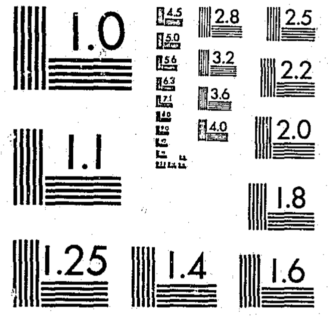


CONSISTENCY OF SOME INTUITIVE ESTIMATORS OF THE PREVALENCE OF VICTIMIZATION

by

Diane Griffin

DEPARTMENT
OF
STATISTICS

Carnegie-Mellon University
PITTSBURGH, PENNSYLVANIA 15213

CONSISTENCY OF SOME INTUITIVE ESTIMATORS OF THE PREVALENCE OF VICTIMIZATION

by

Diane Griffin

TECHNICAL REPORT NO. 271

Department of Statistics
Carnegie-Mellon University
Pittsburgh, PA 15213

January, 1983

This material is based on work performed under Grant 81-IJ-CX-0087 from the National Institute of Justice in the Office of Justice Assistance Research and Statistics, U.S. Department of Justice. Points of view and opinions stated herein are those of the author and do not necessarily represent official positions or policies of the Department of Justice. Thanks are due to Luke Tierney for suggesting the argument for deriving a model under which an estimator is consistent and to Steve Fienberg and Mark Schervish for helpful comments along the way.

1. INTRODUCTION

Several methods of estimating the probability of a housing unit being crime-free during a given year using the National Crime Survey data have been discussed by Eddy, Fienberg, and Griffin(1981,1982), Griffin(1981), and Alexander(1981). Some of these estimators are not based on assumed models but rather are intuitive yet ad hoc in nature. While considering only housing units with no nonresponse and assuming that housing units are victimized independently, it is possible to derive models for which these estimators are consistent.

If we consider the housing units in the survey to be a random sample from an infinite population, then the consistency of an estimator implies, roughly, that as the number of observations tends to infinity, the value of the estimator tends to the value of the parameter it estimates. For further discussion of consistency, see Cox and Hinkley(1974).

In particular, the ad hoc estimator of Eddy, Fienberg, and Griffin is found to be consistent only under a model which assumes that the probability of reporting a victimization does not depend on the number of months of information that a housing unit contributes during the year of interest. The modified version of this estimator is found to be consistent under a model which seems, at least for a small sample, to fit the data well. The models under which the Bureau of Justice Statistics' estimator and Griffin's RNEW are consistent are similar to the model for the ad hoc estimator in that the probability of victimization in one half of the year must be the same as the probability of victimization in an entire year.

Throughout this paper, the term housing unit is abbreviated HU and θ represents the probability that an HU is crime-free during a given year.

2. THE SURVEY AND THE DATA

The National Crime Survey (NCS), designed and executed by the U.S. Bureau of the Census, is based upon a stratified multistage cluster sample. The first stage consists of dividing the United States into 1931 primary sampling units (PSU's) comprised of counties and groups of contiguous counties. The PSU's are then divided into 376 strata, 156 of which are self-representing. From the remaining 220 strata one PSU is selected from each stratum with probability proportional to population size. Within each of the 376 PSU's selected, a systematically chosen group of enumeration districts is selected, and then clusters of approximately four

HU's are chosen within each enumeration district. This method produces a self-weighting probability sample of dwelling units and group quarters within each chosen PSU.

This basic sample is then divided into six rotation groups, each of which contains about 9,000 HU's. Every six months a new rotation group enters the sample and the "oldest" existing rotation group from the previous sample is dropped. Each rotation group is divided into six panels with panel 1 being interviewed in January and July, panel 2 in February and August, etc. This process spreads the workload of the field staff. Each rotation group remains in the survey for a total of seven interviews and is then rotated out.

At each interview NCS respondents provide victimization information on the preceding six months. To actually determine if an HU has been victimized in a particular year it is, in principle, necessary to examine all of the interviews of the occupants of the HU that contain information for some part of the year in question. Typically this will mean that we need information from a respondent for three successive interviews to reconstruct the victimization profile for a single year.

The NCS victimization data are publicly available through the Inter University Consortium for Political and Social Research (ICPSR) at the University of Michigan. These data are grouped into quarterly collection files which include records of all the interviews completed by the U.S. Bureau of the Census for a particular three-month period. Because the occupants of a specific HU are interviewed every six months, each quarterly collection file contains the records for at most one interview for that HU. Since we need information from as many as three successive interviews to determine whether or not an HU has been victimized during a given year, the data will need to be matched or linked in some longitudinal format. Professor Albert Reiss of Yale University has produced longitudinal files from the cross-sectional files for the period from July 1, 1972 to December 31, 1976. These files have been used for the analysis performed in this paper.

In practice, we do not get to see a complete longitudinal record for every HU. When an HU enters or leaves the sample during the year, part of the desired data will be missing. Similarly data for six-month intervals can be missing due to non-interviews. In addition, because of errors in the data, it was not possible to link some of the records to previous records for the same HU. Thus some missing data occurs because of matching difficulties. This report examines the problem of

estimating θ, the probability that an HU is victimized in a year, when the only cause of missing data is the rotation scheme.

3. THE AD HOC ESTIMATOR

The ad hoc estimator of θ discussed by Eddy, Fienberg, and Griffin(1981,1982) is

$$\hat{\theta}_1 = \frac{\text{\# of interview months in crime-free HU's}}{\text{\# of interview months}} .$$

We shall assume that there is no nonresponse and that the matching has been completed without error and thus the only missing data are those due to the rotation scheme. Let n_i be the number of HU's that contribute exactly i interview months in the year (i = 1, ..., 12) and let

$$X_{ij} = \begin{cases} 1 \text{ if the } j^{th} \text{ HU that contributes exactly i interview months is} \\ \quad \text{crime-free} \\ 0 \text{ otherwise} \end{cases}$$

for j = 1, ..., n_i and i = 1, ..., 12 . Suppose that the X_{ij} are independent random variables and that X_{i+} has a Bernoulli distribution with parameter $f(i,\theta)$ where $f(i,\theta)$ is any function of i and θ. Thus $f(i,\theta)$ is the probability that an HU that contributes exactly i months of information will be crime-free in those i months. Note that the independence is between HU's and not between months within an HU.

