

**A Model-Based Approach for
Estimating Prevalence of Hard to Reach Populations**

William Rhodes

Ryan Kling

Patrick Johnston

Abt Associates, Inc.



Friday, October 01, 2004

Please send correspondence to William Rhodes, principal scientist and fellow at Abt Associates, Inc., 55 Wheeler Street, Cambridge, MA., USA 02138-1168. Telephone: 618-349-2731; Fax: 617-349-2610; e-mail: bill_rhodes@abtassoc.com. Ryan Kling is a scientist at Abt Associates. Patrick Johnston is a senior scientist at Abt Associates.

Research was done under contract to the National Institute of Justice and under a McGillis grant from Abt Associates, Inc. The authors thank Stephen Kennedy and Dana Hunt, both of Abt Associates, for their helpful comments on earlier drafts.

**A Model-Based Approach for
Estimating Prevalence of Hard to Reach Populations**

Abstract

Public policy is often concerned with the size and characteristics of special populations that are difficult to reach in household surveys. As an illustration, chronic drug users often live outside of traditional households. An alternative to household surveys is to sample and question such special populations where they congregate – jails, treatment programs, and shelters, for example. Using such opportunistic data for prevalence estimation raises difficult problems for statistical modelers. This paper presents a model-based method for estimating the number of chronic drug users in a county and in the nation based on data collected at jails where offenders are booked, sampled and interviewed following arrest. First it develops a modified Poisson mixture model used to estimate the stochastic process presumed to account for how drug users get arrested. Second it uses that model and statistical reasoning to derive a confidence interval for the number of chronic drug users in a county. Third, it illustrates that approach by providing estimates for thirty-nine counties using sixteen-quarters of data. Finally it provides national estimates of chronic drug use. Although chronic drug users are used to illustrate, the methodology is applicable to other problems where special population are difficult to reach using standard survey methods.

Key words:

chronic drug use; hard-to-reach populations; endogenous stratification; Poisson mixture models; ratio estimation; model-based estimation.

1.0 Introduction

Using traditional surveys to estimate the prevalence of hard-to-reach populations – such as chronic drug users, the mentally ill and the homeless – is challenging. Sampling frames can be difficult or impossible to construct; low population prevalence may require large samples; and once sampled, respondents may deny behaviors that are stigmatized or illegal. An alternative approach is to sample members of these hard-to-reach populations where they congregate – in jails, in treatment programs, and so on. This solves the problem of low prevalence. It increases veracity, because people at those collection points have already self-identified (e.g., those who seek treatment) or been identified by others (e.g. those who are arrested); furthermore, objective tests (e.g. urine tests for recent drug use) can be confirmatory. Estimation faces a new impediment, however. Samples drawn at collection points are not random samples from the general population. Estimation consequently requires a model-based approach, where modeling seeks to explain the probabilities that members of the hard-to-reach population in the community arrived at the collection points where the samples are drawn. The inverse of that selection probability provides the means to weight the opportunistic sample to reflect the population in the community.

This paper develops and demonstrates a model-based approach to estimating the prevalence of one hard-to-reach population: chronic drug users. Estimates of prevalence and trends in chronic drug use are instrumental for public policy. It might be surprising therefore to learn that probability-based prevalence estimates do not exist (Manski, Pepper and Petrie, 2001). Some might consider this an overstatement because the Nation has a premier survey that asks respondents about their drug use – the National Survey on Drug Use and Health (NSDUH). However, many chronic drug users either live outside traditional households, or they are seldom at home when interviewers call, or they refuse

to answer truthfully (Manski, Pepper and Petrie, 2001; Fendrich et. al., 1999; Harrell et. al., 1986). Some researchers have used ratio methods to adjust for undercounting (Epstein and Gfoerer, 1996), but others remain skeptical, because ratio-based estimates are inconsistent with other indicators, including counts of the number of chronic users entering substance abuse treatment and estimates of the amount of cocaine and heroin entering the country (Rhodes et. al., 2002). Empirically grounded public policy therefore requires alternative ways to estimate the prevalence of chronic drug use.

The approach discussed in this paper uses data collected where drug users congregate – at booking facilities across thirty-nine urban counties. Following this introduction, the second section provides an informal explanation of the approach. The third section explains the estimation methodology formally, the fourth section discusses diagnostic tests from statistical modeling, and the fifth section presents estimates. The last section offers comments and suggests extensions.

2.0 An Informal Explanation

Data come from the Arrestee Drug Abuse Monitoring (ADAM) program, a probability sample of arrests and bookings in thirty-nine counties. Within each site, interviewers question a probability sample of adult male arrestees about their drug use and previous arrests; they also request urine specimens. Analysis of a urine specimen indicates whether or not an arrestee has recently used one of ten different illicit drugs, and thus provides confirmation of truthful reporting.

This study seeks to estimate the number of adult male chronic drug users in the general population for counties that house ADAM programs and then to extend county-specific estimates to the nation by using a ratio estimator. A chronic user might be one who uses illicit drugs at greater than a threshold level; or, he might be one deemed to need

treatment. Within limits identified later, the estimation methodology does not depend on the definition of chronic drug use.

The principal analytic problem is to make inferences about chronic drug users in the community based on data about chronic drug users who appear in a sample of chronic users who are arrested and booked into jail pending a judicial hearing. The arrestees are not a random sample of chronic drug users. Therefore we must weight the sample so that arrestees with high predicted arrest rates receive a smaller weight than arrestees with low predicted arrest rates. The key step is to develop a valid prediction for arrest rates and then apply those predictions to weight the sample of arrestees.

To illustrate the basic approach, suppose there were H adult male chronic drug users in the community and that during a given year they experience A arrests. A is observable in theory, but estimated in practice, for two reasons. The first reason is that the data comprise a probability sample of arrests; so estimating A requires the application of sampling weights. The second reason is that some chronic drug users deny their status as chronic drug users when questioned in a booking facility. Dealing with denial and underreporting requires adjustment for the rate of truthful reporting, which itself must be estimated. Application of sampling weights and adjustment for truthful reporting yield \hat{A} , an estimate of A .

