NATIONAL CRIME SURVEY
Empirical and Model-Based Estimators for the Proportion of Households Victimized in a Year

Charles H. Alexander and Michael J. Roebuck

Section I: Overview

Using the National Crime Survey (NCS) to estimate the proportion of households victimized during a year presents a problem in longitudinal imputation. It takes two NCS interviews, with their six-month reference period, to cover a twelve-month period of time. For some sample units, one of the two interviews is missing. However, it still may be possible to use the information from the available interview. To do so requires some assumptions about the joint distribution of victimizations during the two interview periods.

One approach is to assume that the probability distribution for the units with missing data is exactly the same as the distribution for those with complete data. Adjustment factors calculated from the complete observations are then applied to the incomplete observations, in such a way that the resulting "empirical" estimator is consistent if the assumption is correct. The present published estimates of the proportion of households victimized in a year are calculated using this approach.

A potentially superior approach is to assume that the joint distributions of the missing data and the complete data are members of the same family of distributions but may have different parameters. The parameters can be estimated from the available data on incomplete cases and used to adjust for the missing data.

This model-based approach stems from work by Eddy, Fienberg, and Griffin [1]. Three specific families of distributions were suggested for this problem by those authors. They used a slightly different approach from the one used here, in that they assumed the same parameters for the missing and complete data; they showed how the entire available data could then be combined to estimate the parameters in an efficient manner, assuming the model is correct.

A question about the modelling approach is whether use of an inappropriate model for crime data can result in misleading estimates and, if so, how to determine what models give good results. Crime is a complex phenomenon which is unlikely to fit a simple model exactly. Victimization is known to exhibit seasonal fluctuations, is known to have some correlation for neighboring households and correlations from interview to interview for the same household, and is known to occur at very different rates for different kinds of households. The models considered in this paper reflect some of these aspects of victimization, but only in a crude way.

This paper proposes a comparison of these models by examining their fit to the NCS data. It also examines the effect of using an estimator based on one model when the data in fact fit a different model.

Some of the NCS data needed for these comparisons and tests of fit were not available when the paper was prepared. Consequently, the question about the effect of using the wrong model has been addressed using hypothetical parameter values which are consistent with the available data. The results indicated that the model-based estimators may do much worse than the "empirical" or "ad hoc" estimators if the wrong model is used. Thus a test for fit of the model is essential. Suggestions for future work are presented.

Section II gives a summary of the problem. Section III and IV present the models. Section V discusses problems in applying these models to the NCS. Section VI gives the results of the comparison of the models, Section VII discusses the results, and Section VIII gives recommendations for further work.

The models presented here could be applied more generally to any characteristic of interest. Examples may be a person attending a sporting event in a given month, purchasing an article of clothing, or going to a concert. Some recreation-related supplements to the NCS might also use estimators based on models similar to those shown here. However, the models examined here focus on modelling crime data.

We wish to thank Jenell P. Avent, Joan E. George, and Donna M. Littleford, our secretaries, without whose infinite patience and assistance this paper would never have been finished.

Section II: Summary

The following is a simplified statement of the problem. Define two random variables, Z_1 and Z_2, for which $Z_1 = 1$ if the selected housing unit (HU) has the characteristic of interest (for example, some type of victimization) in the i^{th} half year, and $Z_1 = 0$ if not.

The population is divided into three groups, A, B, and C, for which (Z_1, Z_2) may have different unknown joint probability distributions P_A, P_B, and P_C. (For for any event E, $P_A(E)$ denotes the probability that the event occurs for an individual selected from group A.) A sample of HUs is selected from each group. It is assumed that the sample design is ignorable. For sample HUs from group A, both Z_1 and Z_2 are observed. Only Z_1 is observed for HUs from group B, and only Z_2 is observed for group C HUs.

