ESTIMATING THE SIZE OF THE CRIMINAL POPULATION USING AN OPEN POPULATION APPROACH

Michael A. Greene, The American University

I. INTRODUCTION

In Greene and Stollmack (1981), a method was proposed for estimation of the size of the criminal population from arrest history records. The conceptual framework involved in that paper was (1) the arrest histories represented realizations from a common stochastic process which was truncated at zero and (2) the underlying population was closed, i. e. not subject to changes in size during the sampling period. A compound Poisson distribution was fit to the data, which allowed the estimation of the probability of zero arrests which is then used to estimate the population size (see Sanathanan, 1977 for a general discussion of this methodology). The model was fitted to data on approximately 6000 males arrested at least once during 1974-75 in Washington, D. C. We estimated that there were about 24,000 additional individuals who were at-risk to be arrested, yielding a criminal population size of about 30,000.

The present research constitutes a reanalysis of that data, relaxing the assumption of population closure. This new research then allows estimation of the parameters associated with rate of change of the population size as well as the size of the population at any given time. In addition to changes resulting from general population mobility, sources of increases in the criminal population include individuals beginning criminal careers while decreases result from deaths, termination of criminal careers and incarceration. These related parameters, relating to the length of criminal careers, the rate at which new individuals enter the population and the manner in which the individual arrest rate changes as people age have been of considerable interest in the criminological literature. The singular advantage of the open population methodology employed herein is that parameters associated with career termination and entrance, and arrest rates are estimated jointly. Thus within the limitations of the model, the estimates of each of these parameters include the influence of the others. The connection between this present research and the criminological literature are discussed in section II of this paper.

The methodology in the present study involves treating individual arrest histories as if they were observations from a multiple recapture census. The two year period January 1, 1974 to December 31, 1975 is partitioned into 4 contiguous non-overlapping periods of six months each. We record whether or not an individual has been arrested in each such period; such records are aggregated into frequency counts on the fifteen possible patterns of arrest or non-arrest. The estimation methodology utilized is Cormack's (1981) adaptation to the open population problem of Fienberg's (1972) log-linear contingency table model. This model is briefly discussed in section III of this paper and at length in the Appendix.

Data analysis and conclusions are found in section IV.

II. THE CRIMINOLOGICAL LITERATURE

Research on the size of the criminal population, the criminal career length, the rate at which new individuals enter the criminal population and the individual arrest rate has generally be associated with characterizing criminal behavior from an incapacitation or deterrence perspective. Regardless of the perspective, the general notion is that the more punishment, primarily imprisonment, the less crime, although the mechanismby which crime is reduced is somewhat different. In the incapacitation viewpoint crime is reduced by subtracting time from individual criminal careers by temporarily removing some individual from society (Blumstein, Cohen and Nagin, 1978). Deterrence theorists argue that the individual's probability of entering a criminal career, his length in that career and the rate at which he commits crimes are all inversely proportional to the level of punishment, i. e. the probability and length of imprisonment (Becker, 1967).

The empirical research on deterrence is mainly devoted to cross sectional or longitudinal estimation of aggregate crime rates as a function of numerous explanatory variables including punishment levels. As such, it does not explicitly concern itself with individual arrest rates, entrance to- or termination of the criminal career. In contrast, estimates for the parameters governing these processes are critical in the incapacitation studies to assess just how much crime is reducted by the subtractive effects of imprisonment. The research associated with incapacitation and the criminal career is discussed below.

Arrest Rate Studies

Empirical studies of arrest rates usually begin with the data on the spacings between an individual's arrests. Such studies include the following: Stollmack and Harris (1974) for a sample of individuals released from halfway houses in Washington, D. C.; Greenberg (1975) and Shinnar and Shinnar (1975) from national data; Clarke (1975) from the Philadelphia birth cohort data; Greene (1977) and Blumstein and Cohen (1979) from FBI data on individuals arrested in Washington, D. C. in 1973; Van Dine, Conrad and Dinitz (1979) with Franklin County, Ohio data; and Peterson and Braiker (1980) from self-reported crime of California prison inmates. The methodology in these papers is generally similar, namely the assumption of some stochastic process governing the interarrest intervals or arrest frequencies, followed by parameter estimation. The data differ as to whether they are (1) prospective, that is identifying individuals who are arrested as of a given date and following them forward in time, (Stollmack and Harris), (2) retrospective, namely, identifying individuals as of a given date and then collecting records of their prior activity, (Peterson and Braiker, Blumstein and Cohen), (3) age cohort based (Clarke), where both individuals who are subsequently arrested and those who are not are included in the data. Regardless of the method of collecting the data, the findings are remarkably similar, namely that arrest rates seem to lie in the interval of one-third to one arrest per year per person, and these rates tend to be inversely proportional to individuals' ages.

