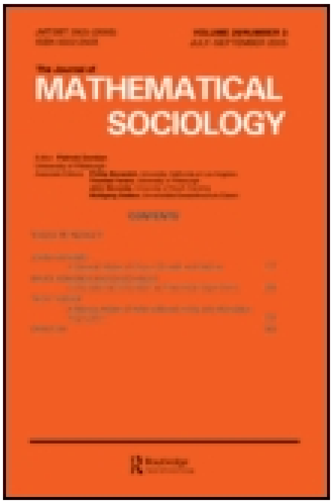

The Journal of Mathematical Sociology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gmas20

The estimation of the prevalence of delinquency: Two approaches and a correction of the literature

Robert A. Gordon [a] & Leon Jay Gleser [a b]
[a] The Johns Hopkins University
[b] Department of Statistics , Purdue University
Published online: 26 Aug 2010.

Journal of Mathematical Sociology
1974, Vol. 3, pp 275-291

THE ESTIMATION OF THE PREVALENCE OF DELINQUENCY: TWO APPROACHES AND A CORRECTION OF THE LITERATURE†

ROBERT A. GORDON and LEON JAY GLESER‡
The Johns Hopkins University

The index "the proportion of a cohort that have become delinquent by a given age," here called the "prevalence of delinquency," is an important social indicator. In the present paper, we indicate methods by which this index can be estimated from data, and correct errors in previous sex- and race-specific prevalence estimates published by Monahan (1960) for the city of Philadelphia. The difference between the sexes and between the races shown by these corrected prevalence estimates are of sufficient magnitude to render suspect any comparisons of prevalences of delinquency among cohorts which do not take account of the sex and race compositions of the cohorts to be compared.

INTRODUCTION

In searching for an index that will give some idea of how widespread the social problem of juvenile delinquency is in a given (age-specific) cohort, the index defined as "the proportion P of the cohort that have become delinquent by a given age" immediately comes to mind as a useful social indicator. This index has been used in the literature of juvenile delinquency, but under a variety of names borrowed from the public health literature. For example, Monahan (1960) has employed the word "incidence" to refer to two different kinds of index within the same paper—one of these corresponds to the proportion under discussion here. Ball, Ross, & Simpson (1964) distinguished between these same two indices, calling one "incidence" and the other "prevalence." Wolfgang, Figlio, & Sellin (1972) empirically determined an index that corresponds to our so far unnamed proportion and to Ball *et al.*'s "prevalence"; however, they did not commit themselves to any label. No rationale was given in any of these publications for the terminological decisions taken.

In the public health literature, the "prevalence" of a disease in a certain group

† The work of Robert A. Gordon on this paper was supported in part by Research Grant MH 13951 from the National Institute of Mental Health, and by National Science Foundation Grant GS-29873. The work of Leon Jay Gleser was supported in part by N.I.H. Training Grant No. GM-1237-07 awarded to the Department of Statistics, The Johns Hopkins University. This paper in whole or in part may be reproduced for any purpose of the United States Government.

‡ Now in the Department of Statistics, Purdue University.

refers to how widespread the disease is in the group at a particular instant of time, and is defined as the proportion of group members currently (at that instant) suffering from the disease. The "incidence" of a disease on the other hand indicates the experience of the group over a given *interval* of time, and is defined as the ratio, n/N, of the number, n, of contractions (even if recurrent) of the disease by group members during the time interval to the total number, N, of individuals in the group (diseased or otherwise). Depending on how we characterize delinquency as a "disease" and on what aspect of a cohort's experience with delinquency we want to emphasize, either term ("prevalence" or "incidence") could serve to describe the index P which we have defined above.

The choice of a verbal label for the index P is a matter of discretion. The index tself is mathematically well defined, and therefore unambiguous. However, since it is convenient to have a word to describe the index P, we have chosen to label this index "the prevalence of delinquency," thus following the practice of Ball *et al.* We do so for two reasons: (1) The formal definition of incidence encourages the impression that one is dealing with ephemeral events ("contractions") and that the count may embody repetitions of these events instead of being based on persons. The formal definition of prevalence, on the other hand, emphasizes that it is persons who are being counted, and it more clearly recognizes that serious delinquency is often a persisting phenomenon, more like a chronic disease than an acute one (without committing itself concerning exact duration). Realistically, delinquents usually manifest delinquent "symptoms" over a period of time, and not just during the act for which they are officially apprehended, although our count may depend on their cumulative exposure to detection (i.e., "diagnosis") over a sufficiently long period. (2) A variety of delinquency rates already appear in the delinquency literature, labeled as "incidences." Thus, using the term "incidence" for the index P would tend (and has tended) to create confusion.

It is our hope that readers who would have preferred other labels for the index P will accept the terminology "prevalence" at least for the duration of this paper. We believe that general adoption of this terminology would eliminate much of the present confusion in terminology within the delinquency literature. In any case, a quibble over labels only serves to emphasize that in dealing with social indicators, our ultimate reliance must be upon explicit mathematical definitions, rather than upon ambiguous verbal terms borrowed from other disciplines or from common use.

EARLIER ESTIMATES OF THE PREVALENCE OF DELINQUENCY —MONAHAN'S DATA

A small number of estimates of the prevalence of official juvenile delinquency for different segments of the general population, employing various criteria, now exist, (Monahan, 1960; Ball *et al.*, 1964; Hathaway and Monachesi, 1963; Wolfgang *et al.*, 1972). Until the recent appearance of Wolfgang *et al.*'s race-specific prevalence of arrest estimate for a single cohort of males, Monahan's estimate, based on having had a juvenile court experience by age 18, for Philadelphia, was the only published source which presented *both* race-and-sex-specific delinquency prevalence estimates. The high standard of the statistics kept by the Municipal Court of Philadelphia, plus the facts that Monahan's data series span a period of years, and that his estimates show great stability over the years, contribute to our sense of the quality of the data in this otherwise underappreciated paper. Although Monahan's paper used the word "incidence," and did not actually employ the word "prevalence" at all, his paper was concerned with estimating the proportion of juveniles destined to appear in court by a given age.

