A note on estimation in incompletely observed inhomogeneous Poisson processes with application to HIV-testing

Sven Ove Samuelsen
Department of Mathematics, University of Oslo
P.O. Box 1053 Blindern, 0316 Oslo, Norway

email: osamuels@math.no.uio

Abstract

This note is devoted to estimation of rates in inhomogeneous Poisson processes when the only obtained information is time (year) of first jump, time (year) of last jump and the total number of jumps in a specified time interval. An EM-algorithm for obtaining the MLE of the year specific rates is derived. The EM-algorithm is applied to data on HIV-testing from the Norwegian study of sexual behavior. Standard errors of the estimated rates are obtained by jackknifing, bootstrapping and a Monte Carlo version of Louis' method and the techniques are compared. The rates to the first HIV-test are also computed both by the actuarial methods and as MLE's by the EM-algorithm.

1 Introduction

In a survey study of sexual behavior in Norway (Stigum et al., 1995) questions regarding testing for HIV were included. However one did not obtain the complete history of testing for the individuals, but only the year of the first test, the year of the last test and the total number of test from 1985 when HIV-testing was introduced to 1992 when the survey was conducted. One problem to be addressed was the development in the rates of HIV-testing in the Norwegian population (Magnus et al., 1994).

Due to the incomplete recorded individual histories of HIV-testing the likelihoods of the observed data can be quite complicated under many models. In section 2 of this note we point out how the EM-algorithm (Dempster et al., 1977) may be applied to obtain the maximum likelihood estimates (MLE) under the model of an inhomogeneous Poisson process with rates constant on each calendar year. In section 3 the EM-algorithm for estimation of rates of first HIV-test is presented and compared to the commonly applied actuarial method. In the last section it is also discussed how one may generalize the approach to heterogeneity models and models where the previous number of tests are predictive of further tests.

2 Population rates for HIV-testing

2.1 The model and the complete data specification

Let $N_i(t)$ equal the number of HIV-tests for individual i in the period $[0,t]$. We suppose that the $N_i(t)'s$ are independent inhomogeneous Poisson processes with common rate $\lambda(t)$, thus the $N_i(t)'s$ are independent and Poisson distributed with expectation $\Lambda(t) = \int_0^t \lambda(s)ds$.

Since our knowledge of the time of the HIV-tests is restricted to the year tests took place we will assume that the $\lambda(t) = \lambda_j$ when $j-1 < t \leq j$, i.e. $\lambda(t)$ is piecewise constant. Letting $X_{ji} = N_i(j) - N_i(j-1)$, i.e. the number of tests for individual in year j we get that the X_{ji} are independent and Poisson distributed with expectation λ_j. On complete recorded data the maximum likelihood estimates of the rates are the averages $\hat{\lambda}_j = (1/n)\sum_{i=1}^n X_{ji}$.

2.2 The observed incomplete data and the EM algorithm

The individuals in the survey study were asked to record the total number of test during the period 1985 to 1992. We label the years as $j = 1$ corresponding to 1985 to $j = 8$ corresponding to 1992. The total number of tests for each individual $Y_i^1 = N_i(8) = \sum_{j=1}^8 X_{ji}$ was observed. The year of first test $Y_i^2 = j$ if $N_i(j-1) = 0 < N_i(j)$ and the year of last test $Y_i^3 = j$ if $N_i(j-1) < N_i(j) = N_i(k)$ for $k > j$ was also obtained for each individual.

Letting $\mathcal{Y}_i = (Y_i^1, Y_i^2, Y_i^3)$ and $\mathcal{X}_i = (X_{1i}, ..., X_{8i})$ we see that the $\mathcal{Y}_i's$ are functions of the $\mathcal{X}_i's$. Thus the $\mathcal{Y}_i's$ are incomplete observations of the $\mathcal{X}_i's$ in the sense of Dempster et al. (1977) and an EM algorithm may be applied to estimate the $\lambda_j's$. Furthermore the complete data $\mathcal{X} = (\mathcal{X}_1, ...\mathcal{X}_n)$ stems from an exponential class and the standard type of the EM algorithm may be applied.

