Estimation of offending and co-offending using available data with model support

Ove Frank[1]
Peter J. Carrington[2]

Submitted: February 21, 2003

Running head: Estimation of offending and co-offending

Corresponding author:

Peter J. Carrington
Department of Sociology
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

Phone: 519 888-4567 x3961
Fax: 519 743-1126
Email: pjc@UWaterloo.ca

Acknowledgment: Preparation of this paper was supported by Social Sciences and Humanities Research of Council of Canada Research Grant 410-2000-036

[1] Department of Statistics, Stockholm University, Stockholm, Sweden

[2] Department of Sociology, University of Waterloo, Waterloo, Canada

Estimation of offending and co-offending using available data with model support

Abstract

Police data under-report the numbers of incidents and of offenders, the numbers of offenders participating in individual criminal incidents (incident sizes) and the numbers of incidents in which individual offenders participate (offender activity). We show that co-offending is a concept which underlies, and unifies, all of these phenomena, so that the numbers of incidents and of offenders, and incident size distributions and offender activity distributions, can all be derived from the criminal participation matrix. Two related statistical models are presented which permit the estimation of numbers of incidents and offenders, incident size distributions, offender activity distributions, and co-offending distributions, from police-reported crime data, and data on the reporting of crime to police. These models rely on simple probability mechanisms for incident reporting and identification of offenders. The models are estimated, using data from the Canadian Uniform Crime Reporting Survey and national victimization surveys for the period 1995-2001. The implications of the results of fitting the models are discussed, as are other data which might be used.

Key words: models of incident reporting; models of offender identification; estimation of incident size; estimation of offender activity; co-offending; criminal participation matrix; Uniform Crime Reporting Survey.

Estimation of offending and co-offending using available data with model support

1 Introduction

1.1 Dark figures of crime

The term "dark figure of crime" usually refers to the amount by which the reported amount of crime, based on police records, underestimates the "real" amount of crime, because of criminal incidents that do not appear in police records. To some extent, the problem of this dark figure has been solved, or at least made more manageable, by the use of victimization surveys. For crimes that have victims, a survey of the victimization experience of a representative sample of the general population provides another estimate of the number of incidents per capita, or crime rate. In countries such as the USA and Britain, which have regular national victimization surveys that are fairly comprehensive with respect to types of crime, the victim-reported victimization rate has, to a considerable extent, supplanted the police-reported crime rate as the principal indicator of "the" national crime rate (Bureau of Justice Statistics, 2002; Kershaw et al., 2000).

Another dark figure arises in studies of criminal careers that use police data, such as the pioneering research of Wolfgang, Figlio, and Sellin (1972), and work which has followed in this tradition (Wolfgang et al. 1987; Tracy et al., 1990; Tracy and Kempf-Leonard, 1996). Here, the basic interest is in the distribution over offenders of numbers and types of crime, and their timing; that is, in the criminal activity of offenders. The practical difficulties of collecting self-reported data on criminal careers (and, particularly, on the timing of offenses) make the use of police data very attractive (Farrington, 1997, p. 367; Tracy and Kempf-Leonard, 1996, p. 75). However, police data on a person's criminal career have two types of omissions: unreported incidents in which the person was involved, and reported incidents for which the police were unaware of the person's involvement. Most research in criminal careers acknowledges these omissions, but then proceeds as though they were not important; e.g. Tracy and Kempf-Leonard (1996, p. 75): "We are well aware that multiple methods of crime measurement are preferable to single methods…[but] resources are not often available to allow multiple indicators…[therefore] this is a study of the official delinquent and criminal careers…". Attempts have been made to estimate from police-reported criminal careers to "real" criminal careers, using multipliers derived from self-reports of criminal activity (Cohen, 1986). Not much progress has been made.

A third dark figure—that of co-offending—has received practically no attention, or even acknowledgment. In research on co-offending, the basic interest is in the distribution over incidents of numbers and types of co-offenders. Although studies of co-offending on a small scale—for example, research on crime by "youth gangs"—often use self-report or even observational data (Bursik and Grasmick, 1993), attempts to estimate the amount and characteristics of co-offending in larger domains, such as cities, states or nations, usually rely on police data (Sarnecki, 1986; 1990; 2001; Reiss, 1988; Carrington, 2002). However, police data omit, to an unknown extent, some co-offenders, who are not identified, in reported incidents. This omission is in addition to incidents which are omitted in their entirety (the first dark figure).

These three dark figures are generally addressed in isolation from one another, to the extent that they are addressed at all. In this paper, we show how the three estimation problems can be solved simultaneously, using simple probabilistic models of police crime data, in conjunction with a *criminal participation matrix*: a matrix whose rows represent individual

criminal incidents and whose columns represent individual offenders. We begin by showing how police data on criminal incidents, offender activity, and co-offending can be conceptualized in a criminal participation matrix. Next, we describe our probabilistic models of police data. Then, we introduce our police data, and the resulting estimates. Finally, we comment on the kinds of data which might be used in this estimation, the implications of different models of police data, and some new kinds of research that might be done, using the criminal participation matrix.

1.2 Criminal size and activity statistics

Consider criminal incidents of a certain kind in a specified geographical region during a specified time period. Assume that m incidents were reported to the police, and that the police identified $x_1,\ldots,x_m$ offenders in these incidents. (For the sake of simplicity, we assume that reporting to the police implies recording, and reporting, by the police.) The frequencies of identified offenders are called the identified sizes of the criminal incidents. For some incidents no offenders might be identified but other incidents with at least one identified offender are called cleared by the police. The identified size distribution of the reported incidents is given by the frequencies m_i of reported incidents having i identified offenders for i=0,1,… . The proportion of cleared incidents is $(m-m_0)/m$ and their average identified size is

$$(m_1+2m_2+\ldots)/(m-m_0)=(x_1+x_2+\ldots+x_m)/(m-m_0).$$

The average identified size of reported incidents is given by

$$(m_1+2m_2+\ldots)/m=(x_1+x_2+\ldots+x_m)/m$$

and, obviously, the ratio between the averages of identified sizes among reported and cleared incidents equals the proportion of cleared incidents. Not all criminal incidents are reported to the police and not all offenders are identified by the police so it is not clear in what way these statistics give any information about the true size distribution of all reported and unreported incidents.

The same offender can be identified as a participant in different incidents, and it is of interest to know how many distinct offenders there are and how many incidents they are involved in. Assume that n different offenders were identified and that $y_1,\ldots,y_n$ are the numbers of incidents among the m reported incidents in which they were identified as participants. The identified activity distribution of the offenders is given by the frequencies n_j of offenders identified as participants in j reported incidents for j=1,2,… . The average identified activity is

$$(n_1+2n_2+\ldots)/n=(y_1+y_2+\ldots+y_n)/n.$$

However, true activity is not the same as identified activity, and therefore identified activity statistics like identified size statistics need to be handled with care and supported by appropriate model assumptions in order to be useful for estimating true size and activity distributions.

1.3 Participation records and co-offending statistics

The statistics about identified sizes $x_1,\ldots,x_m$ of m reported incidents and identified activities $y_1,\ldots,y_n$ of n different participating offenders can be considered as row and column summary statistics of an underlying identified participation matrix

$$Z=(Z_{uv}: u=1,\ldots,m; v=1,\ldots,n)$$

with m rows and n columns and elements Z_{uv} equal to 1 or 0 according to whether or not reported incident u involves the identified offender v. Thus

$$x_u=Z_{u1}+\ldots+Z_{un} \text{ and } y_v=Z_{1v}+\ldots+Z_{mv}.$$

The participation matrix Z contains more co-offending information than just the size and activity statistics. For instance, two matrices X=ZZ' and Y=Z'Z can be obtained that provide information about specific pair-wise co-offending statistics. In fact, X is an m by m matrix with elements that count the numbers of identified offenders participating in each pair of reported incidents, and Y is an n by n matrix with elements that count the numbers of reported incidents involving each pair of identified offenders. In particular, the main diagonals in matrices X and Y are given by the initial size and activity statistics, i.e. $X_{uu}=x_u$ and $Y_{vv}=y_v$. For any two distinct reported incidents u and v, the corresponding X-element is

$$X_{uv}=Z_{u1}Z_{v1}+\ldots+Z_{un}Z_{vn}$$

which counts the number of identified offenders participating both in incident u and incident v. Thus the sizes of the overlaps of the sets of identified co-offenders in any two reported incidents are given in the matrix X. For any two identified offenders u and v, the corresponding Y-element is

$$Y_{uv}=Z_{1u}Z_{1v}+\ldots+Z_{mu}Z_{mv}$$

which counts the number of reported incidents with both offender u and offender v identified as participants. Thus the identified joint activities or co-offending frequencies of any two offenders are given in the matrix Y.

