PARAMETRIC MODELS FOR INCUBATION DISTRIBUTION IN PRESENCE OF LEFT AND RIGHT CENSORING

Arni S. R. Srinivasa Rao[1], *Srabashi Basu*[2], *Ayanendranath Basu*[3] *and Jayanta K. Ghosh*[2].

[1]Centre for Ecological Sciences, Indian Institute of Science, Bangalore 560 012, INDIA

[2]Theoretical Statistics and Mathematics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 108, INDIA.

[3]Applied Statistics Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 108, INDIA.

Abstract : When both left and right censoring are present in the data simultaneously, estimating the incubation distribution becomes difficult. In this paper we have introduced several parametric models and have estimated the parameters by the method of moments. A numerical study explores the efficacy of the method.

Key words : HIV infection, AIDS, left and right censoring, survival data.

Mathematics subject Classification 2000: 62N01, 62P10

1. Introduction

In this paper we attempt to develop parametric models to estimate incubation distributions of diseases with long incubation periods when only partial information on survival is available. Incubation period is the time interval between the onset of infection and development of the disease in question. Inference problems regarding the distribution of long incubation periods is made even more complicated by the fact that in many situations only a segment of the incubation period is followed up, particularly in developing countries. Either the point of time when the infection is acquired, or the

point of time when the disease is fully developed, or in some cases both time points may fall outside the study period. In this paper we try to handle incomplete survival data of this type. The methods of this paper are general enough to be applicable to incomplete data from any disease with a long incubation period. However, since our primary interest happens to be in AIDS (acquired immuno–deficiency syndrome), this paper focuses mainly on the AIDS incubation distribution, *i.e.*, the period between the onset of HIV infection to development of full blown AIDS for an adult person.

In developing countries, identified but uninfected risk group cohorts are typically not followed up for a long period of time. A more likely scenario is a relatively short interval of study during which subjects suspected of HIV infection are referred to clinics. Referred subjects are tested for HIV as well as AIDS. If the person has AIDS he is not followed up any more; the same is true for those testing negative for HIV, unless they are sent back to the clinic with future referrals. Only when a person has HIV but has not developed AIDS, the individual is followed up till he/she has developed AIDS or till the end of the study period whichever is earlier. Hence the observed data can be both left and right censored. Also because of the very nature of the protocol followed all data are left censored in the sense that for all individuals who are followed up, the HIV infection has occurred prior to his/her being included in the study.

Given the incomplete nature of the data it is argued in the appendix that common nonparametric techniques (see, *e.g.*, Klein and Moschberger 1997) do not apply. Hence it becomes necessary to develop parametric models. We present two models for the incubation distribution. The assumptions of the second model are perhaps more realistic than the first, but the first model can also be appealing to readers and users on account of its simplicity. The second model is studied more extensively in the paper; however we also indicate how the first model can be fitted.

For the second model we discuss in detail how model parameters can be estimated. We also indicate how one can calculate the theoretical quantiles to compare with empirical quantiles and examine the linearity of the quantile–quantile plot to assess goodness of fit. Standard errors for the estimated parameters can also be estimated in a fairly straightforward way.

Estimating the AIDS incubation distribution is of great practical importance and

can lead to great benefits to the society. Several authors have contributed to the development of incubation distributions for AIDS. Lui et al. (1988) considered estimating the incubation distribution of pediatric patients in USA using a truncated Weibull distribution. De Gruttola and Lagakos (1989) used nonparametric methods to determine incubation distributions for doubly censored data; their technique does not apply to the scenario under consideration in the present paper. To obtain the incubation distribution Taylor et al. (1990) took an imputation approach in which they imputed the time of AIDS diagnosis for the seroconverter cohort. Bacchetti and Jewell (1991) estimated incubation distribution with unknown infection times using external data. Kuo et al. (1991) parametrically modeled the joint distribution of the date of infection and incubation time. Artzouni (1992) developed a series of cohort specific density functions that take into account the increasing impact of new therapies. Tan et al. (1996) generated Monte Carlo data from different conditions and compared the fitting of HIV incubation distribution by some parametric and nonparametric methods. Rao et al. (2000) applied a random correction to account for left censoring to generate incubation distributions for the Indian scenario.