The current arrest is the one that caused a respondent to get into the ADAM sample. Respondents are questioned about arrests during the year before the current arrest using the “calendar” portion of the ADAM questionnaire (Hunt and Rhodes, 2001). Analysis of responses to the questions about arrests during the previous year yields an estimate of R , the average rate at which chronic users get arrested and booked into jails and lockups, where the arrestees are sampled and interviewed. By definition:

$$A = H \cdot R$$

so an estimate for H is:

$$\hat{H} = \frac{\hat{A}}{\hat{R}}$$

The fundamental problem posed and solved in the next section is deriving point and interval estimates of A , R and H . This requires specifying a stochastic model of the arrest process, estimating the parameter of that model conditional on the fact that data come from a sample of arrests (not a sample of chronic drug users in the community), and using results from that model to estimate confidence intervals for \hat{H} .

The above procedure gives estimates for each county that houses an ADAM program. To extend those county-specific estimates to the Nation, we assume that chronic drug use is proportional to the number of substance abuse treatment admissions in counties across the United States. Then the ratio of treatment admissions across the United States to treatment admissions within the counties that house ADAM programs provides the means to inflate county-specific estimates for the ADAM sites to a national estimate of chronic drug use.

3.0 The Estimator

This section, which explains the estimator used in this study, has five subsections:

- The first subsection (3.1) proposes a stochastic model of the arrest process and shows how the parameters of that model can be estimated conditional on the fact that the sample was drawn when an arrest occurred. That is, this subsection discusses how we estimate the parameters of the arrest process in the face of endogenous stratification of the sample.
- Given the parameter estimates from the first subsection, the second subsection (3.2) uses those parameter estimates to develop an estimator for the number of

chronic drug users in the community. That estimator assumes that the ADAM sample comprises the *population* of chronic users arrested and booked during a specified time.

- Subsection 3.3 then shows how estimates for chronic drug users are adjusted to account for ADAM’s sampling design.
- Subsection 3.4 explains adjustments for underreporting and for the fact that ADAM is limited to adult males while we seek estimates for all adults.
- The last subsection explains the ratio estimation technique used to extend chronic drug user estimates from the ADAM counties to the Nation.

3.1 Modeling the Arrest Process in a County

As noted, the Arrestee Drug Abuse Monitoring (ADAM) system provides a probability sample of adult male arrestees booked in thirty-nine urban counties.¹ Within each jail, interviewers ask a random sample of arrestees to complete a twenty-minute interview and to provide a urine sample, used to test for recent drug use. The interview includes a retrospective “calendar” instructing respondents to recall arrests and other events on a monthly basis during the year before their instant arrest (the window period). We use a Poisson mixture model to account for the rate at which chronic drug users are arrested and booked during that window. Estimation of model parameters is complicated by endogenous stratification – the fact that the sample used to estimate the model parameters is selected conditional on the occurrence of an arrest.² This section first specifies the Poisson mixture model and then derives the likelihood function used to estimate the model’s parameters given endogenous stratification.

We begin by specifying the Poisson mixture model that generates arrests. Components of the model are defined below. Greek letters represent population parameters (e.g. β) and

estimates (e.g. $\hat{\beta}$) and Roman letters represent variables (e.g. X). The exceptions is γ , which represents the realization of a random process. Let:

N_i This represents the number of arrests during the one-year window period for the i^{th} drug user in the general population. The term *general population* is meant to distinguish between the population of drug users and the subset of them who get arrested. In this study, the general population refers to those drug users who reside in the county where arrests can occur. Only some of them get arrested.

X_i, β X_i is a row vector of K *measured* exogenous factors (including a constant) that affect N_i , and β is a conformable parameter vector. Exogenous means that X affects N , but N does not affect X .

γ_i This is an unobserved random variable representing *unmeasured* exogenous factors that affect N . Based on testing reported later, we assumed that γ is distributed as lognormal $f(\gamma)$ with mean $e^{(1/2)\sigma^2}$ and variance $e^{\sigma^2}(e^{\sigma^2} - 1)$.

Assume that in the general population of drug users, arrests occur at a constant rate λ (the hazard) during the window period, where λ is a function of X and γ . Then for the i^{th} drug user, N is generated by a Poisson process conditional on X and γ , such that the probability of N events occurring during a one-year window period is:

$$P(N_i | X_i; \gamma_i) = \frac{e^{-\lambda(X_i; \gamma_i)} \lambda(X_i; \gamma_i)^{N_i}}{N_i!} \quad [1]$$

$$\lambda(X_i; \gamma_i) = \gamma_i e^{X_i \beta} \quad [2]$$

Assuming that λ is log-linear assures that $\hat{\lambda}$ will be positive, a necessary condition given that λ is a rate and is supported by the data. As noted γ is identically and independently distributed as lognormal across individuals.

The intuition behind [2] is that the arrest rate depends both on factors that we can potentially observe and other factors that we cannot observe. We will ultimately remove the conditioning on the unobservable factors through integration and write the arrest rate as a function of X alone:

$$\lambda(X) = \int_{\gamma=0}^{\infty} \gamma e^{X\beta} f(\gamma) d\gamma = e^{X\beta + (1/2)\sigma^2} \quad [3]$$

Equation [3] is the expected number of arrests during the window period for drug users in the general population who have characteristics X . We seek to estimate this rate, because it will enter into our chronic user estimates. However, estimating β and σ^2 is complicated by endogenous stratification (Cosslett, 1993), which occurs because drug users must have been arrested and booked to get into the sample frame.

To estimate the parameters for [3] given endogenous stratification, the likelihood function must recognize that the distribution of γ is different for *arrested* drug users than for the general population of drug users. We need an expression for $g(\gamma | \textit{arrest})$, the density for γ for drug users who get arrested. Using Bayes theorem, we write the density of γ conditional on being arrested during a short period of time as:

$$g(\gamma | \textit{arrest}) = \frac{\textit{prob}(\textit{arrest} | \gamma) f(\gamma)}{\textit{prob}(\textit{arrest})} \quad [4]$$

Defining Δt to be a period of sufficiently short duration that the probability of more than one arrest is negligible, the probability of being arrested during Δt , conditional on γ , is the arrest hazard rate times that short time interval:

$$prob(arrest | \gamma) = \gamma e^{x\beta} \Delta t \quad [5]$$

and the marginal probability of an arrest comes from integrating over γ :

$$prob(arrest) = \int_{\gamma=0}^{\infty} prob(arrest | \gamma) f(\gamma) d\gamma \quad [6]$$