In the course of examining the manifest differences between whites and blacks in official rates of delinquency,† it was noted that, on their face at least, Monahan's original data showed prevalence estimates that were so different between the races as to reduce any prevalence index that failed to control for race to a meaningless linear combination of underlying, but invisible, parameters.‡ No time series, and no comparison of prevalences of delinquency from place to place, would be exempt from this serious ambiguity whenever the relative proportions of blacks and whites in the population were subject to variation, as is usually the case. Yet, *the rational assessment of the performance of all law enforcement agencies depends on exactly this sort of comparison.*

Unfortunately, it was also discovered that Monahan's estimates of prevalence are flawed by the employment of an incorrect computational paradigm. This mistake makes it difficult to employ these published results in connection with any hypothesis.§ This is most regrettable, for with the suppression of categorization according to race that is now in force for many source statistics, it is impossible to redo this research, which would be an expensive undertaking in any case. Monahan himself had to terminate his own data series at 1954 for this very reason. Therefore, if these rare data

† The present paper is a first step in reviewing all available prevalence data, in order to assess their relevance for theory. A second paper (Gordon, 1973) gives race-and-sex-specific data for the nationwide prevalence at age 18.0 of an even more serious degree of delinquency, as defined by commitment to a training school, for the period around 1964. Between them, these two papers bracket the points on a continuum of severity for delinquency that should be of most interest to the criminologist. The first deals with delinquents that come to the broad notice of the juvenile court, whose function it is to sort them out for various dispositions. The second deals with the most severe category of official delinquent, since few juveniles are committed who have not been involved in *more than one* serious offense. Beyond this category in severity there is only the very small percentage—in cities, about one case for every 50 referred to juvenile court (Lunden, 1964, Table 77), or about 0.3 percent of an urban male cohort—who are referred to criminal court; and of these, the smaller percentage convicted and incarcerated.

‡ We are, of course, aware that the intent to publish crime statistics by race and by sex flies in the face of recent expressions of sentiment against the production of exactly this sort of information (Geis, 1965; Van den Berghe, 1971). However, we do not feel that the legitimate interests of anybody are served by this avoidance of *full* awareness of *all* the facts concerning any serious social problem. If by some chance the prospects for reducing crime rates among blacks are linked to an accurate understanding of the cause for their being higher than among whites, these prospects will never be realized if the relevant information is suppressed (see, for example, Durkheim, 1951, p. 41).

§ It should be emphasized that even if the mistake is found to be small, as long as it is known to exist the original findings *remain useless* until its magnitude has been properly evaluated. Correspondence with the author has established that this evaluation will have to be accomplished without access to important components of the original data.

are to be salvaged for scientific purposes, it is necessary to evaluate and, if possible, correct the mistake that has been made.†

Besides the necessity of correcting Monahan's data, it has become apparent to us after reading the literature that there is confusion among criminologists about how to estimate prevalence, and about the use and interpretation of various measures (rates) involved with such estimates. Thus, a general discussion of this subject seems to be needed, so that mistakes such as the one made by Monahan have less chance to happen in the future. This general discussion is presented in the next two sections, leading the way to (and providing theoretical background for) correction of Monahan's own estimates.

THE CALCULATION OF PREVALENCE

We assume that delinquency is operationally defined by some explicit criterion. Then, if a cohort of individuals is followed from birth to some desired age, usually the 18th birthday for delinquency, we define the prevalence of delinquency to be the proportion of original cohort members who meet the criterion at least once by the end of the period. It is convenient to refer to the prevalence of delinquency, so defined, by coupling the rate with the time point (in this case, the birthday) marking the end of the period, such as "prevalence at age 18.0." It should be emphasized that the criterion is a formal one, and hence is completely flexible.

The most straightforward way to determine prevalence would be to follow the given cohort from birth to the desired birthday, noting down the fact whenever a cohort member becomes a delinquent (according to the criterion) for the first time. If at the end of the period of observation, we have observed M first delinquencies, and if the original cohort has N members, the prevalence, P, of delinquency is given by the fraction

$$P = \frac{M}{N}. \quad (1)$$

If the criterion of delinquency involves action by an official agency, and records are good, it would be equivalent to do a retrospective search of the records. In any case, it is usually desirable to identify the cohort membership at the beginning, rather than the end, of the period in question so that *differential mortality* (due to death) can be taken into account. Mortality due to death plays such a small role during some age periods that it can often be disregarded with little effect on the obtained rates. However, if mortality is especially high in the interval from birth to say age 7.0, when exposure to risk begins to occur for delinquency, the identification of the cohort size at the time of birth would inflate the denominator and produce estimates that are

† Since until just recently his was the only estimate in existence for large cities, there is every reason to believe that Monahan's results are the basis for all references to typical official prevalence indices mentioned throughout the criminological literature, as for example in the influential report by the President's Commission on Law Enforcement and Administration of Justice (1967, p. 55), although this is seldom acknowledged. If for no other reason, it is essential that these estimates be at least formally correct.

lower than what members of the society might regard as phenomenologically pertinent. For this reason, it is usually best to identify cohort size at the beginning of the period in which individuals are at risk, as determined both by the nature of social reality and by an appropriate formal criterion. As will be seen later, this is the practice followed by Monahan and the practice which we also follow. It should additionally be noted that although an individual can have multiple delinquencies, each delinquent is recorded only once (the first delinquency) for purposes of calculating the prevalence of delinquency. (This produces no loss of generality since if a greater degree of severity is of interest the criterion of delinquency can easily be made to consist of multiple events: e.g., two arrests, four court appearances, two incarcerations, etc.)

