguire, Pastore, and Flanagan 1993).

3.1. Using a single population count

Let Nt be the estimated count of the number who received treatment for a drug problem during the past year derived from the NDATUS. The count was computed by multiplying the number of treatment slots times the average number of persons treated per year per slot, and includes an adjustment for multiple episodes by the same individual. NDATUS is an annual national universe of publicly and privately funded drug abuse and alcoholism treatment and prevention facilities. Approximately 82 per cent of active eligible facilities responded to the 1991 survey. Since nonrespondent facilities tended to be small, the per cent of clients represented in this survey is greater than the 82 per cent facility response rate. The estimated count adjusts for facility nonresponse.

Na is the estimated count of the number of persons arrested during the past year. Na is calculated by taking the 1991 (latest available at that time) FBI Uniform Crime Report estimated number of arrests, 14,211,900, and dividing this by the average number of arrests per person arrested calculated from the NHSDA, 1.46, resulting in an estimate of 9,734,178. (The average number of arrests per person arrested has been calculated for a number of NHSDA survey years, and the results are very consistent.) Based on recent trends, the 1992 estimate would be expected to be slightly higher than the 1991 estimate.

The typical use of a ratio estimate occurs when the outside source fully overlaps the population of interest. For example, if we wanted to estimate the number of students in

school and had information on the total number of teachers, then this source overlaps the population of interest since every student has a teacher. Then, if we developed an estimate of the average number of students per teacher, we could multiply the total number of teachers times the average number of students per teacher to obtain an estimate of the total number of students.

Our situation is slightly different in that neither our treatment nor arrest population counts fully overlap the population of hard-drug users. However, we can construct counts from the US Census Bureau estimates for 1992 that are so precise at the national level that we can consider them to be population counts. With these we can develop counts that cover the population.

Given the Census estimate of the number of noninstitutionalized persons twelve and older – $N = 205{,}713{,}000$ (for July 1, 1992, the count at the midpoint of data collection for the NHSDA target population) – we can form two pairs of counts that cover the population.

Number of persons in treatment $Nt = 1{,}789{,}000$

Number of persons not in treatment $N - Nt = 203{,}924{,}000$

Number of persons arrested $Na = 9{,}722{,}671$

Number of persons not arrested $N - Na = 195{,}990{,}329$.

From these counts, two estimates of the number of persons using heroin during the past year (1992) are possible.

$$H(t) = r(t)^{*}Nt + r(t)^{*} * (N - Nt) = (139{,}003/834{,}702) * 1{,}789{,}000$$
$$+ (171{,}136/200{,}656{,}309) * 203{,}924{,}000 = 471{,}844.$$
$$H(a) = r(a)^{*}Na + r(a)^{*} * (N - Na) = (184{,}277/4{,}743{,}706) * 9{,}722{,}671$$
$$+ (125{,}861/196{,}747{,}306) * 195{,}990{,}329 = 503{,}070$$

where

$r(t) = h_t/t$ = the estimated rate of hardcore drug use (heroin, in this example) in population Nt. For heroin, it is the estimated number (from the NHSDA sample) in treatment and using heroin in the past year divided by the estimated number (from the sample) in treatment centers for drug use in the past year.[2]

$r(t)^{*}$ = the estimated rate of hardcore drug use in population $N-Nt$. For heroin, it is the estimated number (from the sample) not in treatment but using heroin in the past year divided by the estimated number (from the sample) not in treatment centers for drug use during the past year.

$r(a) = h_a/a$ = the estimated rate of hardcore drug use in population Na. For this example, it is the estimated number (from the sample) arrested and booked and using heroin in the

[2] The heroin variable used from the 1992 NHSDA file to code if the respondent used heroin in the past year was HERREC; the treatment variable was the union of three variables: TRMTTRMT (received treatment for drug use in a treatment or rehab facility in the past year), TRMTHOSP (received treatment for drug use in a hospital in the past year), and TRMTMHC (received treatment for drug use in a mental health center in the past year). The treatment variable was defined in this way in order to be as consistent as possible with the data collected in NDATUS.

past year divided by the estimated number (from the sample) arrested and booked in the past year.[3]

$r(a)^*$ = the estimated rate of hardcore drug use in population $N-Na$. It is the estimated number (from the sample) not arrested and not booked but using heroin during the past year divided by the estimated number (from the sample) neither arrested nor booked in the past year.