Under these assumptions, X_{i+} (the number of HU's that contribute exactly i interview months and are crime-free) has a Binomial $[n_i,f(i,\theta)]$ distribution. Then, $\hat{\theta}_1$ can be written as

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{12} iX_{i+}}{\sum_{i=1}^{12} in_i} = \frac{\sum_{i=1}^{12} in_i\bar{X}_{i+}}{\sum_{i=1}^{12} in_i} .$$

where $\bar{X}_{i+} = X_{i+}/n_i$. Letting $a_i = n_i/N$ (where $N = \sum_{i=1}^{12} n_i$) we have

$$\hat{\theta}_1 = \frac{\sum_{i=1}^{12} ia_i\bar{X}_{i+}}{\sum_{i=1}^{12} ia_i} .$$

By letting N→∞ while the a_i remain constant, and using the strong law of large numbers, we get as the limit

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{12} i\alpha_i \bar{X}_{i+}}{\sum_{i=1}^{12} i\alpha_i} = \frac{\sum_{i=1}^{12} i\alpha_i f(i,\theta)}{\sum_{i=1}^{12} i\alpha_i} . \qquad (1)$$

If $\hat{\theta}_1$ is to be consistent, then $\hat{\theta}_1$ must converge in probability to θ. This, together with (1), implies that

$$\frac{\sum_{i=1}^{12} i\alpha_i f(i,\theta)}{\sum_{i=1}^{12} i\alpha_i} = \theta . \qquad (2)$$

We would like to use this estimator $(\hat{\theta}_1)$ for any $(\alpha_1, \alpha_2, ..., \alpha_{12})$ that the design might specify, or at least for any $(\alpha_1, \alpha_2, ..., \alpha_{12})$ in some neighborhood in the hyperplane given by $\sum_{i=1}^{12} \alpha_i = 1$. In other words, we do not want the values of $f(i,\theta)$, the probability that an HU that contributes exactly i months of data is crime-free in those i months, to depend on the design. The following lemma will help us to find values of $f(i,\theta)$ that satisfy (2) for any $(\alpha_1, \alpha_2, ..., \alpha_{12})$.

Lemma 1: Suppose that

$$\frac{\sum_{i=1}^{12} y_i\alpha_i}{\sum_{i=1}^{12} z_i\alpha_i} = \theta$$

for all $(\alpha_1, \alpha_2, ..., \alpha_{12})$ in some neighborhood on the hyperplane given by $\sum_{i=1}^{12} \alpha_i = 1$. Then $y_i/z_i = \theta$ for $i = 1, ..., 12$.

Proof: Rewriting θ in terms of $\alpha_1, ..., \alpha_{11}$, we get

$$\frac{\sum_{i=1}^{11} y_i\alpha_i + y_{12}(1 - \sum_{i=1}^{11} \alpha_i)}{\sum_{i=1}^{11} z_i\alpha_i + z_{12}(1 - \sum_{i=1}^{11} \alpha_i)} = \theta .$$

Letting ν represent the numerator of the above expression and δ the denominator and taking the derivative with respect to α_j, we get

$$\frac{\delta(y_j - y_{12}) - \nu(z_j - z_{12})}{\delta^2} = 0$$

$$\frac{y_j - y_{12}}{\delta} - \left(\frac{\nu}{\delta}\right) \frac{z_j - z_{12}}{\delta} = 0$$

$$\frac{(y_j - y_{12}) - \theta(z_j - z_{12})}{\delta} = 0$$

$$y_j - \theta z_j = y_{12} - \theta z_{12} .$$

Next, let $c = y_{12} - \theta z_{12}$. Thus $y_j - \theta z_j = c$ and

$$\frac{\sum_{i=1}^{12} (c + \theta z_i)\alpha_i}{\sum_{i=1}^{12} z_i\alpha_i} = \theta .$$

$$\frac{c + \theta \sum_{i=1}^{12} z_i\alpha_i}{\sum_{i=1}^{12} z_i\alpha_i} = \theta$$

$$\frac{c}{\sum_{i=1}^{12} z_i\alpha_i} = 0$$

$$c = 0$$

Therefore $y_i - \theta z_i = 0$ and hence $y_i/z_i = \theta$. ■

Applying Lemma 1 to expression (2) we find that

$$\frac{if(i,\theta)}{i} = \theta ,$$

or

$$f(i,\theta) = \theta \quad i = 1, ..., 12 .$$

Thus the only functions $f(i,\theta)$ which make $\hat{\theta}_1$ consistent for θ are $f(i,\theta) = \theta$. This implies that the probability that an HU that contributes exactly i months of information will be crime-free during those i months is constantly θ, regardless of i — not a very reasonable model! Thus we have shown that, if we are willing to assume that HU's are victimized independently, the only model under which $\hat{\theta}'_1$ is consistent is not sensible.

4. THE MODIFIED AD HOC ESTIMATOR

We can go through an argument similar to that in Section 3 to find the functions $f(i,\theta)$ which make Eddy, Fienberg, and Griffin's modified ad hoc estimator $(\hat{\theta}'_1)$ consistent for θ. In the modified version of the ad hoc estimator, HU's that were victimized are treated as though they had contributed 12 months of information since, regardless of what information would have been gathered in the 12 - i months for which they were not in the sample, we would still treat the HU as a victimized HU. Thus the modified version of $\hat{\theta}_1$ is

$$\hat{\theta}'_1 = \frac{\text{\# interview months in crime-free HU's}}{12(\text{\# victimized HU's}) + (\text{\# interview months in crime-free HU's})} .$$

Using the same notation and assumptions as in the previous section, we have

$$\hat{\theta}'_1 = \frac{\sum_{i=1}^{12} i\alpha_i\bar{X}_{i+}}{12[\sum_{i=1}^{12}\alpha_i(1-\bar{X}_{i+})] + \sum_{i=1}^{12} i\alpha_i\bar{X}_{i+}} .$$

Letting N→∞ as the α_i's remain constant, and using the strong law of large numbers, we get as the limit

$$\lim_{N\to\infty} \hat{\theta}'_1 = \frac{\sum_{i=1}^{12} i\alpha_i f(i,\theta)}{12\sum_{i=1}^{12}\alpha_i[1-f(i,\theta)] + \sum_{i=1}^{12} i\alpha_i f(i,\theta)} . \tag{3}$$

If $\hat{\theta}'_1$ is to be consistent for θ, then $\hat{\theta}'_1$ must converge in probability to θ. This, along with (3), implies that

$$\frac{\sum_{i=1}^{12} i\alpha_i f(i,\theta)}{\sum_{i=1}^{12}\{12[1-f(i,\theta)] + if(i,\theta)\}\alpha_i} = \theta$$

where $\sum_{i=1}^{12}\alpha_i = 1$. Using Lemma 1 of the previous section, we have

$$\frac{if(i,\theta)}{12[1-f(i,\theta)] + if(i,\theta)} = \theta .$$

Then solving for $f(i,\theta)$, we find

$$f(i,\theta) = \frac{12\theta}{(12-i)\theta + i} .$$

Thus $\hat{\theta}'_1$ is consistent for θ under the model where the X_{ij} are independent and X_{ij} has a Bernoulli distribution with parameter $12\theta/[(12-i)\theta + i]$. If the parameter takes any other form, $\hat{\theta}'_1$ will not be consistent under this independent Bernoulli model. From equation (3) we see that under this model $\hat{\theta}'_1$ is in fact strongly consistent for θ.