Our is goal to find consistent estimators for $P_B(Z_1 = 1 \text{ or } Z_2 = 1)$ and $P_C(Z_1 = 1 \text{ or } Z_2 = 1)$. A consistent estimator for $P_A(Z_2 = 1 \text{ or } Z_2 = 1)$ is already available from the corresponding sample proportion for group A. Since we can obtain consistent estimators for the proportion of the population in each group, this will allow us to find a consistent estimator of the overall probability that an HU had the characteristic of interest at some time during the year.

As applied to the National Crime Survey's estimate of the proportion of households victimized in a year (the so-called "touched by crime" estimator), this is an oversimplified

statement of the problem. Section V discusses the actual problem in greater detail. The estimator now in use imputes the missing data for groups B and C by assuming $P_A = P_B = P_C$. (Evidence from the survey suggests that P_A, P_B, and P_C are fairly similar.) Section III describes this "ad hoc" estimator and also provides an alternative estimator which is consistent under the weaker assumptions that $\rho_A(Z_1, Z_2) = \rho_B(Z_1, Z_2) = \rho_C(Z_1, Z_2)$

(where, for example, ρ_A denotes denotes the correlation between Z_1 and Z_2

in group A), and that

$$\frac{P_A(Z_2 = 1)}{P_A(Z_1 = 1)} = \frac{P_B(Z_2 = 1)}{P_B(Z_1 = 1)} = \frac{P_C(Z_2 = 1)}{P_C(Z_1 = 1)}.$$

Another alternative suggested by Griffin [2] is also described. These are method of moments estimators which rely on the fact that a uniformly continuous function of a sample proportion is a consistent estimator of the same function of the corresponding probability.

A more interesting approach is inspired by Eddy, Fienberg, and Griffin [1]:

(1) Define X_1, X_2, X_3, and X_4 to be zero-one random variables indicating respectively whether a HU had the characteristic in each of the four calendar quarters. Hence $Z_1 = 1$ if $X_2 = 1$ or $X_1 = 1$ and $Z_2 = 1$ if $X_3 = 1$ or $X_4 = 1$.

(2) Assume a parametric model, $f(x_1, x_2, x_3, x_4; \theta)$ for the discrete discrete probability function for P_A, P_B, or P_C, with a vector of parameters $\theta = \theta_A$, θ_B, or θ_C, respectively.

(3) Examine the fit of the model for Group A by comparing $f(x_1, x_2, x_3, x_4; \hat{\theta}_A)$ with the empirical distribution, for all values of x_1, x_2, x_3 and x_4, where $\hat{\theta}_A$ is the maximum likelihood estimator (MLE) for group A. If the fit is acceptable, proceed to step (4):

(4) Obtain the MLE $\hat{\theta}_B$ based on the observations of X_1 and X_2 for Group B. Then find P_B by applying the relationship $P_B(Z_1 = 1 \text{ or } Z_2 = 1) = 1 - f(0, 0, 0, 0; \theta_B)$. Similarly, $\hat{\theta}_C$ is estimated from the observations of X_3 and X_4, and P_C is found by $P_C(Z_1 = 1 \text{ or } Z_2 = 1) = 1 - f(0, 0, 0, 0; \theta_C)$.

(5) The estimates for groups B and C are then combined with an estimate for group A, which is the observed sample proportion $P_A^*(Z_1 = 1 \text{ or } Z_2 = 1)$

The model-based approach we present here differs from that of [1] in several respects. Their work uses data for each of the twelve months rather than four quarters. We have used quarters for this initial study to simplify the computational problem. They also estimate $f(0, 0, 0, 0; \theta)$, which is the proportion of HU in the population which did not have the characteristics during the year. As the problem applies to the NCS, $f(0, 0, 0, 0; \theta)$ is the "cheery" probability (proportion of HUs not touched by crime), rather than $1-f(0, 0, 0, 0; \theta)$, which is the "touched-by-crime" probability.