The critical difference between the three types of studies is the manner of assigning opening and closing intervals from which to begin tabulating arrest spacings. With retrospective data, the arrest which includes an individual in a particular sample can be taken as evidence that his criminal career was still in progress at the point of data collection, however the point at which the career began is not available. It is typical in such studies to assume that the adult career began at age 18 (see Blumstein and Cohen, 1979 for example), although the problem that such an assumption generates is the underestimation of individual arrest rates for careers which begin later. With prospective studies, it is not possible to determine whether the length of time between the last arrest and the end of the sampling period represents time in the career or not. Again, various assumptions have to be made, usually assuming that the career remains in progress. This also results in an underestimate of the individual arrest rate, when there is some career termination. Both problems are found with birth cohort data, especially for juveniles, where the age of onset is not known, nor again can the end point of the career, should it occur be determined. The inability to obtain the entrance and/or exit points from the career in all three categories of data generally results in underestimates of the individual arrest rate.

Criminal Career Lengths

The prevailing methodology for studying criminal career lengths was developed in Avi-Itzhak and Shinnar (1973). The idea behind the method is to assume that the age distribution of individuals arrested during a fixed sampling frame represents a realization of a backward recurrence times in a renewal process (see Cox and Miller, 1970 or Bartholomew, 1967). The method for doing this is to postulate a distribution for the complete career length, then formulate that distribution for the backward recurrence time and estimate parameters. Empirical estimates of the parameters for the distribution of the criminal career length were first obtained in Greene (1977) using this methodology and then later improved upon in Blumstein and Average criminal career lengths using these approaches were estimated to vary between 5 and 12 years, depending upon the partitioning of the data, the model and the various correction factors applied.

There are three problems with this approach, similar to the problems with arrest rates. First, the approach applies a longitudinal model to cross sectional data, thus heterogeneity in different birth cohorts may affect the estimates for the career lengths. Second, there are problems in fixing the beginning point of the criminal career (usually assumed to be age 18, but often perhaps later) and in dealing with the fact that low arrest rates of older individuals may be confused with increasing career termination. Blumstein and Cohen(1982) attempt to separate between these effects by separating the data by the types of crimes of first arrest, however the interaction between age of onset, arrest rate and career length are difficult to separate in a convincing manner. The third problem with this approach is that it requires the assumption that the offender population is stationary (i. e. the number of individuals terminating careers is equal to the number initiating careers).

Entrance to Criminal Careers

The research on entrance probabilities is the least well developed of any of these topics. The general conclusion seems to be that entrance to the criminal career tends to occur during an individual's teen age years. This conclusion is primarily based on the peaking of age stratified arrest frequencies. The probability distribution of the age until first arrest has been utilized to estimate the number of individuals who will ultimately be arrested (Christensen,1967), which uses but is different from the rate of growth of the offender population.

Avi-Itzhak and Shinnar (1973) present a model for estimating the growth rate of the offender population by age category, but to the best of our knowledge, it has never found empirical use. Like the other analyses with arrest data, it assumes that arrest rates and entrance rates are independent of each other.

Estimates of the Size of the Offender Population

Greenberg (1975) and Shinnar and Shinnar (1975) both estimate the size of the offender population by dividing the total number of crimes committed by an estimate for the average individual crime rate. Such estimates must be given within broad range in view of the extreme difficulties in estimating crime rates. For example, Greenberg estimated the criminal population in the United States to lie between 2.5 and 16.7 million (approximately 2 - 15 per cent of the US male population). Shinnar and Shinnar estimated the New York State criminal population to lie in the range of 40,000 to 80,000 (0.5 to 1 per cent of the state's male population), however this estimate was based on an estimate of the crime rate within a range that was probably too narrow considering the lack of precision of their model and the approximateness of the data. The problems in these two studies revolve around the significant difficulties in obtaining the number of crimes (this differs perhaps quite substantially from the number of reported crimes) and also in developing accurate estimates of the individual crime rates.