Thus, if we believe that $12\theta/[(12-i)\theta + i]$ is a reasonable form of the parameter, we have found a model under which $\hat{\theta}'_1$ is a reasonable estimator. We might now want to know if $\hat{\theta}'_1$ is also a maximum likelihood estimator under this model. The likelihood of θ given the data is

$$\begin{aligned} \text{lik}(\theta \mid \underset{\sim}{x}_{ij}) &= \prod_{i=1}^{12}\prod_{j=1}^{n_i}\left(\frac{12\theta}{(12-i)\theta + i}\right)^{x_{ij}}\left(\frac{i(1-\theta)}{(12-i)\theta + i}\right)^{1-x_{ij}} \\ &= \prod_{i=1}^{12}\left(\frac{12\theta}{(12-i)\theta + i}\right)^{x_{i+}}\left(\frac{i(1-\theta)}{(12-i)\theta + i}\right)^{n_i-x_{i+}} . \end{aligned}$$

Taking logarithms

$$\begin{aligned} \text{loglik}(\theta \mid \underset{\sim}{x}_{ij}) &= \sum_{i=1}^{12}\{x_{i+}\log 12\theta - x_{i+}\log[(12-i)\theta+i] + (n_i-x_{i+})\log[i(1-\theta)] \\ &\qquad - (n_i-x_{i+})\log[(12-i)\theta+i]\} \\ &= \sum_{i=1}^{12}\{x_{i+}\log 12\theta + (n_i-x_{i+})\log[i(1-\theta)] - n_i\log[(12-i)\theta+i]\} . \end{aligned}$$

Taking derivatives with respect to θ yields

$$\begin{aligned} &\sum_{i=1}^{12}\left[\frac{x_{i+}}{\theta} + (n_i-x_i)\left(\frac{-1}{1-\theta}\right) - n_i\,\frac{12 - i}{(12-i)\theta + i}\right] \\ &= \frac{x_{++}}{\theta} - \frac{N-x_{++}}{1-\theta} - \sum_{i=1}^{12} n_i\,\frac{12-i}{(12-i)\theta + i} \\ &= \frac{x_{++} - N\theta}{\theta(1-\theta)} - \sum_{i=1}^{12}\frac{n_i(12-i)}{(12-i)\theta + i} . \end{aligned}$$

Thus the MLE, $\hat{\theta}$, is specified by

$$\frac{x_{++} - N\hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = \sum_{i=1}^{12}\frac{n_i(12-i)}{(12-i)\hat{\theta} + i} . \tag{4}$$

Note that expression (4) depends on the data only through x_{++}, the total number of crime-free HU's, and not on the number of months of information that they contribute, that is, not on the x_{i+} as we might have suspected.

In general, we would have to solve for $\hat{\theta}$ using some iterative procedure, but in a few special cases, expression (4) can be solved explicitly as seen in the following examples.

Example 1: Suppose $n_1 = \ldots = n_{11} = 0$ and thus $n_{12} = N$. In this case we only observe data in full years. Then (4) becomes

$$\frac{x_{++} - N\hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = 0$$

and

$$\hat{\theta} = \frac{x_{++}}{N}.$$ ■

Example 2: Suppose $n_j = N$ and $n_i = 0$ for $i \neq j$. Note that this is the only case where we can expect $\hat{\theta}$ to equal $\hat{\theta}'_1$ since $\hat{\theta}'_1$ depends on the x_{i+} and $\hat{\theta}$ depends only on x_{++}. Rewriting expression (4) we get

$$\frac{x_{++} - N\hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = \frac{N(12-j)}{(12-j)\hat{\theta} + j}$$

or

$$x_{++}(12-j)\hat{\theta} + x_{++}j - N\hat{\theta}^2(12-j) - Nj\hat{\theta} = N(\hat{\theta} - \hat{\theta}^2)(12-j)$$

or

$$x_{++}(12-j)\hat{\theta} + x_{++}j = 12N\hat{\theta}$$

or

$$\hat{\theta} = \frac{jx_{++}}{12N - (12-j)x_{++}},$$

which is $\hat{\theta}'_1$. Thus $\hat{\theta}$ and $\hat{\theta}'_1$ agree in this special case. ■

Example 3: Suppose $n_6>0$, $n_{12}>0$, and $n_i=0$ for $i\neq 6$ or 12. Thus we observe full years and half years of data. The estimator that the Bureau of Justice Statistics (BJS) uses to produce estimates of θ (to be discussed in the following section) treats the data as though they were in this form. In this case, expression (4) becomes

$$\frac{x_{++} - N\hat{\theta}}{\hat{\theta}(1-\hat{\theta})} = \frac{6n_6}{6\hat{\theta} + 6}.$$

After some simple algebra, this equation can be rewritten as

$$(N-n_6)\hat{\theta}^2 + (N+n_6-x_{++})\hat{\theta} - x_{++} = 0,$$

or

$$n_{12}\hat{\theta}^2 + (n_{12}+2n_6-x_{++})\hat{\theta} - x_{++} = 0.$$

Thus

$$\hat{\theta} = \frac{-(n_{12}+2n_6-x_{++}) \pm \sqrt{(n_{12}+2n_6-x_{++})^2+4n_{12}x_{++}}}{2n_{12}}.$$

Note that:

1. The quantity under the square root is positive

2. Since $\sqrt{(n_{12}+2n_6-x_{++})^2 + 4n_{12}x_{++}}$ is greater than or equal to $n_{12}+2n_6-x_{++}$ and since $n_{12}+2n_6-x_{++}$ is positive, the numerator of $\hat{\theta}$ will be nonnegative iff we add rather than subtract at the ± sign. Since the denominator is always positive, $\hat{\theta}$ will be nonnegative if we replace the ± sign with a + sign. Thus
$$\hat{\theta} = \frac{-(n_{12}+2n_6-x_{++}) + \sqrt{(n_{12}+2n_6-x_{++})^2+4n_{12}x_{++}}}{2n_{12}}. \qquad (5)$$

3. When $x_{++}=0$, $\hat{\theta} = 0$ and when $x_{++} = n_6 + n_{12}$ (=N), $\hat{\theta} = 1$.

4. We can show that the first derivative of $\hat{\theta}$ is positive and thus $\hat{\theta}$ increases in x_{++}. Hence $\hat{\theta}$ as given by equation (5) is real-valued and lies in the interval [0,1].

5. Furthermore the second derivative of $\hat{\theta}$ is positive. Figure 1 shows $\hat{\theta}$ plotted as a function of x_{++}. This curve lies below the line $\hat{\theta} = x_{++}/N$. Thus $\hat{\theta}$ is always less than or equal to the observed proportion of crime-free HU's, a reasonable result since some of those HU's would have reported a victimization if we would have been able to obtain the full year's data for them. ■

Thus we see that the MLE of θ for the model of independence between HU's and $f(i,\theta) = 12\theta/[(12-i)\theta + i]$ is not in general equal to $\hat{\theta}'_1$. In some cases an explicit solution for $\hat{\theta}$ can be found. Otherwise it must be evaluated using some iterative procedure as described in the appendix.

The fit of this model can be tested by means of a χ^2 goodness-of-fit statistic. This statistic has been calculated for each of the years 1973-1975 from the 1% sample of HU's from the Reiss data described in Eddy, Fienberg, and Griffin(1981). The MLEs were evaluated using the program HYBRD1 from the MINPACK package (see Appendix) and the results are shown in Table 1. Only HU's for which there was no nonresponse during the year of interest were used in these calculations and these HU's were assumed to be correctly matched. We see that, since we have 10 degrees of freedom, the fit is fairly good for the data from each of the three years.

