Pergamon

Socio-Econ. Plann. Sci. Vol. 31, No. 4, pp. 281–291, 1997

Printed in Great Britain

PII: S0038-0121(97)00011-6

0038-0121/97 $17.00 + 0.00

Snapshot Samples

EDWARD H. KAPLAN*

Yale School of Management, and Yale School of Medicine, Box 208200, New Haven, Connecticut 06520-8200, U.S.A.

Abstract—We consider a coverage model where an initial event that occurs at some point in time triggers an activity of random duration that leads to some subsequent event. A *snapshot sample* is constructed at a fixed point in chronological time either by sampling only subjects where the initial event has occurred but the subsequent event has yet to occur (*active subjects*), or by sampling only subjects where both the initial and subsequent events have occurred (*inactive subjects*). The biases inherent in snapshot sampling can be neatly characterized by the properties of two random variables: the *history* $\mathscr{H}$ (defined as the time the initial event occurs as measured into the past from the chronological point of sampling), and the *active time* $\mathscr{A}$ (defined as the length of time between the initial and subsequent events). Though snapshot samples are biased, recognizing the biases enables correct inferences to be drawn from snapshot-sampled data. Considering only the case where $\mathscr{H}$ and $\mathscr{A}$ are independent random variables, this paper presents the probability models associated with snapshot sampling, demonstrates the problems that can occur, offers procedures for overcoming these problems, and applies the methods to interesting data sets. © 1997 Elsevier Science Ltd

INTRODUCTION AND MOTIVATION

Biased samples from dynamic coverage processes are common in the social, policy and management sciences, among other disciplines. For examples, AIDS epidemiologists have conducted studies of already diagnosed AIDS patients with the hope of gaining insights applicable to all those infected with the AIDS virus HIV [1–3], economists have surveyed the currently unemployed with the hope of learning about the ever unemployed [4–6], and management scientists have observed existing customers in service systems with the intent of arriving at general statements about all customers [7, 8]. These three examples share two important features. First, in each case, events of interest occur at an initial point in time (HIV infection, unemployment, arrival to the service system) that trigger an interval of random duration (AIDS incubation time, time out of the work force, service time) leading to some subsequent event (AIDS diagnosis, new employment, service completion). Second, the samples drawn are either of subjects for whom the random interval in question has expired as of the date of sampling (HIV infected persons with an AIDS diagnosis), or the samples are of subjects at an arbitrary point within the interval as of the date of sampling (the currently unemployed, or customers currently in service). Consequently, these *snapshot samples*, taken at a fixed point in chronological time, contain biases that distort whatever inferences one might wish to draw regarding *all* members of the population. However, understanding these biases enables their correction. Indeed, correcting sampling biases is an important application of probability modeling [9, 10].

In the sections that follow, we present our basic coverage model and characterize the biases inherent in snapshot samples. We present examples throughout to illustrate how these biases operate. We then discuss procedures for inferring the correct probabilities of interest from such biased samples, and illustrate with data sets constructed naturally via snapshot sampling.

THE COVERAGE MODEL

We condition our entire discussion on the construction of a *snapshot sample* at a fixed point in time of a dynamic coverage process that is occurring over time. There are two random variables

*Author for correspondence: *Phone*: 203-432-6031. *Fax*: 203-432-9995. *e-mail*: edward.kaplan@yale.edu

of interest. The first of these is the *history*, which we denote by $\mathscr{H}(\mathscr{H} > 0)$. Random variable $\mathscr{H}$ reports the time an initial event occurs measured into the past from the time of sample construction. Note that the histories of all subjects for whom the initial event has occurred by the time of sampling, and not just those included in the snapshot sample, are independently and identically distributed as random variable $\mathscr{H}$. We denote the probability density function, cumulative distribution, and survivor distribution of $\mathscr{H}$ by $h(x)$, $H(x)$, and $\bar{H}(x)$, respectively (so the survivor distribution $\bar{H}(x) = \Pr\{\mathscr{H} > x\} = \int_x^\infty h(u)du$, for example).

The second random variable we require is the *active time*, which we denote by $\mathscr{A}(\mathscr{A} > 0)$. The active time is the length of time from the occurrence of the initial event until the subsequent event (e.g. the time between HIV infection and AIDS). The active times of all subjects (and not just sampled subjects) are independently and identically distributed as random variable $\mathscr{A}$. Also, note that random variable $\mathscr{A}$ is not affected by the date of sample construction. We denote the probability density function, cumulative distribution, and survivor distribution of $\mathscr{A}$ by $a(x)$, $A(x)$, and $\bar{A}(x)$, respectively.

Other than assuming that the functions we have discussed thus far are well defined, our only additional assumption is that the random variables $\mathscr{H}$ and $\mathscr{A}$ are independent. This assumption is equivalent to assuming that the distribution of active time remains stable over chronological time. In the service system scenario, for example, this implies that the distribution of service times does not change over chronological time, a reasonable assumption for many service systems (and a standard assumption in queueing theory). In the AIDS example, the assumption implies that the incubation time distribution has not changed over calendar time. This is an assumption that was reasonable for the first several years of the epidemic, but is now more questionable given the advent of new life-prolonging medical treatments. While more general models could be developed that dispense with the assumed independence of $\mathscr{H}$ and $\mathscr{A}$, the insights provided by the case of independence are of sufficient interest to warrant consideration in their own right.