Substituting [5] and [6] into [4], recalling that $f(\gamma)$ is lognormal with mean $e^{(1/2)\sigma^2}$, and canceling terms in the numerator and denominator, $g(\gamma | arrest)$ can be rewritten as:

$$g(\gamma | arrest) = \frac{(\gamma e^{x\beta} \Delta t) f(\gamma)}{\int_{\gamma=0}^{\infty} (\gamma e^{x\beta} \Delta t) f(\gamma) d\gamma} = \frac{\mathcal{F}(\gamma)}{\int_{\gamma=0}^{\infty} \mathcal{F}(\gamma) d\gamma} = \frac{\mathcal{F}(\gamma)}{E(\gamma)} = \frac{\mathcal{F}(\gamma)}{e^{(1/2)\sigma^2}} \quad [7]$$

Equation [7] gives us the distribution of γ for *arrested* drug users. The mean arrest rate for arrested drug user will be larger than the mean arrest rate for drug users in the community. To see this, follow [7A] through [7D] to compute the mean arrest rate and its variance for arrested drug users as:

$$E(\gamma | arrest) = \int_{\gamma=0}^{\infty} \gamma g(\gamma | arrested) d\gamma \quad [7A]$$

Substitute [7] into [7A] gives:

$$E(\gamma | arrest) = \frac{\int_{\gamma=0}^{\infty} \gamma^2 f(\gamma) d\gamma}{e^{(1/2)\sigma^2}} = \frac{\int_{\gamma=0}^{\infty} \gamma^2 f(\gamma) d\gamma}{E(\gamma)} \quad [7B]$$

Because γ is distributed as lognormal in the general population, its variance can be written:

$$VAR(\gamma) = e^{\sigma^2} (e^{\sigma^2} - 1) = \int_{\gamma=0}^{\infty} \gamma^2 f(\gamma) d\gamma - E(\gamma)^2 \quad [7C]$$

Solving [7C] for $\int \gamma^2 f(\gamma) d\gamma$, substituting into [7B], and performing some algebraic manipulation yields:

$$E(\gamma | arrest) = \frac{e^{\sigma^2} (e^{\sigma^2} - 1) + E(\gamma)^2}{E(\gamma)} = \frac{e^{\sigma^2} (e^{\sigma^2} - 1) + e^{\sigma^2}}{e^{(1/2)\sigma^2}} = e^{(3/2)\sigma^2} \quad [7D]$$

$$VAR(\gamma | arrest) = e^{3\sigma^2} (e^{\sigma^2} - 1)$$

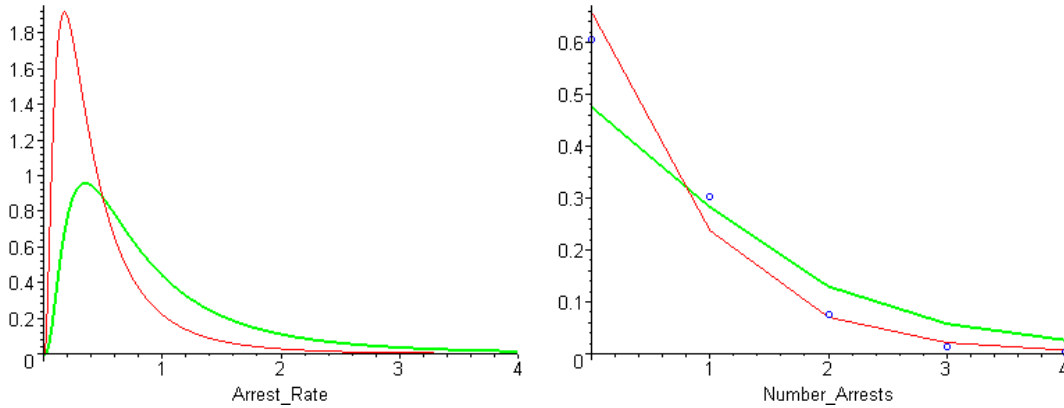
In other words, conditional on X, the expected value of the number of arrests during the window period is $e^{X\beta} e^{(1/2)\sigma^2}$ for drug users *in the community* and $e^{X\beta} e^{(3/2)\sigma^2}$ for drug users *in the sample*, a ratio of 1 to 2.72. Thus, given unmeasured heterogeneity, even after controlling for X the average arrest rate for arrested drug users has upward bias as an estimate for the arrest rate of drug users in the community, an observation that emphasizes the need to take endogenous stratification into account.

We use [7] in conjunction with equation [1] to specify the distribution for N conditional on being arrested at the time that the sample was drawn:

$$P(N_i | X_i; arrest) = \int_0^{\infty} P(N_i | X_i; \gamma) g(\gamma | arrest) d\gamma \quad [8]$$

Integration removes the conditioning on γ . The arrest count excludes the current arrest.³

Figure 1, which represents a stylized version of drug user arrest rates, helps cement ideas. The picture segment on the left shows, for small γ , $f(\gamma)$ as the higher curve and $g(\gamma | arrest)$ as the lower curve. Inspection of the curves shows that the expected value of γ is higher for a sample of arrested drug users than it is for the population of drug users in the community. The picture segment on the right shows, for small N , $P(N | X)$ as the higher curve (where $N=0$) and $P(N | X; arrest)$ as the lower curve. As would be expected, the probability of zero arrests during the window period is lower for the sample of drug users than it is for the population of drug users. For a random sample drawn from the community, the observable data would correspond to $f(\gamma)$ and $P(N | X)$, but for the sample of arrestees, the observable data conform to $g(\gamma | arrest)$ and $P(N | X; arrest)$. Given use of the arrestee sample to estimate β and σ , estimation requires a likelihood based on these conditional distributions.



The product of [8] over all arrested drug users is the likelihood contribution of the i th arrestee in the sample of arrestees:

$$L = \prod_{i \in \text{arrestee sample}} P(N_i | X_i; \text{arrest}) \quad [9]$$

Estimation of β and σ proceeds by maximizing [9], a procedure that requires numerical integration since $P(N_i | X_i; \text{arrest})$ has no closed form expression. We then use the estimates of β and σ to parameterize $P(N_i | X_i)$, the process that generates arrests for drug users in the community. The next subsection explains how knowledge of $P(N_i | X_i)$ leads to estimates of the number of chronic drug users in the county.