A more fundamental difference is that [1] assumes throughout that $\theta = \theta_A = \theta_B = \theta_C$. Under this assumption, their work gives gives an asymptotically efficient estimator of $f(0, 0, 0, 0; \theta)$. The "ad hoc" method-of-moments estimators are not efficient, although they are all consistent if $\theta_A = \theta_B = \theta_C$. By constrast, the present paper assumes that the parameters may differ.

A drawback of the model-based approach is that if the wrong model is used, the estimator may not be consistent, and with a very bad model the estimates may so biased as to be meaningless. Therefore, we feel that it is very important to examine the fit of whatever model is used. We also favor the more cautious approach of using a model only when it is necessary - to impute the missing data for groups B and C - and not to use a model to try to improve on the observed sample proportion in Group A.

For the estimators presented throughout this paper, it is assumed that the nonresponse mechanism is ignorable within certain noninterview cells for all completely noninterviewed house holds (those for which no interview is available). Under this assumption, these cases are properly dealt with by the usual NCS noninterview adjustment.

All of the estimators presented here are biased for small samples, as are many estimators used in survey work. Our concern is with asymptotic (or "large sample") bias. Thus we seek an estimator which is consistent, not necessarily one which is unbiased. Variance is of less concern than possibly large asymptotic biases; for the NCS we are dealing with very large samples.

Section III: Empirical or "Ad-Hoc" Estimators

While we differentiate these estimators from the later "model-based" estimates, these "ad hoc" estimators are in fact based on models about the distribution of Z_1 and Z_2 in the three groups. These models have, however, not featured explicitly in the development

of the estimators.

For a given probability P, which is to be estimated from the sample, let P* denote the corresponding sample proportion, and $\hat{P}$ denote any particular estimator for P. The estimators will be described for P_B in all cases: P_C, where not presented, is determined by keeping in mind that the known proportion in group C is $P^*_C\ (Z_2 = 1)$.

Estimator 1 (E1): The Original "Ad Hoc" Estimator.

This is the estimator currently used to estimate the proportion of HUs touched by crime in a year.

$$(1)\quad \hat{P}_B\ (Z_1 = 1 \text{ or } Z_2 = 1) = \frac{P^*_A\ (Z_1 = 1 \text{ or } Z_2 = 1)}{P^*_A\ (Z_1 = 1)} \cdot P^*_B\ (Z_1 = 1)$$

This estimator is consistent if $P_A = P_B = P_C$, since each sample proportion P* converges in probability to the corresponding population value P. We further restrict the sufficient conditions for consistency by noting that

$$(2)\quad P_B\ (Z_1 = 1 \text{ or } Z_2 = 1) = P_B\ (Z_1 = 1)\ [1 + \frac{P_B\ (Z_2 = 1)}{P_B\ (Z_1 = 1)} - P_B\ (Z_2 = 1 \mid Z_1 = 1)]$$

However, examining the right-hand side of (1)

$$(3)\quad \frac{P_A\ (Z_1 = 1 \text{ or } Z_2 = 1)}{P_A\ (Z_1 = 1)} \times P_B\ (Z_1 = 1) = P_B\ (Z_1 = 1)\ [1 + \frac{P_A\ (Z_2 = 1)}{P_A\ (Z_1 = 1)} - P_A\ (Z_2 = 1 \mid Z_1 = 1)]$$

To be equal, it is sufficient that

$$(4)\quad \frac{P_B\ (Z_2 = 1)}{P_B\ (Z_1 = 1)} = \frac{P_A\ (Z_2 = 1)}{P_A\ (Z_1 = 1)}$$

and

$$(5)\quad P_B\ (Z_2 = 1 \mid Z_1 = 1) = P_A\ (Z_2 = 1 \mid Z_1 = 1)$$

Condition (4) corresponds to a constant multiplicative "seasonal" effect, and seems reasonably mild. The second condition (5) seems to be stronger. For example, if Z_1 is independent of Z_2, (5) implies that $P_B\ (Z_2 = 1) = P_A\ (Z_2 = 1)$, i.e., that the proportions with the characteristic in groups A and B are identical for the second half of the year. Conditions (4) and (5) are presented for comparison with E2 below.