Riccio and Finkelstein (1977) estimated the number of burglars in Montgomery County, Maryland from arrest frequency data. They partitioned the sample into two groups, those with 5 or fewer arrests (denoted as the "P-regime") and those with more ("the D-regime"). This approach was taken because of the appearance of extreme heterogeneity in the sample. The authors fitted two separate Poisson distributions to the data (the method of fitting to a sample where one of the distributions is truncated at zero and over 5 and the other is truncated under 5 is not given in the text). Their estimate for the size of the burglar population was 2200 individuals, of whom only 722 or 35 per cent had been apprehended at least once. These authors assumed that the population was closed (without career termination or entrance), merely that the rates at which individuals were arrested was different. The same assumption of closure was used in our previous analysis (Greene and Stollmack, 1981),

Conclusion

In general, the results from the literature seem to be at some degree of conflict with each other. If the criminal population is closed, then estimates of the criminal career length and entrance rates do not make much sense. On the other hand, if the population is open, then it is necessary to correct for entrance and career termination when estimating arrest rates. It is probably unsatisfactory to develop correction factors for career termination with arrest rates on the basis of the age distribution, then to turn around and use that same distribution to estimate career length.

The result is likely to be underestimates of arrest rates, entrance probabilities and overestimates of career length, depending upon how strong these factors tend to be.

III. THE JOLLY-SEBER METHOD

The Jolly-Seber method has often been applied to estimating the size of a hidden population where the number of individuals in that population can change over time. The usual assumptions of this model are as follows (see Seber, 1973 and Cormack, 1981):

a) Every individual in the population has the same probability p_i of being caught in the i th sample (e.g. arrested duringf the ith time period), given that the individual is alive and in the population when that sample was taken.

b) Every individual in the population at sample i has probability ϕ_i of being in the population at time i+1. This represents a survival (and/or non-migration) probability.

c) At time or sample i, there are N individuals in the population.

d) The number of new individuals joining the population at i and surviving to i+1 are denoted as B_i (births and/or immigrants). Cormack (1981) conveniently expresses this as a "growth rate" ψ_i, where

$$\psi_i = 1 + B_i/N(1-p_i)\ \Pi_{j=1}^{i-1}\ (1-p_j)\phi_j\ \psi_j$$

To obtain estimates for these parameters, we used a variation of Cormack's modification of Fienberg's loglinear model. The principal advantage of this method over the method developed by Jolly and Seber, is that this method facilitates hypothesis testing as to whether the best model is a closed population (ψ , ϕ =1), a birth only model (ϕ =1), a death only model (ψ=1) or the complete open model. Details of this method are in an appendix available from the author.

IV. RESULTS AND DISCUSSION

Data Preparation

The data were derived from the 1974-1975 PROMIS public access tapes, which are available from the Inter-university Consortium for Political and Social Research (ICPSR). These tapes contain the universe of arrests of adults for both felonies and misdemeanors for Washington, D.C., for 1974 and 1975. Individuals are identified by an unique identification number. There were a total of 25,441 individuals with 36,427 arrests. The individuals were predominately male and non-white (both approximately 85 per cent).

We separated the data by age, primarily because experiments with the pooled database produced extremely poor fits to the model (G^2 over 100), which could possibly result from heterogeneity (different parameter values by age). Females were discarded. Individuals were counted as appearing in a period if they experienced an arrest for one of the following offenses: aggravated assault, robbery, burglary, larceny, auto theft, possession of weapons, narcotics, fraud, receiving stolen goods, or simple assault. Homicide and forcible rape were not included among the offenses because the likeliehood of imprisonment is extremely high for these offenses, with the result that the model will confound imprisonment and career termination.