3.2 Estimating the Number of Chronic Drug Users in a County

Assume the availability of estimates of the β and σ parameters (and their covariance matrix V) based on maximizing the likelihood function for the Poisson lognormal mixture (equation [9]). Continuing to ignore ADAM's sampling design, this section discusses how substituting $\hat{\beta}$ and $\hat{\sigma}$ into $P(N_i | X_i)$ leads to point estimates and confidence intervals for the number of chronic drug users in the county.

As before, let:

H This is the total number of chronic drug users in the community. In this application, the community is the county that houses the ADAM program. We seek to estimate H .

It is convenient to put these H chronic drug users into J groups such that within any group, every chronic user has an identical value for X . Then the i^{th} chronic drug user within the j^{th} group generates N_{ij} arrests during a one-year window period, such that:

$$\mu_j = E(N_{ij}) = e^{X_j \beta + (1/2) \sigma^2} \quad [10]$$

where μ_j , the expected value for a lognormal Poisson mixture $P_N(X_j)$, is a succinct representation of the expected value of the arrest rate for a randomly selected member of the j^{th} group. Multiplying the H_j chronic users in group j by μ_j gives the expected value of the number of arrests generated by members of the j^{th} group.

$$E(N_j) = H_j \mu_j \quad \text{where } N_j = \sum_i N_{ij} \quad [11]$$

Hence an estimate for the number of chronic drug users in group j is:

$$\hat{H}_j = \frac{E(N_j)}{\hat{\mu}_j} \quad [12]$$

This is the number of chronic drug users in group j who were arrested (which is observable from booking data) divided by the expected number of arrests by chronic drug users in group j (which is estimated from the ADAM data). An estimate of chronic drug users in the entire population is:

$$\hat{H} = \sum_j \hat{H}_j \quad [13]$$

Equation [13] provides a point estimate for the number of chronic drug users in the community.

To approximate the sampling variance for \hat{H}_j , we use a first-order Taylor expansion about $E[N_j] = H_j \mu_j$ and $1/\mu_j$:

$$\begin{aligned}\hat{H}_j &\approx H_j + \frac{1}{\mu_j} (N_j - H_j \mu_j) + H_j \mu_j \left(\frac{1}{\hat{\mu}_j} - \frac{1}{\mu_j} \right) \\ \text{VAR}(\hat{H}_j) &\approx \sum_j \left[\frac{1}{\mu_j} \right]^2 \text{VAR}(N_j) + (H_j \mu_j)^2 \text{VAR} \left(\frac{1}{\hat{\mu}_j} \right)\end{aligned}\tag{14}$$

where:

$$\text{VAR}(N_j) = \sum_i \text{VAR}(N_{ij}) = H_j \text{VAR}(N_{ij}) = H_j \mu_j [1 + \mu_j (e^{\sigma^2} - 1)]$$

and

$$\text{VAR} \left(\frac{1}{\hat{\mu}_j} \right) = d_j' V d_j$$

To derive $\text{VAR}(1/\hat{\mu}_j)$, we note that this variance term arises from the sampling variance for $\hat{\beta}$ and $\hat{\sigma}$, and the uncertainty about these parameters is captured in the parameter covariance matrix V . The term $d'Vd$ follows from using a Taylor-series approximation where V is the parameter covariance matrix and d_j is a row vector with dimension equal to the K elements in X plus 1, and with typical terms:

$$\frac{\partial}{\partial \beta_k} \frac{1}{\mu_j} = -\frac{x_{jk}}{\mu}$$

$$\frac{\partial}{\partial \sigma} \frac{1}{\mu_j} = -\frac{\sigma}{\mu}$$

To derive $VAR(N_{ij})$, above, we used the decomposition of variance theorem:

$$VAR(N_{ij}) = VAR_{\varepsilon}[E[N_{ij} | \gamma]] + E_{\varepsilon}[VAR[N_{ij} | \gamma]] = \mu_j \left[1 + \mu_j (e^{\sigma^2} - 1) \right]$$

This is the variance for the Poisson lognormal mixture $P(N_j | X_j)$. $VAR(N_j)$ then follows from observing that arrests are generated independently across the drug users in the j th group.

Given [14], the variance for \hat{H} is the sum of the variances for \hat{H}_j :

$$VAR(\hat{H}) \approx \sum_j \frac{H_j}{\mu_j} \left(1 + \mu_j (e^{\sigma^2} - 1) \right) + \sum_j (H_j \mu_j)^2 d_j' V d_j \quad [15]$$

In practice this would be estimated as:

$$VAR(\hat{H}) = \sum_j \frac{N_j}{\hat{\mu}_j^2} \left(1 + \hat{\mu}_j (e^{\sigma^2} - 1) \right) + \sum_j N_j^2 d_j' V d_j \quad [16]$$

Equations [13] and [16] do not take into account the fact that ADAM is a probability sample, but otherwise, they are the estimators used in this study. The next subsection introduces a correction for sampling.

3.3 Dealing with the Sampling Process

ADAM uses post-sampling stratification to assign sampling probabilities (Hunt and Rhodes, 2001). The sampling probability is the ratio of sampled arrests in a stratum to the total number of arrests in that stratum; the sampling weight is the inverse of the

sampling probability. Note that every arrestee within a stratum has the same sampling weight, an observation that simplifies some of the algebra presented here.

Sampling introduces another level of uncertainty for the chronic user estimates. To explain how this additional level of uncertainty gets taken into account, return to equations [13] and [16], but now presume that these estimates are specific to the s^{th} stratum. We rewrite them with one additional variance term, which is defined and derived in this subsection. That additional variance term results from ADAM's sampling design.

$$\hat{H}_s = \sum_j \hat{H}_{sj} = \sum_j \frac{N_{sj}}{\hat{\mu}_j} \quad [17]$$

$$\text{VAR}(\hat{H}_s) \approx \left[\sum_j \left(\frac{1}{\hat{\mu}_j} \right)^2 \text{VAR}(N_{sj}) + (H_{sj} \mu_{sj})^2 \text{VAR} \left(\frac{1}{\hat{\mu}_j} \right) \right] + \sigma_s^2 \quad [18]$$

To account for sampling, we now treat N_{sj} as the number of arrests for the j^{th} group of drug users within the s^{th} stratum. Estimate for the entire population are the sums of the estimates over the strata, so summation of the stratum-specific estimates provided by (17) and (18) are sufficient to provide estimates of chronic drug users in the county.