The above estimates can be compared to the published "simple expansion" estimate for the number of past year users of heroin in 1992 of 323,000. (Actually, the published estimate is really more than a "simple expansion estimator" since NHSDA makes adjustments for nonresponse and benchmarks to known population Census totals.)

3.2. *Using two known population counts*

Having the two separate estimates based on treatment and arrest counts raises the question, "Is there an alternate method that would make simultaneous use of both the treatment and arrest counts?"

The ideal situation when one has two variables, such as the number receiving treatment in the past year and the number arrested and booked in the past year, is to use known counts for the interior cells. In other words, we can make consistent estimates that make use of ratio estimation for each of the cells if we have the following matrix:

	Treatment	No treatment
Arrested	$N(11)$	$N(12)$
Not arrested	$N(21)$	$N(22)$

where $N(11)$ is the known count of the number in treatment and arrested and booked in the past year, etc.

In the earlier description of the calculation for heroin, all that was known at the national level were the marginals Nt, the number in treatment in the past year, and Na, the number arrested and booked in the past year. In the absence of known population counts, we proceed by developing *independent sample estimates* of one cell to estimate the remaining cells. To estimate one of the interior cells, we will use sample data from the 1990 Drug Services Research Survey (DSRS) in conjunction with data from the NHSDA and the estimated marginal count from NDATUS.

DSRS is a national sample survey of treatment centers and records of discharged clients. In 1990 a stratified random sample of 1,183 drug treatment facilities was selected from respondents to the 1990 NDATUS file. In the summer of 1990 a stratified subsample of 118 combined drug and alcohol treatment facilities was selected. Within each facility a sample of client discharge records (discharged during the twelve months between September 1, 1989, and August 31, 1990) was selected. The total sample included over 2,000 records. The Phase I facility response rate was 81 per cent; the Phase II facility

[3] The estimated rate of hardcore drug use in population *Na*. For this example, it is the estimated number (from the sample) arrested and booked and using heroin in the past year divided by the estimated number (from the sample) arrested and booked in the past year (the union of the variable NOBOOKYR with BKLARCNY, BKBURGL, BKAGASLT, BKSMASLT, BKMVTHFT, BKROB, BKRAPE, BKMURDER, BKARSON BKDRVINF, BKDRUNK, BKDRUG, BKPROS, BKVANDAL, and BKOTHOF).

response rate was 82 per cent; and the Phase II discharge completion rate of eligible records was 98 per cent. Both the Phase I and the Phase II facility samples were adjusted for nonresponse, and the discharge sample was adjusted for noncompletion (SAMHSA 1992, 1993c).

We wish to estimate the number of persons in treatment during 1992 who also were arrested and booked – $N(11)$. Since we have the number of persons in treatment Nt, all we have to do is estimate the per cent of those in treatment who also were arrested and booked in 1992.

From DSRS we estimate that the per cent of those in treatment who were ever arrested was 77 per cent. (This number is based on those records having information on arrests and on the number of episodes of treatment in the past year. Approximately 30 per cent of the records were missing one or more of these variables. The estimate from the 1990 DSRS survey has been assumed to hold true for our target year, 1992.) From NHSDA we estimate that the per cent of those in treatment and ever arrested who were also arrested and booked in the past year was 52 per cent. Therefore, we multiply .77 by .52 times the number in treatment 1,789,000 to obtain 716,315, the number in treatment and arrested and booked in the past year.

We can now obtain counts for the interior cells $N(11)$, $N(12)$, $N(21)$, $N(22)$ that are consistent with the marginal counts used earlier.