Using matrix X, a graph can be drawn with vertices representing reported incidents and edges with attached frequencies showing how many identified offenders various pairs of reported incidents have in common. Similarly matrix Y yields a graph with vertices representing identified offenders and edges with attached frequencies showing how many reported incidents various pairs of identified offenders are known to have been involved in together. The first graph shows the identified joint participation relationship between reported incidents, while the second graph shows the known co-offending relationship between identified offenders. Both these relationships are induced by the identified participation relationship between reported incidents and identified offenders as given by the matrix Z.

1.4 Model support

The anticipated discrepancy between incidents and reported incidents and between participants and identified participants has an impact not only on size and activity statistics but on all kinds of co-offending statistics. Its impact on pair-wise co-offending statistics might even be expected to be more severe than on the size and activity statistics. If the reporting and identification processes can be appropriately modeled and incorporated into the statistical analysis of available data to draw inference about size and activity statistics, we could also hope to be able to use the same models to gain information about pair-wise and other aspects of the co-offending structure. The importance of such information will be discussed in the final section when we comment on the need and use of data on co-offending.

Frank (2001) introduced and analyzed a statistical model of co-offending without confronting it with real data. Here this model and a new companion model are used for estimation with available crime statistics on identified size and activity distributions for different types of crime. The two models used are supposed to describe an optimistic and a pessimistic approach to offender identification with the truth somewhere in between. Both models rely on a simple mechanism for incident reporting with no bias due to incident size or specific offender participation. Some of these assumptions can easily be relaxed but others are more difficult to dispense with. Our purpose is to show what results can be obtained and to discuss strengths and shortcomings of our approach. We also point to issues we think are important for future research and for future release of criminal statistics. The next section describes a basic model of co-offending with a simple mechanism for incident reporting and two different extreme mechanisms for offender identification.

2 Model descriptions

2.1 Criminal incidents and offenders

The criminal incidents of a certain type of crime occurring in a specified geographical region can be considered to be a stochastic point process in time. Let M be the number of incidents in a fixed period of time. Each incident involves a number of co-offenders that is assumed to vary between 1 and a finite upper bound a. Let $a_1,\ldots,a_M$ be the numbers of co-offenders involved in the M incidents. These numbers are called the sizes of the incidents. Let $M_1,\ldots,M_a$ be the frequencies of the possible sizes. If the criminal incident point process and the size distributions are stationary, the size distribution is estimated by the proportions $M_1/M,\ldots,M_a/M$ and the expected value and the variance of the size is estimated by using this distribution.

Let $A_1,\ldots,A_M$ denote the sets of co-offenders in the M incidents and let N denote the number of different offenders in the union of these sets. If the offenders are labeled by integers $1,\ldots,N$ it is possible to introduce a participation matrix

$$C=(C_{uv}: u=1,\ldots,M; v=1,\ldots,N)$$

with elements C_{uv} indicating whether or not incident u involves offender v as a participant. Let $B_1,\ldots,B_N$ denote the sets of incidents in which the N different offenders are participants. The numbers of incidents in these sets are denoted $b_1,\ldots,b_N$ and are called the activities of the offenders. Obviously u belongs to B_v if and only if v belongs to A_u, and this is indicated by $C_{uv}=1$ so that

$$a_u=C_{u1}+\ldots+C_{uN} \text{ and } b_v=C_{1v}+\ldots+C_{Mv}$$

for u=1,…,M and v=1,…,N. The sum of all incident sizes is equal to the sum of all offender activities, and this sum is the total number of participations denoted by

$$R=a_1+\ldots+a_M=b_1+\ldots+b_N.$$

The total number of participations can also be expressed in terms of the size frequencies M_i and the activity frequencies N_j according to

$$R=M_1+2M_2+\ldots=N_1+2N_2+\ldots.$$

The three numbers M, N, R are basic summaries of the participation matrix. The ratios R/M, R/N, and R/(MN) are the average numbers of participations per incident, per offender, and per incident and offender. The first two of these ratios represent the mean size of the incidents and the mean activity of the offenders. The ratio R/(MN) is a participation rate, which represents both the mean relative size of the incidents and the mean relative activity of the offenders. The ratios might be useful for temporal and regional comparisons.

2.2 Incident reporting to the police

The reasons why not all incidents are reported to the police are complex. A simple approach is to assume that no bias is introduced by considering the reporting mechanism for a certain type of crime to be satisfactorily described by uncontrolled randomness. This means that the selection of those incidents that are reported is governed by factors beyond our control, and we choose a simple stochastic model for it. Let S_u be an indicator that is 1 or 0 according to whether or not incident u is reported. We assume that there are no false reports of incidents. We also assume that $S_1,\ldots,S_M$ are independent with a common probability α of being equal to 1. This implies that the number of reported incidents is given by a random variable

$$m=S_1+\ldots+S_M$$

that is binomially distributed with parameters M and α. We write this as

$$m=\text{bin}(M,\alpha).$$

According to well-known properties of the binomial distribution, the deviation between the proportion m/M and the probability α is likely to be at most $1/\sqrt{M}$, so if M is estimated from other sources, α can be well estimated by m/M. Conversely, if α is estimated from other sources, M can be estimated to be within $(m\pm2\sqrt{m})/\alpha$.

2.3 Offender identification by the police

Offenders participating in reported incidents may be identified by the police. If an identified offender is also identified as a participant in another incident which was not previously reported, we consider it as being instantaneously reported. Thus by definition, identified involvement refers to reported incidents. We assume that no false identifications are made. If an offender is identified as a participant in a reported incident, this does not necessarily mean that any other, or all, co-offenders in this incident are also identified, or that the participation of this offender in other reported incidents is identified. The complicated mechanisms that underlie identification of offenders make it natural to try to set up probabilistic models for the identifications. Two such models are given in the next two sections.

2.4 Model 1: Co-offenders are independently identified

The participations of offenders in incidents are assumed to be identified independently and with equal chances for different reported incidents and for different offenders. The participation identifications can be indicated by

$$Z_{uv}=S_uS_{uv}C_{uv}$$

where S_u indicates that incident u is reported, S_{uv} indicates that offender v is identified as a participant in u, and C_{uv} indicates that this identification is correct. The indicators S_{uv} are assumed to be independent with a common probability β of being equal to 1. This means that the selection of those offenders that are identified as participating in an incident are made with equal chances for all offenders and all incidents. It follows that the identified number of co-offenders in incident u is given by

$$x_u=S_u(S_{u1}C_{u1}+\ldots+S_{uN}C_{uN})$$

which is a mixture of 0 and $\text{bin}(a_u,\beta)$ with mixing probabilities $1-\alpha$ and α. The identified sizes x_u are independent for the m reported incidents.

The identified activities of different offenders are given by

$$y_v=S_1S_{1v}C_{1v}+\ldots+S_MS_{Mv}C_{Mv}.$$

An identified activity $y_v=0$ means that offender v is not identified in any incident. The identified activities are $\text{bin}(b_v,\alpha\beta)$ and they are not independent. There are n identified offenders corresponding to $y_v>0$. The expected number of identified offenders is given by

$$En=\sum_v[1-(1-\alpha\beta)^{b(v)}]$$

where $b(v)=b_v$. This expected value is approximately equal to

$$\alpha\beta(b_1+\ldots+b_N)= \alpha\beta R,$$

where R is the total number of participations. Conditionally on the reported incidents, the identified activities are independent bin(c_v,β) where

$$c_v=S_1C_{1v}+\ldots+S_NC_{Nv}$$

is the true activity among reported incidents. Conditionally on the reported incidents the number n of identified offenders is therefore given as a sum of independent Bernoulli variables with probabilities

$$1-(1-\beta)^{c(v)}$$

for $c(v)=c_v$ and v=1,…,N. The sum can be approximated by a Poisson variable with mean

$$\beta(c_1+\ldots+c_N)=\beta(S_1a_1+\ldots+S_Ma_M).$$

This mean is distributed as βbin(R,α) and is therefore close in distribution to αβR. Thus n should be reasonably approximated by a variable that is Poisson(αβR).

2.5 Model 2: Co-offenders are identified together

In Model 1, we made the extreme assumption that the identification of one co-offender in an incident has no impact on the probability of the identification of any other co-offenders. This seems somewhat unrealistic, since identification of at least one co-offender implies (a) that the crime has been at least partially "solved"; that is, at least something is known about the circumstances of the incident; and (b) police now have access to one co-offender, who might be used a source of information about the others. Rather than trying to model by how much the identification of one co-offender might increase the probability of the identification of the others, we specify a second model, whose assumption concerning co-identification is at the opposite extreme. We then anticipate that the results obtained from the two extreme models would give us an indication of in what range the outcomes could vary during further specified conditions.