Let:

M_s the population of arrestees in the s^{th} stratum

m_s the sample size for the s^{th} stratum

$\hat{\mu}_{sm}$ the estimated arrest rate for the m^{th} observation in stratum s , derived using the procedures described in the previous section.

σ_s^2 the part of the sampling variance that is attributable to sampling.

Within stratum s , we have m_s estimates of μ for arrestee in that stratum-- $\hat{\mu}_{s1}, \hat{\mu}_{s2} \dots \hat{\mu}_{sm_s}$. We treat these m_s estimates as a sample of the M_s estimates that we would have gotten if we had data for all M_s arrestees in the stratum. Then σ_s^2 represents the additional uncertainty that arises from this sampling process.

Within the s^{th} stratum, we have m_s estimates of μ , from which we can estimate a stratum mean and the sampling variance for that mean:

$$MEAN\left[\frac{1}{\hat{\mu}_s}\right] = \frac{\sum_{m=1}^{m_s} \frac{1}{\hat{\mu}_{sm}}}{m_s} \text{ is the stratum mean.}$$

$$VAR\left[\frac{1}{\hat{\mu}_s}\right] = \frac{\sum_{m=1}^{m_s} \left(\frac{1}{\hat{\mu}_{sm}} - MEAN\left[\frac{1}{\hat{\mu}_s}\right]\right)^2}{(m_s - 1)m_s} \text{ is the stratum variance for } MEAN\left[\frac{1}{\hat{\mu}_s}\right]$$

Then:

$$\sigma_s^2 = [M_s - m_s]^2 VAR\left[\frac{1}{\hat{\mu}_s}\right] \text{ which is substituted into [18].}$$

Estimators that take ADAM's sampling design into account are given by:

$$\hat{H} = \sum_s \hat{H}_s \tag{19}$$

$$VAR(\hat{H}) = \sum_s VAR(\hat{H}_s) \tag{20}$$

The estimator works when ADAM uses a stratified sample from a single jail or from several jails. The estimate requires some additional manipulation for ADAM sites that use a stratified cluster sample of jails, where an additional variance component arises because jails are sampled. We do not consider that complication here.⁴

Equations [19] and [20] would provide the estimators used in this study, except that they do not yet account for underreporting of chronic drug use. This paper turns to that issue next.

3.4 Adjusting for Underreporting

Not all chronic drug users identify themselves as drug users when interviewed in a jail or lockup, so without an adjustment for underreporting, estimates of chronic drug users would be biased downward. Unfortunately, beyond a urine test for recent drug use (useful for a two to three day period), ADAM lacks any special provisions for estimating the proportion of chronic users who self-identify.

This study developed an adjustment for underreporting that relied on some crucial assumptions, the results from the urine tests, and responses to questions about drug use during the month before being arrested. Specifically, for each ADAM site, we identified arrestees who tested positive for cocaine, heroin or methamphetamine. (We did separate estimates for each of the three drugs.) We determined the proportion of them who admitted to use of at least one of these three drugs during the last month. That proportion was used as an estimate of truthfulness.

Discounting a small rate of false positives, a urine test is definitive proof of recent drug use for cocaine, heroin or methamphetamine. Thus, anyone who tested positive is definitely a user, and we can ask whether or not that users would admit to his or her use.

However, we are not interested in use during the last two or three days, but rather, we are interested in use during the last thirty days.⁵ Therefore we treat a “truth teller” as someone who will admit to use during the last thirty days conditional on being a recent drug users.

Let:

- \hat{R} the estimated rate of truthful reporting.
- $\sigma_{\hat{R}}^2$ the estimated sampling variance for \hat{R} .

Assuming a Bernoulli process, an estimate of the sampling variance is given by $\sigma_{\hat{R}}^2 = \hat{R}(1 - \hat{R}) / m$, where m is the number of drug users who tested positive for cocaine, heroin, or methamphetamine.

Then equation [19] is modified to be:

$$\hat{H} = \frac{1}{\hat{R}} \sum_s \hat{H}_s \quad [21]$$

To derive the sampling variance, we again apply a Taylor approximation to get:

$$VAR(\hat{H}) \approx \left[\frac{1}{\hat{R}} \right]^2 VAR\left(\sum_s \hat{H}_s \right) + \frac{\sigma_{\hat{R}}^2}{\hat{R}^4} \left[\sum_s \hat{H}_s \right]^2 \quad [22]$$

A final problem when developing estimates for a county is that estimates apply to adult males. We seek to adjust the estimates to account for adult females. Let:

- A_1 represents the total number of adult female arrests in the county divided by the total number of adult male arrests in the county. Arrest counts come from the Uniform Crime Reports of the Federal Bureau of Investigation.
- A_2 represents the proportion of adult female arrestees who are chronic users divided by the proportion of adult male arrestees who are chronic users. The proportion of male arrestees who are chronic users comes from the ADAM data. The proportion of female arrestees who are chronic users comes from the female counterpart of ADAM, which unlikely the male ADAM is not a probability sample.
- A_3 represents the adult male arrest rate divided by the female arrest rate. The arrest rates comes from the male and female ADAM calendar.

Then to adjust for females, we multiply [21] by $1+A_1A_2A_3$ and we multiply [22] by $[1+A_1A_2A_3]^2$. We have not attempted to adjust the sampling variance for this additional uncertainty.

3.5 Extending Estimates to the Nation

To this point, we have derived estimates of chronic drug users for each ADAM county. Summing across the ADAM counties provides an estimate of the total number of users in counties that house ADAM programs. We used a ratio estimator to extend the estimates from the 39 counties that housed ADAM to the rest of the United States.

For this purpose, we presume that the number of admissions into substance abuse treatment in a county is proportional to the number of chronic drug users in that county. The proportionality is only approximate because the availability of substance abuse services varies across the United States. Where substance abuse services are readily

available, we would expect a relatively low proportion of chronic users to treatment admissions. Where substance abuse services are comparatively unavailable, we would expect a higher proportion of chronic users to treatment admissions. Furthermore, reporting practices vary from county-to-county.