In the next section we introduce both the models and the following two sections discuss estimation and model fitting. In Section 5 we generate simulated data and estimate the incubation distribution by fitting the second model. For a realistic follow–up data size of about four hundred used in conjunction with other available information like how many people had already developed AIDS at the time of the first visit the method of estimation worked reasonably well. In the last section some concluding remarks are given. Finally, a small appendix provides some arguments and counter examples to establish that nonparametric models do not work for the problem considered in this paper.

2. Models

Let us assume that the study begins at time C_1 and terminates at time C_2. Let T, X and Z be random variables defined as follows : T is the point at which HIV infection occurs for a referred person, $T + X$ is the point of time at which the individual visits

the clinic and is diagnosed as being HIV seropositive for the first time and $T + Z$ is the point of time when the individual develops full blown AIDS. Time may be calendar time or in any scale with an arbitrary origin. Thus X is the period between onset of infection and the first subsequent clinic visit where the individual is tested for HIV, and Z is the incubation period. The person is followed up if the following conditions hold simultaneously:

$$C_1 \leq T + X \leq C_2, \tag{1}$$

$$X < Z. \tag{2}$$

For such a person one observes

$$\begin{aligned} Y' &= (T + Z) \wedge C_2 - (T + X) \\ &= Z - X \text{ if } T + Z \leq C_2, \textit{i.e.,} \text{ if } Z \leq C_2 - T \\ &= C_2 - T - X \text{ if } T + Z > C_2, \textit{i.e.,} \text{ if } Z > C_2 - T \end{aligned}$$

If (2) holds, we define Y as $Z - X$, the time to AIDS after the clinic visit where the infection is first diagnosed. Y is not defined for $Z < X$. The observation $Y' = Y \wedge (C_2 - T - X)$ is right censored if $Y' = C_2 - T - X$ and not right censored (i.e. the end point is observed) if $Y' = Y$. We define the following three indicator variables

$$I^{(1)} = \begin{cases} 1 \text{ if equation (1) holds} \\ 0 \text{ otherwise,} \end{cases}$$

$$I^{(2)} = \begin{cases} 1 \text{ if equation (2) holds} \\ 0 \text{ otherwise,} \end{cases}$$

$$I^{(3)} = \begin{cases} 1 \text{ if } Y' = Y \\ 0 \text{ if } Y' = C_2 - T - X. \end{cases}$$

The observed data consist of $(Y_i', I_i^{(3)}), i = 1, \ldots, n$ for those individuals who satisfy (1) and (2). We refer to these cases as CASE I. We also assume that we know the number of cases where (1) is true but (2) fails, *i.e.*, individuals who were not followed up because AIDS had developed already. We refer to these as CASE II and let the number of such

cases in the data set be m. Let J be the indicator for CASE II, *i.e.* $J(T, X, Z) = 1$ iff (1) holds but not (2). We assume $(Y_i', I_i^{(3)})$'s are *i.i.d.* with same distribution as that of $(Y', I^{(3)})$.

To proceed further we now have to specify the joint distribution of T, X and Z. We first consider the independence assumptions.

It seems natural to assume T and (X, Z) are independent. In MODEL I we also assume $Z = X + Y$ where X and Y are independent.

Perhaps a more realistic assumption is that X and Z are independent, which is assumed under MODEL II. Clearly MODEL I and MODEL II cannot hold simultaneously.

It is now only necessary to specify the marginal distributions of T, X and Z. We assume the AIDS epidemic is in its early stage so that a geometric growth rate may be assumed from the beginning of the study C_1 till C_2 where the study terminated. In view of the nature of the data we do not need the density beyond C_2 or previous to C_1. Hence the density of T may be expressed as

$$\begin{aligned} f^T(t) &= k \exp(\rho t) \text{ if } C_1 \leq t \leq C_2 \\ &= 0 \text{ if } t < C_1 \text{ or } t > C_2. \end{aligned} \tag{3}$$

where $\rho > 0$, and k is an appropriate normalising constant. The above may be viewed as the conditional density for T where T is constrained to lie between C_1 and C_2.