Table 1

Fit of model $f(i,\theta) = 12\theta/[(12-i)\theta + i]$

1973 $\hat{\theta}'_1 = 0.694$ $\hat{\theta} = 0.692$

months (i)	#HU's	#vict HU's	#cf HU's	$f(i,\theta)$	expected #cf HU's	residual	contribution to χ^2
1	19	2	17	.96429	18.322	-1.322	.09532
2	22	2	20	.93104	20.483	-0.483	.01139
3	15	0	15	.90001	13.500	1.500	.16663
4	19	3	16	.87098	16.549	-0.549	.01819
5	19	3	16	.84376	16.032	-0.032	.00006
6	29	7	22	.81820	23.728	-1.728	.12581
7	6	2	4	.79414	4.765	-0.765	.12276
8	4	1	3	.77145	3.086	-0.086	.00239
9	8	3	5	.75002	6.000	-1.000	.16672
10	5	2	3	.72975	3.649	-0.649	.11535
11	10	6	4	.71055	7.105	-3.105	1.35727
12	550	161	389	.69233	380.782	8.218	.17734

$\chi^2 = 2.3592$

1974 $\hat{\theta}'_1 = 0.685$ $\hat{\theta} = 0.682$

months (i)	#HU's	#vict HU's	#cf HU's	$f(i,\theta)$	expected #cf HU's	residual	contribution to χ^2
1	36	1	35	.96260	34.656	0.346	.00346
2	39	4	35	.92789	36.188	-1.188	.03899
3	30	7	23	.89560	26.868	-3.868	.55687
4	30	1	29	.86549	25.965	3.035	.35487
5	42	10	32	.83733	35.168	-3.168	.28533
6	41	8	33	.81094	33.249	-0.249	.00186
7	43	9	34	.78617	33.805	0.195	.00112
8	31	14	17	.76287	23.649	-6.649	1.86935
9	33	6	27	.74091	24.450	2.550	.26597
10	35	15	20	.72017	25.206	-5.206	1.07527
11	26	8	18	.70057	18.215	-0.215	.00253
12	451	129	322	.68201	307.584	14.416	.67561

$\chi^2 = 5.1312$

continued

1975 $\hat{\theta}'_1 = 0.685$ $\hat{\theta} = 0.716$

months (i)	#HU's	#vict HU's	#cf HU's	$f(i,\theta)$	expected #cf HU's	residual	contribution to χ^2
1	24	2	22	.96802	23.233	-1.233	.06539
2	40	2	38	.93801	37.521	0.479	.00611
3	30	5	25	.90983	27.295	-2.295	.19297
4	39	4	35	.88329	34.448	0.552	.00884
5	30	5	25	.85824	25.747	-0.747	.02169
6	37	5	32	.83458	30.880	1.120	.04065
7	21	4	17	.81219	17.056	-0.056	.00018
8	39	7	32	.79097	30.848	1.152	.04304
9	34	3	31	.77083	26.208	5.792	.87612
10	33	10	23	.75169	24.806	-1.806	.13144
11	29	6	23	.73347	21.271	1.792	.14058
12	452	132	320	.71612	323.688	-3.688	.04202

$\chi^2 = 1.5690$

5. THE BJS ESTIMATOR

The Bureau of Justice Statistics(1980,1981a,1982) has published estimates of the proportion of households touched by crime (i.e., $1-\theta$) for each of the years 1975-1981. To calculate these estimates, Alexander(1981) of the Bureau of the Census developed two ad hoc estimators whose form is similar to the estimators discussed above. Rates for 1980 were to be published before the end of March 1981, and thus, since it would be necessary to have the information from interviews through June of 1981 in order to calculate 1980 rates, data for HU's that would have been interviewed after January of 1981 were imputed from their corresponding 1980 interviews as shown in Figure 2. For instance, for an HU in panel 3, information collected at the March 1980 and September 1980 interviews would be used as the information for 1980 although those 2 interviews actually cover the period of September 1970 through August 1980. Note that, in this way, it is possible to estimate the proportion of HU's victimized in a year by considering only two records for each HU as opposed to the three records which would normally be used for HU's in panels 2 through 6.

In order to calculate the BJS estimates, each HU is first classified according to the number and type of noninterview. The Census Bureau (undated) separates household noninterviews into three types:

Type A 1) no one is at home in spite of repeated visits

2) the entire household is temporarily away during all of the interview period

3) the household refuses to give any information

4) the unit cannot be reached due to impassable roads

5) interview is not conducted due to a serious illness or death in the family

Type B 1) unit is a vacant regular housing unit

2) unit is vacant and used for storage

3) unit is occupied by persons usually residing elsewhere

4) unit unfit for habitation or to be demolished

5) unit under construction and not ready for occupancy

6) unit temporarily converted to business or storage

7) address identifies an unoccupied tent or trailer sight

8) permit granted, but construction not started

Type C 1) no address was listed on the sample line of the listing sheet

2) unit demolished by time of enumeration

3) house or trailer has been moved

4) unit converted to permanent business or storage

5) unit has been merged with another unit

The classifications of the HU's are then:

group a both records are interviews

group b only the first record is an interview--the second record is missing because the HU was rotated out of the sample, or the second interview is a type A noninterview.

group c only the second record is an interview--the first record is missing because the HU had not yet been rotated into the sample or the first record was a type A noninterview.

group d the first record is an interview, the second record is a type B or C noninterview

group e the first record is a type B or C noninterview, the second is an interview

group f neither record is an interview

From these groups, the following quantities are computed:

H_a = # of HU's in group a

H_b = # of HU's in group b

H_c = # of HU's in group c

H_d = # of HU's in group d

H_e = # of HU's in group e

C_a = # of HU's in group a that report at least 1 victimization in either interview

C_{a_1} = # of HU's in group a that report at least 1 victimization in the first interview

C_{a_2} = # of HU's in group a that report at least 1 victimization in the second interview

C_b = # of HU's in group b that report at least 1 victimization in the first interview

C_c = # of HU's in group c that report at least 1 victimization in the second interview

C_d = same as C_b but for group d

C_e = same as C_c but for group e

Instead of using actual counts in the above quantities, the Census Bureau uses the weights associated with the appropriate HU's. In the calculations and analysis presented here, the above quantities are treated as actual counts of HU's.