Treatment admission data are inexact. Our data source for treatment admissions was the N-SSATS, which reports the number of admissions for substance abuse treatment in each treatment program in every county in the United States. The ratio estimator requires *adult* treatment admissions for substance abuse *excluding alcohol*, so we needed to adjust the treatment data. We multiplied total treatment admissions by the percentage of adult clients, and we multiplied the result by the percent of treatment admissions that were for drugs or drugs/alcohol combined. (N-SSATS provided the required proportions by facility.) That is, we excluded treatment admissions that were for alcohol and those that were for juveniles. The N-SSATS data are available for 2000 and 2002. We averaged treatment admissions for the two years.

The ratio estimator also requires treatment admission by drug type. To estimate the number of treatment admissions that were for cocaine, heroin and methamphetamine, we used drug-specific estimates from TEDS. TEDS reports treatment admissions for publicly funded treatment programs in large Metropolitan Statistical Areas. From the TEDS data, we estimated the proportion of treatment admissions that were for cocaine, heroin or methamphetamine by MSA. (That is, one of these three drugs had to be the primary substance of abuse, or it had to be the secondary substance of abuse when alcohol was the primary substance of abuse. In fact, methamphetamine was included as amphetamine in the N-SSATS data.) We assume that the drug-specific proportions for the MSA are the same as the proportions for the county, and applied these proportions to the N-SSATS county-level data. The result was an estimate of the number of adult treatment admissions for cocaine, heroin and methamphetamine in each county within the United States.

Define δ to be expected value of the ratio of chronic drug users to treatment admissions. To estimate δ , we estimated the weighted least squares regression:

$$\frac{\hat{C}_i}{\sqrt{\hat{T}_i}} = \delta \sqrt{\hat{T}_i} + e_i \quad [23]$$

where \hat{C}_i is the chronic users estimate for the i th county and $\sqrt{\hat{T}_i}$ is the square root of the estimated number of treatment admissions in the i th county. Both C_i and T_i are the averages over the years for which we had data. Treatment admissions are the average admissions for 2000 and 2002. For the chronic use estimates, they are the average over the years for which a county reported ADAM data. Not all counties reported for all four years.

This regression specification requires justification, because an ordinary least squares regression $\hat{C}_i = \delta \hat{T}_i + e_i$ would seem more natural absent knowledge that the variance of e is proportional to T . However, an OLS regression will yield an inconsistent estimate of δ when T is measured with error, while [23] will be consistent (Cheng and Van Ness, 1999). The resulting estimate of δ is:

$$\hat{\delta} = \frac{\sum \hat{C}_i}{\sum \hat{T}_i} \quad \text{where the sum is over all counties that have estimates for C.}$$

There is no reason to assume that the distribution of e is homoscedastic, so we used a robust variance (sandwich) estimator to estimate the sampling variance for $\hat{\delta}$.

Following the argument in Valliant, Dorfman and Royall (2000) the national estimator for chronic drug use is:

$$\hat{H}_N = \hat{\delta} \sum_{i \notin ADAM} \hat{T}_i + \sum_{i \in ADAM} \hat{C}_i \quad [24]$$

where the first summation is the sum of treatment admissions over all counties that do not have ADAM programs and the second summation is the sum of chronic user estimates over all counties that do have ADAM programs. The logic here is that we use the chronic user estimates for those counties where the estimates are available. We use the ratio estimator for counties that do not have ADAM programs.

The sampling variance for the National Estimate is:

$$VAR(\hat{H}_N) = \sigma_{\delta}^2 \left[\sum_{i \notin ADAM} T_i \right]^2 + VAR\left(\sum_{i \in ADAM} \hat{C}_i \right) \quad [25]$$

Here σ_{δ}^2 is the sampling variance for the slope coefficient in the regression [23].

Equations [24] and [25] provide the final estimators used in this study.

4.0 Generalization, Diagnostics and Adjustments

To this point, estimators have relied on specific assumptions about model specification. We explored more general alternative specifications. The first generalization was to introduce time-varying covariates into the model. This is a straightforward extension that continues to assume that the Poisson process operates throughout the period of interest, but that variables that affect the arrest rate can change over time. We eventually adopted

a two-period model, which allowed the arrest rate to differ between the first and second half of the window period.

The second potential generalization was to substitute a flexible variance function for the scalar parameter σ^2 . However, after testing, we concluding that a flexible variance function made no material improvement on the estimation, so we continued to assume that σ^2 was a scalar.

The third potential generalization was to substitute a more flexible mixing distribution for the log-normal mixture. A model based on a gamma mixture did not converge, so we tried a mixture based on a polynomial approximation to a density function. This polynomial approximation made no material improvement to the estimation, so we continued to assume that the log-normal mixture was acceptable.

We plotted and examined deviance residuals. Based on inspection of these plots, we identified and excluded a few outlier observations, but the inspection did not reveal any serious specification errors in the modeling. A plot of predicted and observed frequencies showed that for most sites there were typically more zero arrests than could be predicted by the model. Separate investigation suggests that this could be the result of the mixing distribution being slightly different from the lognormal. In any event the underprediction of zeros was modest and in the absence of a corrective, we assumed that this bias had a minor effect on our estimates.

A few observations had extreme values for arrests during the window period. Our approach was to censor reports of arrests exceeding one-per-month. The likelihood was adjusted to take that censoring into account.

The theoretical derivations presented above would lead to chronic user estimates with probability-based confidence intervals, but real-world data limitations introduce nonsampling variation into the final estimator.

The ADAM sampling weight pertains to an eight-week sampling period comprising four two-week quarterly periods. We adjusted the weight by $52/8$ to account for the rest of the year. Sometimes the ADAM site collected for periods that were longer or shorter than two weeks; sometimes an ADAM site failed to collect for all four quarters. When that was the case, we prorated accordingly.

ADAM sometimes failed to account for all jails that were in the county. When those jails that were included in the sample accounted for n arrestees and a total of N arrestees were booked, we inflated the sampling weights by N/n . The principal problem is with ADAM sites that use a stratified cluster sample, because NIJ's contractors have not provided variance adjustments for that complex sampling design. Our estimates of the sampling variances will be underestimated for places that use stratified cluster samples.