Since X is an arrival time we assume for simplicity that X has an exponential distribution with density

$$f^X(x) = \lambda \exp(-\lambda x), \ x > 0. \tag{4}$$

Finally for Z we assume a Weibull or a Gamma distribution. The Weibull distribution has been extensively used in previous studies with effective results, e.g. Medley et al. (1987), Brookmeyer and Gail (1988). The Gamma distribution allows a simple implementation of model fitting, particularly for MODEL I. This completes the specification of the two models. Formally these two models can be written as:

MODEL I

1. T and (X, Z) are independent

2. X and Y are independent

3. The marginal distributions of T, X and Z are as given above.

MODEL II

1. T and (X, Z) are independent

2. X and Z are independent

3. The marginal distributions of T, X and Z are as given above.

MODEL I proposed here is essentially an extension of a method proposed in similar contexts by Rao et al. (2000), which is one of the earliest attempts to analyze data of the above kind with specific focus on the Indian scenario. In the above mentioned paper, the distribution of X was considered known, unlike the approach in the present work.

Since Y is a basic random variable we write its density under MODEL II as

$$f^Y(y) = \frac{\int\int f^T(t) f^X(x) f^Z(x+y) dt dx}{\int_0^\infty \int\int f^T(t) f^X(x) f^Z(x+y) dt dx dy} \tag{5}$$

where the two integrals on the numerator of the above expression as well as the two inner integrals in the denominator are for values of (t, x) over the range where $C_1 \leq t + x \leq C_2$.

If one has to calculate moments or a probability event of Y lying in an interval, it is convenient to substitute $z = x + y$ or $x = z - y$ and integrate out t and z, leaving a one-dimensional integral involving y which has to be integrated numerically. For MODEL I this method runs into a difficulty because to get the density of Y from that of X and Z one needs some form of deconvolution. However this can be done explicitly if Z has a Gamma density with same scale parameter λ as that of X.

Here we briefly indicate why left censoring cannot be handled nonparametrically in this context. We confine our attention to Y and X, ignoring right censoring through C_2. There are two difficulties in handling this through an approach based on Kaplan–Meier (Kaplan and Meier, 1958). In the first place one would need a model of the form $Y = Z \wedge W$, where W is an independent censoring random variable. Such a representation does not hold in the present problem. Secondly, even if one tries to apply a Kaplan– Meier type estimate, one would get a trivial result because all the data are left censored.

We conclude this section by discussing conditions (1) and (2) in more detail. If $T + X > C_2$, we interpret this to mean that the person's first post infection visit to the clinic came after the study was over. Similarly if $T + X < C_1$, we mean that the person was referred to the clinic before the study began and so was lost. However, one might take a different point of view and start with the random variables actually observed, namely T', which is the time between C_1 and C_2 when a person was first checked at the clinic for HIV and AIDS after acquiring the infection. In case such a person had HIV but not AIDS, $T'' > T'$, where T'' is defined as

$$\begin{aligned} T'' &= C_2, && \textit{if AIDS developed after } C_2 \\ &= \text{time before } C_2 \text{ when AIDS developed} && \textit{otherwise} \end{aligned}$$

In terms of our original random variables $T' = T + X$ and $T'' = (T + Z) \wedge C_2$ if (1) and (2) hold. Thus T' is constrained to lie between C_1 and C_2 whereas $T + X$ is initially not restricted in this way but this restriction is imposed later through (1). An alternative way of modeling would be to take T, T' and Z as the primary variables instead of T, X and Z. A full model can be developed assuming T, T' and Z to be independent. One may assume T' to be an exponentially distributed random variable conditioned to lie between C_1 and C_2. In this model one would define X as $T' - T$ on the set where $T' > T$.

3. Inference for Model II

We essentially estimate our parameters by the method of moments. One can, in principle, also use the method of maximum likelihood, but in our case the numerical computation becomes very messy. Moreover the method of moments is likely to be more robust than maximum likelihood because the presence of latent, i.e. unobservable random variables makes the model similar to a mixture model.