Given our notation, there are two types of snapshot samples that can be taken at a fixed point in chronological time. Thus, a subject is defined to be *active* if $\mathscr{A} \geq \mathscr{H}$, and *inactive* if $\mathscr{A} < \mathscr{H}$. One can therefore construct a sample of active subjects, or a sample of inactive subjects. An active sample consists of subjects for whom the initiating event has already occurred but the subsequent event has yet to occur, while an inactive sample involves those for whom both the initiating and subsequent events have already occurred. In the AIDS context, for example, a sample of active subjects would consist of HIV positives who have yet to progress to AIDS, while a sample of inactive subjects would consist of HIV positives who already have AIDS.

SAMPLING PROBABILITIES FOR SNAPSHOT SAMPLES

We wish to compute the probability that a subject is eligible for sampling. As discussed, there are two cases corresponding to samples of active or inactive subjects. Consider first samples of active subjects. The probability σ_a that a subject can be included in a snapshot sample of active subjects is given by

$$\sigma_a = Pr\{\mathscr{A} \geq \mathscr{H}\} = \int_0^\infty h(x)\bar{A}(x)dx = E(\bar{A}(\mathscr{H}))\,. \tag{1}$$

Alternatively,

$$\sigma_a = \int_0^\infty a(x)H(x)dx = E(H(\mathscr{A}))\,. \tag{2}$$

Both of these expressions for the sampling probability will prove useful later on.

At the point in time a snapshot sample is constructed, all subjects in the population are either active or inactive. Thus, the probability σ_i that a subject can be included in a snapshot sample of inactive subjects is given simply by

$$\sigma_i = Pr\{\mathscr{A} < \mathscr{H}\} = 1 - \sigma_a\,, \tag{3}$$

which also equals

$$\sigma_i = E(A(\mathscr{H})) = E(\tilde{H}(\mathscr{A})) \tag{4}$$

as in eqns (1) and (2).

SAMPLED ACTIVE TIMES

Let $\mathscr{A}^*$ denote the active time for subjects included in a snapshot sample. What are the properties of $\mathscr{A}^*$ and how do they compare to the properties of $\mathscr{A}$? For example, in the AIDS context, how will the distribution of AIDS incubation times in a snapshot sample compare to the true probability distribution of AIDS incubation times? We denote the probability density, distribution, and survivor functions of $\mathscr{A}^*$ by $a^*(x)$, $A^*(x)$, and $\bar{A}^*(x)$, respectively.

ACTIVE SUBJECTS

We first consider samples of active subjects. In such samples, a subject with active time equal to x can only appear in the sample if its history is at most equal to x; this occurs with probability $H(x)$. Given the independence of $\mathscr{A}$ and $\mathscr{H}$, the probability density function describing the active time of sampled active subjects equals

$$a^*(x) = \frac{a(x)H(x)}{\sigma_a} = \frac{a(x)H(x)}{E(H(\mathscr{A}))} \quad \text{for } x > 0\,. \tag{5}$$

Eqn (5) concisely captures the bias of snapshot samples of active subjects, in that the likelihood of sampling a subject with an active time of x is proportional not only to the fraction of all subjects with active time x (which is $a(x)$), but also to the fraction of all subjects with a history of at most x (which is $H(x)$). As $H(x)$ is an increasing function of x, we see from eqn (5) that the snapshot sample is biased towards longer active times than occur in nature.

EXAMPLE: LENGTH-BIASED SAMPLING

Suppose that we allow $\mathscr{A}$ to follow an arbitrary distribution, but we specify

$$h(x) = \begin{cases} \frac{1}{\tau} & 0 < x \leq \tau \\ 0 & x > \tau \end{cases} \tag{6}$$

and, hence,

$$H(x) = \begin{cases} \frac{x}{\tau} & 0 < x \leq \tau \\ 1 & x > \tau \end{cases} \tag{7}$$

We see from eqn (2) that

$$\sigma_a = E(H(\mathscr{A})) = \int_0^\infty a(x)\frac{x}{\tau}\,dx + \bar{A}(\tau) = \frac{1}{\tau}E(\mathscr{A}|\mathscr{A} \leq \tau)A(\tau) + \bar{A}(\tau) \tag{8}$$

and, consequently, from eqn (5)

$$a^*(x) = \frac{xa(x)}{E(\mathscr{A}|\mathscr{A} \leq \tau)A(\tau) + \tau\bar{A}(\tau)} \quad \text{for } 0 < x \leq \tau \tag{9}$$

and $a^*(x) = a(x)/\sigma_a$ for $x > \tau$. Taking the limit of eqn (9) as $\tau \to \infty$, we recover

$$a^*(x) = \frac{xa(x)}{E(\mathscr{A})} \quad \text{for } x > 0\,, \tag{10}$$

which is, of course, the density function for active time that would follow from *length-biased sampling* [11], also known as *random incidence* [12]. Thus, we see how the bias in snapshot samples of active subjects generalizes length-biased sampling.