We need to calculate the $\hat{X}_{ji}(\lambda) = E[X_{ji}|\mathcal{Y}_i]$ as a function of $\lambda = (\lambda_1, ..., \lambda_8)$. Note that when the total number of tests is zero (i.e. $Y_i^1 = 0$) then all $X_{ji} = 0$, when $Y_i^1 = 1$ then $X_{ij} = 1$ when $Y_i^2 = Y_i^3 = j$ and zero otherwise and when the total number of test equals 2 with first and last test in different years (i.e. $Y_i^1 = 2$ and $Y_i^2 < Y_i^3$) then $X_{ji} = 1$ when $Y_i^2 = j$ and $Y_i^3 = j$. Furthermore if $Y_i^1 > 1$ with all test performed in the same year we have $X_{ji} = Y_i^1$ for $j = Y_i^2 = Y_i^3$ and zero otherwise. Also in general $X_{ji} = 0$ for $j < Y_i^2$ and $j > Y_i^3$. In all these cases $\hat{X}_{ji}(\lambda) = X_{ji}$ no matter what the value of λ might be.

We thus only need to calculate $\hat{X}_{ji}(\lambda)$ for $Y_i^1 > 2$ and $Y_i^2 < Y_i^3$. Note that $X_{1i}, ..., X_{8i}$ given Y_i^1 is multinomial with Y_i^1 trials and probabilities $\lambda_j / \sum \lambda_j$. It turns out that given $\mathcal{Y}_i = (y, j, k)$ then $X_{ji}, ..., X_{ki}$ has a multinomial distribution with y trials and probabilities $p_{ir} = \lambda_r / \sum_{r=j}^{k} \lambda_r$ that is truncated to $X_{ji} \geq 1$ and $X_{ki} \geq 1$. Now let

$$p_i^0 = P(X_{ji} \geq 1, X_{ki} \geq 1 | Y_i^1 = y, X_{ri} = 0 \text{ for } r < j, r > k). \tag{1}$$

If $k = j + 1$ then $p_i^0 = 1 - p_{ij}^y - p_{ik}^y$ and we have

$$\hat{X}_{ji}(\lambda) = y \frac{p_{ij} - p_{ij}^y}{p_i^0} \tag{2}$$

and of course $\hat{X}_{ki} = y - X_{ji}$. If on the other hand $k > j + 1$ then $p_i^0 = 1 - (1 - p_{ij})^y - (1 - p_{ik})^y + (1 - p_{ij} - p_{ik})^y$ and

$$\hat{X}_{ji}(\lambda) = \begin{cases} y p_{ir} \dfrac{1 - (1 - p_{is})^{y-1}}{p_i^0} & \text{if } r = j \text{ and } s = k \\ & \text{or if } r = k \text{ and } s = j \\ y p_{ir} \dfrac{1 - (1 - p_{ij})^{y-1} - (1 - p_{ik})^{y-1} + (1 - p_{ij} - p_{ik})^{y-1}}{p_i^0} & \text{if } j < r < k \end{cases} \tag{3}$$

Now the EM algorithm may be stated by the following E(stimation)- and M(aximation)-steps.

E-Step: Given estimates $\lambda^{(r)} = (\lambda_1^{(r)}, ..., \lambda_8^{(r)})$ and observed incomplete data $\mathcal{Y} = (\mathcal{Y}_1, ...\mathcal{Y}_n)$ calculate estimates of X_{ji} by $\hat{X}_{ji}(\lambda^{(r)})$.

M-Step: Update estimates for $\lambda = (\lambda_1, ..., \lambda_8)$ by $\lambda_j^{(r+1)} \doteq (1/n) \sum_{i=1}^{n} \hat{X}_{ji}(\lambda^{(r)})$.

Starting with arbitrary values of $\lambda_j^{(1)} > 0$ we have $\lambda_j^{(r)} \to \lambda_j^*$ where $\lambda^* = (\lambda_1^*, ..., \lambda_8^*)$ is the MLE of λ (if the likelihood of the incomplete data $\mathcal{Y}$ has only one stationary point). Thus we were able to obtain the MLE without deriving the fairly complicated likelihood of $\mathcal{Y}$.