Each of the indicators $I^{(1)}$, $I^{(2)}$ and $I^{(3)}$ are functions of T, X and Z. We estimate the four parameters (ρ, λ, and the two parameters of the Gamma or the Weibull distribution for Z) by solving the set of the following four equations.

$$E(I^{(2)}|I^{(1)} = 1) = \frac{n}{m+n} \tag{6}$$

$$E(I^{(2)}(1 - I^{(3)})|I^{(1)} = 1) = \frac{n_1}{m+n} \tag{7}$$

where n_1 is the number of Y_i''s (among the n observations satisfying (1) and (2)) which are censored on the right and the expectations are with respect to the joint distributions of T, X, and Z. Further

$$E(I^{(2)}I^{(3)}Y'|I^{(1)}=1) = \frac{\sum_{i\in A_2} Y_i'}{m+n}, \tag{8}$$

$$E(I^{(2)}(1-I^{(3)})Y'|I^{(1)}=1) = \frac{\sum_{i\in A_1} Y_i'}{m+n}, \tag{9}$$

where A_1 is the index set of all censored Y_is, A_2 is the index set of all uncensored Y_is, and n_2 is the number of Y_is which are uncensored, $n = n_1 + n_2$.

Alternatively we can define the relations

$$E(I^{(2)}I^{(3)}Y'|I^{(1)}I^{(2)}I^{(3)}=1) = \frac{\sum_{i\in A_2} Y_i'}{n_2}, \tag{10}$$

$$E(I^{(2)}(1-I^{(3)})Y'|I^{(1)}I^{(2)}(1-I^{(3)})=1) = \frac{\sum_{i\in A_1} Y_i'}{n_1}, \tag{11}$$

It is easily seen that the relations (10) and (11) are equivalent to (8) and (9) in the presence of equations (6) and (7). However, equations (10) and (11) are easier to understand than (8) and (9), and later in our numerical examples (Section 5) we solve equations (6), (7), (10) and (11) to estimate the four parameters.

As explained earlier in Section 2 for a given set of parameters, we can calculate the left hand sides as one dimensional integrals in y which can be integrated numerically. For a discrete grid of parameters one can calculate the set of four theoretical values and choose the set for which we are closest to corresponding observed values as representing the estimates of the parameters.

Standard errors and confidence intervals based on large sample theory are easy to calculate.

Using the estimated parameter values in the distribution of Z, one can estimate the median and other quantiles of the incubation distribution.

Finally to assess model validity, one can compare $P_{\hat{\theta}}(Y' < a)$ with its nonparametric analogue $1/n \sum I_{Y_i'<a}$ for various values of a as in quantile–quantile plot.

4. Inference for MODEL I

To estimate the parameters of MODEL I, we use the simplification mentioned above, namely, we assume Z has a Gamma distribution and X and Z have the same scale parameter. In this case Y has a Gamma distribution, using the independence of X and Y we can avoid having to use the distribution of T as follows. Since $Y' = (Z \wedge (C_2 - T)) - X = Y \wedge (C_2 - T - X)$, Y and (T, X) are independent. So based on data in $(Y_i', I_i^{(3)})$ we can use the Kaplan–Meier estimate for the distribution of Y and hence estimate moments of Y using only data on $(Y_i', I_i^{(3)})$. Let $\widehat{\mu}_1', \widehat{\mu}_2'$ be the nonparametric estimates of $E(Y^r), r = 1, 2$ (Bose and Sen, 2001). We now solve

$$\begin{aligned} E(Y) &= \widehat{\mu}_1' \\ E(Y^2) &= \widehat{\mu}_2'. \end{aligned}$$

Thus we can estimate the relevant parameters and hence the distribution of Z. We cannot use the Kaplan–Meier estimate under MODEL II because there Y is not independent of (T, X) (hence the variable under study and the censoring variable do not have independent distributions).

In MODEL I our assumption is that X is exponential, Z is Gamma, and X and Y are independent, $X + Y = Z$. For convenience, we assume that X and Z have the same scale parameter, from which it is easy to deduce Y is also Gamma with the same scale parameter.