The BJS estimators are then given by

$$R_1 = \frac{2C_a + C_d + C_e + (C_b+C_c)\left(\frac{2C_a}{C_{a_1}+C_{a_2}}\right)}{2H_a + H_b + H_c + H_d + H_e}$$

and

$$R_2 = \frac{2C_a + C_d + C_e + C_b\left(\frac{C_a}{C_{a_1}}\right) + C_c\left(\frac{C_a}{C_{a_2}}\right)}{2H_a + H_b + H_c + H_d + H_e}$$

Each of these estimators scales the observed variables to account for the missing information for HU's in groups b and c (that is, for type A noninterviews and for HU's that are only in the sample for part of the year), but not for HU's in groups d and e (that is, for type B and C noninterviews). An example of the calculation of R_1 may help clarify the estimator. Consider the following hypothetical data:

	first record	second record	either
group a			
total HU's	$630=H_a$	$630=H_a$	$630=H_a$
victimized HU's	$90=C_{a_1}$	$80=C_{a_2}$	$134=C_a$
groups b and c			
total HU's	$120=H_b$	$130=H_c$	??? use $125 = \frac{120+130}{2} = \frac{H_b+H_c}{2}$
victimized HU's	$15=C_b$	$18=C_c$	??? use $26 = [15+18] \times \left[\frac{139}{90+80}\right]$ $= [C_b+C_c] \times \left[\frac{C_a}{C_{a_1}+C_{a_2}}\right]$
groups d and e			
total HU's	$30=H_d$	$20=H_e$	??? use $25 = \frac{30+20}{2} = \frac{H_d+H_e}{2}$
victimized HU's	$5=C_d$	$3=C_e$	??? use $4 = \frac{5+3}{2} = \frac{C_d+C_e}{2}$

Aggregating the above quantities yields

$$\text{total HU's} = 630 + 125 + 25 = 780$$
$$= H_a + ½(H_b + H_c + H_d + H_e)$$
$$\text{total victimized HU's} = 134 + 26 + 4 = 164$$
$$= C_a + [C_b+C_c] \times \left[\frac{C_a}{C_{a_1}+C_{a_2}}\right] + \frac{C_d+C_e}{2}$$

and R_1 is just the ratio of total victimized HU's to total HU's. Note that the sum of the victimized HU's in groups b and c has been scaled down by the factor $C_a/(C_{a_1}+C_{a_2})$ which is calculated from group a to account for HU's that would have reported a victimization at both interviews had both interviews taken place. The sum of the victimized HU's in groups d and e is divided by 2 to reflect the fact that these HU's are only contributing one half of the year's data. It is assumed that no victimizations occurred at these HU's during the period covered by the missing record since the noninterview was a type B or C and hence no one was living in the unit (at least at the scheduled interview time). Since

$$1 \geq \frac{C_a}{C_{a_1}+C_{a_2}} \geq \frac{\max(C_{a_1}+C_{a_2})}{C_{a_1} + C_{a_2}} \geq \frac{1}{2} ,$$

the scaling factor for groups b and c is greater than or equal to the scaling factor for groups d and e. Thus HU's that have type A noninterviews or are out of sample for part of the year are weighted more heavily in R_1 than HU's that have had a type B or C noninterview.

In order to apply a consistency argument similar to the one we used on the ad hoc estimators $\hat{\theta}_1$ and $\hat{\theta}'_1$ in Sections 3 and 4, we need to develop additional notation. Let

$$X_{a_1j} = \begin{cases} 1 & \text{if } j^{th} \text{ HU in group a reports a victimization at the first interview} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{a_2j} = \begin{cases} 1 & \text{if } j^{th} \text{ HU in group a reports a victimization at the second interview} \\ 0 & \text{otherwise} \end{cases}$$

$$X_{bj} = \begin{cases} 1 & \text{if } j^{th} \text{ HU in group b reports a victimization at the first interview} \\ 0 & \text{otherwise} \end{cases}$$

and define X_{cj}, X_{dj}, X_{ej} similarly. Then the quantities used to calculate R_1 are

$$C_a = \sum_{j=1}^{Ha} I(X_{a_1j} + X_{a_2j} > 0)$$

$$C_{a_1} = \sum_{j=1}^{Ha} X_{a_1j} \qquad C_{a_2} = \sum_{j=1}^{Ha} X_{a_2j}$$

$$C_b = \sum_{j=1}^{Hb} X_{bj}$$

$$C_c = \sum_{j=1}^{Hc} X_{cj}$$

$$C_d = \sum_{j=1}^{Hd} X_{dj}$$

$$C_e = \sum_{j=1}^{He} X_{ej}$$

R_1 is then given by

$$R_1 = \frac{2\sum_{j=1}^{Ha} I(X_{a_1j}+X_{a_2j}>0) + \sum_{j=1}^{Hd} X_{dj} + \sum_{j=1}^{He} X_{ej} + \left(\sum_{j=1}^{Hb} X_{bj} + \sum_{j=1}^{Hc} X_{cj}\right)\left[\frac{2\sum_{j=1}^{Ha} I(X_{a_1j}+X_{a_2j}>0)}{\sum_{j=1}^{Ha} X_{a_1j} + \sum_{j=1}^{Ha} X_{a_2j}}\right]}{2H_a + H_b + H_c + H_d + H_e}$$

For each i = b, c, d, e, suppose that

$$X_{ij} \sim \text{iid Bernoulli}[g(i,\theta], \ i = 1, \ldots, H_i.$$

Thus $g(i,\theta)$ is the probability that an HU in group i reports a victimization. Suppose also that X_{a_1j} and X_{a_2j} have the following marginal distributions

$$X_{a_1j} \sim \text{iid Bernoulli}[g_1(a,\theta)], \ j=1, \ldots, H_a$$

and

$$X_{a_2j} \sim \text{iid Bernoulli}[g_2(a,\theta)], \ j=1, \ldots, H_a .$$

Note that the joint distribution of X_{a_1j} and X_{a_2j} is then

$$\begin{array}{cc|c|c|c} & & \multicolumn{2}{c}{x_{a_2j}} & \\ & & 0 & 1 & \\ \hline x_{a_1j} & 0 & \theta & 1-g_1-\theta & 1-g_1 \\ \hline & 1 & 1-g_2-\theta & g_1+g_2+\theta-1 & g_1 \\ \hline & & 1-g_2 & g_2 & \end{array} \qquad (6)$$

Letting $N = 2H_a + H_b + H_c + H_d + H_e$, we can rewrite R_1 as

$$R_1 = \frac{2H_a}{N}\left(\frac{\sum_{j=1}^{Ha} I(X_{a_1j}+X_{a_2j}>0)}{H_a}\right) + \frac{H_d}{N}\left(\frac{\sum_{j=1}^{Hd} X_{dj}}{H_d}\right) + \frac{H_e}{N}\left(\frac{\sum_{j=1}^{He} X_{ej}}{H_e}\right)$$

$$+ \left[\frac{H_b}{N}\left(\frac{\sum_{j=1}^{Hb} X_{bj}}{H_b}\right) + \frac{H_c}{N}\left(\frac{\sum_{j=1}^{Hc} X_{cj}}{H_c}\right)\right]\left[\frac{2\{\sum_{j=1}^{Ha} I(X_{a_1j}+X_{a_2j}>0)\}/H_a}{\sum_{j=1}^{Ha} X_{a_1j}/H_a + \sum_{j=1}^{Ha} X_{a_2j}/H_a}\right].$$

Let $\alpha_a = 2H_a/N$ and $\alpha_i = H_i/N$ for i = b, c, d, e. Then $\sum_{i=a}^{e} \alpha_i = 1$ and if we hold the α_i's constant as N tends to ∞ we find that

$$\lim_{N\to\infty} R_1 = \alpha_a(1-\theta) + \alpha_d g(d,\theta) + \alpha_e g(e,\theta)$$

$$+ [\alpha_b g(b,\theta) + \alpha_c g(c,\theta)]\left[\frac{2(1-\theta)}{g_1(a,\theta) + g_2(a,\theta)}\right]. \qquad (7)$$

In order for R_1 to be consistent, expression (7) must equal $1-\theta$. To find the forms of $g(i,\theta)$, i=a, ..., e for which (7) is equal to $1-\theta$, we need the following lemma.