ADAM is a sample of male arrestees, but we wanted to include women in the estimates. A female ADAM program interviews woman in most of the ADAM sites, but it is not a probability-based sample, and it does not provide a basis for estimating the total number of female arrests. Nevertheless, we presume that the ADAM data are representative, and we use female ADAM to estimate the proportion of chronic users among female arrestees, and the arrest rate for female arrestees, just as we did for the male ADAM data. We used the Uniform Crime reports to estimate the proportion of female arrests to male arrests. Let: A represent the proportion of female arrests to total arrests; let B represent the proportion of female chronic drug users to male chronic drug users; and let C represent the ratio of the average arrest rate for male chronic users (based on the ADAM calendar) to the average arrest rate for female arrestees. If D is the estimated number of

male chronic drug users in the county, then $Dx(1+AxBxC)$ is an estimate of the total number of chronic drug users in the county.

Finally, with respect to nonsampling variance, we note that some sites reported for fewer than the four years of ADAM's existence. When the ADAM sites did not exist for year T, we used the average estimate for years T-1 and T+2 when these existed. Otherwise, if we had a year T-1 estimate, we used that; or if we had a year T+1 estimates, we used that in place of the missing year T estimate.

5.0 Estimates

For this paper, we define a chronic drug user as someone who used cocaine, crack cocaine, heroin, methamphetamine or marijuana on at least 4 of the past 30 days before his current arrest. Table 1 presents estimates of the number of chronic adult drug users in the United States for 2000 through 2003 by drug: cocaine, heroin, methamphetamine and marijuana. Details regarding estimation are discussed elsewhere (Rhodes et. al., 2004).

[Table 1 here]

Interpretation is straightforward. The point estimates suggest that there were about 2.6 million chronic users of cocaine during any month between 2000 and 2003. Considerable uncertainty surrounds these point estimates. For example, during 2003, we estimate that there were between 2.3 million and 3.2 million chronic cocaine users. The table reports comparable estimates for the other drugs.

6.0 Comments

The chronic user estimation methodology has other applications. For example, we have used variations on this technique to estimate the proportion of the general population with

specified infectious diseases who come into contact with the criminal justice system (Hammett, Harnon and Rhodes, 2002), and we have used other variations to support the estimation of illicit drug prices and expenditures by drug users (Rhodes et. al., 2002). The approach would seem to be useful for studying illicit drug markets and for estimating the number of career criminals in a population, among other applications.

There are limitations. This technique does not require that all chronic drug users get arrested. In fact, many chronic users can go through entire drug use careers without an arrest yet still be represented by the estimates. Treating police as samplers, an analyst has no more need to assume that police arrest all chronic drug users than a traditional sampler has to interview everybody for the resulting “sample” to represent the population.

Nevertheless, a crucial assumption is that chronic drug users run an *appreciable* risk of arrest. If this were untrue, then our method of inflating the number of chronic drug users in an arrestee population by the rate at which chronic users get arrested would provide statistics with intolerably high sampling variation, because we would be dividing by a very small number (i.e. the estimated arrest rate) with a high sampling variance for chronic users with low arrest rates. The National Survey on Drug Use and Health provides confirmation that chronic drug users do have an appreciable risk of being arrested. From the NSDUH for 2001 and 2002, we extracted the records of occasional and chronic users and tabulated their annual arrest rates.⁶ Table 2 reports estimates.

[Table 2 here]

The category “all respondents” includes everyone included in the NSDUH sampling frame, not just drug users. Clearly the arrest rate for drug users is higher than the arrest rate for those who do not use drugs. These statistics have a considerable margin of error, but they are sufficiently accurate to imply that chronic drug users run an appreciable risk of arrest, justifying the use of our chronic user estimation methods.

There may be subsets of the population, however, who have so little risk of being arrested that they would not (as a practical matter) be represented by the police sample. Television stars, for example, might avoid the criminal justice system – although the arrest of Robert Downey Jr. would seem to make the point that even movie stars lack immunity. Star athletes may avoid the justice system – although Daryl Strawberry might argue otherwise. Nevertheless, a subset of chronic drug users may avoid arrest entirely or have such a small probability of being arrested that estimates of the prevalence for that subset would be so imprecise as to be useless. What can be said about this potential residual?

One pragmatic position is that such a subset of the population that is immune to arrest is small or otherwise of marginal importance to policy makers. This is not a group that makes heavy demands on the criminal justice system, or heavy demands on the publicly-financed treatment system, or on the public health system. Perhaps from a policy standpoint it is sufficient to estimate the prevalence of chronic drug users who run an appreciable risk of arrest.

References

- Cameron, A. and Trivedi, P. (1998) *Regression Analysis of Count Data*. Cambridge University Press. Cambridge, United Kingdom.
- Cheng, C. and Van Ness, J. (1999) *Statistical Regression with Measurement Error*. Oxford University Press, New York, N.Y.
- Cohen, J. (1992) *Incapacitation Effects of Incarcerating Drug Offenders, Final Report submitted to the National Institute of Justice*.
- Cosslett, S. (1993) *Estimation from Endogenously Stratified Samples*. In *Handbook of Statistics, Volume 11*, G. Maddala, C. Rao and H. Vinod, editors. Elsevier Science Publishing.
- Englin, J. and Shonkwiler, J. (1995) *Estimating Social Welfare Using Count Data Models: An Application to Long-Run Recreational Demand Under Conditions of Endogenous Stratification and Truncation*. *The Review of Economics and Statistics*:104-112.
- Fendrich, M., T. Johnson, S. Sudman, J. Wislar and V. Spiehler, "Validity of Drug Use Reporting in a High-Risk Community Sample: A Comparison of Cocaine and Heroin Survey Reports with Hair Tests," *American Journal of Epidemiology* 149(10): 955:62, 1999.
- Hammet, T., Harmon, P. and Rhodes, W. (2002) *The Burden of Infectious Disease Among Inmates of and Releasees from US Correctional Facilities*. *American Journal of Public Health*, 92:1789-1794.
- Harrell, K. Kapsak, I. Caisson, and P. Wirtz, "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, December 1986.
- Hser, Y. (1993) *Population Estimates of Illicit Drug Users in Los Angeles County*, *Journal of Drug Issues* 23(2): 323-334.
- Hunt, D. and Rhodes, W. (2001) *Methodology Guide for ADAM*. Report prepared by Abt Associates Inc. for the National Institute of Justice.