2.3 Estimation of the covariance matrix

The covariance matrix of λ^* is not obtained directly from the EM algorithm. Three approaches for estimating this matrix have been applied. Firstly by jackknifing, secondly by bootstrapping and finally by Louis' method (Louis, 1982, Tanner, 1993).

In the jackknifing approach estimates λ^*_{-i} is calculated by the EM algorithm from incomplete data $\mathcal{Y}_{-i} = \mathcal{Y} \backslash \mathcal{Y}_i$. The empirical covariance matrix of the λ^*_{-i} is now a good approximation of the true covariance matrix of λ^* (See e.g. Efron & Tibshirani, 1993).

Similarly, with bootstrapping (Efron & Tibshirani, 1993) "data" $\mathcal{Y}^{(r)} = (\mathcal{Y}_1^{(r)}, ..., \mathcal{Y}_n^{(r)})$ are sampled with replacement from $\mathcal{Y} = (\mathcal{Y}_1, ..., \mathcal{Y}_n)$. On $\mathcal{Y}^{(r)}$ the EM-algorithm was applied to obtain estimates $\lambda^*_{(r)}$. By repeating the resampling R times and calculating the empirical covariance matrix one also obtains an estimate of the covariance matrix. The variance-covariance estimates obtained from jackknifing and bootstrapping is not likely to differ much.

The third approach, Louis' method, makes more directly use of the assumed model. Let $S_\mathcal{X}$ and $I_\mathcal{X}$ be, respectively, the score function and the observed information matrix of the complete data $\mathcal{X}$ and similarly $I_\mathcal{Y}$ the observed information matrix of $\mathcal{Y}$. Generally one then have (Louis, 1982, Tanner, 1993) that

$$I_\mathcal{Y} = E[I_\mathcal{X}|\mathcal{Y}] - VAR(S_\mathcal{X}|\mathcal{Y}) \tag{4}$$

where $VAR(.)$ denotes a covariance matrix.

In this particular case we have that the j-th component of $S_\mathcal{X} = (S_{\mathcal{X}1}, ..., S_{\mathcal{X}8})$ equals

$$S_{\mathcal{X}j} = \sum_{i=1}^{n} (\frac{X_{ji}}{\lambda_j} - 1)$$

and that $I_\mathcal{X}$ is a diagonal matrix with terms $\sum_{i=1}^{n} X_{ji}/\lambda_j^2$. It turns out that $E[I_\mathcal{X}|\mathcal{Y}]$ is also a diagonal matrix with terms $\sum_{i=1}^{n} \hat{X}_{ji}(\lambda)/\lambda_j^2$ which may be estimated simply by substituting λ with λ^*.

Furthermore with v_{jk} being the jk-th term of $VAR(S_\mathcal{X}|\mathcal{Y}]$, we get that

$$v_{jk} = \sum_{i=1}^{n} \frac{\mathrm{Cov}(X_{ji}, X_{ki}|\mathcal{Y}_i)}{\lambda_j \lambda_k}$$

When $Y_i^1 \leq 2$ or $Y_i^2 = Y_i^3$ we know all X_{ji}'s and so these variances / covariances are all zero. Otherwise it should in principle be possible to derive expressions corresponding to (2) and (3) for $\mathrm{Cov}(X_{ji}, X_{ki}|\mathcal{Y}_i)$. Instead of doing so one may apply simulation techniques.

Thus for each i with $\mathcal{Y}_i = (y, j, k)$ we draw from the multinomial distribution with y trials and probabilities $\lambda_l^* / \sum_{l=j}^{k} \lambda_l^*$ until we have a draw with at least one outcome both in year j and in year k. By repeating the draws the $\mathrm{Cov}(X_{ij}, X_{ik}|\mathcal{Y}_i)$ may estimated by empirical covariances. The number of draws needed is determined by observing when the estimates appears stable. Now estimates of the v_{jk} are obtained by substituting λ with λ^*. Inserting the estimates of $E[I_{\mathcal{X}}|\mathcal{Y}]$ and $VAR(S_{\mathcal{X}}|\mathcal{Y}]$ into (4) we thus get an estimate $I_{\mathcal{Y}}^*$ of $I_{\mathcal{Y}}$ and then inverting this matrix we have an estimate of the covariance matrix of λ^*.