	Treatment	No treatment	Total
Arrested	716,315	9,006,356	9,722,671
Not arrested	1,072,685	194,917,644	195,990,329
Total	1,789,000	203,924,000	205,713,000

The estimate based on all four cells rather than the marginals is 587,966 calculated as follows:

$$H(t,a) = (88{,}495/287{,}200) * 716{,}315 + (50{,}507/547{,}502) * 1{,}072{,}685 + (95{,}781/4{,}456{,}505) * 9{,}006{,}356 + (75{,}354/196{,}199{,}804) * 194{,}917{,}644 = 587{,}966.$$

The estimated rates for each cell are calculated from the NHSDA sample.[4] The above estimate is larger than either of the estimates based on marginals. By analogy with stratification and poststratification, the interior-cells estimate is generally to be preferred to marginal estimates, especially if there are large differences in the usage ratios among the cells.

Some comments about interpretation of the above numbers are in order. The estimated number of persons arrested and/or treated in the past year from the NHSDA have each been underestimated by a similar amount, a factor of approximately 2 to 2.5. The estimated number of those arrested and in treatment from the sample is only 287,200, while

[4] To calculate the ratios $r(t)$, $r(a)$, etc., only those cases that had no missing data for any of the variables mentioned above were used. In other words, each variable was coded 1, 0, or missing. Over three variables there are 27 combinations; but only the eight combinations of 1's and 0's were used. The implicit assumption is that the item nonresponse is at random. About 1,000 cases had one or more variables missing, but out of a sample of 28,000, this only represents about three per cent.

the estimated number in the universe in 716,315, a factor of 2.5. The estimated number of those arrested and booked in the past year but not in treatment is 4,456,505, but the independent count of 9,006,356 is larger by a factor of 2.0. Finally, for those in treatment but not arrested, the underestimate is by a factor slightly less than 2 (547,502 versus 1,072,685). The similarity of underreporting among the three cells is reassuring in that it implies a stability in the underlying process and in the magnitude of underreporting (or undercoverage). As would be expected, the underreporting is greatest in the arrested and treated cell where the stigma would be greatest. The estimated counts of heroin users in each of these cells from the NHSDA will be inflated by the respective factors, implying that persons in those cells underreported their usage by a factor of 2.0–2.5 times.

The estimated count of those not in treatment and not arrested in the past year is very close to the independent count (the factor is .99), so that the effect of the ratio adjustment on the number of heroin users in this cell is minimal. That is not to say that this cell does not contribute very much to the overall estimate of drug usage in general. That depends on the prevalence rate in that cell of the drug being measured. Since 96 per cent of all persons fall into this cell, even a small prevalence can result in a large contribution.

In order to obtain the most accurate estimate, one should utilize independent population counts that are believed to have the best coverage and are most closely related to the variable of interest. Both of these are somewhat subjective. One can also calculate the estimated variance of alternative estimates, preferring the one with the smallest variance, other things being equal.

The mean squared error of estimates based on the four cells (given that the population counts are known) will typically be smaller than using just two cells – as long as the cell sample sizes are sufficiently large and certain relationships between the numerator and denominator of cell prevalence rates hold.

4. Assumptions

Given that we expect underestimation of these hard drug populations from the household sample, a basic assumption being made in the estimated prevalence rates in these cells is that both numerator and denominator are being similarly underestimated. This would be the case, for example, if drug users underreport their drug use (or it is undercovered) at the same rate as the treatment population underreports its treatment and arrestees underreport being arrested. Taking the estimate of $r(t)$ for the earlier two-cell case, the estimate of h_t is assumed to equal $c_{ht} * H_t$, where c_{ht} is some constant and H_t is the true value. Similarly, the estimated value of t is assumed to equal $c_t * Nt$, so that the expected value of $r(t)$ equals approximately H_t/Nt when $c_{ht} = c_t$. That is, the estimate is unbiased when the proportion of underreporting is the same for both variables. (Another possible assumption is that $c_{ht} < c_t$ because some will assert that the NHSDA coverage of the heaviest users is poorest.)