Estimator 2 (E2): Griffin's Estimator.

This estimator was described by Griffin [2].

$$(6)\quad \hat{P}_B\ (Z_1 = 1 \text{ or } Z_2 = 1) = P^*_B\ (Z_1 = 1) + P^*_B\ (Z_1 = 0) \cdot P^*_A\ (Z_2 = 1 \mid Z_1 = 0)$$

This estimator is again consistent if $P_A = P_B = P_C$. More precisely, a necessary and sufficient condition for consistency is that

$$(7)\quad P_A\ (Z_2 = 1 \mid Z_1 = 0) = P_B\ (Z_2 = 1 \mid Z_1 = 0)$$

These assumptions appear to be less restrictive than (4) and (5) in that conditions (7) are similar to (5), while no condition like (4) is needed. However, it may be that violations of condition (7) have a greater effect on E2 than violations of (5) have on E1. Thus, in practice the relative merits of Estimator 1 and Estimator 2 are unclear. (For the actual NCS data, E1 is almost identical to E2, so we have not investigated E2 very carefully.) Under the assumption $P_A = P_B = P_C$, E2 is a maximum likelihood estimator for Group B (ignoring Group C) and Group C (ignoring Group B). Consequently, E2 warrants serious consideration as an alternative to E1.

Estimator 3 (E3): Equal Correlation Model

This estimator imputes the missing data for groups B and C as follows.

$$(8)\quad \hat{P}_B\ (Z_1 = 1 \text{ or } Z_2 = 1) = \hat{P}_1\ [1 + \hat{R}_1 - \hat{R}_1\hat{P}_1 - \hat{\rho}\ \sqrt{\hat{R}_1} \cdot \sqrt{(1 - \hat{P}_1)\ (1 - \hat{R}_1\hat{P}_1)}]$$

where

$$\hat{P}_1 = P^*_B\ (Z_1 = 1)$$

$$\hat{P}_2 = P^*_C\ (Z_2 = 1)$$

$$\hat{R}_1 = \frac{P^*_A\ (Z_2 = 1)}{P^*_A\ (Z_1 = 1)}$$

$$\hat{R}_2 = \frac{P^*_A\ (Z_1 = 1)}{P^*_A\ (Z_2 = 1)} = 1/\hat{R}_1,$$

Sufficient conditions for E3 to be consistent are that both the ratio $R_1 = \frac{P\ (Z_2 = 1)}{P\ (Z_1 = 1)}$ and the correlation $\rho\ (Z_1, Z_2)$ are the same for groups A, B, and C. Indeed, in the case of (8), letting

$P_1 = P_B(Z_1 = 1)$ and $P_1' = P_B(Z_2 = 1)$,

$$P_B(Z_1 = 1 \text{ or } Z_2 = 1) = P_1 + P_1'$$
$$- P_B(Z_1 = Z_2 = 1)$$
$$= P_1(1 + R_1)$$
$$- [\rho \sqrt{P_1(1 - P_1) P_1'(1 - P_1')} + P_1 P_1']$$
$$= P_1 [1 + R_1 - R_1 P_1 - \rho \sqrt{R_1(1 - P_1)(1 - R_1 P_1)}]$$

Since under the assumptions, $\hat{R}_1$, $\hat{\rho}$, and $\hat{P}_1$ are consistent estimators of R_1, $\rho(Z_1, Z_2)$ and P_1, respectively then (8a) gives a consistent estimator of $P_B(Z_1 = 1 \text{ or } Z_2 = 1)$. The argument for $\hat{P}_C$ is similar.

Section IV: Model-Based Estimators

The model-based estimators are inspired by the work of Eddy, Fienberg and Griffin [1]. The basic idea is that more information is used by an estimator based on the monthly experience of each household, or in our work, on the quarterly experience. However, to use this information a specific model is required.