To define Model 2, Model 1 is modified deterministically in the following way. If at least one offender is identified as a participant in an incident, then all the co-offenders in this incident are also identified. This assumption implies that an incident u is identified as involving offender v if and only if u is reported, v participated in u, and at least one co-offender w who participated in u is identified. Formally the participation identifiers can be given as

$$Z_{uv}=S_uC_{uv}\max_wS_{uw}C_{uw}$$

where all indicators S_u and S_{uv} are defined as for Model 1. According to Model 2 the identified sizes are given by

$$x_u=S_ua_u\max_wS_{uw}C_{uw}$$

for the m reported incidents with $S_u=1$. Thus the identified size is equal to the size a_u with probability $1-(1-\beta)^{a(u)}$ where $a(u)=a_u$, and equal to 0 otherwise if $S_u=1$. The identified sizes are independent for the reported incidents.

The identified activities of different offenders are given by

$$y_v=Z_{1v}+\ldots+Z_{Mv}$$

which is a sum of $M-b_v$ zeros and b_v independent indicators that are 1 with probability

$$\alpha[1-(1-\beta)^{a(u)}]$$

for u in B_v. It follows that $y_v>0$ with probability

$$1-\Pi_{u\in B(v)}[1-\alpha+\alpha(1-\beta)^{a(u)}]$$

where $B(v)=B_v$, and this probability is approximately equal to

$$\alpha\beta\sum_{u\in B(v)}a_u = \alpha\beta(a_1C_{1v}+\ldots+a_MC_{Mv}).$$

Hence the expected number of identified offenders is approximately equal to

$$En=\sum_v P(y_v>0)=\alpha\beta(a_1^2+\ldots+a_M^2)=\alpha\beta\sum_i i^2M_i.$$

Now

$$\alpha\beta\sum_i i^2M_i=\alpha\beta[R+\sum_i i(i-1)M_i].$$

Unless no incident has more than one offender, this number is obviously considerably larger than the corresponding value

$$\alpha\beta R=\alpha\beta\sum_i iM_i$$

obtained according to Model 1. This reflects that the participation identification mechanism of Model 2 is likely to identify more offenders than that of Model 1. We note that the approximate expected values

$$\alpha\beta R \text{ and } \alpha\beta\sum_i i^2M_i$$

obtained with the two models can also be given as

$$\alpha\beta\sum\sum C_{uv} \text{ and } \alpha\beta\sum\sum B_{uv}$$

where B_{uv} are the elements in matrix $B=C´C$ counting joint activities of pairs of offenders.

2.6 Clearance rates

A reported incident is considered to be cleared by the police if at least one of the co-offenders is identified as a participant in that incident. According to Model 1 incident u is cleared if and only if $S_u \max_v S_{uv} C_{uv}=1$ which occurs with probability $\alpha[1-(1-\beta)^{a(u)}]$. Different incidents are independently cleared. The expected number of cleared incidents is

$$\alpha\sum_u[1-(1-\beta)^{a(u)}] = \alpha\sum_i M_i[1-(1-\beta)^i].$$

The expected number of cleared incidents divided by the expected number of reported incidents is the clearance rate, denoted γ, and according to Model 1 it equals

$$\gamma=\sum_i(M_i/M)[1-(1-\beta)^i]=1-\sum_i(M_i/M)(1-\beta)^i.$$

The clearance rate is simply estimated by the proportion of reported incidents that are cleared, that is by $(m-m_0)/m$ where m_0 is the number of reported incidents with no identified offender. This proportion is usually referred to as the clearance rate, or the empirical clearance rate if it is important to distinguish it from the theoretical clearance rate.

Model 2 has the same clearance indicators as Model 1, so the clearance rates are given by the same formula in the two cases. However, according to Model 2 all or no co-offenders are identified for each reported incident, and this affects the probabilities of identifying co-offending and the probabilities of identifying offenders. Consequently the clearance rate is the same for Model 1 and Model 2, but the probabilistic properties of its estimator are different due to the different identification mechanisms in the two models.

2.7 Reported incidents and identified offenders

The identified participation in incident u by offender v is indicated by Z_{uv} for u=1,…,M and v=1,…,N. The M by N matrix Z is the observed part of the participation matrix C. In matrix Z, the row total x_u is zero for any unreported incident but is also zero for reported incidents which are not cleared; i.e. for which no participants are identified. The column total y_v in matrix Z is zero if and only if the offender is not identified as a participant in any incident. The matrix Z could be, but is not necessarily, confined to the m reported incidents and the n identified offenders. The M incidents and the N offenders could be, but are not necessarily, relabeled so that the reported incidents are labeled 1,…,m and the identified offenders are labeled 1,…,n. Generally this is not assumed and we include the unknown numbers (dark figures) of M-m unreported incidents and N-n unidentified offenders in the formal treatment of Z. Special care is needed to distinguish between unreported and uncleared incidents, since they are not distinguished by their row totals.

2.8 Incidence rates and size distributions

The number of criminal incidents of a certain type in a specified region during a specified period of time is M. The incidence rate is the number of incidents per time unit. (We use this term rather than "crime rate", because "crime rate" is generally understood to be normalized by population.) When there is no need to use a particular time unit, such as year or five-year period, we let the time period used be the unit so that M is the incidence rate.

The size distribution is given by $(M_1,\ldots,M_a)$ in absolute frequencies and by $(M_1/M,\ldots,M_a/M)$ in relative frequencies. The relative frequencies together with the incidence rate provide a convenient specification of the incidents. The mean value and the standard deviation of the incident size distribution are denoted by μ_{size} and σ_{size}.

The identified size distribution $(m_0,\ldots,m_a)$ gives the frequencies of reported incidents with different numbers of identified co-offenders. The identified size distribution counts incidents u with $S_u=1$ and $x_u=i$ for i=0,…,a. Other incidents have $S_u=0$ and have frequency M-m. Thus we distinguish a+2 categories of incidents.

According to Model 1 the identified size distribution is given by a sum of identified size distributions calculated separately for each true size. Among the M_k incidents of size k, the identified size distribution is multinomially distributed with parameters M_k and $(p_{k0},\ldots,p_{ka})$ where p_{ki} is the probability that an incident of size k is reported and obtains i identified offenders. This probability is α times the probability that a bin(k,β)-variable is equal to i. We write the identified size distribution as

$$(m_0,\ldots,m_a)=\sum_{k=1,\ldots,a}\text{mult}(M_k,p_{k0},\ldots,p_{ka})$$

where

$$p_{ki}=\alpha f(i,k,\beta) \text{ and } f(i,k,\beta)=[k!/i!(k-i)!]\beta^i(1-\beta)^{k-i}.$$

It follows that the expected value of m_i is given by

$$Em_i= \sum_k M_k\alpha f(i,k,\beta) \text{ for } i=0,\ldots,a.$$

This equation system can be inverted to express the size frequencies as linear combinations of the expected identified size frequencies (See Frank (1980) for details.) We replace Em_i by m_i for i=0,…,a and interpret the corresponding solutions M_k for k=1,…,a as moment method estimators. This leads to the following equation system:

$$\alpha M_k=m_k f(k,k,1/\beta)+\ldots+m_a f(k,a,1/\beta) \text{ for } k=1,\ldots,a$$

$$\sum_{i=0,\ldots,a} m_i(1-1/\beta)^i=0.$$

The final equation can be used to estimate β and then the other equations give estimates of the relative size frequencies. It is not possible to separate α from the absolute frequencies though, so we don't get an estimate of the incidence rate M in this way. The equation to use for finding the

estimated β can be manipulated to show that it is in fact an alternating series in the m_i that should be set to zero:

$$\sum_{i=0,\ldots,a} m_i(-1)^i[(1-\beta)/\beta]^i = m_0 - m_1(1-\beta)/\beta + m_2(1-\beta)^2/\beta^2 - \ldots = 0.$$