For the complementary cell in the two-cell estimate, the numerator of the ratio $r(t)^*$, the number of heroin users that are not in treatment, is probably underestimated. But the sample estimate of the denominator will generally not be an underestimate because it will include those persons who say they have not been treated, but really have been treated. We conclude that the proportion of underreporting in the numerator is not

equal to that in the denominator, so that the estimate of total users is not unbiased, but is an underestimate.

Similarly, in the four-cell estimate, the estimate of those who have not been in treatment or arrested and booked in the past year is very similar to the independent count for that cell. Here also, we expect that the assumption of equal underestimating of the numerator and denominator would not hold. Therefore, the ratio estimate still would underestimate this cell for any measure of hard drug use. The overall estimate, $H(t, a)$, is thus expected to be an underestimate.

The effect, generally, of underreporting both being arrested and treated is that these persons are misclassified into the no arrest and no treatment cell. This gives rise to a slightly inflated sample estimate of the true number of persons in this cell. In other words, the constant multiplier in the denominator is greater than 1.

A minor issue is the extent to which our sample ratios cover the entire population (noninstitutionalized individuals age twelve and over). By design, the NHSDA only covers those who spend some portion of the year in a household, dormitory, or shelter. It does not cover the population that is comprised of persons who are in a residential treatment facility for a full year. These occur relatively infrequently since most treatment is outpatient. For those arrested and booked, the group that we miss in our household survey is the group of prisoners that are in prison for the full twelve months because they *cannot* be in our household sample. We are also missing the "permanently" homeless (those who do not appear in a household or a shelter for an entire year).

5. Estimates of Other Measures of Hardcore Drug Use (Table 1)

The following measures of hardcore drug use are derived from the NHSDA using the methodology that was used for heroin. The individual cell rates come from the NHSDA.

5.1. Weekly cocaine use[5]

To extend this discussion to the number of users of cocaine who use it on a weekly basis, we change the variable of interest from heroin to cocaine.

The corresponding estimates of the ratios are then

	Treatment	No treatment	Marginal avg.
Arrested	.259	.016	.031
Not arrested	.108	.002	.0022
Marginal avg.	.160	.0023	.0031

Therefore, the estimates are

using the treatment marginals: $C(t) = 750{,}504$.

using the arrested and booked marginals: $C(a) = 742{,}202$.

using the four interior cells: $C(t, a) = 829{,}017$.

The simple weighted estimate of the number of persons using cocaine is 642,221.

[5] Variable COCWKF.

Table 1. Comparison of ratio estimate to standard estimate for various drugs

Drug usage	Ratio estimate		Standard estimate	
	Treatment/ arrest ratio	Treatment marginal ratio	Arrest marginal ratio	Simple expansion estimator
Past year heroin	587,966	471,844	503,070	323,000
Weekly use of cocaine	829,017	750,504	742,202	642,221
Past year dependence on any drug	2,869,242	2,467,074	2,635,084	2,104,508
Past year needle use	1,019,165	755,977	960,773	659,292
Past year use of marijuana	19,461,280	17,734,861	19,197,855	17,400,273

5.2. *Dependence on any illicit drug*

To estimate the number of persons dependent on any illicit drug in the past year, we have developed an algorithm to approximate the DSM3R criteria (Epstein et al. 1995).

The corresponding estimates of the ratios are then

	Treatment	No treatment	Marginal avg.
Arrested	.519	.093	.118
Not arrested	.348	.0066	.0076
Marginal avg.	.406	.0085	.0102

Therefore, the estimates are

using the treatment marginals: $D(t) = 2{,}467{,}074$.

using the arrested and booked marginals: $D(a) = 2{,}635{,}084$.

using the four interior cells: $D(t, a) = 2{,}869{,}242$.

The simple expansion estimator for number of persons dependent on any illicit drug in the past year is 2,104,508.