Results

The results from applying the model to our data are shown in Table 1. G^2 likeliehood ratio statistics are presented for goodness of fit measures. The age of individuals (column 2) are keyed to their year of birth, thus for example, the youngest individuals shown in the table, those born in 1955, would have had their 19th birthday during 1974. Younger individuals were excluded because those individuals becoming 18 during 1974 would not have any recorded arrests before their 18th birthday. Individuals born in 1936 or earlier (38th birthday during 1974) were pooled together.

We tested four models for each age group, namely the open model with both entrance and exit parameters, a model with death only, a model with birth only and finally a closed model, which allowed the sampling or capture probability to vary over the four intervals. Acceptable fits were found for all but three age categories, the youngest group (the first row of the table), those born in 1949 (age 25) and the pooled oldest group (the last row of the table). Scanning down the column for the open population model, one can see an interesting pattern in the G^2 statistic, namely that the fit seems to be the best for birth years and/or ages that are not divisible by 5. For example very good fits are recorded for birth years 1953, 51, 48, 47, 46, 43, 42, 41, and 37. The exception to this is 1950 and several others. One can speculate as to why this occurs, probably because the recording of individual ages and birth years in the data is apparently done from the individual's own testimony as to his age and is usually unverified. It is entirely possible that individuals are rounding either their ages or birth years to the nearest 5. This may be introducing some additional heterogeneity into the data.

TABLE 1
G^2 STATISTICS
FOR DIFFERENT MODELS AND AGES

Model and Degrees of Freedom

BIRTH YEAR	AGE	OPEN 6	DEATH ONLY 8	BIRTH ONLY 8	CLOSED 10
55	19	35.67	44.71	45.08	48.38
54	20	15.41	34.26	35.47	41.27
53	21	10.84	18.94	17.53	20.81
52	22	16.84	35.57	32.69	39.96
51	23	5.94	14.96	18.39	21.03
50	24	6.85	9.85	13.28	14.44
49	25	40.94	52.08	55.97	57.83
48	26	14.83	31.82	36.50	40.38
47	27	9.17	11.02	19.04	19.21
46	28	7.62	21.54	11.32	22.10
45	29	16.17	20.09	18.09	20.39
44	30	12.69	15.02	24.54	24.57
43	31	11.11	14.48	15.36	17.83
42	32	6.48	11.97	16.23	17.50
41	33	12.29	22.24	27.79	29.44
40	34	16.37	22.57	27.53	28.73
39	35	12.79	15.84	19.70	22.32
38	36	15.64	16.62	19.61	19.62
37	37	7.14	9.90	12.01	13.54
<37	>37	58.30	70.85	74.30	79.05

Tail Area Probabilities for the Chi-Square Distribution by DF

p	6	8	8	10
.10	10.65	13.36	13.36	15.99
.05	12.60	15.50	15.50	18.30
.025	14.40	17.50	17.50	20.50
.01	16.80	20.10	20.10	23.20

From the plethora of statistics in table 1, we selected the best single model for each age group by subtracting the G^2 statistics and testing at the appropriate degrees of freedom (note: the difference between chi-square statistics is distributed as a chi-square with degrees of freedom equal to the difference of the degrees; see Cormack, 1981). The best single model for each age group is shown in table 2.

To review, the interpretation of the parameters is as follows. The quantities denoted as p are the probability that an individual who is in the criminal population is sampled during a particular six month period. Thus, for example, the chance for a 40 year old to be sampled during the second six month period is .22. The next set of columns over are the values of ψ and ϕ denoting the growth rate (more specifically, the ratio of the unsampled individuals at the start of the i+1st period to the unsampled individuals in the i th period). The next set of columns contains the survival probabilities, defined as the proportion of individuals, both sampled and unsampled in the i th period who are available