The proportion P defining the prevalence of delinquency at some age k can also be calculated by finding the numbers, m_i, of individuals in the cohort who become delinquent for the first time during the i-th year (age i to age $i+1$) of their life, $i = 0,1,\ldots,k-1$. The sum $\sum_{i=0}^{k-1} m_i$ is then the total number of individuals in the cohort who have ever become delinquent during the period from birth (age 0) to their k-th birthday, and

$$P_k = \frac{\sum_{i=0}^{k-1} m_i}{N} = \sum_{i=0}^{k-1} \frac{m_i}{N} \tag{2}$$

is the prevalence of delinquency at age k. For example, $P_{16} = \sum_{i=0}^{15} (m_i/N)$ is the prevalence of delinquency at age 16.0, and $P_{18} = \sum_{i=0}^{17} (m_i/N)$ is the prevalence of delinquency at age 18.0.

If we define $a_i = m_i/N$, then the quantities a_i are "age-specific first-incident rates of delinquency" for the given cohort. From Equation (2), we obtain

$$P_k = \sum_{i=0}^{k-1} a_i, \tag{3}$$

so that the prevalence, P_{16}, of delinquency at age 16.0 equals $\sum_{i=0}^{15} a_i$, and the prevalence, P_{18}, of delinquency at age 18.0 equals $\sum_{i=0}^{17} a_i$.

Often, it is inconvenient either to follow a cohort for such a long period, or to define a cohort for a retrospective record search. In these cases, and for other reasons of convenience as well, an estimate of prevalence can be obtained from age-specific first-incident rates of delinquency for the age range in question as of a given calendar year, when these rates are available from already existing records. (For example, the numbers of first-delinquents of a given age can be found from court records, and base populations for the various ages can be determined from census or school records.) For the given calendar year, we determine the number n_i of juveniles of age i at the beginning of the year who became delinquent for the first time during the year. We then divide n_i by the number N_i of juveniles who were of age i at the beginning of that calendar year, thus obtaining an estimate $\hat{a}_i = n_i/N_i$ of the age-specific rate of first delinquency for age i (and that calendar year). The prevalence P_k of delinquency at age k can then be estimated by substituting $\hat{a}_0, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_{16}$, and $\hat{a}_{17}$ for $a_0, a_1, a_2, \ldots, a_{16}, a_{17}$, respectively, in Equation (3). The resulting estimate of prevalence is then

$$\hat{P}_k = \sum_{i=0}^{k-1} \hat{a}_i = \sum_{i=0}^{k-1} \frac{n_i}{N_i} \tag{4}$$

and in particular,

$$\hat{P}_{16} = \sum_{i=0}^{15} \frac{n_i}{N_i}, \qquad \hat{P}_{18} = \sum_{i=0}^{17} \frac{n_i}{N_i}. \tag{5}$$

Although estimates of prevalence determined in this way do not apply to any cohort of real persons, it is often reasonable to assume that they would apply at about that period to a hypothetical cohort passing through the ages of interest. Since a separate prevalence estimate can be generated this way for each of a series of adjacent calendar years in which there have been records kept of age-specific first-incident rates, and of base populations, it is a simple matter to compare the age-specific rates and prevalence estimates from year to year to gain an indication of the stability of the incidence and of the prevalence of delinquency.

The calculation of an estimate of the prevalence of delinquency through estimates of age-specific rates of first delinquency may be illustrated by using the data obtained by Monahan (1960, Table 4). These data, which appear in Table 1 below,† include both a tabulation of first offenders by age (column 3) and a tabulation of base populations for those ages (column 2).‡ Note that the tabulation in this table begins with the

† Except for minor changes in labeling, our Table 1 and Table 4 in Monahan (1960) are identical.

‡ Although Monahan reported that he obtained his data on base populations by year of age from Annual Reports of the Board of Public Education, School District of Philadelphia, and although he refers to the base population in his Table 4 (see our Table 1) as the "school population," an examination of these reports for the years 1946–54 and 1957 (the years of interest to readers of his paper) shows that the count is never for just the in-school population alone, and that Monahan's specimen figures for the base population in his Table 4 (see our Table 1) in fact refer to *all* members of the age-specific cohort, both enrolled and not enrolled in school. The statistics given in these Annual Reports are always comprehensive, and in addition to listing the public, parochial, and private school populations, the tables also present the population "not enrolled in school," both "employed" and "not employed," and the totals, by race and by sex. We have compared the age-specific rates by race and by sex given in the Annual Report for 1950, with similar data from the 1950 U.S. Census for Philadelphia (based upon a 20 percent sample), and the correspondence between these two sources is excellent (see The Board of Public Education, School District of Philadelphia, 1950, Table 9; U.S. Bureau of Census, 1952, Tables 51 and 63). This comparison affords no evidence of cases being lost due to school-leaving at ages 16 and 17, or of cases being lost differentially by race. Finally, throughout the Annual Report series, there is good continuity between the subpopulation sizes at, say, age 15 or 16 in a given year and the size of the same subpopulation at ages 16 and 17, respectively, in the next later year. The quality of the data concerning age-specific base populations is therefore judged to be extremely good throughout the period in question, throughout the age range in question, and for all of the subpopulations in question. Any slight differences in timing between the years of age as defined by the Board of Public Education and as defined by the Municipal Court would have only a negligible effect on the final rates, because they would entail only small transfers between adjacent age cohorts; these transfers tend to compensate on the one side what a cohort gives up on the other side.

7th birthday. This is because the number of first delinquencies prior to age 7 is defined to be zero; here, as is usual, the fact of no delinquencies before age 7 results from the legal definition of the lower age bound of culpability.

To estimate prevalence of delinquency at age 18.0 from Table 1 and Equation (5), we divide the number n_i of first offenders of age i (column 3) by the number N_i in the base (school age) population for that age (column 2), for $i = 7, 8, \ldots, 17$. The resulting age-specific rates $\hat{a}_i$ of first delinquency, expressed as percentages, appear in column 4 of Table 1. One one-hundredth of the column sum of column 4, namely

$$\hat{P}_{18} = \frac{1}{100}(.086+.219+\ldots+2.749+2.218) = 0.16297,$$

gives an estimate of the prevalence of delinquency at age 18.0 for the hypothetical cohort represented by the data in Table 1.