Lemma 2: If $\sum_{i=a}^{e} y_i\alpha_i = 1-\theta$ for all $(\alpha_a, \ldots, \alpha_e)$ in some neighborhood in the hyperplane specified by $\sum_{i=a}^{e} \alpha_i = 1$, then $y_i = 1-\theta$, i=a, ..., e.

Proof: Rewriting the expression in terms of $\alpha_1, \ldots, \alpha_{1,1}$ we find that

$$\sum_{i=a}^{d} y_i\alpha_i + y_e(1 - \sum_{i=a}^{d} \alpha_i) = 1 - \theta .$$

Taking derivatives with respect to α_k yields

$$y_k + y_e(-1) = 0$$

$$y_k = y_e .$$

Thus $y_a = y_b = \ldots = y_e = y$, and

$$\sum_{i=a}^{e} y\alpha_i = 1 - \theta$$

$$y\sum_{i=a}^{e} \alpha_i = 1 - \theta$$

$$y = 1 - \theta . \quad ■$$

Applying Lemma 2 to equation (7), we have

i) $g(d,\theta) = 1 - \theta,$

ii) $g(e,\theta) = 1 - \theta$,

iii) $\frac{2g(b,\theta)(1-\theta)}{g_1(a,\theta) + g_2(a,\theta)} = 1 - \theta$, and thus $g(b,\theta) = \frac{g_1(a,\theta) + g_2(a,\theta)}{2}$,

iv) $g(c,\theta) = \frac{g_1(a,\theta) + g_2(a,\theta)}{2}$.

The restrictions on $g(d,\theta)$ and $g(e,\theta)$ are clearly unreasonable at least from an intuitive point of view. The probability that an HU that is interviewed once and is then demolished reports a victimization at that interview should not be the same as the probability that an HU that is interviewed twice reports a victimization at either of those interviews.

Referring to the hypothetical example, notice that 134/630 seems to be a reasonable estimate of $1 - \theta$ and so does 26/125 . But 4/25 = $(C_d+C_e)/(H_d+H_e)$ seems to estimate the probability of being victimized in one half of the year and it is not being combined with the previous two quantities in a way that reflects this fact. The problem here is that we need to model the relation between the probability of being crime-free in half a year and the probability of being crime-free in a whole year (θ) if we are to be able to use the data from half a year in an estimate of θ.

In a similar analysis for R_2, we find that restrictions i) and ii) remain unchanged while iii) and iv) become $g(b,\theta) = g_1(a,\theta)$ and $g(c,\theta) = g_2(a,\theta)$, respectively. Thus in either case we have a restriction which is intuitively unreasonable. This same problem exists for the estimator RNEW suggested by Griffin(1981). In practice, this problem may not greatly affect the numerical results since the HU's in groups d and e are those which have had either a type B or C noninterview and there are relatively few of these. Still, the impact of restrictions i) and ii) should be carefully examined.

We also see from this analysis that, in the case of R_2, the probability of victimization in half a year for an HU that is in the sample for only part of the year or that has a type A noninterview (groups b and c) is considered to be the same as the probability that an HU that contributes a full year's information is victimized in half a year. We know that, due to rotation group biases, HU's that are in sample for the first time are more likely to report a victimization than those that have been in the sample longer. In addition, HU's that have had at least one type A noninterview seem to be more likely to report a victimization than HU's that have not. Thus

careful consideration should be given to the plausibility of restrictions iii) and iv).

Another problem may be the definition of θ. We have been taking it to be the probability of an HU being crime-free in a given year. That is, it is the probability that an HU drawn at random from the population is crime free in a particular year. This definition is used explicitly in equation (6). Alternatively, θ might have been considered to be the probability that an HU drawn at random from the HU's contributing information for a particular year reports a victimization as having occurred in that year. In this case, θ would be something like

$$\prod_{i=1}^{12} P(\text{HU reports a victimization} \mid \text{HU in survey } i \text{ months}) \times \alpha_i$$

where α_i is the proportion of HU's contributing i months of information, and an analysis different from the preceding one would be necessary. The former definition seems to be more intuitive and easily understood, but in any case it is necessary to be explicit about the definition of θ before we can discuss the advantages and disadvantages of estimators of θ.

6. SUMMARY

Several of the previously proposed estimators of the proportion of HU's victimized in a given year have been studied and models under which these estimators are consistent have been derived. Some of these models require that the probability of an HU reporting a victimization be independent of the number of months of information that the HU contributes. We have seen, in these cases, the need to model the relation between the probability of being crime-free in any fraction of the year and θ, the probability of being crime-free in the entire year. The model under which the modified version of the ad hoc estimator is consistent does not require the probability of reporting a victimization to be independent of time in sample. A 1% sample of the data seems to fit this model fairly well.

APPENDIX

Evaluating the maximum likelihood estimator, $\hat{\theta}$, specified by equation (4) requires the use of an iterative procedure. The MINPACK package, written by B. S. Garbow, K. E. Hillstrom, and J. J. More of the Argonne National Laboratory, includes several programs to find a zero of a system of N non-linear equations in N variables by a modification of the Powell hybrid method. The HYBRD1 program, which estimates the Jacobian by a forward-difference approximation, was used to obtain the three values of θ given in Table 1. The program took about 0.5 seconds of CPU time on a DEC20 to compute all three of these values. Although we could have used the HYBRJ1 program, in which the user specifies the Jacobian, the fast convergence of the HYBRD1 program indicated that the estimated Jacobian was adequate.

The log likelihood for the 1973 data is plotted in Figure 3 as a function of θ. It is clearly unimodal and so the result of the iterative procedure is the global maximum. The plot seems to be rather flat near the maximum, possibly indicating a large variance for $\hat{\theta}$, but the log likelihood can be misleading in this respect. By adding the value of the log likelihood at $\hat{\theta}$ to the log likelihood and then exponentiating, we can compute a multiple of the likelihood of θ. (Note that by simply exponentiating the log likelihood we would have numbers that were too small to be handled by the computer.) This multiple of the likelihood is plotted as a function of θ in Figure 4. We see that the likelihood actually has a fairly sharp peak despite the apparent flatness of the log likelihood.

The log likelihood and likelihood for the data for each of the years 1974 and 1975 are very similar to those for the 1973 data.

REFERENCES

Alexander, C.H. (1981) "Plans to produce National Crime Survey Estimates of 'residences touched by crime'". Unpublished Census Bureau Memorandum.

Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*, London: Chapman and Hall.

Eddy, W.F., Fienberg, S.E., Griffin, D., and Trader, D. (1981). "NCS Data Files Available at CMU." Working Memorandum NCS-1, Department of Statistics, Carnegie-Mellon University.

Eddy, W.F., Fienberg, S.E., Griffin, D.L. (1981). "Estimating Victimization Prevalence in a Rotating Panel Survey." Bulletin of the International Statistics Institute, 42nd Session.