TABLE 2
PARAMETER ESTIMATES FROM SELECTED MODELS

AGE	CAPTURE PROB		GROWTH RATE		SURVIVAL PROB		SAMPLE SIZE	POP. EST.
	p_2	p_3	ψ_2	ψ_3	ϕ_1	ϕ_2		
19	.	.	.	.	.	.	.	.
20	.22	.25	1.42	1.85	.71	.71	1253	1503
21	.20	.21	1.41	1.41	.85	.76	1120	1568
22	.25	.22	1.74	1.67	.68	.80	1113	1166
23	.23	.19	1.59	1.39	.70	.75	1005	1294
24	.12	.12	.	.	.	.	1041	2411
25	.	.	.	.	.	.	.	.
26	.24	.29	1.60	2.04	.70	.58	828	952
27	.13	.16	.	.	.87	.66	763	1789
28	.21	.13	1.51	1.65	.	.	540	735
29	.14	.09	.	.	.	.	503	1296
30	.11	.22	.	.	1.11	.48	456	1204
31	.11	.11	.	.	.	.	420	1091
32	.07	.13	.	.	1.17	.50	417	1507
33	.23	.40	1.26	.32	.68	.42	326	424
34	.27	.26	2.03	1.57	.67	.54	336	362
35	.08	.19	.	.	1.70	.46	293	1005
36	.14	.09	.	.	.	.	259	664
37	.12	.12	.	.	.	.	241	553

NOTES

The model selected can be determined from the pattern of omitted (.) estimates.

POP. EST. = Estimated size of the offender population at the start of the second period.

to be sampled in the i+1st period. The last two columns contain the sample size (N) and the estimate of the size of the criminal population during the second period (July - December of 1974). The quantities p and ϕ are bounded in the interval (0 - 1), where ϕ =1 indicates that everybody survives (no mortality or career termination). ψ is lower bounded at 1.0 denoting no growth in the population.

In our judgement, only the fits and parameter estimates for age groups 20-23 are acceptable. The appropriate model for age 24 seems to be the closed population model, which appears to be implausible in view of the results for ages 20-23. No model fits at all for age 25. The parameter estimates for age 26 are not bad, but the growth rate in the third period seems to be extremely high. Age 27 has a death only model which results in the lowest arrest probability estimates (except for age 24) and a much higher survival probability in period 1 than seen elsewhere. The estimates for the age 28 group result from the single birth only model, with again a very low arrest probability. The models for older individuals seem to suffer from some of these defects as well, and also some out of range parameter estimates (see ψ_3 for age 33 and ϕ_1 for ages 30 and 32). It should be noted that the loglinear model does not constrain the parameter estimates to be within the required range, thus when they do fall in the range, there is some indication of the validity of the model.

Table 3 shows some additional statistics for the 4 age groups with parameter estimates that we consider trustworthy. Both the arrest rate and the survival probability are assumed to follow an exponential distribution, with the annualized value of the arrest rate and the average criminal career length shown. The arrest rates are within the range found in the literature (usually between one third and one arrest per year), however the average career lengths tend to be much shorter than what has usually been reported to be between 4 and 10 years (see Greene, 1977; Blumstein and Cohen, 1982, table 4,page 60). The growth rate in the offender population (found in table 6) between 41 (shown as 1.41) and 85 per cent is seems quite large, however there is no literature for comparisons. The general picture emerging of the criminal population for individuals in their early 20's is that of considerable turbulence resulting from a large number of individuals terminating and initiating criminal careers, however, the net growth in the population seems to be relatively small (this may be obtained by comparing ϕ with $1/\psi$).

For individuals in their early 20's, the size of the population ranges from 4 per cent (age 22- computed as the population estimate divided by the sample size - 1; see Table 6) to about 40 per cent averaging 23 per cent larger over the four age groups. This estimate is considerably lower than our previous work with the assumption of a closed population where estimated the population to exceed the number sampled by a factor of 4.8. Large discrepancies between the number sampled and the estimate for the population size are apparent in table 5 for the closed and death only models.

In conclusion, we believe that the open population approach is far superior to modelling the offender population as closed. The estimates for population size, growth and career termination rates seem to be reasonable, and at least for the latter two, somewhat contrary to the literature. The model seems to show a population with considerable underlying turbulence.

TABLE 3
ESTIMATES OF ARREST RATES
AND CRIMINAL CAREER LENGTHS
(AGES 20-24)

Age	Annualized Arrest Rate (by period)		Average Career Length in Years (by period)	
20	.50	.57	1.47	1.47
21	.44	.47	3.13	1.85
22	.57	.50	1.29	2.27
23	.52	.42	1.48	1.75

ACKNOWLEDGEMENT

This research was supported by Grant No. 81-IJ-CX-0001 from the National Institute of Justice. The conclusions drawn in this paper are those of the author and not to be attributed to the U. S. Department of Justice. The author is grateful for the support.