It is possible to avoid the assumption of the equality of the scale parameters and still estimate all the parameters. In this case we cannot find the density of Y explicitly. However the characteristic function of Y can be easily calculated as:

$$\phi(t) = E(e^{itY}) = E(e^{itZ})/E(e^{itX}),$$

at any t for which the denominator is not zero. The function $\phi(t)$ can be written down explicitly in terms of t and the three parameters appearing in the density of X and Z (namely scale parameter of X, scale and shape parameters of Z). Now choose three values of t, say t_1, t_2, t_3 and solve

$$\phi(t_i) = \int_0^\infty e^{it_i y} d\hat{F}(y),$$

$i = 1, 2, 3$, where $\hat{F}$ is the Kaplan-Meier estimate of the distribution function of Y. One can use values of $\phi(t)$ at other values of t to check goodness of fit.

5. Numerical Studies

We selected parameters that are plausible for the AIDS epidemic in a city in India – in fact we had Mumbai in mind. We assume that the epidemic started in 1986 with the growth rate $\rho = 0.25$ and the parameters of the Weibull distribution for Z are $\alpha = 8.836$ and $\beta = 2.022$. These parameter values have been adapted from Rao and Kakehashi (2001). For illustration we take the starting point of the study to be the year 1986, and the end point to be the year 2000. However we shift the origin so that C_1 equals zero (and hence C_2 equals 14).

The density for T is

$$f^T(t) = k \exp(0.25t), \quad 0 \leq t \leq 14.$$

We do not need values of t beyond $C_2 = 14$. The constant k is chosen such that $\int_0^{14} f^T(t)dt = 1$.

The density of Z is

$$f^Z(z) = \frac{\beta}{\alpha}\left(\frac{z}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{z}{\alpha}\right)^{\beta}\right\}, \ z > 0, \ \alpha > 0, \ \beta > 0.$$

Finally we take $\lambda = 0.25$, so that the density of X is

$$f^X(x) = 0.25 \exp(-0.25x), \ x \geq 0.$$

This means a person takes four years time on an average to be diagnosed for the first time as HIV positive after the HIV infection.

We generate about 1000 samples (T, X, Z) from the above distributions and equations (1) and (2) in Section 2 are satisfied in n cases. This provides the basic data $(Y_i', I_i^{(3)})$, $i = 1, 2, \ldots, n$. In n_1 of these cases $I_i^{(3)} = 0$, and in the remaining $n_2 = n - n_1$ cases $I_i^{(3)} = 1$. In an additional m cases (not belonging to the above n) (1) holds, but equation (2) is violated. The value of m is used only via the fact that $n/(m+n)$ is an estimate of $P(Z > X | I^{(1)} = 1)$.

Let A_1 be the set of n_1 indices for which Y_i is censored to the right, and let A_2 represent the set of indices for the n_2 uncensored observations. The theoretical expression for the left hand side of, say, equation (11) can be written as

$$\frac{\int_0^{14}\int_{14-z}^{14}\int_0^{14-t}(14-t-x)g(x,t,z)dxdtdz + \int_{14}^{\infty}\int_0^{14}\int_0^{14-t}(14-t-x)g(x,t,z)dxdtdz}{\int_0^{14}\int_{14-z}^{14}\int_0^{14-t}g(x,t,z)dxdtdz + \int_{14}^{\infty}\int_0^{14}\int_0^{14-t}g(x,t,z)dxdtdz}$$

where $g(x,t,z) = f^X(x)f^T(t)f^Z(z)$. Although the integrals look complicated, both the numerator and the denominator can be reduced to one dimensional integrals which can then be solved numerically – an extremely convenient simplification, the method would not have been feasible otherwise. Similarly the theoretical expressions for the equations (6), (7), and (10) can also be written down. Solving these equations for the generated data set the estimates came out to be $\alpha = 10$, $\beta = 2$, $\rho = 0.3$ and $\lambda = 0.35$. The values n, n_1, and m were respectively 397, 216 and 90 for these data.

For a second example, the initial values were chosen to be $\rho = 0.25$, $\alpha = 5.0804$, $\beta = 2.286$, $\lambda = 0.25$. The values of α and β correspond to the parameters of the Weibull distribution for US data as presented in Brookmeyer and Gail (1988). Here the final estimated parameters based on a sample of size 1000 are $\rho = 0.25$, $\alpha = 5.0$, $\beta = 3$ and $\lambda = 0.3$.