This section presents five such models. The first, which we call model 4 because it corresponds to estimator E4, assumes that the characteristic of interest occurs independently and with equal probability for all quarters and all households. The other models are aimed more specifically at representing the way crime is thought to behave. Model 5 reflects the fact that some households tend to have repeated or "multiple" victimizations, by assuming that a subgroup of the population have an unusually high probability of being victimized in every quarter. Model 6 assumes that the probability of victimization in a given quarter depends on whether there was a victimization in the previous quarter. Model 7 reflects multiple victimizations by starting with an independent model, and randomly adding one extra victimization to a proportion of the victimized households. Model 8 allows different households to have different probabilities of victimization.

In each model (X_1, X_2, X_3, X_4) denotes the random vector describing whether or not there has been an occurrence in each quarter. For example $X_1 = 1$ if there has been an occurrence in the first quarter and $X_1 = 0$ if not.

For reasons of space, we have abbreviated the description of several of the models. The full description is contained in a longer version of the paper, available from the authors.

Estimator 4(E4): Complete Independence Model

This model corresponds to the assumptions that X_1, X_2, X_3 and X_4 are mutually independent with equal probabilities of occurrence and that this probability is the same for all households. The monthly version of this model was used in [1].

This model for the joint distribution is

$$(9) \quad f(X_1, X_2, X_3, X_4; p) = p^{\sum X_i} (1-p)^{4 - \sum X_i}$$

Here $\theta = p$, the probability of victimization in a given quarter.

Let $[(X_{1_j}, X_{2_j}, X_{3_j}, X_{4_j}), j=1, \ldots, n_A]$ be an i.i.d. sample from this distribution, describing the n_A HUs in Group A. Let $\underset{\sim}{x} = (x_{ij})$ denote the observed values in the sample. The likelihood function is written as

$$L(p; \underset{\sim}{x}) = p^{n_1} (1 - p)^{n_0}$$

where, for Group A, $n_1 = \sum_{j=1}^{n_A} \sum_{i=1}^{4} x_{ij}$ and $n_0 = 4n_A - n_1$ denote respectively the total number of HU-quarters with and without the characteristic of interest. (Each HU in group A contributes 4 HU-quarters.) The maximum likelihood estimator based on $\underset{\sim}{x}$ for the proportion of HUs with the characteristic in a quarter is $\hat{p}_A = \frac{n_1}{4n_A}$ for Group A HUs. For Group B, only (x_1, x_2) is observed. For an i.i.d. sample selected from Group B $(j=1, \ldots, n_B)$, the likelihood function is the same, except that

$n_1 = \sum_{j=1}^{n_B} \sum_{i=1}^{2} x_{ij}$, $n_0 = 2n_B - n_1$, and $\hat{p}_B = \frac{n_1}{2n_B}$.

A similar change is made for Group C households.

The estimate of the probability that an HU had the characteristic at some time during the year is then

$$(10a) \quad \hat{P}_B(Z_1 = 1 \text{ or } Z_2 = 1) = \hat{P}_B(X_1 = 1 \text{ or } X_2 = 1 \text{ or } X_3 = 1 \text{ or } X_4 = 1) = 1 - f(0,0,0,0;\hat{p}_B)$$
$$= 1 - [\hat{p}_B^0 (1-\hat{p}_B)^4] = 1 - (1-\hat{p}_B)^4$$

$$(10b) \quad \hat{P}_C(Z_1 = 1 \text{ or } Z_2 = 1) = 1 - f(0,0,0,0;\hat{p}_C)$$
$$= 1 - (1-\hat{p}_C)^4$$

For group A, $\hat{P}_A(Z_1 = 1 \text{ or } Z_2 = 1)$ is the proportion of the n_A sample cases which have either $Z_1 = 1$ or $Z_2 = 1$. The estimator E4 is calculated as

(11) $\hat{R}_A\ \hat{P}_A(Z_1 = 1 \text{ or } Z_2 = 1)$

$+ \hat{R}_B\ \hat{P}_B(Z_1 = 1 \text{ or } Z_2 = 1) + \hat{R}_C\ \hat{P}_C(Z_1 = 1 \text{ or } Z_2 = 1)$,

where $\hat{R}_A$, $\hat{R}_B$, and $\hat{R}_C$ are estimates of the proportion of the population in groups A, B, and C.