EXAMPLE: TRANSIENT BEHAVIOR OF AN M/G/∞ QUEUE

Suppose that an infinite server queueing system with service times given by $\mathscr{A}$ and Poisson arrivals with rate λ began operations with no customers in service τ time units ago. Let $m(\tau)$ denote the average number of customers in service now. As customers in service comprise a snapshot sample of all customers who have arrived since the system began operations, and, since the average number of arrivals over τ time units equals $\lambda\tau$, we have

$$m(\tau) = \lambda\tau\sigma_a\,, \tag{11}$$

where σ_a is the sampling probability. Viewed as a snapshot sample of active subjects, the history corresponds to the time in the past at which customers arrived, where $h(x)$ will follow the uniform density of eqn (6) as a result of the homogeneous Poisson process assumed. Consequently, σ_a is given by eqn (8). The average number of customers in service thus equals

$$m(\tau) = \lambda[E(\mathscr{A}|\mathscr{A} \le \tau)A(\tau) + \tau\bar{A}(\tau)]\,, \tag{12}$$

which increases from 0 to $\lambda E(\mathscr{A})$ as τ grows. Also, the probability density describing the (elapsed plus remaining) service times of customers currently in service corresponds to $a^*(x)$, as given in eqn (9). Eqn (12) shows how the average number of customers present in an $M/G/\infty$ queue evolves over time, and is equivalent to equation (20) in [8].

More generally, consider an $M_t/G/\infty$ queue that began operations with an empty system τ time units ago. Let $\lambda(x)$ denote the Poisson arrival rate x time units into the past, set $h(x) = \lambda(x)/\int_0^\tau \lambda(u)du$ for $0 < x \le \tau$ and 0 otherwise, and define $\bar{\lambda}(\tau) = 1/\tau \int_0^\tau \lambda(u)du$. The average number of customers currently in the system is then given by

$$m(\tau) = \bar{\lambda}(\tau)\tau\sigma_a = \bar{\lambda}(\tau)\tau E(H(\mathscr{A}))\,, \tag{13}$$

which is equivalent to equation (3) in [8]. The density $a^*(x)$ from eqn (5) continues to correspond to the density of (elapsed plus remaining) service times for those customers currently in service.

EXAMPLE: AN EXPONENTIAL EPIDEMIC WITH EXPONENTIAL ACTIVE TIME

Suppose that we specialize $h(x) = re^{-rx}$ for $x > 0$. This would characterize the history in an exponentially growing epidemic with growth rate given by r; the times of infection viewed into the past from the point of sampling follow an exponential distribution with mean $1/r$. Similarly, set $a(x) = \mu e^{-\mu x}$ for $x > 0$; that is, assume that persons remain active for exponentially distributed lengths of time following infection. This would correspond to a disease with a constant death rate, for example, or, alternatively, to a latent infection where persons face an exponential amount of time from infection to the development of symptoms. From (5), we have

$$a^*(x) = \frac{\mu e^{-\mu x}(1 - e^{-rx})}{r/(r + \mu)} \quad \text{for } x > 0\,, \tag{14}$$

which is clearly not exponential. Note that if r is very large, so that $H(x) = 1 - e^{-rx} \approx 1$, then $a^*(x) \to a(x)$, which *is* unbiased. This means that the epidemic is rising so rapidly that virtually all those ever infected became infected very recently. Thus, the snapshot sample corresponds to a sample of the *newly* infected (which will have the same properties as a sample of the ever infected).

On the other hand, if r is so small that $H(x) \approx rx$ and $r + \mu \approx \mu$, then $a^*(x) \to \mu^2 x e^{-\mu x}$, which is an Erlang distribution of order two. This is exactly the density one obtains from length-biased sampling of exponential random variables; for, if r is very small, then the incidence of infection is approximately constant over time.

MOMENTS OF SAMPLED ACTIVE TIME FOR ACTIVE SUBJECTS

Equation (5) enables computation of the moments of $\mathscr{A}^*$ for active subjects. For the mean, we thus have

$$E(\mathscr{A}^*) = \int_0^\infty x a^*(x)dx = \frac{E(\mathscr{A}H(\mathscr{A}))}{E(H(\mathscr{A}))} \geq E(\mathscr{A}) \tag{15}$$

with the last inequality following from the fact that $\mathscr{A}$ and $H(\mathscr{A})$ are positively correlated (for $H(x)$ is a distribution function and, hence, monotonically increasing). Thus, the mean sampled active time will be at least as large as the mean active time occurring in the population. More generally, we have

$$E(\mathscr{A}^{*k}) = \frac{E(\mathscr{A}^k H(\mathscr{A}))}{E(H(\mathscr{A}))} \geq E(\mathscr{A}^k)\,. \tag{16}$$

EXAMPLE: LENGTH-BIASED SAMPLING

Let us apply eqn (16) to the length-biased sampling example discussed earlier. Referring to eqn (9), we have

$$E(\mathscr{A}^{*k}) = \frac{E(\mathscr{A}^{k+1}|\mathscr{A} \leq \tau)A(\tau) + E(\mathscr{A}^k|\mathscr{A} > \tau)\tau\bar{A}(\tau)}{E(\mathscr{A}|\mathscr{A} \leq \tau)A(\tau) + \tau\bar{A}(\tau)} \overset{\tau\to\infty}{\to} \frac{E(\mathscr{A}^{k+1})}{E(\mathscr{A})} \tag{17}$$

as is well known.