2.4 Application to HIV-test data

Among the 4760 individuals responding to the survey on sexual behavior a total of $n = 4667$ individuals gave complete answers to the questions regarding HIV-testing (Magnus et al., 1994). Among these 650 reported to have been tested for HIV. The total number of reported test was 1115. A total of 422 had been tested exactly once, 139 had been tested exactly two times and 89 more than two times. Due to this we know the year the test was carried out in 878 cases. In addition the year of another 6 tests were known because 2 individuals had taken all tests in the same year. The maximum numbers of tests reported was 23.

The HIV-test was introduced in in Norway in July 1985, thus the individuals had only a six months follow-up that year. Similarly the survey was carried out late November / early December in 1992, giving a follow-up time per individual of 11/12 of a year. When estimating rates based on the number of HIV-tests carried out one thus divides by $2n$ and $(12/11)n$ respectively. Thus a minor modification to the (M-step of the) EM-algorithm described in the previous section had to be made.

The maximum likelihood estimates of the rate (per 100 person-years) along with their jackknifed estimates of the standard errors are given in Table 1.

Table 1. Maximum likelihood estimates

Year	Estimated number of tests	Rate per 100 person-years	Standard error
1985	34.3	1.47	0.36
1986	39.7	0.85	0.14
1987	83.4	1.79	0.22
1988	144.7	3.10	0.31
1989	162.4	3.48	0.34
1990	217.0	4.65	0.42
1991	207.2	4.44	0.40
1992	225.9	5.28	0.46

The incidence in HIV-testing in Norway appears to have dropped from 1985 to 1986, but from 1986 and on the incidence is essentially estimated to increase.

An intuitive approach to this estimation problem would be to distribute the tests for which the exact year is not known evenly over the years that they can have been taken. This will not differ much from the estimates after one iteration of the EM-algorithm if one starts out with all rates equal. As start values for the EM-algorithm we used $\lambda_j^{(1)} = 0.032 \approx 1115/34613$, i.e. the average test-rate over the period. The first iteration gave the following results

Table 2. Estimates after the first iteration

Year	Estimated number of tests	Rate per 100 person-years
1985	36.9	1.58
1986	44.4	0.95
1987	91.0	1.95
1988	148.2	3.17
1989	166.5	3.57
1990	206.2	4.42
1991	205.0	4.39
1992	216.8	5.07

Comparing Table 1 and Table 2 we see the same trends. In this dataset one iteration was thus sufficient to discover the tendency. However the fully iterated estimates gives a somewhat more pronounced picture of what has been going on. The algorithm converged (to 4th decimal) after 5 iterations.

2.5 Comparison of the estimates of standard errors and correlations

In subsection 2.2 it was discussed how to obtain covariance matrices of the estimated rates by jackknifing, bootstrapping and by Louis' method. Of course, one obtains estimates of standard errors by taking square roots of the diagonal of the covariance matrices and estimates of the correlations by dividing the estimated covariances by the the standard errors. With jackknifing one formally gets 4667 estimates of λ although for the 4017 individuals that had never taken a HIV-test one will get the same estimates. The bootstrapping procedure was repeated 2000 times. This is clearly more than is necessary to obtain acceptable standard errors, but for the correlation coefficients a quite high number of resamples were required. Luckily computational time was not inhibiting. For Louis' method 80 draws were taken for each of the 89 individual with more than 2 tests in different years. With this choice the correlations seemed to have two valid decimals and the standard errors were estimated to second valid digit.

In the following table the estimated standard errors are given.

Table 3. Standard errors of estimated rates per 100 years by different methods

Year	jackknifing	Bootstrapping	Louis' method
1985	0.36	0.37	0.26
1986	0.14	0.15	0.14
1987	0.22	0.22	0.21
1988	0.31	0.32	0.29
1989	0.34	0.35	0.31
1990	0.42	0.44	0.36
1991	0.40	0.43	0.34
1992	0.46	0.48	0.37

The agreement between the standard errors obtained by jackknifing and bootstrapping was, as anticipated, quite good. However Louis' method generally gave smaller standard errors than the two other methods and they are considerably smaller for 1985 and 1990-1992.