5.3. *Past year use of needles*[6]

The corresponding estimates of the ratios are

	Treatment	No treatment	Marginal avg.
Arrested	.238	.056	.067
Not arrested	.068	.0014	.0016
Marginal avg.	.127	.0026	.0032

Therefore, the estimates are

using the treatment marginals: $I(t) = 755{,}977$.

using the arrested and booked marginals: $I(a) = 960{,}773$.

using the four interior cells: $I(ta) = 1{,}019{,}165$.

The simple expansion estimate is 659,292.

[6] Variable NEDYR3.

6. Implications of Components of the Estimator

It is instructive to look at the relative size of certain components to gain insight into how the estimate might be improved and where its weaknesses are. Note that in the above assumptions, the magnitude of the numerators and denominators that make up the ratios $r(1)$ and $r(2)$ are not really important to the size of the estimates H and C, respectively, if one believes that the underreporting of the numerator and denominator are similar. On the other hand, the estimate for H would be larger if one believes that there has been more underreporting of heroin users in treatment, than there has been of the total persons in treatment.

The next components to look at are the denominators of the ratios as compared to the independent counts (actually, partly estimated). The sample-based estimate from the 1992 NHSDA for the number of persons in treatment during the past year is 834,702. The universe estimate from NDATUS (adjusted for certain double counting and for non-response, but not adjusted for any undercoverage from having an incomplete frame) is 1,789,000. This would imply that the NHSDA survey undercounts the NDATUS number by 53 per cent (or more). Similarly, the sample-based estimate for the number arrested and booked in the past year is 4,743,706, while the estimate of the number of persons arrested in the past year derived from the UCR is 9,722,671. The sample-based estimate "underestimates" the universe estimate of 9,722,671 by 51 per cent. (Here, we put "underestimates" in quotes because the two numbers may not be completely analogous. The sample estimate is of persons arrested and booked, while the universe count is for those arrested. In most arrests, however, if one is arrested, one is usually "processed" or booked.)

Both of these estimates confirm the amount of underestimation discussed in the methodology section for the four-cell estimate. The inverses of these rates of underestimation are approximately 2.0 – very similar to the range of factors of underestimation of from 2.0 to 2.5 in the four-cell model. These factors adjust upward the estimated number of drug users. Again, only in the not arrested/not treated cell do we expect the assumption of similar rates of underreporting not to hold. So, the overall estimate of drug prevalence using this methodology is still expected to be a lower bound of the true value.

What is the effect of this kind of estimation on other non-hardcore drugs? Generally speaking, it is not as dramatic as with the above hardcore drugs. We calculated the effect on use of marijuana in the past year.[7] The simple expansion estimator was 17,400,273, while the ratio estimate based on four cells was 19,461,280. The latter estimate is only twelve per cent larger.

The reason for this is that marijuana is used more widely in the population, and most of the users fall in the not arrested/not treated cell. Therefore, this cell (13,644,235 users) dominates the estimate. The relative differences in prevalence rates among cells is not as dramatic as with the hardcore drugs. The cell rates are as follows

	Treatment	No treatment
Arrested	.91	.46
Not arrested	.45	.07

[7] Variable MRJYR.

7. Precision of Estimates

Estimates of variance were calculated for each of the above estimates using SUDAAN – a package that calculates the variance of complex sample surveys using Taylor series (Research Triangle Institute 1992). We used the ratio estimation procedure with post-stratification weights. It was assumed that the independent counts were estimated without error.

The estimated standard error for the ratio estimate of heroin was very similar to the estimated standard error for the simple expansion estimator, approximately 106,000. Even though the cell count of those who have neither been in treatment nor arrested and booked in the past year is the largest of the four cells, its contribution both to the estimate and to the estimate of variance was relatively small because of the estimated low incidence in that cell. The 95 per cent (2σ) confidence interval for the estimate of past year users of heroin was 587,966 plus or minus 212,000.

For the other variables, the standard errors for the simple expansion estimates were similar to those for the ratio estimates. Since the standard errors of the estimates have remained similar while the estimates themselves have increased, the coefficients of variation (the relative precision of the estimates) have been somewhat improved.