6. Concluding Remarks

The standard nonparametric methods for handling right and left censoring do not apply to the incubation data that are normally available in India (and possibly in other developing countries) – see the discussion in the appendix. We propose two parametric models and indicate how one can estimate parameters and can check goodness of fit. It appears from some examples that even with a moderate sample size of a few hundreds one gets reasonably accurate results.

Parametric models make the best use of currently available data sets but it would be prudent to strive for better data. Since the major source of complexity in modelling comes from left censoring, future observational studies need to eliminate this. If a high

risk group of uninfected people are followed up for a reasonable period, we would have a better idea about the time of HIV infection. Observations on time from HIV to AIDS would still be subject to right censoring. But then standard methods based on Weibull or Kaplan-Meier will be applicable.

Once HIV is detected ethical principles would require that remedial drugs to delay AIDS be used. If the study provides treatment facilities this can act as an incentive to visit the centre and hence reduce right censoring.

However the use of drugs will have changed the incubation distribution. Hopefully the new Z will be stochastically larger than the current Z. So both the currently available data and the methods for analyzing them will continue to be historically important.

As a final comment, we mention that it is possible that there is an extra source of censoring if individuals who have not developed full blown AIDS drop out of the study before C_2. We have not considered this in this paper, but it can be handled by making the model and the analysis somewhat more complex. Choosing Y'' to be the minimum of the three variables Y, $C_2 - T - X$, and W, where W is another independent censoring variable, the following approach can be taken to handle the problem of interest. Although Y' is no longer observable and hence the empirical estimates occuring in equations (10), (11), etc. are no longer available, one can estimate the distribution function of Y' by a Kaplan-Meier estimate based on Y_i''s. Then one can also estimate the moments of Y' as in Bose and Sen (2001). So the equations (10), (11), etc. can be replaced by equating theoretical and empirical values for quantities which can be moments or values of the distribution function of Y' evaluated at certain points. The empirical estimates will be obtained through the Kaplan Meier estimates based on Y'', while the theoretical values can be written as ratios of two univariate integrals as in the section on numerical studies.

Appendix

In this paper we have proposed parametric models for tackling the problem of interest. Here we demonstrate that one could not have used nonparametric methods for handling this problem. Note that unless the distribution of Y can determine the distribution of Z, there is no hope of getting a consistent nonparametric estimate of the distribution of Z like the Kaplan-Meier in the right censored case.

A single parametric counterexample is enough to show that one cannot estimate the distribution of Z nonparametrically just having data on Y. Here we will present several counterexamples. To set up the background, let Z and X be independent non negative random variables. Then, on the set $Z \geq X$, Y is defined as $Z - X$. Y is not defined elsewhere. For $y > 0$,

$$P[(Y \text{ is defined}) \text{ and } (Y > y)] = P[Z > X + y] = \int_0^\infty \bar{F}^Z(x+y) f^X(x) dx \ . \qquad \text{(A1)}$$

where $\bar{F}^Z = 1 - F^Z$, the survival function of Z. In particular $P[Z > X] = \int_0^\infty \bar{F}^Z(x) f^X(x) dx$ gives the probability that Y is defined. Assume that both X and Z have an upper bound M. Then the right hand side of (A1) becomes equal to

$$\int_0^{M-y} \bar{F}^Z(x+y) f^X(x) dx. \qquad \text{(A2)}$$

Assume $f^X(x) = ax^r, 0 < x < M$, and $\bar{F}^Z(x) = b(M-x)^m, 0 < x < M$. Then (A2) equals

$$ab \int_0^{M-y} (M - x - y)^m x^r dx$$

which, by repeated integration by parts, is equal to

$$c \int_0^{M-y} (M - x - y)^m x^r dx = d(M-y)^{m+r+1}.$$

Hence from (A1),

$$P[Y > y | Y \text{ is defined}] = P[(Y \text{ is defined}) \text{ and } (Y > y)] / P[Y \text{ is defined}] = \frac{(M-y)^{m+r+1}}{M^{m+r+1}}.$$

If we take another pair (X, Z) with $f^X(x) = const \times x^{r'}$, $\bar{F}^Z(z) = const \times (M-z)^{m'}$ where $m' + r' = m + r$, we would get the same distribution as that of Y, i.e. Y cannot determine the distribution of Z.