The estimate of prevalence based on the data in Table 1 which was reported by Monahan (1960, Table 3) is, however, (1/100)th of the bottom entry in column 7 of Table 1; namely, .15194. This value, as an estimate of prevalence, is thus off by an error of .01103—an error of approximately 6 percent of the proper value. The magnitude of this error is small. However, it would be a serious mistake, as we shall see, to assume that the errors in Monahan's other estimates are always this small in magnitude.

Although the internal evidence in Monahan's paper leads us to believe that he was

TABLE 1

Monahan's Table 4, entitled,
"Incidence of Delinquency—Philadelphia: 1957"

Age (1)	School population (2)	First offenders (3)	Rate per 100 (4)	Survivors age 6 = 100,000 (5)	Becoming delinquent in age (6)	Percent delinquent to end of age (7)
7	34,933	30	.086	99,914	86	.1
8	34,183	75	.219	99,695	219	.3
9	34,876	165	.473	99,223	472	.8
10	37,688	232	.616	98,612	611	1.4
11	28,622	274	.957	97,668	944	2.3
12	27,705	400	1.444	96,258	1,410	3.7
13	28,819	602	2.089	94,247	2,011	5.8
14	32,976	763	2.314	92,066	2,181	7.9
15	28,194	883	3.132	89,182	2,884	10.8
16	26,445	727	2.749	86,730	2,452	13.3
17	24,797	550	2.218	84,806	1,924	15.2
Sum			16.297		15,194	

NOTE: The "school population" here refers to the entire cohort, enrolled in school or not. See p. 280, fn. ‡.

Source: Monahan (1960).

interested in estimating prevalence as we have defined it, using the indirect method of estimation outlined in Equations (4) and (5) above, it might be argued that Monahan was actually estimating some other indicator of the severity of delinquency. To answer this argument, we will more closely examine Monahan's own "specimen calculation" (as displayed in his Table 4; our Table 1) to see what quantity actually was being computed. However, for the sake of clarity it is helpful to discuss first still another way of estimating prevalence from data of the sort given in Table 1 which, although it always gives the same estimate $\hat{P}_k$ as the method described in Equation (4), is computationally different from our previous method.

DELINQUENCY RATES VERSUS "DELINQUENCY RATES FOR NONDELINQUENTS"

Recall that the quantities a_i are the proportion of juveniles committing their very first incidents of delinquency at age i in a given calendar year as compared to a base of *all* juveniles who are of age i during the given calendar year. Even though a child who became delinquent before age i clearly cannot become delinquent for the *first* time at age i, it is important to recognize that such a child is still considered part of the base population against which the rate

$$a_i = \frac{\textit{number of juveniles of age } \text{i and } \textit{delinquent for first time}}{\textit{number of juveniles of age } \text{i}} \tag{6}$$

is calculated.

In contrast, we might consider the proportion p_i of juveniles committing their very first incidents of delinquency at age i as compared to a base of juveniles of age i who were *not* previously delinquent. Although both of the fractions a_i and p_i have the same numerators, they differ in the base populations in their denominators. Whereas the quantity a_i is an unconditional rate (or probability) in that it refers to the *entire* cohort in question, the quantity p_i is a *conditional* rate (or probability), referring only to that restricted part of the cohort which has not become delinquent prior to age i. To estimate p_i when a cohort of N individuals is being observed over a period of years (from birth to age 18.0), we would first need to determine the number of individuals who had not become delinquent prior to age i, by subtracting the number D_i who already had from the total N. The number not yet delinquent *prior* to age i is $N-D_i = N-\sum_{j=0}^{i-1} m_j$. The quantity p_i is then the quotient of the number m_i of individuals who become delinquent for the first time and $N-D_i$. Thus, $p_0 = m_0/(N-0) = m_0/N = a_0$, and

$$p_i = \frac{m_i}{N-D_i} = \frac{a_i}{1-\sum_{j=0}^{i-1} a_j} \tag{7}$$

since $D_i = \sum_{j=0}^{i-1} m_j$ and $m_i = Na_i, i = 0, 1, \ldots, 17$.

Equation (7) expresses each p_i as a function of $a_0,a_1,\ldots,a_i$. We can invert these relationships and express each a_i in terms of $p_0,p_1,\ldots,p_i$. First note that $a_0 = p_0$. Then, making repeated use of Equation (7), we can show that $a_1 = p_1(1-p_0)$, $a_2 = p_2(1-p_0)(1-p_1)$, and in general that

$$a_i = p_i[\prod_{j=0}^{i-1}(1-p_j)]. \tag{8}$$

Thus, using Equation (3), we see that an alternative formula for the prevalence P_k of delinquency at age k is

$$P_k = \sum_{i=0}^{k} p_i[\prod_{j=0}^{k-1}(1-p_j)]. \tag{9}$$

The quantities p_i are customarily of interest in actuarial contexts where, instead of delinquency, mortality is of concern. In such contexts, an individual who meets the criterion (death) during a given age i is removed from the cohort for all later ages $j, j \geq i+1$. Calculating the rate of incidence of the criterion relative to a base consisting only of those individuals who have not previously met the criterion (i.e., calculating the p_i's) makes sense in such contexts, since records usually give base populations only of *survivors* at any given age. The quantities p_i also are meaningful in medical contexts where infection by a disease (such as measles) confers immunity against future infections; the base population of people who have not caught the disease is then the population still at risk for contracting the disease. For this reason, the quantities p_i might perhaps be called age-specific "risks" (of first delinquency).† In studies of product reliability (where failure or destruction of a product is of interest), the quantities p_i are sometimes called (age-specific) "hazard rates," which more clearly distinguishes the conditional rates p_i from the unconditional rates a_i. However, terminology suitable to the present context is preferable; thus, we will call the p_i's the "age-specific rates of first delinquency among nondelinquents."‡