8. Comparison with Estimates from Other Methods

Previous national estimates of hardcore drug use have used widely varying methods. Estimates of heroin prevalence published by the National Institute on Drug Abuse in the 1970s relied on a small number of locally derived prevalence estimates that were projected to the entire nation using available heroin problem indicators available in other locations (Person et al. 1977). These estimates of the number of heroin addicts ranged from 584,000 in 1974 to 420,000 in 1979. However, these are not comparable to estimates of any past year heroin use, because they do not include casual heroin users.

A nominative method of estimating heroin prevalence from the NHSDA produced an estimate of 1.9 million past year heroin users in 1982 (Miller 1985). A recent estimate of 658,000 weekly heroin users in 1990 was derived from a "synthetic estimation" procedure that involved combining multiple data sources under various assumptions (Rhodes 1993). This same methodology was used to derive an estimate of 2.1 million weekly cocaine users in 1991. These recent synthetic estimates represent the most rigorous attempts to utilize multiple sources of data in estimating hardcore drug use prevalence.

While there are many differences between the synthetic estimation model and the NHSDA ratio estimation, the large discrepancies in estimates from the two methods are largely explained by the assumptions made regarding the arrestee population. The synthetic model relied heavily on drug prevalence data from the Drug Use Forecasting System (not a representative sample of arrestees), resulting in an estimated 1.8 million weekly cocaine users (more than 80 per cent of the total estimate) and 500,000 weekly heroin users (more than 80 per cent of the total estimate) among arrestees. By contrast, the ratio estimation method relies more heavily on NHSDA drug prevalence data for arrestees, and resulted in an estimated 329,626 weekly cocaine users and 414,265 past year heroin users among arrestees.

A complete evaluation and comparison of the ratio estimation procedure with other

methods of estimating hardcore drug use is beyond the scope of this article. However, we can make some overall statements about ratio estimation.

- Ratio estimation does not fully account for underreporting and undercoverage in the NHSDA. In particular, for the population not arrested and not in treatment, the method does not adjust for underreporting at all. Thus, we consider these estimates of hardcore drug use to be improvements on the generally published NHSDA estimates (using the simple expansion estimator) but still conservative estimates.
- Because ratio estimation can be looked at as an adjustment to the NHSDA analytic weights (which are based on a probability-based sample design), it provides analytic capabilities that are not possible in any of the previously used methods. While other methods essentially focus obtaining the "bottom line" estimate of the number of hard core drug users, by constructing estimates within the framework of the NHSDA data set we can extend the estimation to population subgroups such as by region, gender, race/ethnicity, income, etc., taking advantage of the multitude of data collected in the NHSDA.

  We must acknowledge, however, that there are limitations to these secondary applications of the ratio estimation procedure which have yet to be determined. Because the procedure is designed to improve national hardcore drug use estimates, it may not be appropriate (without modification) for certain other estimation, such as for some subgroups and for other drug use measures (e.g., casual use).
- The ratio estimation model, as applied in this case, relies primarily on regularly updated and consistently collected data from the NHSDA, NDATUS, and UCR, and a relatively small number of easily understood assumptions. Thus, it is likely to be able to provide more reliable trend information (given constant levels of underreporting) than the previously used methods which rely more heavily on assumptions that could change over time.
- Because it relies primarily on the NHSDA sample design and weighting, it is possible to develop estimates of the variances of ratio-adjusted estimates. This is generally not possible in the methods previously used.

9. Possible Future Research/Applications

There are three primary areas for further investigation. One is in the population counts. Another is the assumptions made about the ratios used. The third involves a search for "unbiased" methods to estimate the ratio.

1. It would be useful to explore the development of more accurate estimates for the four cell counts or of alternative counts based on different variables. Estimating the counts used in this article necessitated using multiple sources to make the counts comparable to what is collected by NHSDA. Generally, this is best accomplished by coordinating the questions on the NHSDA and other surveys with those systems used to develop administrative counts so that the definitions are as consistent as possible. Coordination of item wording among surveys will at a minimum make it possible to compare estimates across surveys. For 1994, the NHSDA question on

being in treatment has been changed to agree exactly with the definition used in NDATUS.