EXAMPLE: EXPONENTIAL EPIDEMIC WITH EXPONENTIAL ACTIVE TIME

Once again, let $h(x) = re^{-rx}$ and $a(x) = \mu e^{-\mu x}$. From eqn (15), the mean active time is given by

$$E(\mathscr{A}^*) = \frac{E(\mathscr{A}(1 - e^{-r\mathscr{A}}))}{E(1 - e^{-r\mathscr{A}})} = \frac{1/\mu - \mu/(r+\mu)^2}{r/(r+\mu)} = \frac{r + 2\mu}{\mu(r+\mu)}\,. \tag{18}$$

Again, we see how this particular example can generate results that range from unbiased (when r becomes large and $E(\mathscr{A}^*)\to 1/\mu$), to length-biased (when r becomes small and $E(\mathscr{A}^*)\to 2/\mu$).

SAMPLED ACTIVE TIMES: INACTIVE SUBJECTS

Samples of inactive subjects are biased towards shorter active times than occur in nature. Analogous to the reasoning behind eqn (5), an inactive subject with an active time equal to x can only be included in a snapshot sample if $\mathscr{H} > x$. Given that $\mathscr{A}$ and $\mathscr{H}$ are independent, we have

$$a^*(x) = \frac{a(x)\bar{H}(x)}{\sigma_i} = \frac{a(x)\bar{H}(x)}{E(\bar{H}(\mathscr{A}))} \quad \text{for } x > 0 \tag{19}$$

for the density of sampled active times, with moments given by

$$E(\mathscr{A}^{*k}) = \frac{E(\mathscr{A}^k\bar{H}(\mathscr{A}))}{E(\bar{H}(\mathscr{A}))} \leq E(\mathscr{A}^k)\,. \tag{20}$$

EXAMPLE: UNIFORM HISTORY

Consider again the uniform history density of eqn (6) and an arbitrary distribution of active time. Recall that the probability a subject is eligible for inclusion in a snapshot sample of *active* subjects was found in eqn (8). The sampling inclusion probability for *inactive* subjects thus equals

$$\sigma_i = 1 - \sigma_a = 1 - \left(\frac{1}{\tau}E(\mathscr{A}|\mathscr{A} \leq \tau)A(\tau) + \bar{A}(\tau)\right) = A(\tau)\left(1 - \frac{1}{\tau}E(\mathscr{A}|\mathscr{A} \leq \tau)\right)\,. \tag{21}$$

Instead of a length-biased sample, this snapshot of inactive subjects yields

$$a^*(x) = \frac{a(x)(\tau - x)}{A(\tau)(\tau - E(\mathscr{A}|\mathscr{A} \le \tau))} \quad \text{for } 0 < x \le \tau\,, \tag{22}$$

and $a^*(x) = 0$ otherwise. As $\tau \to \infty$, we have $a^*(x) \to a(x)$ and the sample becomes unbiased.

EXAMPLE: EXPONENTIAL EPIDEMIC WITH EXPONENTIAL ACTIVE TIME

As before, we take $h(x) = re^{-rx}$ and $a(x) = \mu e^{-\mu x}$. Application of eqn (19) yields the formula

$$a^*(x) = \frac{\mu e^{-\mu x} e^{-rx}}{\mu/(r+\mu)} = (r+\mu)e^{-(r+\mu)x} \quad \text{for } x > 0\,. \tag{23}$$

eqn (23) shows that the sampled active time also follows an exponential distribution, but with a mean of $1/(r+\mu)$ instead of $1/\mu$. In fact, one would get the same result if an epidemic was growing exponentially with rate μ and subjects remained active for exponentially distributed times with mean $1/r$. This example shows how two completely different situations can combine to yield the same sampled active time distribution, again demonstrating the general danger of interpreting active time distributions directly from snapshot samples.

SAMPLED HISTORIES: ACTIVE SUBJECTS

Now, consider the snapshot sampled history for active subjects. In the AIDS context, this history corresponds to the time since HIV infection for those who have yet to progress to AIDS. Note that this time is typically unknown for individuals in the sample. We denote the sampled history by $\mathscr{H}^*$, and, similar to our earlier notation, we let $h^*(x)$, $H^*(x)$ and $\bar{H}(x)$ denote the density, cumulative distribution, and survivor function, respectively, for random variable $\mathscr{H}^*$. Given the independence of $\mathscr{A}$ and $\mathscr{H}$, the density function $h^*(x)$ equals

$$h^*(x) = \frac{h(x)\bar{A}(x)}{\sigma_a} = \frac{h(x)\bar{A}(x)}{E(\bar{A}(\mathscr{H}))} \quad \text{for } x > 0\,. \tag{24}$$

As is clear from eqn (24), the sampling process leads to a shorter collection of histories than has occurred in nature. Thus, in the AIDS setting, those in the sample will tend to have been infected for *shorter* periods of time than those in the population at large.

EXAMPLE: LENGTH-BIASED SAMPLING

Again, we allow $\mathscr{A}$ to follow an arbitrary distribution while $h(x)$ is defined as in eqn (6). From eqn (24), we obtain

$$h^*(x) = \frac{\bar{A}(x)}{E(\mathscr{A}|\mathscr{A} \le \tau)A(\tau) + \tau\bar{A}(\tau)} \quad \text{for } 0 < x \le \tau \tag{25}$$

and $h^*(x) = 0$ for $x > \tau$. As $\tau \to \infty$, we see that $h^*(x) \to \bar{A}(x)/E(\mathscr{A})$, which we recognize as the backwards recurrence time density for an equilibrium renewal process with interarrival times given by $\mathscr{A}$ [11]. Clearly, this sampled density bears little resemblance to the true $h(x)$. Note that the density $h^*(x)$ would also result from sampling the arrival times for customers currently present in an $M/G/\infty$ queue, assuming that the queueing facility began operations with an empty system τ time units ago.