An alternative is to apply the model to all three groups, not just groups B and C. Then under the assumption that $P_A = P_B = P_C$, a maximum likelihood estimate of p may be calculated as

$$\hat{p} = \frac{\text{total number of observed HU-quarters with a victimization}}{\text{total number of observed HU-quarters}}$$

The estimator E4' calculated under this assumption is defined to be $1 - (1 - \hat{p})^4$.

Estimator 5 (E5). This is the Tallis distribution model described in [1], using a distribution from [3].

Estimator 6 (E6). This is the Markov Model from [1]. For group B or C, a closed-form expression of the MLE can be easily derived.

Estimator 7 (E7) (Independent with additional occurrences)

For a fixed probability of occurence, the model corresponds to complete independence.

This model allows more multiple victimizations than the independent model. Households which have been victimized are given a certain probability of receiving an extra victimization. The idea is similar to model 5, but the model principally increases the proportion of units with $\Sigma X_i = 2$ rather than those with $\Sigma X_i = 4$.

The distribution may be generated in the following way. Each household is first victimized in the four quarters according to the independent model with parameter p. Each household which has $\Sigma X_i > 0$ then, with probability π, has a chance of receiving one extra victimization. The extra victimization is equally likely to occur in any quarter. It may occur in a quarter in which there is already a victimization.

For this model,

$$f(0, 0, 0, 0) = (1-p)^4$$

$$f(X_1, X_2, X_3, X_4) = p(1-p)^3 [(1-\pi) + 1/4\ \pi] \quad \text{if } \Sigma X_i = 1$$

$$f(X_1, X_2, X_3, X_4) = p^2(1-p)^2 [(1-\pi) + 2/4\ \pi] + 2\pi\ p(1-p^3)/4; \quad \text{if } \Sigma X_i = 2$$

$$f(X_1, X_2, X_3, X_4) = p^3(1-p) [(1-\pi) + 3/4\ \pi] + 3\pi p^2(1-p)^2/4, \quad \text{if } \Sigma X_i = 3$$

$$f(1, 1, 1, 1) = p^4 + \pi p^3 (1-p)$$

Estimator 8 (E8). Beta-binomial distribution. This assumes that there are many probabilities of occurence in the population, but that there probabilities are distributed as a beta distribution with unknown (but estimable) parameter α and β.

Section VI: Comparison of Estimators

This section compares the estimators described above by seeing how well they perform under the various models which have been proposed. We are able to compute estimators E1, E2, E3, E4, and E6, for which closed-form expressions exist; the computer programming necessary the calculate the other maximum likelihood estimators has not been completed. Estimator E4 appears in two versions, first applied only to group B and C and second applied to all three groups. The latter version is called E4'. Except for E4', each estimator is used only to impute the missing data; for group A, $f(0,0,0,0;\theta_A)$ is estimated by the corresponding sample proportion. The final estimate is computed using (11).

For these comparisons, no group C was used. For our models this is equivalent to assuming that groups B and C have identical parameters; this is because for these models

$f(x_1, x_2, x_3, x_4; \theta) = f(x_4, x_3, x_2, x_1; \theta)$

for all values of x_1, x_2, x_3, x_4. This assumption that groups B and C are identical was made for simplicity. It is not supported by NCS data and it should be relaxed in future work. The comparisons were made by assuming a particular model, selecting fairly realistic parameter values, and assuming that the data would correspond exactly to the probability distribution. Each estimator was calculated under this assumptions, using each model. Detailed quarterly data for NCS were not available in time for this study. Consequently parameter values were chosen to correspond to the available data for NCS total households touched by crime in 1982. For 1982 we have the following estimates.