We also notice that the estimated relative frequencies are given by linear combinations of the identified relative frequencies for equal or higher sizes:

$$M_k/M = \sum_{i=k,\ldots,a}(m_i/m)f(k,i,1/\beta) \text{ for } k=1,\ldots,a.$$

The formulas obtained can be used to derive estimates for the mean value and the standard deviation of incident size according to

$$\mu_{size} = (\gamma/\beta)\mu_{idsize}$$

$$\sigma_{size} = (\gamma^{1/2}/\beta)[\sigma_{idsize}^2 + (1-\gamma)\mu_{idsize}^2 - (1-\beta)\mu_{idsize}]^{1/2}$$

where μ_{idsize} and σ_{idsize} refer to the mean value and standard deviation of the distribution of identified size among the cleared incidents. The mean value and the standard deviation of the distribution of identified size among reported incidents, $\mu_{idsizerep}$ and $\sigma_{idsizerep}$, can be used instead of the statistics for cleared incidents since they are simply related according to

$$\mu_{idsizerep} = \gamma\mu_{idsize},$$

$$\sigma_{idsizerep} = [\gamma\sigma_{idsize}^2 + \gamma(1-\gamma)\mu_{idsize}^2]^{1/2},$$

$$\mu_{size} = \mu_{idsizerep}/\beta$$

$$\sigma_{size} = [\sigma_{idsizerep}^2 - (1-\beta)\mu_{idsizerep}]^{1/2}/\beta.$$

According to Model 2 the identified size distribution is given by a sum of identified size distributions calculated separately for each true size just as for Model 1. But now among the incidents of true size k, there is a trinomial identified size distribution distinguishing between the unreported incidents of size k and the reported incidents of identified sizes 0 and k. If we introduce notation m_{k0} for the number of uncleared reported incidents of size k, we can write

$$(m_{k0}, m_k) = \text{trin}(M_k, \alpha(1-\beta)^k, \alpha - \alpha(1-\beta)^k)$$

for k=1,…,a and these trinomial distributions are independent. It follows that

$$m_k = \text{bin}(M_k, \alpha - \alpha(1-\beta)^k)$$

for k=1,…,a and m_0 is a sum of independent

$$m_{k0} = \text{bin}(M_k, \alpha(1-\beta)^k).$$

Hence

$$Em_k = M_k\alpha[1-(1-\beta)^k]$$

for k=1,…,a and

$$Em_0=\sum_{k=1,\ldots,a}M_k\alpha(1-\beta)^k.$$

We see that α cannot be separated from M_k but the relative size distribution is easily obtained according to

$$M_k/M=m_k/[1-(1-\beta)^k]m$$

for k=1,…,a. The remaining equation yields β as the solution to the equation

$$\sum_{k=1,\ldots,a}m_k/[1-(1-\beta)^k]m=1.$$

If we use the approximation

$$1-(1-\beta)^k=\beta k$$

we get

$$M_k/M=m_k/m\beta k \text{ and } \beta=\sum_{k=1,\ldots,a}m_k/km.$$

The estimate of β is the inverted harmonic mean of the identified sizes of the cleared incidents times the empirical clearance rate. An upper bound to the harmonic mean is the arithmetic mean, and it follows that β is at least equal to the empirical clearance rate divided by the average identified size among cleared incidents or, equivalently, the squared empirical clearance rate divided by the average identified size among reported incidents. Thus

$$\beta\geq(m-m_0)^2/m(m_1+2m_2+\ldots+am_a).$$

According to Model 2, the estimated value of β is given by the empirical clearance rate divided by the harmonic mean of the identified sizes of the cleared incidents. Moreover, the number of incidents of size k is approximately proportional to the number of cleared incidents of size k divided by k. This result is considerably more explicit than the result obtained for Model 1 which requires that some equation system is solved.

The formulas for the mean value and standard deviation estimates under Model 2 are given by

$$\mu_{size}=\gamma/\beta \text{ and}$$

$$\sigma_{size}=[(\gamma/\beta)(\mu_{idsize}-\gamma/\beta)]^{1/2}=[\mu_{idsizerep}/\beta-(\gamma/\beta)^2]^{1/2}.$$

2.9 Number of offenders and activity distributions

According to Model 1 the identified activities $y_v=\text{bin}(b_v,\alpha\beta)$ are not independent for v=1,…,N. Conditional on the incident reporting indicators $S_1,\ldots,S_M$ they are independent $\text{bin}(c_v,\alpha\beta)$ where $c_v=\sum_u S_u C_{uv}$ is the true activity of v among reported incidents. Let n_j be the number of offenders of identified activity j for j=1,…,b where b is some specified finite upper limit. The identified activity distribution is given by $(n_1,\ldots,n_b)$. Conditional on the incident reporting, the identified activity distribution is a sum of b independent multinomial distributions corresponding to true activity 1,...,b among reported incidents. The expected value of n_j is

$$En_j=\sum_k N_k f(j,k,\alpha\beta)$$

for j=1,…,b, and by using the same technique as employed in the previous section we find the estimated true activity frequencies

$$N_k=\sum_j n_j f(k,j,1/\alpha\beta)$$

for k=1,…,b. In particular, the total number of offenders is estimated by

$$N=\sum_k N_k=\sum_j n_j[1-(1-1/\alpha\beta)^j].$$

This means that the dark figure N-n is estimated by an alternating series

$$n_1\theta-n_2\theta^2+n_3\theta^3-\ldots$$

where $\theta=(1-\alpha\beta)/\alpha\beta$. Thus the number of offenders and their activity distribution can be estimated if $\alpha\beta$ is provided. In the previous section we found that β can be estimated but for α things look less promising. We could apply a few possible alternatives for α and investigate the consequences for M, N, N_k.

The formulas obtained can be used to derive estimates for the mean value and the standard deviation of the activity distribution, μ_{act} and σ_{act}, from the mean value and standard deviation of the identified activity distribution according to

$$\mu_{act}=\delta\mu_{idact}/\alpha\beta,$$

$$\sigma_{act}=(\delta^{1/2}/\alpha\beta)[\sigma_{idact}{}^2+(1-\delta)\mu_{idact}{}^2-(1-\alpha\beta)\mu_{idact}]^{1/2},$$

where $\delta=n/N$ is the proportion of identified offenders. We note that an estimator of δ can be obtained from an estimate of N that requires estimation of $\alpha\beta$ as noted above.

According to Model 2 the identified activity y_v is given as a sum of independent Bernoulli variables for all the b_v incidents in which v is involved. If u is such an incident, then the corresponding Bernoulli variable has expected value $\alpha[1-(1-\beta)^{a(u)}]$. It follows that the probability that $y_v>0$ is given by

$$1-\Pi_u[1-\alpha+\alpha(1-\beta)^{a(u)}]^{C(u,v)}$$

and consequently the expected number of identified offenders is equal to

$$En=N-\Sigma_v\Pi_u[1-\alpha+\alpha(1-\beta)^{a(u)}]^{C(u,v)} \text{ where } C(u,v)=C_{uv}.$$

Approximation leads to

$$En=N-\Sigma_v\Pi_u(1-\alpha\beta a_u)^{C(u,v)}=N-\Sigma_v(1-\alpha\beta\sum_u a_u C_{uv})=\alpha\beta\sum_u a_u{}^2=\alpha\beta\sum_i i^2 M_i.$$

By similar reasoning we get that

$$En_j=\sum_v P(y_v=j)=\sum_k N_k f(j,k,\alpha\beta R/M)$$

where the last equality follows by approximating $\sum_u a_u C_{uv}$ with kR/M when $b_v=k$. This is an approximation that requires fairly equal incident sizes. Assuming that to be the case, we have an equation system for $j=1,\ldots,b$ which can be solved by the same technique as applied for the first model. It follows that the number of offenders of activity k can be estimated by

$$N_k=\sum_j n_j f(k,j,M/\alpha\beta R)$$

for $k=1,\ldots,b$. The total number of offenders is estimated by

$$N=\sum_k N_k=\sum_j n_j[1-(1-M/\alpha\beta R)^j].$$

Just as for Model 1 we notice that Model 2 allows us to estimate β and the relative size distribution of the incidents but the total number of incidents and the numbers of offenders of different activities require α to be estimated by other means. If α is provided, then we get estimates of the number of incidents of different sizes as well as estimates of the number of offenders of different activities. In particular, for both the models the total number of offenders is estimated by an alternating series in the frequencies of identified offenders of different identified activities. These frequencies are weighted differently according to Model 1 and Model 2, and all the weights depend on the probability α of incident reporting. Under Model 2 the mean value and the standard deviation of the activity distribution are estimated according to

$$\mu_{act}=\gamma\delta m/\alpha\beta n,$$

$$\sigma_{act}=[\delta\sigma_{idact}{}^2+\delta\mu_{idact}{}^2-\delta(1-\alpha\gamma)\mu_{idact}-(\gamma^2\delta m/\beta n)^2]^{1/2}/\alpha\gamma.$$

We note that if γ^2/β can be approximated by r/m, then the formulas for the estimators μ_{act} and σ_{act} according to Model 2 can be obtained by substituting γ for β in the corresponding formulas according to Model 1. In particular, if β, γ, and r/m are close, then the estimators μ_{act} and σ_{act} agree for the two models.