MOMENTS OF SAMPLED HISTORY FOR ACTIVE SUBJECTS

Consider now the moments of the sampled history $\mathscr{H}^*$. For the mean sampled history we have

$$E(\mathscr{H}^*) = \int_0^\infty x h^*(x)dx = \frac{E(\mathscr{H}\bar{A}(\mathscr{H}))}{E(\bar{A}(\mathscr{H}))} \le E(\mathscr{H}) \tag{26}$$

with the inequality following from the negative correlation between $\mathscr{H}$ and $\bar{A}(\mathscr{H})$ (for $\bar{A}(x)$ is a survivor function). More generally, we have

$$E(\mathscr{H}^{*k}) = \frac{E(\mathscr{H}^k \bar{A}(\mathscr{H}))}{E(\bar{A}(\mathscr{H}))} \leq E(\mathscr{H}^k) . \tag{27}$$

EXAMPLE: EXPONENTIAL EPIDEMIC WITH EXPONENTIAL ACTIVE TIME

Again, we take $h(x) = re^{-rx}$ and $a(x) = \mu e^{-\mu x}$. From eqn (26), we have

$$E(\mathscr{H}^*) = \frac{E(\mathscr{H} e^{-\mu\mathscr{H}})}{E(e^{-\mu\mathscr{H}})} = \frac{r/(r+\mu)^2}{r/(r+\mu)} = \frac{1}{r+\mu} \tag{28}$$

which, of course, is the mean of an exponential distribution with rate $r + \mu$. In this example, the sampled history distribution from snapshot samples of active subjects is the same as the sampled active time distribution from snapshot samples of inactive samples (as derived in eqn (23).

SAMPLED HISTORIES: INACTIVE SUBJECTS

The bias in snapshot-sampled histories from inactive subjects serves to overestimate the magnitude of histories relative to their true occurrence in nature. The density and moments for $\mathscr{H}^*$ in snapshot samples of inactive subjects are given by

$$h^*(x) = \frac{h(x)A(x)}{E(A(\mathscr{H}))} \text{for } x > 0 \tag{29}$$

and

$$E(\mathscr{H}^{*K}) = \frac{E(\mathscr{H}^k A(\mathscr{H}))}{E(A(\mathscr{H}))} \geq E(\mathscr{H}^k) . \tag{30}$$

ESTIMATION FROM SNAPSHOT-SAMPLED DATA

The previous sections have demonstrated the sorts of biases one will encounter in snapshot samples. Of course, recognizing these biases provides the basis for correcting them. Given that snapshot samples are often much easier to construct than are true random samples, estimation from snapshot-sampled data can be an attractive approach to inference.

ESTIMATION WITH $a(x)$ OR $h(x)$ KNOWN

Suppose that one seeks to estimate the probability distribution of $\mathscr{H}$ from a snapshot sample given knowledge of $a(x)$. This is easily accomplished via eqn (24) (for active subjects) or eqn (29) (for inactive subjects) after equating the observed fraction of cases in the snapshot sample with histories of x equal to $h^*(x)$ and solving for $h(x)$. In the case of active subjects, this manipulation yields

$$h(x) = \frac{h^*(x)}{\bar{A}(x)} \sigma_a = \frac{h^*(x)/\bar{A}(x)}{\int_0^\infty h^*(u)/\bar{A}(u)du} \quad \text{for } x > 0 , \tag{31}$$

with the latter expression following from the fact that $h(x)$ is a probability density. In practice, one would replace the integral by a sum and estimate $h(x)$ as a discrete probability mass function (and interpret $\bar{A}(x) = \Pr\{A \geq x\}$). Analogous substitutions hold for estimating $a(x)$ from $a^*(x)$ assuming that $h(x)$ is known.

EXAMPLE: AN EPIDEMIC OF INJECTING DRUG USE

The New Haven needle exchange is an intervention designed to slow the spread of the AIDS virus among drug injectors [13, 14]. Participants in this program were asked at entry for both their current age and their age at the time they began injecting drugs. Treating the initiation of drug injection as an initial event, and cessation of drug use as a subsequent event, this survey comprises a snapshot sample of active drug injectors: only persons who initiated drug injection prior to the needle exchange and have yet to cease injecting are eligible for sampling. Each injector sampled provides an observed history (given by the difference between current age and age at initiation of injection), where the frequency distribution of these histories is the sample analogue of $h^*(x)$ for active subjects. Of interest is the 'true' history of the epidemic of injection drug use in New Haven.

Independently of these data, Caulkins and Kaplan [15] modeled the duration of drug injection careers based on national heroin data as following an exponential distribution with a mean of 7.5 y. For this example, we consider the active time as equivalent to a career of drug injection, and equate $a(x)$ to a geometric distribution with a mean of 7.5 y in order to facilitate discrete computations (so $a(x) = (7.5 - 1)^{x-1}/7.5^x$ for $x = 1,2,3\ldots$).