In the next three tables the estimated correlations coefficients by the three methods are given.

Table 4. The jackknifed correlation matrix

Year	1985	1986	1987	1988	1989	1990	1991	1992
1985	1.00	0.06	0.09	0.08	0.07	0.06	0.02	0.03
1986	0.06	1.00	0.05	0.06	0.05	0.05	0.08	0.05
1987	0.09	0.05	1.00	0.15	0.14	0.15	0.18	0.18
1988	0.08	0.06	0.15	1.00	0.20	0.25	0.21	0.19
1989	0.07	0.05	0.14	0.20	1.00	0.18	0.21	0.20
1990	0.06	0.05	0.15	0.25	0.18	1.00	0.27	0.32
1991	0.02	0.08	0.18	0.21	0.21	0.27	1.00	0.26
1992	0.03	0.05	0.18	0.19	0.20	0.32	0.26	1.00

Table 5. The bootstrapped correlation matrix

Year	1985	1986	1987	1988	1989	1990	1991	1992
1985	1.00	-0.02	0.06	0.05	0.07	0.06	0.04	0.02
1986	-0.02	1.00	0.05	0.07	0.08	0.02	0.07	0.05
1987	0.06	0.05	1.00	0.11	0.09	0.13	0.14	0.17
1988	0.05	0.07	0.11	1.00	0.13	0.22	0.17	0.21
1989	0.07	0.08	0.09	0.13	1.00	0.14	0.18	0.17
1990	0.06	0.02	0.13	0.22	0.14	1.00	0.23	0.31
1991	0.04	0.07	0.14	0.17	0.18	0.23	1.00	0.19
1992	0.02	0.05	0.17	0.21	0.17	0.31	0.19	1.00

Table 6. The correlation matrix obtained by Louis' method

Year	1985	1986	1987	1988	1989	1990	1991	1992
1985	1.00	-0.01	-0.00	-0.01	-0.01	-0.00	-0.00	+0.00
1986	-0.01	1.00	-0.01	-0.00	-0.01	-0.01	-0.00	-0.00
1987	-0.00	-0.01	1.00	-0.01	-0.02	-0.02	-0.00	-0.01
1988	-0.01	-0.00	-0.01	1.00	-0.06	-0.05	-0.03	-0.02
1989	-0.01	-0.01	-0.02	-0.06	1.00	-0.07	-0.4	-0.02
1990	-0.00	-0.01	-0.02	-0.05	-0.07	1.00	-0.08	-0.03
1991	-0.00	-0.00	-0.00	-0.03	-0.04	-0.08	1.00	-0.05
1992	+0.00	-0.00	-0.01	-0.02	-0.02	-0.03	-0.05	1.00

If the data had been completely observed the estimated rates would have been uncorrelated. Here, however, one might expect negative correlation between the estimated rates. The reason for this is that the estimated number of test carried out in one year affects the corresponding number in other years. Louis' method gave negative correlations, but both bootstrapping and jackknifing gave positive correlations. This may be explained by considering the resampling plans. Excluding individuals with a high number of tests makes the estimates lower than the average.

Furthermore for most of the tests we know what year they were carried out. One would thus here expect quite small correlations. This holds true for the correlations obtained with Louis' method, but some of the correlations obtained by jackknifing and bootstrapping are fairly large.

2.6 Discussion

The preceding paragraph showed that different methods of obtaining the standard errors and correlations of the estimated rates gave somewhat different results. A likely explanation is that the model does not fit the data very well. If the model was correct then the total number of tests per individual Y_i^1 would be Poisson distributed with expectation $\lambda = \frac{1}{2}\lambda_1 + \sum_{j=2}^{7} \lambda_j + \frac{11}{12}\lambda_8$. The estimate for λ would be $\widehat{\lambda} = \frac{1115}{4667} \approx 0.24$. With this value of λ it would be very unlikely to observe any individual with 20 or more tests. A goodness of fit test on the number of individuals with 0, 1, 2 and 3+ tests given at the beginning of section 2.4 gives a Pearson $X^2 = 1180$ to be compared to a χ^2 distribution with 2 degrees of freedom. Thus there is some sort of heterogeneity in the data. This also shows up in the next section where the rates until first test are estimated to be considerably lower than the total population rates. Such a discrepancy would not occur if all individuals has the same rates of taking tests.