- Kish, L. (1995) *Survey Sampling*. New York, Wiley.
- Lancaster, T. (1990) *The Econometric Analysis of Transition Data*. Cambridge University Press, Cambridge, United Kingdom.
- Lorh, S. (1999) *Sampling: Design and Analysis*. Duxbury Press.
- Maltz, M. (1996) From Poisson to the Present: Applying Operations Research to Problems of Crime and Justice. *Journal of Quantitative Criminology*, Vol. 12 (1): 3-61.
- Manski, C., Pepper, J. and Petrie, C. (2001) *Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us*. Washington, D.C.: National Academy Press.
- Manski, C. (1995) *Identification Problems in the Social Sciences*. Harvard University Press. Cambridge, MA.
- Nagin, D. and Land, K. (1993) Age, Criminal Careers and Population Heterogeneity: Specification and Estimation of a Nonparametric, Mixed Poisson Model." *Criminology* 31: 327-362.
- Rhodes et. al. (2004) *What America's User Spend on Illegal Drugs: 1988-2003*. Report submitted to the Office of National Drug Control Policy, October 2004.
- Rhodes, W., Layne, M., Bruen, A., Kling, R. and Becchetti, L. (2002) *What Americans Users Spend on Illegal Drugs: 1988-2000*. Report submitted to the Office of National Drug Control Policy by Abt Associates, Inc. Winter, 2002.
- Rhodes, W., Hyatt, R. and Scheiman, P. (1996) Predicting Pretrial Misconduct with Drug Tests of Arrestees: Evidence from Eight Settings, *Journal of Quantitative Criminology*, 12(3): 315-348.
- Simeone, R., Rhodes, W., Hunt, D. and Truitt, L. (1997) *A Plan for Estimating the Number of Chronic Drug Users in the United States*. Report submitted to the Office of National Drug Control Policy by Abt Associates Inc. April 1, 1997.
- Valliant, R., Dorfman, A. and Royall, R. (2000) *Finite Population Sampling and Inference*. John Wiley & Sons, New York, N.Y.
- Wish, E., Cuadrado, M. and Martorana, J. (undated) *Drug Abuse as a Predictor of Pretrial Failure-to-Appear in Arrestees in Manhattan*, unpublished paper prepared under Grant 83-IJ-CX-K048 to Narcotic and Drug Research Inc.

Table 1.1**Chronic Users of Cocaine, Heroin, Methamphetamine and Marijuana
(Thousands: 2000–2003)**

	Year							
	2000 Total		2001 Total		2002 Total		2003 Total	
Cocaine	2,580		2,654		2,626		2,743	
<i>Interval estimate</i>	2,144	3,016	2,217	3,091	2,189	3,063	2,307	3,179
Heroin	865		834		854		779	
<i>Interval estimate</i>	617	1,112	587	1,082	606	1,101	532	1,026
Methamphetamine	804		857		881		1,011	
<i>Interval estimate</i>	694	913	748	966	772	990	902	1,121
Marijuana	9,812		9,886		10,054		10,412	
<i>Interval estimate</i>	8,472	11,152	8,546	11,226	8,714	11,395	9,071	11,753

Note: Interval estimates equal plus or minus two standard errors.

Table 2.3**Arrest Rates by Drug Type and Use Level 2001 – 2002
(Averages from 2001 and 2002)**

Drug Type	Use Level	
	Occasional	Chronic
Cocaine	0.30	0.44
Heroin	0.56	0.49
Methamphetamine	0.52	0.32
Marijuana	0.15	0.29
All respondents	0.04	

Notes: All respondents include all respondents to the NSDUH regardless of self-report drug use. Most were not users.

End Notes

¹ The term *arrest* appears interchangeably with the term *booking* in the text, but both terms mean that a person was detained in jail pending a court hearing. The reader should note that an arrest does not always result in an arrestee being booked into jail, because the arresting officer sometimes has authority to release a person after issuing a citation for that person's appearance in court. Conversely a booking does not always require an arrest, because probation and parole officers can bring probationers/parolees to jail in response to technical violations of the conditions of supervision. When the term is used in this paper, an arrest means being booked into a jail. Booking is the process depicted on television police programs when a person is fingerprinted and waits for a magistrate to set bail. Typically this process takes a few hours; sometimes it takes a few days. Because ADAM interviews after a person is booked, arrests that do not result in booking are invisible to the ADAM survey.

² As Manski (1995) notes, endogenous stratification raises the question of identification. We deal with identification by making a strong, but justifiable distribution assumption. Specifically, criminal justice researchers have frequently used a Poisson process to model arrests (see Maltz, 1996), so we adopt a Poisson mixture (generally, see Cameron and Trivedi, 1998) as a provisional stochastic model of the arrest process. Englin and Shonkwiler (1995) provide a likelihood function for the Poisson process with endogenous stratification when λ follows a mean-one gamma distribution. Their approach can be adapted to other mixing distributions, including the log-normal.

³ The arrest that caused the drug user to enter the sample is uninformative about the arrest rate and hence does not enter into estimation. Cameron and Trivedi (1998, page 329) provide a formal explanation. The intuition is that the likelihood can be written as a joint probability of observing N arrests during the window period and observing the N+1st arrest at the time that the sample is drawn. Given the Poisson distribution, the occurrence of the N+1st arrest is independent of the occurrence of the prior N arrests, so the likelihood can be written as the product of the probability of N events and the probability of an additional event at the time that the sample is drawn. However, that second probability is 1 conditional on the fact that an arrest did in fact occur. Hence the likelihood is the probability of observing N prior arrests.

⁴ ADAM does not provide sufficient information to compute a standard error for those counties that sample arrestees according to a stratified cluster sample design. We inflated estimates by N/n where N is the total booking population in the county and n is the total booking population for those jails that were included in the ADAM sample. This approach will understate the sampling variance in those places that use a stratified cluster sample.

⁵ Arrestees are less likely to admit drug use during the two to three days before their arrest than they are to admit drug use during the thirty days before their arrest. Therefore responses to the question about use during the last thirty days understate the level of truthfulness that is important for chronic user estimates.

⁶ We average the reported arrest rates (variable NOBOOKY2) across the 2001 and 2002 surveys. This question was not asked on the 2000 survey. NOBOOKY2 reported the number of arrests during the last year conditional on having ever been arrested. We used responses to the question about whether the respondent had ever been arrested and booked (BOOKED) to convert the conditional rates to unconditional rates.