Eddy, W.F., Fienberg, S.E., Griffin, D.L. (1982). "Longitudinal Models, Missing Data, and the Estimation of Victimization Prevalence." In *Quantitative Criminology; Innovations and Applications* (J, Hagan, ed.), Sage Research Progress Series in Criminology.

Fienberg, S.E. (1978). "Victimization and the National Crime Survey: Problems of Design and Analysis." *Survey Sampling and Measurement* (N.K. Namboodiri, ed.), New York: Academic Press, 89-106.

Fienberg, S.E. (1980a). "Statistical modelling in the analysis of repeat victimization." *Indicators of Crime and Criminal Justice: Quantitative Studies* (S. Fienberg and A. Reiss, Jr., eds.), Washington, D.C.: U.S. Government Printing Office, 54-58.

Fienberg, S.E. (1980b). "The measurement of crime victimization: prospects for panel analysis of a panel survey." *The Statistician 29*, 313-350

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

Reiss, A.L. Jr. (1979). "Population dynamics and attrition bias in panel surveys." Working Paper, Institution for Social and Policy Studies, Yale University, December.

Reiss, A.J. Jr. (1980). "Victim proneness by type of crime in repeat victimization." *Indicators of Crime and Criminal Justice: Quantitative Studies.* Washington, D.C.: U.S. Government Printing Office, 41-53.

U.S. Department of Commerce, Bureau of the Census (undated). *National Crime Survey, National Sample, Survey Documentation.*

U.S. Department of Justice, Bureau of Justice Statistics(1981a). "The Prevalence of

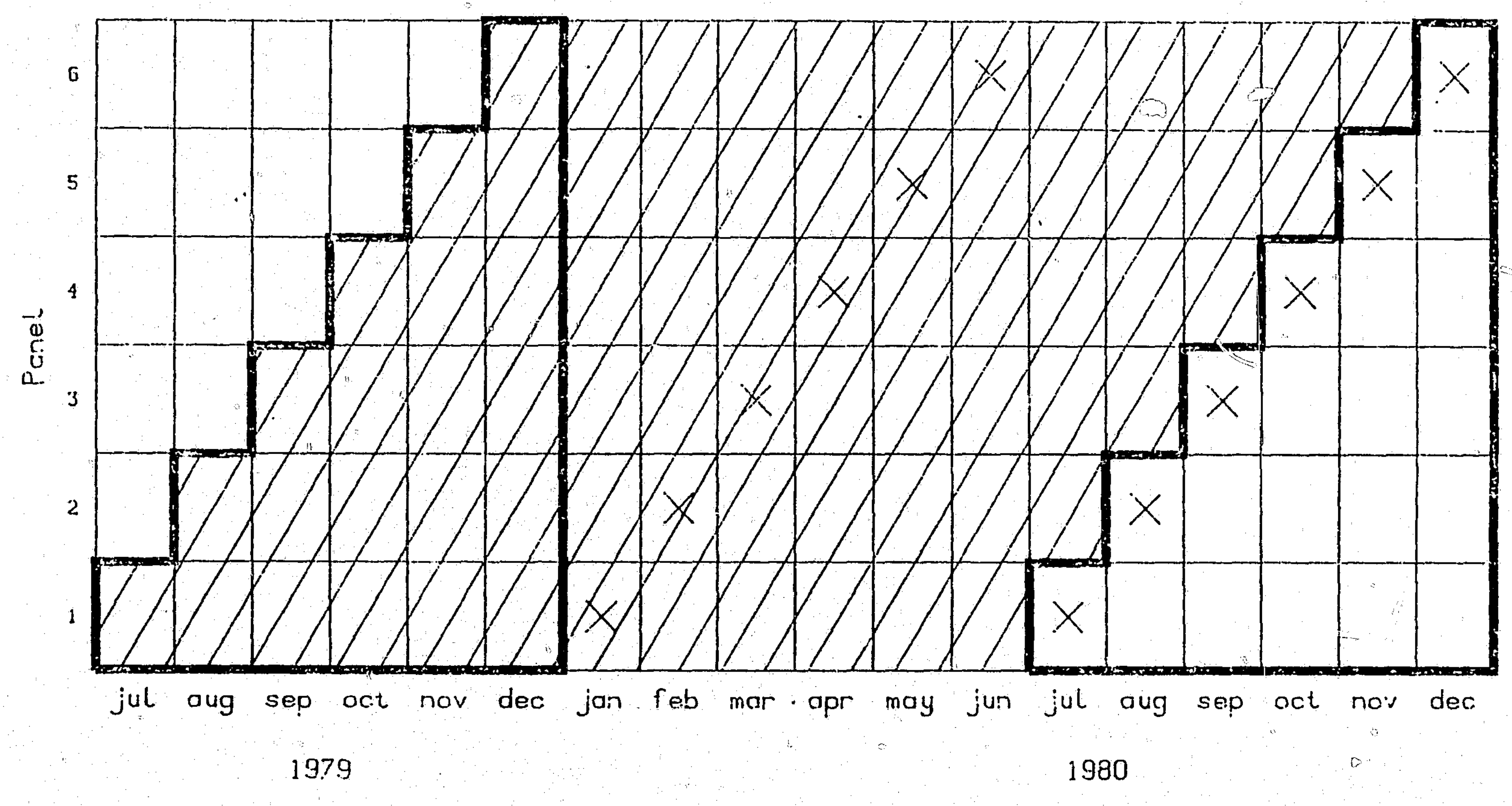
Crime."*Bureau of Justice Statistics Bulletin.* Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Justice, Bureau of Justice Statistics(1981b). *Criminal Victimization in the United States, 1979.* Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Justice, Bureau of Justice Statistics(1981a). "The Prevalence of Crime."*Bureau of Justice Statistics Bulletin.* Washington, D.C.: U.S. Government Printing Office.

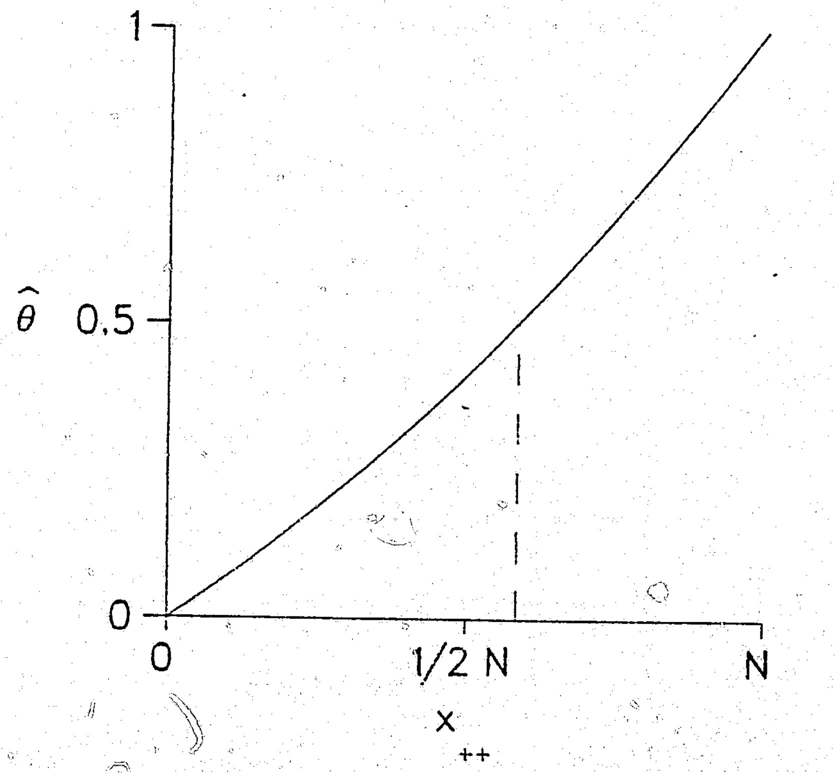
U.S. Department of Justice, Bureau of Justice Statistics(1982). "Households Touched by Crime 1981."*Bureau of Justice Statistics Bulletin.* Washington, D.C.: U.S. Government Printing Office.