Focusing only on men attending the needle exchange, 798 injectors provided data enabling construction of $h^*(x)$ for $x = 1,2,\ldots, 33$. (We are thus technically conditioning on $\mathscr{H} \leq 33$.) The sampled $h^*(x)$ (as well as a statistical smooth) appear in Fig. 1, along with $h(x)$ (and a statistical smooth) as derived via eqn (31) (with the integral replaced by a sum). The differences in the two history distributions are striking: while the raw data would suggest that many injectors began injecting in the recent past (with the mean duration of injection reported equaling 13.7 y), the snapshot-corrected data show that the epidemic of drug injection peaked nearly 25 y before (with a mean duration of injection equal to 22.2 y). The analysis also shows that those drug users who remained active long enough to participate in the needle exchange are rather rare; for (from eqn (31)), the data provide the estimate $\sigma_a \approx 7.3\%$. (Thus, only about 7% of those who initiated drug injection in New Haven within the last 33 y could have been sampled.) Correcting the snapshot bias in these data suggests that, at least in New Haven, the initiation rate of new drug injectors has largely waned.

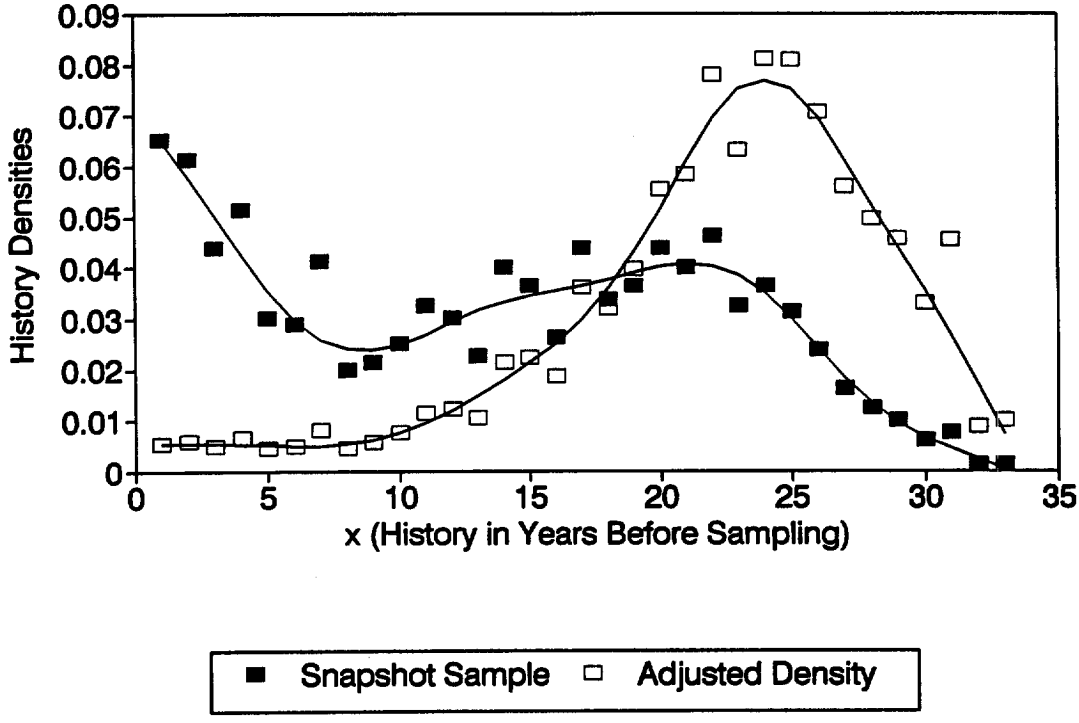

Fig. 1. Initiation of drug injection (source: New Haven Needle Exchange).

BOTH $a(x)$ AND $h(x)$ UNKNOWN

Now, consider a snapshot sample where subject histories and active times are both available, thus enabling the construction of $a^*(x)$ and $h^*(x)$, where the problem is to use the biased sample to estimate $a(x)$ and $h(x)$. Specifically, suppose we have a snapshot sample of *inactive* subjects, and that we have grouped the data (i.e. formed the relevant histograms) to produce $a^*(x)$ and $h^*(x)$. Analogous to eqn (31), we have (from eqns (19),(29)),

$$a(x) = \frac{a^*(x)}{\bar{H}(x)}\sigma_i = \frac{a^*(x)/\bar{H}(x)}{\int_0^\infty a^*(u)/\bar{H}(u)du} \quad \text{for } x > 0 \tag{32}$$

and

$$h(x) = \frac{h^*(x)}{A(x)}\sigma_i = \frac{h^*(x)/A(x)}{\int_0^\infty h^*(u)/A(u)du} \quad \text{for } x > 0\,. \tag{33}$$

Upon replacing the integrals by sums (and interpreting $\bar{H}(x) = \Pr\{\mathscr{H} \geq x\}$ while $A(x) = \Pr\{\mathscr{A} \leq x\}$), eqns (32) and (33) can be solved iteratively by first assuming a mass function that places non-negative weight at all points x for $h(x)$, forming the associated survivor function $\bar{H}(x)$, computing $a(x)$ from (32), using the computed $a(x)$ to form $A(x)$ in (33), and then, finally, using (33) to produce a new estimate of $h(x)$ for use in (32). The convergence of this method is assured as the discrete versions of eqns (32) and (33) are special instances of the more general model of quasi-independence in incomplete two-way tables (where the estimates $\hat{a}(x)$ and $\hat{h}(x)$ resulting at convergence are, in fact, maximum likelihood estimates of $a(x)$ and $h(x)$; see Chapter 5 in [16]).