The above counterexample will not apply when the indicator which defines whether Y is defined or not is known. A second, relatively more trivial counterexample would however still work in that situation. We have presented both counterexamples because the first one is more interesting, and in certain cases one might want to work with Y only. The second counterexample goes as follows: Let Z and X be arbitrary, and $c > 0$. Then if $Z' = Z + c$, $X' = X + c$, then $Y' = Z' - X'$ on $Z' > X'$ has the same distribution

as $Y = Z - X$ on $Z > X$. Moreover $P[Z > X]$ =P[Z' ¿ X'] (which is not true for the previous example). In our paper we assumed that we had data from which we can estimate the distribution of Y, as well as data for estimating $P[Z > X]$. The second counterexample applies in this case.

It seems plausible that if one knows the distribution of X and Y, then the distribution of Z can be found. Even that is hard to prove, and in fact we have been able to prove it only for the bounded lattice case. Suppose X, Z, take positive integral values $\leq M$, where M is a positive integer. Then

$$\begin{aligned} P[Y = M] &= P[Z = M]P[X = 0] \\ P[Y = M-1] &= P[Z = M]P[X = 0] + P[Z = M-1]P[X = 1] \\ &\vdots \\ P[Y = M-r] &= P[Z = M]P[X = r] + \ldots + P[Z = M-r]P[X = 0] \end{aligned}$$

so that the probability distribution of Z can be recovered from the above conditions.

References

Artzouni, M. (1992). A modeled time-varying density function for the incubation period of AIDS. *Journal of Mathematical Biology*, **31**, 73–99.

Bacchetti, P. and Jewell, N. P. (1991). Nonparametric estimation of the incubation period of AIDS based on a prevalent cohort with unknown infection times. *Biometrics*, **47**, 947–960.

Bose, A., and Sen, A. (2001). Asymptotic distribution of the Kaplan-Meier U-Statistics. To appear in *Journal of Multivariate Analysis*.

Brookmeyer, R., and Gail, M. H. (1988). A method for obtaining short term predictions and lower bounds on the size of the AIDS epidemic. *Journal of the Americal Statistical Association*, **83**, 301–308.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS, *Biometrics*, **45**, 1–11.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations, *Journal of American Statistical Association*, **53**, 457–481.

Klein, J. P., and Moschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer-Verlag, New York.

Kuo, J.-M., Taylor, J. M. G. and Detels, R. (1991). Estimating the AIDS Incubation Period From a Prevalent Cohort. *American Journal of Epidemiology*, **133**, 1050–1057.

Lui, K. J., Peterman, T. A., Lawrence, D. N., and Allen, J. R. (1988). A model-based approach to characterise the incubation period of pediatric transfusion-associated acquired immunodeficiency syndrome, *Statistics in Medicine*, **7**, 395–401.

Medley, G. F., Anderson, R. M., Cox, D. R., and Billard, L. (1987). Incubation period of AIDS in patients infected via blood transfusion. *Nature*, **328**, 719–721.

Rao, Arni S. R. Srinivasa, Basu, S., Hira, S. K., Basu, A. and Pal, S. (2000). Estimation of AIDS incubation distribution for the Indian seroconverts. *Technical Report, Stat-Math Division*, 1/2000, Indian Statistical Institute, Calcutta 700 035, India.

Rao, Arni S. R. Srinivasa, and Kakehashi, M. (2001). Projection of AIDS and possible implications in India. *Unpublished Manuscript*.

Tan, W. Y., Lee, S. R., and Tang, S. C. (1996). Characterization of HIV incubation distributions and some comparative studies. *Statistics in Medicine*, **15**, 197–220.

Taylor, J. M. G., Muñoz, A., Bass, S. M., Saah, A., Chmiel, J. S., and Kingsley, L. A. (1990). Estimating the Distribution of Times From HIV Seroconversion to AIDS Using Multiple Imputation. *Statistics in Medicine*, **9**, 505–514.