Other cells could be developed based on alternative variables. For example, the UCR collects information on the number of persons on parole in the past year as does the NHSDA.

Since it is known that age and race are major correlates of the rate of drug use, yet another improvement would be to seek to find a source or a method of estimation that could provide further age/race breakouts to the treatment/arrest cell counts.

2. Another way to address the problem of underreporting, particularly in the no treatment/not arrested cell where we know there is no underreporting, is to apply repeated measures methodology to this cell. For example, one could use the Hui-Walter method to estimate false positive and false negative probabilities based on a number of questions spread throughout the NHSDA which seek to ascertain drug use.
3. In the area of assumptions, where possible one can compare the distributions of persons for a variable used in the cross-classification based on the NHSDA to those of the population frames to see if they are similar. For example, we can compare the distribution of those in treatment from the NHSDA to the distribution of the population values from NDATUS that are available by age, race, and sex.

   Another possibility is the introduction of additional weights reflecting the proportion of the year that a person is in treatment or living in a household. This would serve to increase the size of the populations that are not year-round household residents.
4. With respect to the instrument, one could perhaps try to introduce methodology that would result in less undercounting of the variables that form the ratios: heroin, treatment, arrested and booked, etc., possibly using multiplicity methods or nominative techniques or using some new method, like hair tests if the methodology proves to be feasible, to confirm drug use or nonuse.
5. Other applications: While population counts were not directly available at the national level for each of the interior cells, similar methodology may be useful for small geographic entities such as states, where the interior cell counts may be known. In order to use this methodology, one would have to conduct a prevalence survey including questions as similar as possible to what is collected in existing population counts of related social indicators. Again, more coordination of item wording between sample surveys and administrative records systems would enhance this kind of estimation.

10. References

Biemer, P.P., Groves, R.M., Lyberg, L.E., Mathiowetz, N.A., and Sudman, S. (eds.). (1991). Measurement Errors in Surveys. New York: John Wiley and Sons, Inc.

Brodsky, M.D. (1985). History of Heroin Prevalence Estimation Techniques. In B. Rouse, N. Kozel, and L. Richards (eds.), Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity, 94–103. NIDA Research Monograph 57, Department of Health and Human Services Publication No. (ADM) 85-1402, Washington, DC: U.S. Government Printing Office.

Caspar, R.A., Gfroerer, J.C., Gustin, J., and Lessler, J.T. (1996). Converting an Ongoing National Survey to Computer-Assisted Interviewing: Making the Decision and Implementing the Approach. Paper presented at the International Conference on Computer-Assisted Survey Information Collection, San Antonio, TX, U.S.A.

Duffer, A.P., Lessler, J.T., Weeks, M.F., and Mosher, W.D. (1996). Impact of Incentives and Interviewing Modes: Results From the National Survey of Family Growth Cycle V Pretest. In R. Warnecke (ed.), Conference on Health Survey Research Methods, 147–152. DHHS Publication No. (PHS) 96-1013.

Epstein, J.F. and Gfroerer, J.C. (1995). Estimating Substance Abuse Treatment Need From a National Household Survey. Paper presented at the 37th International Congress on Alcohol and Drug Dependence.

Gfroerer, J.C. (1992). An Overview of the National Household Survey on Drug Abuse and Related Methodological Research. Proceedings of the Survey Research Section of the American Statistical Association, 464–469.

Harrell, A.V., Kapsak, K.A., Cisin, I.H., and Wirtz, P.W. (1986). The Validity of Self-Reported Drug Use Data: The Accuracy of Responses on Confidential Self-Administered Answer Sheets. Paper prepared for the National Institute on Drug Abuse, Contract Number 271-85-8305.