† There may be some interest in using the sum $\sum_{i=0}^{k} p_i$ as a social indicator. From Equation (7) and the fact that the a_i's are numbers between 0 and 1, it follows that $p_i \geq a_i$ for each i, and thus that $\sum_{i=0}^{k} p_i \geq \sum_{i=0}^{k} a_i$, for any age k. Summing the p_i's does not make much sense, however, since each of the component "risks" refers to a different cohort (inasmuch as the base population for each is the changing number of survivors from year to year). This sum can exceed 1, while the prevalence cannot. The result of dividing such a sum by the number of years it spanned could be construed as the "mean risk" for the ages in question—for whatever that is worth.

It might be emphasized that in general the age-specific rate, a_i, is more informative as a social index than the hazard rate, p_i. Both concepts share in the sense of hazard, but the a_i emphasize the hazard to society (of so many members of a cohort becoming delinquent during a given age), whereas the p_i emphasize the hazard to the individual, given that he is not already delinquent.

‡ The failure to clearly distinguish in terminology between the rate a_i of first-incident of delinquency and the rate p_i of first-incident of delinquency for nondelinquents (often called the "risk" of first-incident of delinquency) seems to be endemic in the literature. For example, Ball *et al.* (1964, p. 92) call our age-specific first-incident rates a_i the "annual risks" or "age-specific risks" and call our age-specific first-incident rates p_i for nondelinquents the "age-specific rates," but then define their "age-specific rates" to be the "probability of (first) incident during i-th year of age," thus using the term they have previously used to refer to the population composed solely of nondelinquents to refer now o the entire cohort!

So far, our discussion of the age-specific rates p_i of first delinquency among nondelinquents has been confined to the case where a single cohort of individuals is observed from birth. If instead we try to estimate these quantities in the situation where we have age and delinquency data for a given calendar year, then a reasonable estimate of p_i can be formed by substituting the estimated age-specific rates $\hat{a}_i = n_i/N_i$ of first delinquency into Equation (7). Thus

$$\hat{p}_i = \frac{\hat{a}_i}{1-\sum_{j=0}^{i-1}\hat{a}_j} . \tag{10}$$

If these estimated "hazard rates" $\hat{p}_i$ are substituted into Equation (9), the resulting equation provides an alternate way of estimating the prevalence P. The result obtained in this way will always agree with the result obtained from Equation (4), although care must be taken not to confuse the various equations and the purpose for which they are intended.

Unfortunately, Monahan seems to have been laboring under exactly such a confusion. We have earlier seen how to estimate $\hat{P}_{18}$ from the data in Table 1 correctly. Now, let us examine Monahan's actual calculation, as revealed by Columns 5–7 of Table 1.

Column 5 postulates a hypothetical cohort of $N = 100{,}000$ individuals at age-year 6 (at which age there is as yet no delinquency by the criterion being used here), and it supposedly indicates the number $N-D_i$ of these individuals who have not become delinquent ("survivors") by the end of the age-year i given in column 1 on the same row. Column 6 purports to show the number, D_i, of the original 100,000 in the hypothetical cohort, becoming delinquent for the first time in each age-year i. It should be obvious from our previous discussion that if this were true, the entries in column 6 would be related exactly to the age-specific first-incident rates in column 4, except for a change in the decimal place. In other words, the *i-th* entry in column 6 would give the number newly becoming delinquent by simply applying the *i-th* rate $\hat{a}_{i+6}$ of column 4 to 100,000. The first two entries in column 6 are actually of such a value that they could have been obtained by this means. However, study of the entries in later rows of column 6 reveals that any parallelism with column 4 after the first row is due entirely to rounding-off, and not to the manner of computation. What Monahan is actually doing is applying the rates $\hat{a}_i$ of column 4 to the number of "survivors" as of the end of the preceding year, rather than to the appropriate unconditional base population of 100,000. Hence, after the first row, the number m_i of new delinquents is always underestimated, and the number of "survivors" is overestimated. The result is that instead of summing the numbers $m_i = a_i$ (100,000) of new delinquents to produce the entries in column 6, Monahan is creating entries for column 6 of the following kind: a_7 (100,000), $a_8(1-a_7)$(100,000), $a_9(1-a_8)(1-a_7)$(100,000), etc. In short, Monahan is treating the (estimated) age-specific rates $\hat{a}_i$ of first delinquency as if they were the (estimated) age-specific rates of first delinquency for nondelinquents! It is hard to see how the resulting index, formed by summing column 6 and dividing the sum by 100,000, measures delinquency in any meaningful way, since this index is calculated by mis-applying the (estimated) rates $\hat{a}_i$. In any case, it is clear that the index

$$a_7+\sum_{i=8}^{17}[a_i\prod_{j=7}^{i-1}(1-a_j)] \tag{11}$$

resulting from Monahan's computational method is always an *underestimate* of the true prevalence $P_{18}=\sum_{i=7}^{17}a_i$ of delinquency at 18.0. This underscores the fact that his error is a more subtle one than simply mistaking the overall conditional rate, $\sum_{i=0}^{k-1}p_i$, for the unconditional one, $\sum_{i=0}^{k-1}a_i$, since this would have led to a final "prevalence" that was too high (see p. 283, fn. †).

ASSESSING THE MAGNITUDE OF THE ERRORS

The individual terms

$$b_i=a_i\prod_{j=7}^{i-1}(1-a_j),\ i=8,\ldots,17, \tag{12}$$

other than the first term $b_7\equiv a_7$, in the sum in Equation (11) could be called "pseudo age-specific first-incident rates of delinquency" since these are the numbers which Monahan appears to be using in place of the age-specific first-incident rates of delinquency a_i in his calculations of prevalence. By the same logic, the index which is the sum of these terms (see Equation [11]) might well be called a "pseudo prevalence."