3 Data

To illustrate the application of the model, we use Canadian data for 1995-2001 from the Uniform Crime Reports (UCR) and the national victimization survey. There are some problems

with these data, which are discussed in Section 5 below, where we also provide some comparisons with data from other countries.

3.1 Proportion of crimes reported to the police

The proportion α of crimes reported to the police was estimated from data from the Canadian national criminal victimization survey, which is a module of the General Social Survey (GSS), conducted by Statistics Canada. The victimization module is included in the GSS approximately every five years, and data are available for victimization occurring in 1988, 1993, and 1999 (Sacco and Johnson, 1990; Statistics Canada Housing, Family and Social Statistics Division, 1994; Besserer, 2001). We estimated an overall value for the period 1995-2001 by interpolating between the reported 1993 and 1999 values. Overall, 39% of victimizations were reported to the police in 1999, and 42% in 1993 (Besserer and Trainor, 2000), giving an estimate of 40% for the overall proportion of crimes reported to police for 1995-2001.

Estimation of the model is necessarily limited to crimes for which crime-reporting proportions are available; that is, for which the victimization survey provides data on victim reporting behavior. The 1993 and 1999 surveys were administered to respondents 15 years of age or older, and included seven types of crime: sexual assault, other assault, robbery, break and enter (burglary), vandalism, theft of personal property or household property, and theft of motor vehicles or parts. The issue of analyzing aggregated versus disaggregated types of crime is discussed in Section 5.3. Our purpose is not to try to go into details here but simply to illustrate that there seems to be justification for the inclusion of a random reporting rate in our models.

3.2 Numbers of reported and cleared incidents and identified incident size distribution

The data shown in Table 1 on numbers of incidents were taken from the annual published reports of the Uniform Crime Reporting (UCR) Survey, summed over the period 1995-2001 (Canadian Centre for Justice Statistics, 1996; 1997a; 1998; 1999; 2000; 2001; 2002). Some incidents were excluded from our analysis because they involved types of crime not included in the GSS victimization survey (see Section 3.1). These are mainly offenses with no known victim or a corporate victim, such as fraud, possession of stolen property, weapons possession, public order, morals, drinking-driving, and drug offenses. There were a total of about 19.2 million reported incidents from UCR and about 5.9 million were excluded because no corresponding victimization data were available. Out of the remaining 13.3 million reported incidents, there were about 3.5 million cleared and 9.8 million not cleared, giving a clearance rate of 0.26.

-- Table 1 about here --

Table 2 shows the numbers and proportions of cleared incidents having different numbers of identified offenders. This distribution is derived from custom tabulations of about 1.46 million cleared incidents from the Incident-Based UCR Survey ("UCR2") provided by the Canadian Centre for Justice Statistics. This survey has been in operation in Canada since 1988, but does not yet include all Canadian police forces. During 1995-2001, police forces accounting for approximately 42% of cleared incidents in Canada reported to the UCR2. The proportions of

cleared incidents of different identified sizes obtained from UCR2 can be multiplied by the clearance rate obtained from UCR to provide estimated proportions of reported incidents of different identified sizes. Such estimates are given in the last column of Table 2. The total number of identified participations is about 1.75 million, giving a mean identified size of 1.2 offenders per cleared incident and a standard deviation of 0.53. Using the estimated figures for reported incidents, we get a mean identified size of 0.31 offenders per reported incident with a standard deviation of 0.58. We note that the statistics for cleared and reported incidents are related according to the formulas given in Section 2.8.

-- Table 2 about here --

3.3 Identified offender activity distribution

Table 3 shows the numbers and proportions of offenders identified as participating in different numbers of incidents during the period. This distribution is derived from the same data as the distribution in Table 2, and, therefore, the total number of identified participations is the same. There are about 882 thousand identified offenders for the 1.46 million cleared incidents considered. The mean identified activity over the seven year period is 2.0 incidents per offender with a standard deviation of 2.7. There is a small number of offenders having very large identified activities. In order to reduce their effects on the statistics we can consider the relative frequencies rounded to two decimal places, which might be more representative for this kind of data. For the rounded figures, identified activity is 1, 2, …, 9 for 71, 14, 6, 3, 2, 1, 1, 1, 1 % of the offenders and we get a mean identified activity of 1.7 incidents per offender with a standard deviation of 1.5.

--Table 3 about here –

4 Results

4.1 Estimated incidence rate

Basic incidence rates of interest are the total number M of incidents for the period and region considered and the breakdown of this rate into the numbers M_1, M_2, … of incidents having size 1, 2, …, i.e. incidents involving a single offender, two co-offenders, and so forth. It is convenient to consider the rate M and the proportions M_1/M, M_2/M, … of incidents of different sizes. The size distribution is discussed in the next section. Here we confine ourselves to the estimation of M.

According to our models M can be estimated by m/α where m is the number of reported incidents and α is the probability that an incident is reported by the police. This probability is estimated by the proportion of reported incidents and could be set to $\alpha=0.40$, which might be reasonable according to victimization surveys. Hence M is estimated by 2.5m. Data in Table 1 yield a clearance rate of $(m-m_0)/m=0.26$ so that M is estimated by $M=2.5m=9.6(m-m_0)$, or almost ten times the number of cleared incidents.

4.2 Estimated size distribution

According to Model 1 the proportion of incidents of size k is estimated by

$$M_k/M=\sum_{i=k,\ldots,a}(m_i/m)f(k,i,1/\beta) \text{ for } k=1,\ldots,a.$$

This is a linear combination of the proportions of reported incidents of identified sizes k, k+1, … with coefficients that depend on the participation identification probability β. Probability β is obtained from the equation

$$\sum_{i=0,\ldots,a}m_i(1-1/\beta)^i=0.$$

With data from Table 2 this equation is

$$74-22\theta+3\theta^2-\theta^3=0,$$

where $(1-\beta)/\beta=\theta$. By numerical solution we find $\theta=3.2$ which implies $\beta=0.24$.

From Table 2, the proportions of reported incidents having identified sizes 0, 1, 2, 3 are 0.74, 0.22, 0.03, 0.01, and it follows that the proportions of all incidents having true sizes 1, 2, 3 are given by

$$M_1/M=0.22/0.24-2(0.03)(0.76/0.24^2)+3(0.01)(0.76^2/0.24^3)=1.4$$

$$M_2/M=0.03/0.24^2-3(0.01)(0.76/0.24^3)=-1.1$$

$$M_3/M=0.01/0.24^3=0.7.$$

We see that the estimated proportions are not always obtained as numbers between 0 and 1 even if their sum is always equal to 1. This is due to the inherent random variability or uncertainty, which in this case does not guarantee sufficiently good behavior of the estimators. It would of course be better to have estimators of smaller variances. The conclusion from the present numerical results should be that the size distribution estimators are too poor in this case. We might have to be content with restricting ourselves to an estimate of the average incident size and an estimate of the standard deviation of incident size. According to the formulas presented in Section 2.8 and the data in Section 3.2 we get

$$\mu_{size}=(\gamma/\beta)\mu_{idsize}=(0.26/0.24)1.2=1.3,$$

$$\sigma_{size}=(\gamma^{1/2}/\beta)[\sigma_{idsize}^2+(1-\gamma)\mu_{idsize}^2-(1-\beta)\mu_{idsize}]^{1/2}=1.3.$$

According to Model 2, β is estimated by

$$\beta=m_1/m+m_2/2m+m_3/3m+\ldots$$

and data in Table 2 give $\beta=0.24$, which is the same numerical value as that obtained for Model 1. We feel more confident with the estimate when the two models with different estimator procedures agree on it. With Model 2 the proportions of incidents of sizes 1, 2, 3 are estimated by

$$M_1/M=m_1/\beta m,$$

$$M_2/M=m_2/2\beta m,$$

$$M_3/M=m_3/3\beta m.$$

Thus, according to Model 2 the incidence rate M_k is simply proportional to the number of cleared incidents of identified size k divided by k so that

$$M_1/M=m_1/(m_1+m_2/2+m_3/3+\ldots)=0.924,$$

$$M_2/M=(m_2/2)/(m_1+m_2/2+m_3/3+\ldots)=0.063,$$

$$M_3/M=(m_3/3)/(m_1+m_2/2+m_3/3+\ldots)=0.013.$$

The average size is estimated by 0.924+2(0.063)+3(0.013)=1.1 in accordance with the formula provided in Section 2.8:

$$\mu_{size}=\gamma/\beta=0.26/0.24=1.1.$$

The standard deviation under Model 2 is given by

$$\sigma_{size}=[(\gamma/\beta)(\mu_{idsize}-\gamma/\beta)]^{1/2}=[\mu_{idsizerep}/\beta-(\gamma/\beta)^2]^{1/2}=0.3.$$

For comparison, according to Table 2, the proportions of cleared incidents of identified sizes 1, 2, 3, 4 are 0.85, 0.11, 0.03, 0.01, and the average identified size among cleared incidents is 0.84+2(0.12)+3(0.03)+4(0.01)=1.2. There seems to be good agreement between average identified size among cleared incidents and the estimates of average true size among all incidents provided by the two models. The identified size distribution and the estimated true size distribution are very different though, and the precision of the estimated proportions could be poor, as was found for Model 1 with data from Table 2. We feel more confident with the estimated size distribution under Model 2 which has the advantage of giving the same mean and standard deviation as those estimated from Model 1.