As the model is likely not correct it is not immediately clear what is being estimated. However, what is actually being done is that the likelihood of a misspecified model is maximized. In the sense of Hjort (1992) it can be shown that such an estimator is consistent for the least false parameter value. Furthermore such an estimator will be approximately normally distributed with a covariance matrix that will be estimated by the jackknifing and bootstrapping procedures that have been applied. It thus appears to be better to use the standard errors from one of these approaches than a method that is in meaningful only under a possibly false model.

Also, if the individual number of tests in each year X_{ji} had been known one might very well calculate the $\widehat{\lambda}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ji}$ as summary measure of total test activity regardless of whether the model is correct or not. With the present dataset a high percentage of X_{ji}'s are known and any sensible distributing method to obtain estimates of the remaining will not deviate much from $\widehat{\lambda}_j$. Indeed the estimation method described by (2) and (3) are quite natural under a certain heterogeneity model, namely that X_{ji} are independent and Poisson distributed with expectation λ_{ji} where

$$\lambda_{ji} = Z_i \lambda_j \tag{5}$$

where the Z_i's are independent and follow some common distribution. In this case we have given $\mathcal{Y}_i = (y, j, k)$ that $(X_{ji}, ..., X_{ki})$ has a multinomial distribution with y trials, probabilities $p_{ir} = \lambda_{ri} / \sum_{r=j}^{k} \lambda_{ri} = \lambda_r / \sum_{r=j}^{k} \lambda_r$ that is truncated to $X_{ji} \geq 1$ and $X_{ki} \geq 1$. Thus $\hat{X}_{ji}(\lambda)$ retains the property of being the conditional expectations $E[X_{ji} | \mathcal{Y}_i]$. However as the model for the complete data is not in a exponential class, the previously suggested EM-algorithm will no longer give the MLE.

3 Rates of first HIV-test

For estimation of rates until first test one is essentially faced with a usual right censoring problem. The only complicating factor is that only the year of taking the test is reported. With such data one commonly applies the so-called actuarial method (Cox & Oakes, 1984) giving estimates

$$\hat{\lambda}_{0j} = d_j / (r_j - \frac{1}{2} d_j)$$

where r_j is the number not yet tested in the beginning of year number j and d_j the number of tests in year j. (In general one will also in the actuarial estimator correct for the number of withdrawals in year j, but in the present data there are no withdrawals).

As discussed by Cox & Oakes the actuarial estimator is an approximation to the MLE when the λ_{0j} are small. They derive the full likelihood based on such data. One may, however, also find the MLE by an EM-algorithm. Define T_i as the possibly censored time until taking a test. Given rates $\lambda_{0j} = \lambda$ we find that for T_i we have

$$T(\lambda) = E[T_i - (j-1) | (j-1) \leq T_i < j, \lambda_{0j} = \lambda] = \frac{1 - \exp(-\lambda) - \lambda \exp(-\lambda)}{\lambda(1 - \exp(-\lambda))}.$$

Now the EM-algorithm may be stated

E-Step: Given estimates $\lambda_{0j}^{(r)}$ and observed incomplete data $I(j-1 \leq T_{ij} < j)$ calculate estimates of $T_i - (j-1)$ by $T(\lambda_{0j}^{(r)})$.

M-Step: Update estimates for λ_{0j} by the occurrence-exposure rates

$$\lambda_{0j}^{(r+1)} = \frac{d_j}{r_j - d_j T(\lambda_{0j}^{(r)})}.$$

The connection with the actuarial method is realized when noting that $T(\lambda) \approx \frac{1}{2}$ when λ is small (e.g. $\lambda < 0.1$).

On the present data the actuarial estimator and the MLE differed on the fifth valid digit, so the rates in Table 4 represent both the estimators. The standard error is calculated by the usual estimator $\sqrt{d}_j/(r_j - \frac{1}{2}d_j)$. Table 4 indicates that the rates have been at about 2-3 per 100 person-years from 1988 and on. Furthermore comparing with Table 1 we see that the total rates are higher than the first test rates. A previous HIV-test can thus be taken as an indicator of a new test and it is of interest to estimate the rates of testing after the first test.