Note that the b_i bear the same mathematical relationship to the a_i as do the a_i to the p_i. Therefore, if we are given Monahan's pseudo age-specific rates b_i, $i=7, 8, \ldots, 17$, we can recover the real age-specific rates a_i by solving for the a_i in Equation (12). This yields $a_7=b_7$ and

$$a_i=\frac{b_i}{1-\sum_{j=7}^{i-1}b_j},\ i=8,\ldots,17. \tag{13}$$

Recall again that in the present context $a_0=a_1=\ldots=a_6=0$.

Monahan gives only one table (his Table 4; our Table 1) which is detailed enough so that we can directly obtain the true age-specific first-incident rates a_i of delinquency. However, he does present graphs (his Figure 2) of his pseudo age-specific first-

† To finish our description of Table 1, note that the entry in the row of column 7 corresponding to age k is equal to (1/100)th of the sum of those entries in column 6 parallel to or above that row. Thus, the entry in the row of column 7 corresponding to age 7 is 100 $\hat{a}_7$, and in general for the row of column 7 corresponding to age k is $100[\hat{a}_7+\sum_{i=7}^{k-1}\hat{a}_i\prod_{j=7}^{i-1}(1-\hat{a}_j)]$, $k=8,9,\ldots,18$. The last entry in column 7, when divided by 100, is thus an estimate of the index shown in Equation (11).

incident rates of delinquency b_i for cohorts of white girls, of Negro girls, of white boys, and of Negro boys, and of the total cohort of all juveniles calculated from data obtained from Philadelphia Municipal Court Records and Annual Reports of the Philadelphia Board of Education for the years 1949–54, combined as a composite set of data.† From these pseudo rates, it is possible (within the margin of error resulting from having to read the b_i's from the graphs, and from whatever errors Monahan made in plotting his graphs) to obtain the true rates a_i by use of Equation (13) above. The important results of our calculations appear in Table 2.

The first row in Table 2 gives our estimate of the prevalence of juvenile delinquency at age 18.0 for Philadelphia for a period covering the years 1949–54. The numbers appearing in each column of row 1 are the percentage equivalents of the sum $P_{18} = \sum_{i=7}^{17} a_i$ of the age-specific rates of first-incident of delinquency a_i specific to the particular race-sex category indicated at the top of the column. The a_i's for each column have been calculated from corresponding pseudo age-specific rates b_i read from Figure 2 of Monahan (1960), as described in Equation (13). In row 2 of Table 2 appear the percentage equivalents of the pseudo prevalences $\sum_{i=7}^{17} b_i$ that would have been calculated by Monahan if he had used the values of b_i which we read from his Figure 2, rather than his original tabled values. As a check of the accuracy of our reading of his values of b_i from his Figure 2, in row 5 of our Table 2 appear the percentage equivalents of the pseudo prevalences $\sum_{i=7}^{17} b_i$ which Monahan actually calculated (and which appear in Table 2 of his paper).

Comparing rows 4 and 5 of our Table 2, we see that there is a positive bias in our estimates of Monahan's pseudo age-specific rates (row 6). The values in row 4 always exceed the values in row 5; however, the largest difference between a value in row 4 and a value in the corresponding column of row 5 is 0.23. Since there are 11 different ages (ages 7, 8, 9, . . ., 17) over which each sum is calculated, we conclude that on the average there is a positive bias of no more than 0.02 percentage points in our reading of the pseudo age-specific rates b_i from Monahan's Figure 2. Some calculations (not presented) have convinced us that such a small error should result in errors in our estimate of prevalence in row 1 of at worst 0.4 percentage points.‡ In view of the small magnitude of these graph-reading errors and of their small effect on the final prevalence, no further effort to obtain a more refined reading of Monahan's Figure 2 was made.

† Monahan included cases that were dismissed or adjusted by court probation officers, following an arrest or the filing of an official complaint. Traffic offenses, per se, figured only negligibly in these cases. For additional information concerning the criterion of delinquency he employed, the reader is urged to consult Monahan (1960).

‡ The calculations that lead to this assertion stem from the assumption that the values of Monahan's pseudo age-specific first-incident rates b_i which we read from his Figure 2 always exceed the true values (those actually calculated by Monahan) by a fixed positive bias Δ equal to one-eleventh of the difference between the graphed pseudo prevalence and Monahan's tabled pseudo prevalence. For example, from row 6 of Table 2 (after converting percentages back to decimal form) we find that the Δ for "All juveniles" is $.001 \div 11 = .0001$. From the above assumption, it follows that our calculated

TABLE 2

Corrections obtained from Monahan's graphed 1949–54 data, and their application.

Datum	White girls	Negro girls	All juveniles	White boys	Negro boys
			At age 18.0		
1. Recalculated prevelance	3.53%	15.86%	15.75%	18.10%	50.98%
2. Graphed pseudo prevalence	3.48	14.84	14.71	16.73	40.92
3. Difference:	.05	1.02	1.04	1.37	10.06
4. Graphed pseudo prevalence	3.48	14.84	14.71	16.73	40.92
5. Tabled pseudo prevalence	3.3	14.8	14.6	16.5	40.8
6. Difference:	.18	.04	.11	.23	.12
7. Tabled pseudo prevalence	3.3	14.8	14.6	16.5	40.8
8. Correction	.05	1.02	1.04	1.37	10.06
9. Final prevalence	3.35	15.82	15.64	17.87	50.86
			At age 16.0		
10. Recalculated prevalence	2.43%	11.68%	11.19%	12.53%	36.67%
11. Graphed pseudo prevalence	2.41	11.16	10.69	11.89	31.46
12. Difference:	.02	.52	.50	.64	5.21
13. Graphed pseudo prevalence	—	—	10.69	—	—
14. Tabled pseudo prevalence	—	—	10.6	—	—
15. Difference:	—	—	.09	—	—
16. Tabled pseudo prevalence	2.24[a]	10.60[a]	10.6	11.26[a]	30.52[a]
17. Correction	.02	.52	.50	.64	5.21
18. "Final prevalence"	2.26	11.12	11.10	11.90	35.73
19. Recalculated prevalence	2.43	11.68	11.19	12.53	36.67
20. Prorated reading error	−.15	−.03	−.09	−.19	−.10
21. Final prevalence	2.28	11.65	11.10	12.34	36.57