4.3 Estimated offender and participation frequencies

The number of offenders N counts different individuals, and the participation frequency R counts the different offenders once for each incident in which they participated. Thus N and R can be considered as offender frequencies obtained without and with regard to offender activity. According to Model 1, N is estimated by

$$N=n+n_1\theta-n_2\theta^2+n_3\theta^3-\ldots$$

where n_1, n_2, n_3, … are the numbers of identified offenders of identified activities 1, 2, 3, … and n is the total number of identified offenders. The number $\theta=(1-\alpha\beta)/\alpha\beta$ depends on the reporting rate α for incidents and the participation identification probability β. If we set $\alpha=0.40$ and

β=0.24 as in the previous section, then θ=9.4 and with data from Table 2 we get the number of offenders estimated by

$$N=n[1+0.34(9.4)-0.19(9.4)^2+0.10(9.4)^3-\ldots$$

This estimate is obviously very sensitive to the presence of even a few identified offenders of large identified activity. The estimator is extremely unstable and not fit for calculations in this case.

According to Model 2, a similar formula applies but now with $\theta=(M-R\alpha\beta)/R\alpha\beta$ which can be estimated by $\theta=[m-(m-m_0)\alpha]/(m-m_0)\alpha$=8.6. This implies the same difficulties as for Model 1.

In order to estimate the participation frequency R, Model 1 suggests $R=r/\alpha\beta$=10.4r where r is the identified participation frequency, and Model 2 suggests $R=(m-m_0)/\alpha\beta$=10.4$(m-m_0)$ where $m-m_0$ is the number of cleared incidents. For Model 2 there is also an alternative estimator given by $R=[r(m-m_0)/\alpha^2\beta\gamma]^{1/2}$=10 $[r(m-m_0)]^{1/2}$. For the data in Table 3, this yields an estimated participation frequency of about 18.2 million according to Model 1 and about 9.2 million or 16.0 million according to the two estimators for Model 2. The large discrepancy between the frequency estimates forces us to turn to relative frequencies. In the previous section the average participation frequency per incident, i.e. the mean size R/M was estimated by $r/m\beta$ according to Model 1 and by $(m-m_0)/m\beta$ according to Model 2, where β=0.24 in both the cases. From Table 2, r=0.22+0.03*2+0.01*3=0.31, so $r/m\beta$=0.31/0.24=1.3; and $(m-m_0)/m\beta$=0.26/0.24=1.1, so there is close agreement between the estimated average size according to the two models. Generally, $(m-m_0)/m\beta<r/m\beta$ so Model 2 always gives a smaller estimate than Model 1. This can be intuitively understood because Model 2 assumes that there are no unidentified co-offenders of an identified offender which tends to underestimate participation.

4.4 Estimated activity distribution

The number of offenders of activity k is estimated by

$$N_k=\sum_j n_j f(k,j,1+\theta)$$

for k=1, 2, …, and the total number of offenders is estimated by

$$N=n+n_1\theta-n_2\theta^2+n_3\theta^3-\ldots,$$

with

$$\theta=(1-\alpha\beta)/\alpha\beta$$

according to Model 1 and with

$$\theta=(M-R\alpha\beta)/R\alpha\beta=[m-(m-m_0)\alpha]/(m-m_0)\alpha$$

according to Model 2. For both the models θ is positive, usually much larger than 1, and with $\theta>1$ the estimators are very sensitive to the occurrence of large identified activities. It seems reasonable then to confine the estimation to the mean value and the standard deviation of the activities. The average activity R/N seems to be much harder to estimate than the average size R/M which is estimated by $r/m\beta$ according to Model 1 and by $(m-m_0)/m\beta$ according to Model 2 as discussed in the previous section. In order to get estimates of the average activity and the standard deviation of activity it is convenient to consider them as functions of the offender identification rate δ when N, and consequently $\delta=n/N$, cannot be reliably estimated. According to the formulas given in Section 2.9,

$$\mu_{act}=\delta\mu_{idact}/\alpha\beta,$$

$$\sigma_{act}=(\delta^{1/2}/\alpha\beta)[\sigma_{idact}^{2}+(1-\delta)\mu_{idact}^{2}-(1-\alpha\beta)\mu_{idact}]^{1/2},$$

under Model 1, and

$$\mu_{act}=\gamma\delta m/\alpha\beta n,$$

$$\sigma_{act}=[\delta\sigma_{idact}^{2}+\delta\mu_{idact}^{2}-\delta(1-\alpha\gamma)\mu_{idact}-(\gamma^{2}\delta m/\beta n)^{2}]^{1/2}/\alpha\gamma,$$

under Model 2, we see that mean activity is estimated as linear functions of δ under both the models, and the standard deviation is estimated as the square root of quadratic functions of δ under both the models. Plotting these functions might provide some information about what uncertainty we are facing. It could be argued that the mean activity should be larger than the mean identified activity since not all participations are identified. It could also be argued that the mean activity should be smaller than the mean identified activity since offenders of low activity are less likely to be identified than those with high activity. In fact, mean activity and mean identified activity are equal if and only if $\delta=\alpha\beta$ under Model 1 and $\delta=r\alpha\beta/m\gamma$ under Model 2. Taking this as a tentative suggestion we estimate mean activity by 1.7 with a standard deviation of 5.8 under Model 1 and estimate mean activity by 1.7 with a standard deviation of 6.1 under Model 2. Thus, during the seven year period considered in Tables 2 and 3 both the models estimate mean activity to 1.7 offences per offender with a standard deviation of about 6. This indicates a very skewed distribution, and it holds, for instance, that true activity varies between 1 and 13 for at least 75 % of the offenders and it is 14 or more for at most 25 % of the offenders.

5 Discussion

5.1 Sources of data

Four types of data are needed in order to estimate the models presented above: three from police records, and an additional one. The police-reported data are: the proportion of cleared incidents among the reported ones, the distribution of identified incident sizes among the cleared incidents, and the distribution of identified offender activities among the identified offenders. The fourth data element—the proportion of incidents reported to police—would normally come from a victimization survey.

One advantage of the Canadian data is the existence of a well-developed “national” incident-based Uniform Crime Reporting Survey (the UCR2), which is the source of our distributions of identified incident sizes and offender activities (Tables 2 and 3 above; Canadian Centre for Justice Statistics, 2001). This survey has considerable potential to support elaboration of the models, since it can provide incident size distributions broken down by the type of crime and circumstances of the incident, and the age and sex of the offender(s). The UCR2 does not yet include all Canadian police forces. In 1995, police forces accounting for approximately 43% of reported crime in Canada reported to the UCR2. By 2001, this had risen to 59%. We restricted our sample of police forces to those which reported continuously to the UCR2 during the period 1995-2001. This sample accounted for 42% of cleared incidents in Canada during that period. It

is heavily biased toward police forces in Quebec and Ontario, and is restricted to urban police forces in provinces other than Quebec.

The Canadian UCR2 survey uses a criterion for the inclusion of persons implicated in incidents, and therefore for the coding of incidents as "cleared", which is different from the criterion of "arrest" used in the American UCR Survey and the British register of Recorded Crime. The UCR2 includes persons who meet the criterion of a "charged suspect/chargeable," who is defined as "a person who has been identified by police as being involved in a criminal incident and against whom an information [i.e. a charge] could be laid as a result of sufficient evidence/information." This definition explicitly excludes "…individuals involved only for investigative purposes and subsequently released…" (Canadian Centre for Justice Statistics, 2001, p. 74).