$P_A(Z_1 = 1) = .1735 \qquad \rho_A = .329$

$P_A(Z_2 = 1) = .1677$

Accordingly parameter values for group A were selected so that $P_A(Z_1 = 1)$

$= P_A(Z_2 = 1) = .170$ and $\rho_A = .33$.

(For model 4, $\rho_A = 0$.) Our conjecture is that the required parameter values are unique, although we have not proved this.

For groups B and C, the NCS estimates are

$P_B(Z_1 = 1) = .1768 \qquad P_C(Z_2 = 1) = .1854$

Estimates for ρ_B and ρ_C are not available. Since our group B represents both the real groups B and C, the difference between groups A and B was exaggerated slightly by choosing parameter values for group B which make $P_B(Z_1 = 1) = P_B(Z_2 = 1) = .19$. Many choices of θ_B satify this condition when there are two or more parameters. For definiteness we required that percentage differences between

all the group A and B parameters be the same. For example for model 6 we require that $P_B = (1 + C)P_A$ and $\pi_B = (1 + C)\pi_A$, where the constant C is chosen so that $P_B(Z_1 = 1) = .19$. Models 4, 5, and 7 are are treated similarly. The one exception is model 8, where $\alpha_B = (1 + C)\alpha_A$ but $\beta_B = \beta_A/(1 + C)$. For this model, the correlation between Z_1 and Z_2 varies inversely with β.

Accordingly the following parameters were used:

Model	Group A Parameters	Group B Parameters	$P_A(Z_1 = 1)$	$P_B(Z_1 = 1)$
4	$p = .0890$	$p = .1000$	.170	.190
5	$p = .1135$; $\pi = .4381$	$p = .1340$; $\pi = .5171$	.170	.190
6	$p_0 = .0587$; $p_1 = .5624$	$p_0 = .0627$; $p_1 = .6006$	.170	.190
7	$p = .0739$; $\pi = .4518$	$p = .0818$; $\pi = .5001$	.170	.190
8	$\alpha = .3486$; $\beta = 3.1357$	$\alpha = .3732$; $\beta = 2.9286$	.170	.190

ρ_A	ρ_B
0	0
.330	.412
.330	.360
.330	.355
.330	.340

The probability functions $f(x_1, x_2, x_3, x_4; \Theta)$ were calculated for each of these models. For each model all the estimators E1, E2, E3, E4, E6, and E4' were calculated assuming that the empirical distribution matched the probability function exactly. The results are given in the following table. The estimates have been expressed as a rate per 1000 households.

	4	5
True Rate per 1000	317.8	267.8
E1: Estimate:	318.5	270.8
(Rel. Error)	(0.2%)	(1.1%)
E2: Estimate:	314.5	268.1
(Rel. Error)	(-1.0%)	(0.1%)
E3: Estimate:	317.8	270.3
(Rel. Error)	(0)	(0.9%)
E4: Estimate:	317.8	299.2
(Rel. Error)	(0)	(11.7%)
E6: Estimate:	317.8	270.0
(Rel. Error)	(0)	(0.8%)
E4': Estimate:	314.9	388.7
(Rel. Error)	(-0.9%)	(45.2%)

	Model 6	7	8
	269.3	269.4	270.0
E1:	270.7	270.6	270.8
	(0.5%)	(0.5%)	(0.3%)
E2:	268.1	268.0	268.1
	(-0.5%)	(-0.5%)	(-0.7%)
E3:	270.3	270.2	270.3
	(0.3%)	(0.3%)	(0.1%)
E4:	300.1	284.8	287.9
	(11.4%)	(5.7%)	(6.6%)
E6:	269.3	278.0	276.5
	(0)	(3.2%)	(2.4%)
E4':	400.9	334.5	348.2
	(48.8%)	(24.2%)	(29.0%)

Note that when the right model is used, as in applying E4 to data from model 4 or E6 for model 6, the asymptotic bias is zero.