EXAMPLE: THE SALE AND RESALE OF AUCTION QUALITY ART

Goetzmann [17] reports a study of art and the financial markets over time. A major data source for this study is a snapshot sample of auction quality art that was sold at least twice between 1715 and 1986. The data contain 3329 observations of initial sale date and resale date for such artworks. Considering only paintings with an initial date of sale beyond 1715, these data comprise a snapshot sample of inactive subjects, for, to be included in the sample, a painting must have been resold prior to 1986. Thus, paintings that had yet to resell by 1986 could not be included in the sample (indeed, equating σ_i to $[\int_0^\infty h^*(u)/A(u)du]^{-1}$ shows that only 22.7% of all sale/resale intervals were eligible for sampling).

Application of eqn (32) and (33) in iterative fashion to these data followed by statistical smoothing yields the distributions shown in Figs 2 and 3. Consider, first, the history distributions of Fig. 2. Tabulating the raw data shows that the paintings in the sample were initially sold 74.5 y prior to 1986 on average. (The median history equals 68 y.) However, as discussed earlier in this paper, snapshot samples of inactive subjects lead to overestimates of history. The corrected history distribution is shifted markedly to the left of the snapshot sample, with the result that the mean history for *all* paintings sold after 1715 is estimated at only 35 y (with a median of only 20 y), thus representing a rather large discrepancy.

Figure 3 reports the snapshot sampled and corrected active time distributions. Recall that snapshot samples of inactive subjects lead to underestimates of active time, and the example bears this out. The observed mean time between sales for the paintings in Goetzmann's sample equals 37.5 y (median 27 y), but, correcting for the snapshot bias yields an estimated average of 71.8 y between sales (median 74 y). It is thus apparent that the owners of auction quality paintings retain their art for longer periods of time than the snapshot sample would suggest.

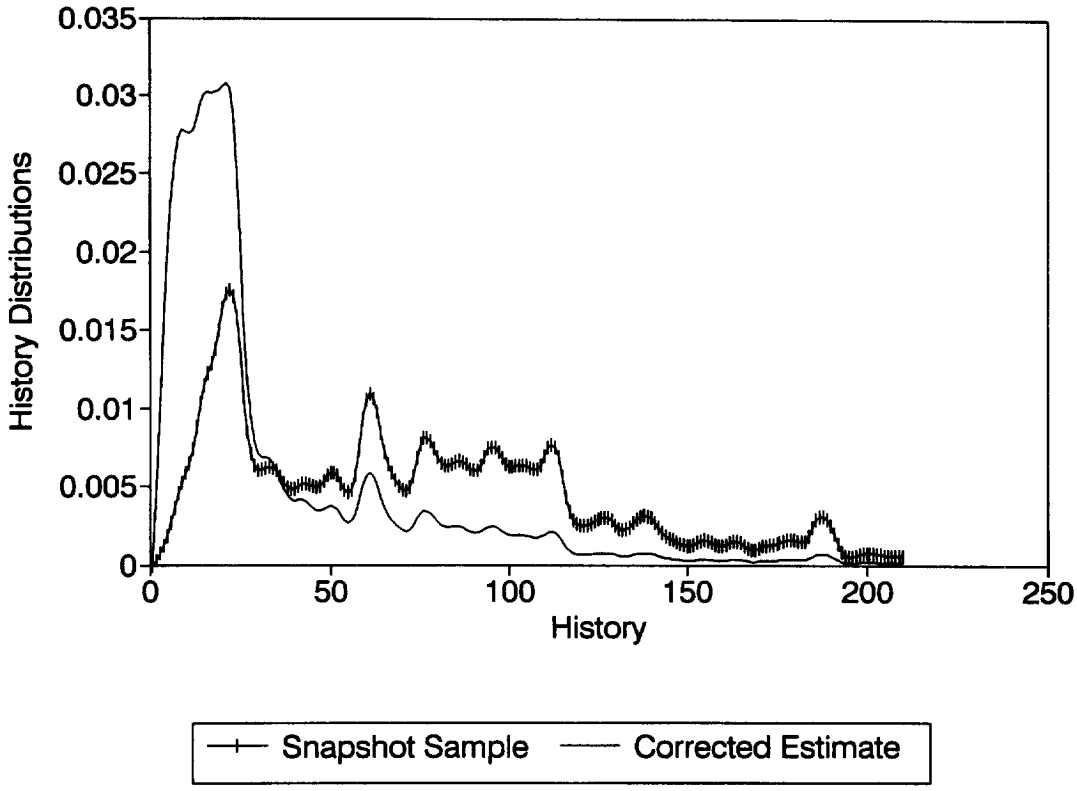

Fig. 2. History of initial art sales.