Table 4. Rates of first HIV-test

Year	Rate per 100 person-years	Standard error
1985	1.16	0.11
1986	0.77	0.13
1987	1.51	0.18
1988	2.39	0.23
1989	2.49	0.23
1990	2.96	0.25
1991	2.77	0.22
1992	2.20	0.20

4 Extended models

In this note it has been developed an EM-algorithm for obtaining the MLE of the rates in inhomogeneous Poisson processes when only the total number of events along with the time of first and last time event is recoreded. Furthermore an EM-algorithm for estimating the rate to the first event is developed and compared to the commonly applied actuarial method. The methods have been applied to data on HIV-testing. Regarding the total HIV-test rates, however, a careful look at the data indicates that an inhomogeneous Poisson process is not an alltogether adequate model for the HIV-test activity in Norway, and so in order to describe the data better there is a need to construct better models.

One such extension, the heterogeneity model (5) has been mentioned. In this model however one will need to specify the distribution of the heterogeneity values Z_i's. Several choices are possible. One might think of the population as consisting of two subpopulations with heterogeneity values $z_1 = 1$, say, and z_2, respectively. This will induce a model with two new parameters z_2 and the proportion p_1 with $Z_i = z_1$. A natural extension of such a model would be to assume that the population consisted of three or more subpopulations with unknown heterogeneity values and / or proportions in the subpopulations. Also in line with a typical heterogeneity model one might assume that the Z_i's have a gamma distribution with $EZ_i = 1$, say, and an unknown variance. Of course other, typically skewed, distributions could be appropriate.

Another way of extending the model is related to the approach in section 3 where it was assumed that all individuals had the same rate of taking their first HIV-test. One may additionally assume that all individuals who have taken one (or more) HIV-tests have the same rate of taking a new test. Thus with T_i being the time of the first test the rate of individual i taking a test at time t would be

$$\lambda_i(t) = \begin{cases} \lambda_{j0} & \text{if } j \leq t < j+1 \text{ and } t \leq T_i \\ \\ \lambda_{j+} & \text{if } j \leq t < j+1 \text{ and } t > T_i \end{cases} \tag{6}$$

In principle one could fit models with a new set of rates whenever a new test was taken, but the data will hardly give reliable estimates under such a model. Another approach with a better chance of success could be to let a function of the previous number of tests be a covariate in a proportional hazard model. Indeed if the covariate was the indicator of a previous test $Z_i^{'}$ a special case of model (6) could be $\lambda_{j+} = exp(\beta Z_i^{'})$.

One could also try to combine and thus distinguish between the two above approaches by letting the rate of a test, now being called $\alpha_i(t)$, depend on a heterogeneity value Z_i through $\alpha_i(t) = Z_i\lambda_i(t)$ where $\lambda_i(t)$ is given by (6). With a small estimated variance of Z_i the approach of previous test would be considered the best and vice versa with estimates $\widehat{\lambda}_{j0} \approx \widehat{\lambda}_{j+}$ the heterogeneity would be the better choice.

References

[1] Cox, DR and Oakes, D (1984) Analysis of Survival Data. Chapman and Hall. London.

[2] Dempster AP, Laird, N, Rubin, DB (1977) Maximum likelihood estimation from incomplete data via the EM algorithm. J Roy Statist Assoc, Ser B 39, 1-38.

[3] Hjort, NL (1992) On inference in parametric survival data models. Int Statist Rev 60, 355-387.

[4] Louis, TA (1982) Finding the observed information matrix when using the EM algorithm. J Roy Statist Soc, Ser B 44, 226-233.

[5] Magnus, P, Samuelsen, SO, Eskild, A and Stigum, H (1994) Omfang av HIV-testing i den norske befolkning sett i sammenheng med seksualadferd. Tidskr Nor Lægeforen 114; 3064-3067 (in Norwegian).

[6] Stigum, H, Magnus, P, Veierød, M and Bakketeig, LS (1995) Impact of sexually transmitted disease spread of increased condom use by young females, 1987-1992. Int J Epid 24, 813-820.

[7] Tanner, MA (1993) Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions. Springer-Verlag. New York.