This criterion is both more and less restrictive than the criterion of arrest. On the one hand, it excludes the many suspects who are arrested but later released because there is insufficient evidence to charge or otherwise officially accuse. On the other hand, it includes individuals who are never arrested, but against whom police have sufficient evidence to lay a charge. This includes situations in which the police use their discretion not to charge, and situations in which police do lay a charge, but use a means other than arrest and detention to compel attendance at court (e.g. summons or appearance notice). Research based on the UCR2 which analyzed the (identified) sizes of co-offending groups in Canada, and compared them with co-offending groups reported in other countries, suggests that the "identification" criterion used in the Canadian UCR2 may be more stringent on the whole – i.e. may include fewer suspects – than the "arrest" criterion used in the American and British official statistics (Carrington, 2002).

Numbers of reported incidents cleared and not cleared for 1995-2001 (Table 1) were obtained from the traditional ("aggregate") UCR Survey, which has approximately 100% coverage of Canadian police forces. Clearance rates by type of incident and year are shown in Table 4, with comparative data from the USA and Britain. (Clearance rates for "victimless" crimes are omitted, since they were omitted from estimation of our models.) There is considerable consistency over time and place, when the types of crime are combined, as they are in the estimation of our models.

-- Table 4 about here --

Comparative data on distributions of offenders' identified activities (the number of incidents in which an offender is identified by police as a participant) are much more difficult to obtain. Offender activity distributions are currently available only as the by-product of criminal careers research on selected birth cohorts, in which the researchers have trawled through police records, matching incidents involving the same offender. The only such Canadian research is the "Montreal study" (LeBlanc and Fréchette, 1989), for which no activity distributions have, to our knowledge, been published. That study is limited to a small sample of males in the city of Montreal. Some published activity distributions from well-known American and British research projects are shown in Table 5. These distributions are very similar.

-- Table 5 about here –

The distributions shown in Table 5 differ considerably from our data (Table 3): in our data, mean identified activity is much lower, and the proportion of offenders with only one identified participation is much greater. There could be several reasons for this discrepancy. One is that the criminal careers data in Table 5 are derived by following a cohort for a considerable number of years, concentrating on the age range during which individuals offend at the greatest rate; whereas, our data represent a cross-sectional window of seven years on an entire

population, including offenders of all ages, from 3 to 89 years old. Thus, many offenders in our data were too young or too old to be offending at a high rate. Second, our data include both male and female offenders; whereas the distributions in Table 5 are based on male offenders. Third, our data are taken from the Canadian UCR, which has a relatively high threshold criterion for inclusion of "identified offenders" (see above); whereas, the distributions in Table 5 are based on data coded by researchers from police files, in which inclusion of alleged offenders was subject to less stringent criteria, such as "arrest" or "police contact" (the exception is the 1955 London data, which used conviction as its criterion of inclusion).

Individuals' official annual offending rates tend to be low. Blumstein et al. (1986) report annual arrest frequencies of 1.09, 0.56, and 0.84, for three American samples of "active" offenders (Table 6). According to American and British data, an official criminal career generally lasts no more than 10 years, and most are in the range of 5-7 years (Blumstein et al., 1986, pp. 91-95; Tarling, 1993, p. 49). In our sample, the mean annual identified offender activity is 2/7 = 0.29 incidents (or 1.69/7=0.24 according to the rounded frequencies), which is considerably lower than the rates reported by Blumstein et al. Although, like our data, the arrest data cited by Blumstein et al. and by Cohen are derived from observation of a "window" period, nevertheless, cohorts of young adults, i.e. high-rate offenders, were selected from the overall population of offenders, and offenders were followed only during the period during which they were "active" (Cohen, 1986, pp. 325-329). Furthermore, the selected arrestees were "almost exclusively males" and "reasonably serious" offenders (Ibid.). In contrast, our estimate of 0.29 incidents per year represents the identified activity of offenders of all ages and both sexes, implicated in all types of crime except victimless crimes.

-- Table 6 about here --

The final data element used in estimation of the models is the proportion of all incidents that are known to the police. In using data from a victimization survey on public reporting of crime to the police, we make several unrealistic assumptions: that the public correctly identify all crime, with neither false positives nor false negatives, that they accurately report their crime reporting behavior to the victimization survey, and that the police faithfully record all public crime reports and transmit them to the UCR Survey. The obvious unreality of the first assumption in particular, concerning the omniscience of the public, requires the usual modification of the claimed scope of our estimates of crime: since our victimization data explicitly omit "victimless" crime and crime with corporate victims, we have, as far as possible, omitted them also from our data on identified incident size and offender activity distributions and clearance rates; therefore, our estimates of offending and co-offending exclude these types of crime, which constitute approximately 30% of recorded crime in Canada (see Table 1).

Our data on victim reporting behavior come from the Canadian victimization surveys of 1988, 1993, and 1999 (Sacco and Johnson, 1990; Statistics Canada Housing, Family and Social Statistics Division, 1994; Besserer, 2001). These surveys are nationally representative, and are therefore consistent with the domain of the UCR data. However, in addition to the exclusions noted above, the victimization surveys omit crime with human victims aged less than 15 years. We were unable to adjust the other data elements to be consistent with this. Thus, we make the additional assumption that the proportion of crimes reported to the police is the same for crimes with victims under 15 years old, as for crimes with victims aged 15 and older.

The estimate of 0.40, averaged over 1995-2001, for the proportion of crime reported to the police in Canada is similar to that reported by the (American) National Crime Victimization Surveys (0.37 for 1996 and 1999; Table 7), and somewhat lower than that reported by the British

Crime Survey (0.46 in 1995, 0.44 in 1999; Kershaw et al., 2001) and by the International Crime Victimisation Surveys of eleven industrialized countries (0.50 to 0.53; Table 7).

-- Table 7 about here –

5.2 Choice of social and temporal domain

Possible domains include the city, or metropolitan area, the region (e.g., state, or province in Canada), and the nation. We chose the nation as domain, primarily in order to maximize the number of identified incidents and offenders, in particular the very small numbers in the tail of the identified incident size distribution (Table 2).

There are disadvantages to using national data, versus data for a city. A city generally represents the jurisdiction of a single police force, making more tenable the model assumptions of equal probabilities of incident clearances and of identification of offenders. A city would also probably be more homogeneous than a state or nation with respect to the probability of reporting of an incident by the public. In our national data, the clearance rate, probability of reporting by the public, and identified incident size distribution, are all aggregated over heterogeneous units. In addition, a national criminal participation matrix will be unnecessarily sparse, since most offenders would be limited to co-offending with others in their own city; thus, the participation matrix would consist of relatively dense areas, representing intra-city co-offending, with large empty areas, representing (non-existent) inter-city co-offending. On the other hand, the use of regional or national data allows one to assess empirically the extent of inter-city offending and co-offending; whereas this is ruled out *a priori* by the use of data limited to a city.

Like the choice of social domain, our choice of a seven year time period for UCR data was motivated primarily by the desire to attain substantial numbers in the tails of the identified incident size and offender activity distributions. Obviously, the more years of data that are aggregated, the more reliable the small numbers in the tails of the distributions will be. We would have preferred to use more than seven years of data, but these were all that were available, due to the recency of the implementation of the UCR2. If incident clearance rates or public crime reporting propensities were trending over time, this would be a reason not to aggregate them over several years, but this does not appear to be the case in Canada (Tables 4 and 7).

5.3 Aggregation or disaggregation of types of crime

Considerations similar to those discussed above apply to the question of the aggregation of types of crime. Aggregation has the advantage of increasing cell frequencies, which is particularly important in view of the small numbers in the tails of the observed incident size and offender activity distributions. Disaggregation improves the precision of estimates by segregating the parameters of heterogeneous phenomena.

Table 7 (above) shows the propensity to report different types of crimes according to various surveys conducted in Canada and the USA. Table 8 shows clearance rates for the same, or similar, types of crime. In Tables 7 and 8, the rows have been arranged from high to low victim reporting rates. There are some obvious differences between types of crime and between countries and also some similarities. Reporting rates are relatively high, but clearance rates are low, for break and enter and theft motor vehicle. Sexual assault has a very low reporting rate, and a relatively high clearance rate. Robbery, vandalism, theft over, and assault have medium victim

reporting rates. The clearance rate for assaults is high, it is medium for robbery, and low for theft over and vandalism. Contact crimes (robbery, sexual assault, and assault) tend to have low to medium victim reporting rates and medium to high clearance rates. Non-contact (property) crimes tend to have medium to high reporting rates but low clearance rates.

-- Table 8 about here --

We were unable to find any published offense-specific police-reported incident size distributions, but Carrington (2002) reports mean sizes, based on Canadian UCR2 data, ranging from 1.05 (std. dev. = 0.96) for sexual assault to 1.50 (std. dev. = 3.37) for break and enter. We were unable to find any published offense-specific official offender activity distributions, but Blumstein et al. (1986) report annual arrest frequencies ranging from 0.13 for aggravated assault to 0.41 for larceny theft in the Philadelphia 1945 cohort (Table 7). Evidently, there is a strong argument from heterogeneity for estimating our models separately by type of crime.