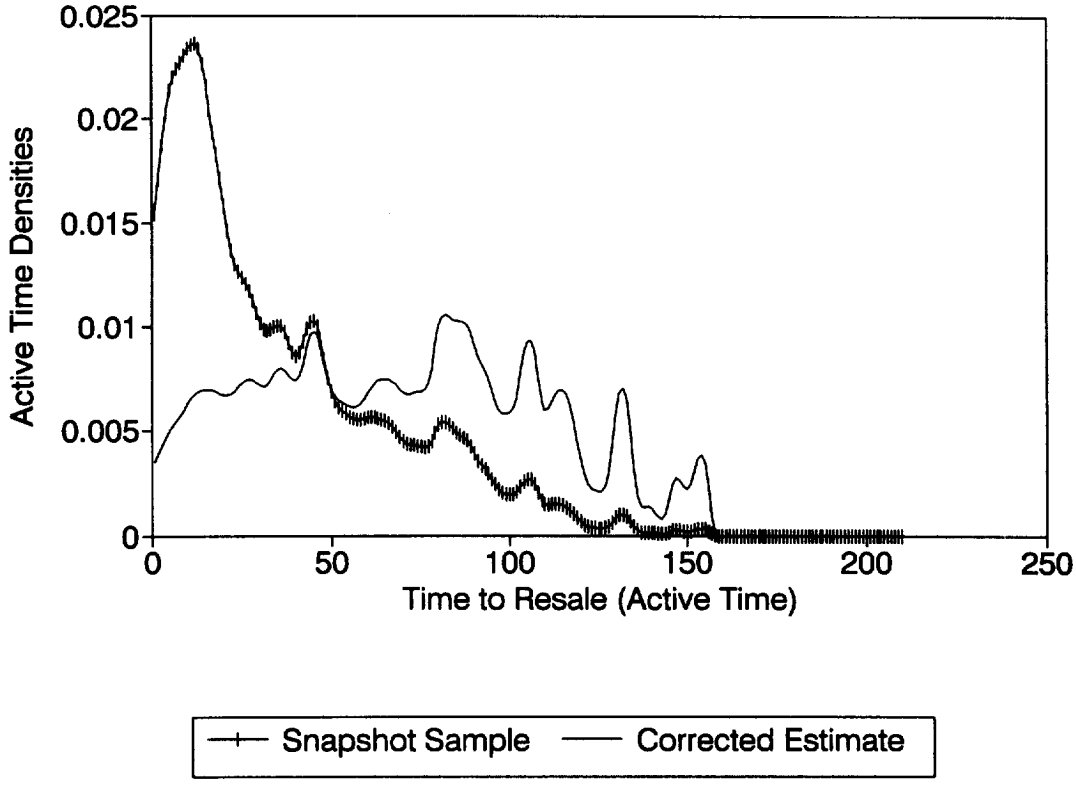

Fig. 3. Time to resale of auction quality art.

CONCLUSIONS

Snapshot samples are easy to construct, but require some care to interpret. Via simple probability models, we have shown how snapshot samples can produce biased views of the underlying random processes. However, we also showed that it is not difficult to correct these biases once they have been recognized. Given this, it is reasonable to deliberately construct snapshot samples when more direct random sampling is not possible or practical, and use the approaches suggested here to properly interpret the data.

Acknowledgements—This research was supported in part by Grant DA09531 from the National Institute on Drug Abuse

REFERENCES

1. Brookmeyer, R. and Gail, M. H., *AIDS Epidemiology: A Quantitative Approach*. Oxford University Press, Oxford, 1994.
2. Jewell, N. P., Non-parametric estimation and doubly-censored data: General ideas and applications to AIDS. *Statistics in Medicine*, 1994, **13,** 2081–2095.
3. Kalbfleisch, J. D. and Lawless, J. F., Inference based on retrospective ascertainment: An analysis of data on transfusion related AIDS. *Journal of the American Statistical Association*, 1989, **84,** 360–372.
4. Bane, M. J. and Ellwood, D., Slipping into and out of poverty: The dynamics of spells. *Journal of Human Resources*, 1986, **21,** 1–23.
5. Flinn, C., Econometric analysis of CPS-type unemployment data. *Journal of Human Resources*, 1986, **21,** 456–484.
6. Torelli, N. and Trivellato, U., Youth unemployment duration from the Italian labour force survey. *European Economic Review*, 1989, **33,** 407–415.
7. Gross, D. and Harris, C. M., *Fundamentals of Queueing Theory* (2nd Edition). John Wiley and Sons, New York, 1985.
8. Eick, S. G., Massey, W. A. and Whitt, W., The physics of the $M_t/G/\infty$ queue. *Operations Research*, 1993, **41,** 731–742.
9. Blumstein, A., Canela-Cacho, J. A. and Cohen, J., Filtered sampling from populations with heterogeneous event frequencies. *Management Science*, 1993, **39,** 886–899.
10. Maltz, M. D. and Pollock, S. M., Artificial inflation of a deliquency rate by a selection artifact. *Operations Research*, 1980, **28,** 547–559.
11. Cox, D. R., *Renewal Theory*, Metheuen and Company Ltd., London, 1962.
12. Drake, A. W., *Fundamentals of Applied Probability Theory*. McGraw Hill, New York, 1967.
13. Kaplan, E. H., Probability models of needle exchange. *Operations Research*, 1995, **43,** 558–569.
14. Kaplan, E. H. and O'Keefe, E., Let the needles do the talking!. *Evaluating the New Haven Needle Exchange*, *Interfaces*, 1993, **23,** 7–26.
15. Caulkins, J. P. and Kaplan, E. H., AIDS impact on the number of intravenous drug users. *Interfaces*, 1991, **21,** 50–63.
16. Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W., *Discrete Multi-variate Analysis: Theory and Practice*. MIT Press, Cambridge, Massachusetts, 1975.
17. Goetzmann, W. N., Accounting for taste: Art and the financial markets over three centuries. *American Economic Review*, 1993, **83,** 1370–1376.