Figure 2



X indicates the month of interview.
Shaded area indicates months covered by those interviews.
Data for the interview months in the bordered section on the loft
are used as the values of the unavailable data for the interview
months in the bordered section on the right.

Figure 1



note that $\hat{\theta}$ = 1/2 when x_{++} = 1/2 N + 1/6 n_6

Figure 3

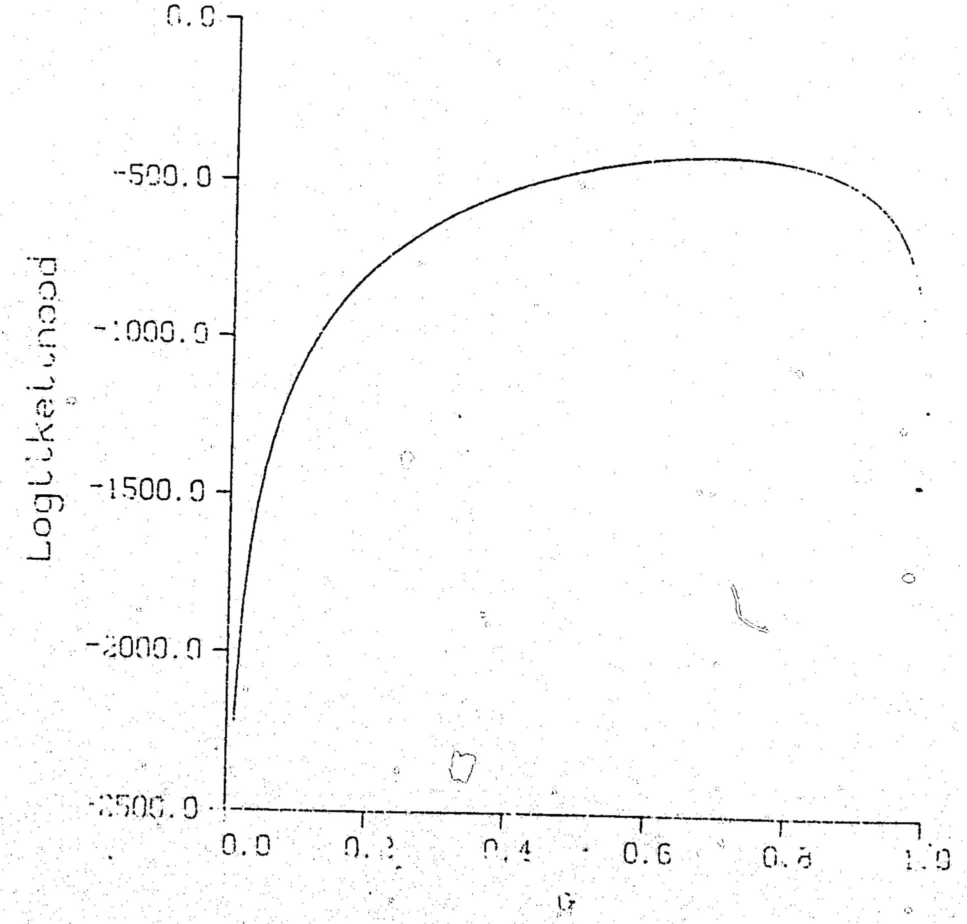
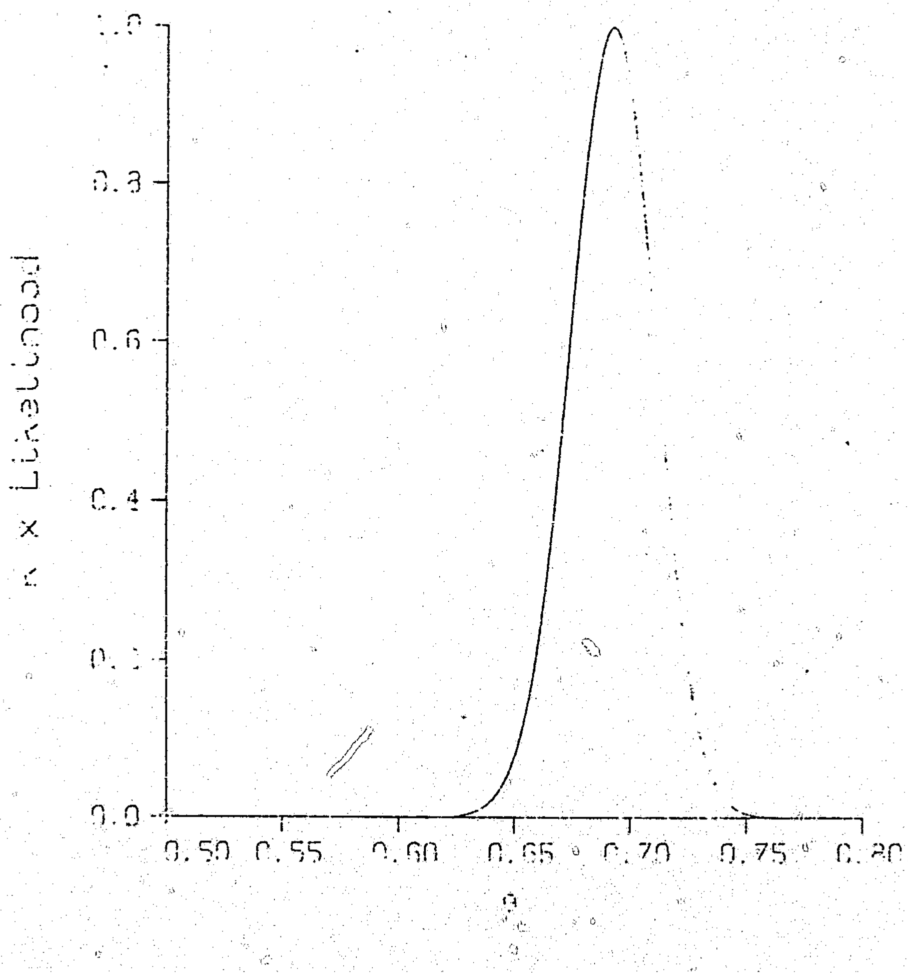


Figure 4



END