We chose to use aggregated data for two reasons: to maximize frequencies in the tails of the incident size and offender activity distributions, and the unavailability of offense-specific identified offender activity distributions.

5.4 Implications for future research

Two innovations are proposed in this paper. One is the use of an explicit probability model to estimate indicators of crime such as the number of incidents and the distributions of offender activities and of co-offending, from police-reported data. The second is the use of the criminal participation matrix to conceptualize and operationalize criminal incidents, criminal careers, and co-offending, simultaneously.

Estimation from observed data with the support of two explicit probability models permits not only a more realistic assessment of the distribution of crime and co-offending in a social domain or an offender's career, but also allows us to set upper and lower bounds on the estimates, and to assess the variability of the estimates.

We have shown how mean values and standard deviations of the true size and activity distributions can be estimated. More detailed estimation of the frequencies of different sizes and activities turned out to be more difficult, partly because of the limitations of the available data, and also perhaps because of the overly simplistic assumptions of the models. Further elaboration of the models, and applications with other data, are needed.

The criminal participation matrix could be the basis for the integrated study of co-offending and criminal careers—a study which was advocated by Reiss (1986; 1988; Reiss and Farrington, 1991), but which has not yet been done because, we suggest, of the lack of an integrative conceptual model and data structure. If the rows (incidents) of the participation matrix are sorted on time, then the sequence of incidents, which is crucial to criminal careers research, would be preserved. This would permit the simultaneous analysis of the development over time of criminal careers and co-offending relationships and structures. Multiplying the transpose of the participation matrix by itself yields the offender-by-offender co-offending matrix, which could be analyzed by conventional social network methods (Wasserman and Faust, 1994). All of these analyses could be elaborated by variables such as the type of crime, gender of the offender, etc., by adding dimensions incorporating these variables to the basic two-dimensional participation matrix.

List of References

Besserer, S. (ed.). (2001). *A Profile of Criminal Victimization: Results of the 1999 General Social Survey*. Catalogue No. 85-553. Canadian Centre for Justice Statistics, Statistics Canada, Ottawa.

Besserer, S., and Trainor, C. (2000). *Criminal Victimization in Canada, 1999*. *Juristat* 20 (10). Catalogue No. 85-002. Canadian Centre for Justice Statistics, Statistics Canada, Ottawa.

Biderman, A. D., and Lynch, J. P. (1989). *Understanding Crime Incidence Statistics*, Springer-Verlag, New York.

Blumstein, A., Cohen, J., Roth, J.A., and Visher, C.A., eds. (1986). *Criminal Careers and "Career Criminals"*, Vol. I. National Academy Press, Washington, D.C.

Bureau of Justice Statistics. (1995). *Sourcebook of Criminal Justice Statistics 1994*, Bureau of Justice Statistics, U.S. Dept. of Justice, Washington, D.C.

Bureau of Justice Statistics. (1998). *Sourcebook of Criminal Justice Statistics 1997*, Bureau of Justice Statistics, U.S. Dept. of Justice, Washington, D.C.

Bureau of Justice Statistics. (2001). *National Criminal Victimization Survey*. Document NCJ 184938. At http://www.ojp.usdoj.gov/bjs/abstract/cvusst.htm. Accessed on Jan. 21, 2002.

Bureau of Justice Statistics. (2002). Crime and victims statistics. At: http://www.ojp.usdoj.gov/bjs/cvict.htm. Accessed on Jan. 21, 2002.

Bursik, R.J., Jr., and Grasmick, H.G. (1993). *Neighborhoods and Crime: The Dimensions of Effective Community Control*, Lexington Books, Lanham, MD.

Canadian Centre for Justice Statistics. (1996). *Canadian Crime Statistics, 1995*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (1997a). *Canadian Crime Statistics, 1996*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (1997b). *An Overview of the Differences Between Police-Reported and Victim-Reported Crime, 1997*. Catalogue No. 85-542. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (1998). *Canadian Crime Statistics, 1997*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (1999). *Canadian Crime Statistics, 1998*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (2000). *Canadian Crime Statistics, 1999*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (2001). *Canadian Crime Statistics, 2000*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Canadian Centre for Justice Statistics. (2002). *Canadian Crime Statistics, 2001*. Catalogue No. 85-205. Statistics Canada, Ottawa.

Carrington, P.J. (2002). Age and group crime. *Canadian Journal of Criminology* 44: 277-315.

Cohen, J. (1986). Research on criminal careers: Individual frequency rates and offense

seriousness. In Blumstein, A., Cohen, J., Roth, J.A., and Visher, C.A. (eds.), *Criminal Careers and "Career Criminals"*, Vol. I. National Academy Press, Washington, D.C., pp. 292-418.

Farrington, D.P. (1997). Human development and criminal careers. In Maguire, M., Morgan, R, and Reiner, R. (eds.), *The Oxford Handbook of Criminology*, 2nd ed., Oxford University Press, Oxford, pp. 361-408.

Frank, O. (1980). Estimation of the number of vertices of different degrees in a graph. *Journal of Statistical Planning and Inference* 4: 45-50.

Frank, O. (2001). Statistical estimation of co-offending youth networks. *Social Networks* 23: 203-214.

Gartner, R., and Doob, A.N. (1994). *Trends in Criminal Victimization, 1988-1993*. *Juristat* 14 (13), Catalogue No. 85-002. Canadian Centre for Justice Statistics, Statistics Canada, Ottawa.

Greenberg, D. F. (1991). Modeling criminal careers. *Criminology* 29: 17-46.

Kershaw, C., Budd, T., Kinshott, G., Mattinson, J., Mayhew, P., and Myhill, A. (2000). *The 2000 British Crime Survey*, Home Office Statistical Bulletin 18/00. Home Office: London.

Le Blanc, M., and Fréchette, M. (1989). *Male Criminal Activity from Childhood Through Youth*, Springer-Verlag, New York.

Mayhew, P., and van Dijk, Jan J. M. (1997). *Criminal Victimisation in Eleven Industrialized Countries*, Ministry of Justice, The Netherlands, Amsterdam.

Reiss, A.J., Jr. (1986). Co-offender influences on criminal careers. In Blumstein, A., Cohen, J., Roth, J.A., and Visher, C.A., eds., *Criminal Careers and "Career Criminals"*, Vol. II. National Academy Press, Washington, D.C., pp. 121-160.

Reiss, A.J., Jr. (1988). Co-offending and criminal careers. In Tonry, M., and Morris, N. (eds.), *Crime and Justice: An Annual Review of Research*, Vol. 10, U. Chicago Press, Chicago, pp. 117-170.

Reiss, A.J., Jr. and Farrington, D.P. (1991). Advancing knowledge about co-offending: Results from a prospective longitudinal survey of London males. *Journal of Criminal Law and Criminology* 82: 360-395.

Rennison, C. M. (2000). *Criminal Victimization 1999*. NCJ 182734. Bureau of Justice Statistics, U.S. Department of Justice, Washington, D.C.

Sacco, V. F., and Johnson, H. (1990). *Patterns of Criminal Victimization in Canada.* General Social Survey Analysis Series. Catalogue No. 11-612E, No. 2. Statistics Canada, Ottawa.

Sarnecki, J. (1986). *Delinquent Networks*, The Swedish Council for Crime Prevention, Stockholm.

Sarnecki, J. (1990). Delinquent networks in Sweden. *Journal of Quantitative Criminology* 6, 31-50.

Sarnecki, J. (2001). *Delinquent Networks: Youth Co-offending in Stockholm.* Cambridge U. Press, Cambridge.

Statistics Canada Housing, Family and Social Statistics Division. (1994). *Tables in Victimization. 1993*. General Social Survey Series. Product No. 12F0042XPE, Statistics Canada, Ottawa.

Tarling, R. (1993). *Analysing Offending*, HMSO, London.

Tracy, P.E., Wolfgang, M.E., and Figlio, R.M. (1990). *Delinquency Careers in Two Birth Cohorts*, Plenum, New York.

Tracy, P.E., and Kempf-Leonard, K. (1996). *Continuity and Discontinuity in Criminal Careers*, Plenum, New York.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications*, Cambridge U. Press, New York.

Wolfgang, M. E., Figlio, R. M., and Sellin, T. (1972). *Delinquency in a Birth Cohort*, U. Chicago Press, Chicago.

Wolfgang, M. E., Thornberry, T. P., and Figlio, R. M. (1987). *From Boy to Man: From Delinquency to Crime*, U. Chicago Press, Chicago.