Section V: Overview of the NCS

This section of the full paper has been omitted for reasons of space.

Section VII: Conclusions

1. The empirical touched by crime estimators perform fairly well The five models allow for a variety of causes for the correlation between Z_1 and Z_2.

   In all cases, E1, E2, and E3 have fairly small biases. Which estimator does best depends on the model. Except for model 5, E3 does the best of the three. For model 5, E2 has the least bias.

2. The model-based estimators can do poorly when the wrong model is used. Estimator E4 has a substantial bias for models 5, 6, 7, and 8, in which there is a correlation between Z_1 and Z_2. Estimator E6, which is based on a model which allows for a correlation, does better. However, it does not do very well for model 7 and 8. The somewhat better performance for E6 suggests that estimator E5, E7, or E8 might also have promise.

   Of course, when the right model is used, the model-based estimator has better variance than the empirical estimators. This is not taken into account in our comparisons, which deal only with asymptotic bias.

3. Adjusting the complete data (Estimator E4') can lead to much larger biases if the model is wrong Applying the model to group A in addition to group B, estimator E4' leads to large overestimates of the proportion of households touched by crime.

   It should be noted that by choosing different models, one can support different conclusions. Some other examples were examined, using arbitrary parameters not consistent with NCS data. If there is a very large difference between the parameters for groups A and B, estimators E1, E2, and E3 may not do very well. E3 did better in such cases except when the parameters were chosen so that ρ_A and ρ_B are quite different. In such cases, the correctly specified model-based estimator does quite well. The incorrectly specified estimators do quite poorly in most cases.

VIII. Future Work The limitations of the present study suggest directions for future work in applying these methods to the NCS. We hope that researchers outside the Census Bureau will be able to improve on our efforts. The following work still needs to be done.

1. Test the fit of the various models to the empirical distribution of the quarterly values (x_1, x_2, x_3, x_4). Use the best-fitting models to compare estimators.

2. Use actual maximum likelihood estimates for Θ_A and Θ_B in the above analysis in place of the hypothetical values which were deduced in Section VII. Extend the analysis to include E5, E7, and E8.

3. Extend the work to situations in which the EX_i are not equal. This might be done by making an adjustment for seasonality and the recall (or "recency") effect before applying the model.

4. a. Incorporate a group C whose characteristics differ from group B.
   b. Further divide the Groups according to available information on covariates, such as race of household head, urban/ rural status.

5. Seek additional models which more realistically represent crime. One approach would be to include several effects by combining model 5 or 7, model 6, and model 8, along with a seasonality adjustment.

6. Extend this work to monthly, rather than quarterly, data.

7. Improve on estimator E3 using models more appropriate for discrete data.

Beyond this work remain the larger questions of how to incorporate the actual NCS sample design into the theory and of how to construct prevalence estimates on a data year basis.

References

[1] Eddy, W.F., S.E. Fienberg, and D.L. Griffin, "Some First Attempts "Estimating Victimization Prevalence in a Rotating Panel Survey." Techincal Report No. 220, Department of Statistics, Carnegie-Mellon University, September, 1981.

[2] Griffin, D.L., "Discussion of Several Estimators of the Proportion of Households Victimized in a Year," Working Memorandum NCS-3, Department of Statistics, Carnegie-Mellon University, June 1981.

[3] Tallis, G.M., "The use of a generalization multinomial distribution in the estimation of correlation in discrete data," J. Roy Statist. Soc. B., 24(1962), pp. 530-4.

[4] Bateman, David and Paul Bettin, "Standard error estimation for the National Crime Survey," paper presented at the annual meetings of the American Statistical Association in Atlanta, Georgia, August 25-28, 1975.