Hser, Y.I., Anglin, M.D., Wickens, T.D., Brecht, M.L., and Homer, J. (1992). Techniques for the Estimation of Illicit Drug-Use Prevalence: An Overview of Relevant Issues. National Institute of Justice Publication No. (NCJ) 133786.

Maguire, K., Pastore, A.L., and Flanagan, T.J. (eds.) (1993). Sourcebook of Criminal Justice Statistics. U.S. Department of Justice, Bureau of Justice Statistics, Washington, DC: USGPO.

Miller, J.D. (1985). The Nominative Technique: A New Method of Estimating Heroin Prevalence. In B. Rouse, N. Kozel, and L. Richards (eds.), Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity, 104–118. NIDA Research Monograph 57, Department of Health and Human Services Publication No. (ADM) 85-1402. Washington, DC: U.S. Government Printing Office.

Person, P.H., Retka, R.L., and Woodward, A.J. (1977). A Method for Estimating Heroin Use Prevalence. DHEW Publication No. (ADM) 77-439. Washington, DC: Superintendent of Documents, U.S. Government Printing Office.

Research Triangle Institute (1992). SUDAAN Users Manual. Research Triangle Park, N.C.: Research Triangle Institute.

Rhodes, W. (1993). Synthetic Estimation Applied to the Prevalence of Drug Use. Journal of Drug Issues, 23, 297–321.

Substance Abuse and Mental Health Services Administration (1992). Drug Services Research Survey (DSRS) Final Report: Phase II. Rockville, MD: U.S. Department of Health and Human Services.

Substance Abuse and Mental Health Services Administration (1993a). National Household Survey on Drug Abuse: Population Estimates, 1992. DHHS Publication No. (SMA) 93-2053. Washington, DC: Superintendent of Documents, U.S. Government Printing Office.

Substance Abuse and Mental Health Services Administration (1993b). National Drug and Alcohol Treatment Unit Survey (NDATUS) 1991 Main Findings Report. DHHS

Publication No. (SMA) 93-2007. Rockville, MD: U.S. Department of Health and Human Services.

Substance Abuse and Mental Health Services Administration (1993c). Drug Services Research Survey (DSRS) Final Report: Phase I Revised. Rockville, MD: U.S. Department of Health and Human Services.

Substance Abuse and Mental Health Services Administration (1994). Preliminary Estimates from the 1993 National Household Survey on Drug Abuse (Advance Report Number 7). Rockville, MD: U.S. Department of Health and Human Services.

Turner, C.F., Lessler, J.T., and Gfroerer, J.C. (1992). Survey Measurement of Drug Use: Methodological Studies. National Institute on Drug Abuse. DHHS Publication No. (ADM) 92-1929.

Turner, C.F., Ku, L., Sonenstein, F.L., and Pleck, J.H. (1996). Impact of ACASI on Reporting of Male–Male Sexual Contacts: Preliminary Results From the 1995 National Survey of Adolescent Males. In R. Warnecke (ed.), Conference on Health Survey Research Methods, 171–176. DHHS Publication No. (PHS) 96-1013.

Warner, S.L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. Journal of the American Statistical Association, 60, 63–69.

Wish, E.D. (1990–1991). Drug Policy in the 1990s: Insights from New Data on Arrestees. International Journal of Addiction, 25, 377–409.

Woodward, J.A., Bonett, D.G., and Brecht, M.L. (1985). Estimating the size of a Heroin Abusing Population Using Multiple-Recapture Census. In B. Rouse, N. Kozel, and L. Richards (eds.), Self-Report Methods of Estimating Drug Use: Meeting Current Challenges to Validity, 158–171. NIDA Research Monograph 57, Department of Health and Human Services Publication No. (ADM) 85-1402. Washington, DC: U.S. Government Printing Office.

Zdep, S.M., Rhodes, I.N., Schwartz, R.M., and Kilkenny, M.J. (1979). The Validity of the Randomized Response Technique. Public Opinion Quarterly, 43, 544–549.

Received July 1995
Revised June 1997