REFERENCES

Avi-Itzhak, B. and R. Shinnar (1973). "Quantitative models in crime control." Journal of Criminal Justice, 1.

Bartholomew, D. J. (1967), Stochastic Models for Social Processes. Wiley, NY.

Becker, G. (1967), "Crime and punishment: an economic approach." Journal of Political Economy, 78(2).

Bishop, Y. M. M., S. E. Fienberg and P. W. Holland (1975), Discrete Multivariate Analysis. MIT Press, Cambridge, MA.

Blumenthal, S., R. C. Dahiya and A. J. Gross (1978), "Estimating the complete sample size from an incomplete Poisson sample." Journal of the American Statistical Association, 73(362).

Blumenthal, S. and R. Marcus (1975), "Estimating population size with exponential failure." Journal of the American Statistical Association, 70(352).

Blumstein, A., J. Cohen and D. Nagin (eds.) (1978), Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates. National Academy of Sciences, Washington, D. C..

Blumstein, A. and J. Cohen (1979), "Estimation of individual crime rates from arrest records." Journal of Criminal Law and Criminology, 70.

Blumstein, A. and J. Cohen (1982), "The duration of adult criminal careers." Report submitted to the National Institute of Justice. School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, PA.

Christensen, R. (1967), "Projected percentage of US population with criminal arrest and conviction records," Appendix J in Task Force Report: Science and Technology. The President's Commission on Law Enforcement and the Administration of Justice, Washington, D. C..

Clarke, S. H. (1974), "Getting 'em out of circulation: does incarceration of juvenile offenders reduce crime?" Journal of Criminal Law and Criminology, 65.

Cormack, R. M. (1980), "Model selection in capture-recapture experiments." Paper No. BU-712-M, University of St. Andrews, St. Andrews, Scotland, UK.

Cormack, R. M. (1981), "Loglinear models for capture-recapture experiments on open populations," in R. W. Hiorns and D. Cook (eds.), The Mathematical Theory of the Dynamics of Biological Populations. Academic Press, London.

Cox, D. R. and H. D. Miller (1970), The Theory of Stochastic Processes. Methuen, London.

Feller, W. H. (1950), An Introduction to Probability Theory and Its Application, Vol 1. Wiley, New York.

Fienberg, S. E. (1972), "The multiple recapture census for closed populations and incomplete contingency tables." Biometrika, 59.

Greenberg, D. F. (1975), "The incapacitative effects of imprisonment: some estimates." Law and Society Review, 9(4).

Greene, M. A. (1977), "The incapacitative effect of imprisonment on crime." Ph. D. Dissertation, School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, PA.

Greene, M. A. and S. Stollmack (1981), "Estimating the number of criminals," in J. A. Fox (ed.), Models in Quantitative Criminology. Academic Press, New York.

Otis, D. L., K. P. Burnham, G. C. White and D. R. Anderson (1978), "Statistical inference from capture data on closed animal populations." Wildlife Monograph No. 62, Wildlife Society, Louisville, KY.

Peterson, M. and H. B. Braiker (1980), Doing crime: a survey of California Prison Inmates. The Rand Corporation, Santa Monica, CA.

Riccio, L. J. and R. Finkelstein (1977), "Using police arrest data to estimate the number of burglars operating in a suburban county." (mimeo).

Roughgarden, J. and P. Heckel (1979), "A technique for estimating the size of lizard populations." Ecology, 60(5).

Sanathanan, L. (1977), "Estimating the size of a truncated sample." Journal of the American Statistical Association, 72(359).

Seber, G. A. F. (1973), The Estimation of Animal Abundance and Related Parameters. Griffin, London, UK.

Shinnar, S. and R. Shinnar (1975), "The effects of the criminal justice system on the control of crime: a quantitative approach." Law and Society Review, 9(4).

Stollmack, S. and C. M. Harris (1974), "Failure rate analysis applied to recidivism data." Operations Research, 22(6).

Van Dine, S., J. P. Conrad and S. Dinitz (1979), Restraining the wicked: the dangerous offender project. Lexington Books, Lexington, MA.