[a] Since there are no tabled values for race- and sex-specific pseudo prevalences at age 16.0 in Monahan's report for the total period 1949–54, these values are the averages, in each case, of the values given for the five separate years, 1949, 1950, 1951, 1952, and 1953, in his Table 3.

value of the prevalence would always exceed the true value of the prevalence by an amount no greater than (and this is a very gross bound)

$$E = \frac{10P'\Delta}{1-P'+10\Delta} + \frac{11\Delta}{1-P'},$$

where P' is the recalculated value of the prevalence (see Table 2, row 1) expressed in decimal form. For "All juveniles," E equals .0015, while for "Negro boys" E equals .0036. Similar calculations for "White girls," "White boys," and "Negro girls" led to the stated assertion. It is worth noting that the error bound E usually greatly overestimates the true error.

However, to (hopefully) cancel the effects of the error involved in reading his Figure 2, we compare our estimates of prevalence in row 1 of our Table 2 to corresponding entries in row 2, rather than to Monahan's actual tabled estimates in row 5. Since the data in rows 1 and 2 all depend upon reading Monahan's graphs, presumably any constant error in this reading will be largely neutralized in the difference between these rows. The differences between the entries in rows 1 and 2 represent our estimate of Monahan's error, and they appear in row 3 of Table 2. These differences, which are always positive, are small for all populations except Negro boys, for whom there is a substantial error of 10.06 percentage points. This last error, of itself, bears out the need to correct Monahan's data.

The number in each column of row 3 of Table 2 can be regarded as a correction constant to be added to Monahan's estimate of "prevalence" which appears in the same column of row 5. These additions are performed in rows 7 and 8 of Table 2, and the revised estimates of prevalence at age 18.0 appear in row 9.

Since in some states juvenile court jurisdiction ends at 16, the prevalence of delinquency at age 16.0 is often of special interest. Monahan also singled out this age for presentation of prevalence rates. In the bottom half of Table 2, we give estimates of prevalence at age 16.0, calculated as before from Monahan's pseudo age-specific rates b_i obtained from his Figure 2. These estimates, which appear in row 10 of Table 2, are compared to the estimates of "prevalence" (row 11 of Table 2) which Monahan would have given if he had used the b_i's which we have taken from his Figure 2, rather than his tabled values, just as for age 18.0. However, our reading of the graph can be checked against his tabled value only at one point at age 16.0 because Monahan's paper gives only one estimate of prevalence at age 16.0 in its tables that corresponds to the time period in his graph (that for all juveniles during 1949–54—taken from his Table 1). Therefore, rows 13, 14 and 15 contain but one entry each. Comparing the one entry in row 13 with the corresponding entry in row 14, we obtain a difference of 0.09 percentage points, corresponding to an average bias of 0.01 percentage points in our reading of each of the pseudo age-specific rates b_i from Monahan's Figure 2. This average bias is exactly the same as the corresponding average bias for reading the pseudo age-specific rates for "all juveniles" over ages 7 through 17 (see row 6) from Monahan's Figure 2, indicating that the reading error is about the same for all ages.

In row 12 of Table 2, we give the difference $100\,(\sum_{i=7}^{15} a_i - \sum_{i=7}^{15} b_i)$ to indicate the errors that would have resulted if Monahan had used his method to estimate race-and-sex-specific prevalences of delinquency at age 16.0 for the period in question. Here again, only the error in calculating prevalence for Negro boys is large: 5.21 percentage points.

These corrections are applied in rows 16, 17 and 18. However, for four of the populations, we had to apply the corrections to pseudo prevalences that consist of the average of yearly values for the five years 1949–53, since Monahan tabled a value only for "all juveniles" for the composite 1949–54. This yields a set of "final prevalences" at age 16.0 in row 18 that are probably every bit as reliable as those at age 18.0 in row 9, although strictly speaking they do not apply to either the same data or the same period, exactly, except for "all juveniles." For four of these values in row 18, it would be better to take as estimates for the period 1949–54 proper the values in row 10, since these contain only the graph-reading error component, which is apparently the smallest of all the error components (the other two being those due to change in time and change in form of data). Finally, the estimates in row 10 can themselves be improved, by prorating the graph-reading errors in row 6 that are for ages 7 through 17 for the ages 7 through 15, and deducting those amounts (9/11's of row 6) from row 10, to obtain a final prevalence that truly applies to the composite data for 1949–54. This last step is accomplished in rows 19–21.

Table 3 presents the cumulative prevalence rates for 1949–54 for all ages, corrected for prorated graph-reading errors. These data are valuable for comparison purposes with data from other states employing different age ranges for defining juvenile court jurisdiction, as well as for showing the age-structure of the rates. Obviously, the column headed "all juveniles" is least valuable for comparison purposes, since it is composed of widely different sex- and race-specific rates. Although the sex-ratio is fairly constant from place to place and from one time to another, racial composition is not, and this would have to be taken into account in any such comparison. Since it is a simple matter to recover the age-specific rates a_i, already corrected for graph-reading errors, from the cumulative entries in Table 3, no special tabulation of these equally important data is presented. To facilitate such a recovery, entries are given to enough decimal places so that no age-specific rate appears to be zero simply as a consequence of rounding errors.

TABLE 3

Cumulative prevalence of delinquency, Philadelphia 1949–54, corrected for prorated graph-reading errors (in percent)

At age:	White girls	Negro girls	All juveniles	White boys	Negro boys
8.0	.006%	.033%	.130%	.145%	.594%
9.0	.016	.096	.386	.490	1.663
10.0	.053	.258	.893	1.137	3.636
11.0	.097	.587	1.614	2.051	6.251
12.0	.145	1.253	2.563	3.249	9.898
13.0	.329	2.464	3.955	4.758	14.834
14.0	.711	4.648	5.789	6.694	21.028
15.0	1.426	7.893	8.288	9.350	28.374
16.0	2.287	11.643	11.105	12.342	36.576
17.0	2.883	14.153	13.553	15.170	44.539
18.0	3.347	15.821	15.645	17.865	50.864

The final row in Table 3 brings home the importance of controlling for race in data of this kind. Given the intrinsic race-specific rates in the last row, we would expect composite prevalences at age 18.0 to range from 3.3 percent to 15.8 percent for girls, and from 17.9 percent to 50.9 percent for boys, as racial composition varied from all white to all black, *simply as a function of this composition alone.*

APPLYING CORRECTIONS TO MONAHAN'S DATA FOR OTHER YEARS

Comparing rows 8 and 9 in Table 2, we see that the relative errors involved in using Monahan's method for calculating prevalence go up from a low of 1.49 percent, 100(.05 ÷ 3.35), for white girls, through a middle range of approximately 6.9 percent for Negro girls, all juveniles, and white boys, to a high of 19.78 percent, 100(10.06 ÷ 50.86), for Negro boys. Seen another way, the relative errors involved in using Monahan's method go up roughly exponentially with the final (or true) prevalence. Thus, there can be no one fixed universal correction factor by which to multiply Monahan's tabled data for individual years. This assertion is further borne out by inspection of the results in Table 2, rows 20 and 21. There is also no single functional relationship between the estimates Monahan presented and the actual prevalence, since the relationship is dependent in part on the structure of the age-specific rates underlying each sum.

The estimates of prevalence presented by Monahan for individual years show a high degree of stability over the time series, however. For example, "prevalences" at age 18.0 over his eight-year (1946 to 1953) series (except for all juveniles, which affords a twelve-year series to 1957) have the following standard deviations, in percentage points:

White Girls	Negro Girls	All Juveniles	White Boys	Negro Boys
0.32	2.25	1.91	1.52	1.88

Thus, the correction constants we have given (Table 2, rows 3 and 12) could probably be added to the "prevalence" of the appropriate population for any of these years with fair security, since the "prevalence" for each population tends to remain at about the same level. This assumes that all of the estimates of "prevalence" which Monahan gives are based on pseudo age-specific rates b_i, the graphs over ages i of which resemble the graphs of corresponding populations in Monahan's Figure 2. The last part of this assumption is a reasonably secure one in view of certain common properties of age-specific rates for delinquency: the tendency to rise gradually during early years, and to peak around age-year 15 and fall off slightly thereafter. All five populations show this pattern in Monahan's Figure 2. Furthermore, our correction constants were derived from Monahan's most important set of data, the composite of years 1949–54; consequently they should be fairly stable themselves, and apply in an average sense to the individual years of this period, as well as to adjacent years, if one desired to correct these data too. Alternatively, one could fit an exponential curve to our constants as a function of pseudo prevalence (separate curves for ages 16.0 and 18.0), and read new corrections off of the curve—but this is probably more trouble than it is worth compared to the method we have suggested here.

SUMMARY

We have discussed the concept of prevalence in a general way, clarified the differences between various rates that are employed in the calculation of prevalence, and have

explicated an error made by Monahan (1960) in estimating prevalence in his valuable paper. The magnitude of Monahan's error has been examined, and it has been found to be substantial in the case of Negro boys. Corrections are given for prevalence estimates in Monahan's most important series of data, the composite for the years 1949–54, and suggestions are made as to how these might be applied to the remainder of his data, for individual years, if these results are required. The corrected data indicate that composite delinquency statistics that fail to control for the characteristics race and sex are apt to be quite meaningless for making comparisons when the proportion of the populations possessing any of these characteristics is subject to appreciable variation.

REFERENCES

Ball, J. C., A. Ross, and A. Simpson (1964) Incidence and estimated prevalence of recorded delinquency in a metropolitan area. *American Sociological Review* 29:90–3.

Durkheim, E. (1951) *Suicide.* The Free Press, New York.

Geis, G. (1965) Statistics concerning race and crime. *Crime and Delinquency* 11:142–50.

Gordon, R. A. (1973) An explicit estimation of the prevalence of commitment to a training school, to age 18, by race and by sex. *Journal of the American Statistical Association* 68: 547–53.

Hathaway, S. R. and E. D. Monachesi (1963) *Adolescent personality and behavior; MMPI patterns of normal, delinquent, dropout, and other outcomes.* University of Minnesota, Minneapolis.

Lunden, W. A. (1964) *Statistics on delinquents and delinquency.* Charles C. Thomas, Springfield, Ill.

Monahan, T. P. (1960) On the incidence of delinquency. *Social Forces* 39: 66–72.

The Board of Public Education (1950) *Annual Report of the Board of Public Education, School District of Philadelphia.*

The President's Commission on Law Enforcement and Administration of Justice (1967) *The Challenge of crime in a free society.* U.S. Government Printing Office, Washington, D.C.

U.S. Bureau of the Census (1952) *U.S. census of population:* 1950. *Vol. II. Characteristics of the population, part 38, Pennsylvania.* U.S. Government Printing Office, Washington, D.C.

Van den Berghe, P. L. (1971) The benign quota: Panacea or Pandora's box. *American Sociologist* 6 (Supplementary Issue): 40–3.

Wolfgang, M. E., R. M. Figlio, and T. Sellin (1972) *Delinquency in a birth cohort.* University of Chicago Press, Chicago.