Analysis of Incomplete Data in Competing Risks Model

Masami Miyakawa
Tokyo Institute of Technology, Tokyo

Key Words—**Competing risks model, Incomplete data, Maximum likelihood estimator, Uniformly minimum variance unbiased estimator, Kaplan-Meier estimator, EM algorithm.**

Reader Aids—
Purpose: Extend state of the art
Special math needed for explanations: Mathematical statistics
Special math needed to use results: Same
Results useful to: Reliability & actuarial statisticians

Abstract—**Parametric and nonparametric estimation methods are proposed for reliability characteristics of one failure mode under competing risks with incomplete data; some of the failure times are observed without observing the cause of failure. The efficiencies of these methods are discussed.**

1. INTRODUCTION

The theory of competing risks has been developed by demographers and biostatisticians who have been concerned with the assessment of a specific risk in the presence of other risks. In reliability theory a series system of components can be considered as an individual who is subject to a number of risks, and each component can be regarded as the failure mode (cause of failure) [3]. In reliability engineering, in order to design a system satisfying the reliability requirements, it is essential to consider reliability characteristics of the components. For this purpose, the reliability characteristics of components can be estimated from life-test and field-life data. In the analysis of field-life data, since the observed values are for the systems, the competing risks model has been used for inferring the characteristics of a particular component. This paper assumes that there are only two failure modes — because once the purpose is to consider the characteristics of a specific failure mode, the model with two failure modes never loses its generality. The competing risks model assumes that the data consist of a failure time and an event indicator denoting the cause of failure, and in the presence of censoring an incomplete observation of the failure time necessarily precludes observation of the indicator. Several studies have been carried out under this assumption for parametric [1, 2, 11], and nonparametric cases [6, 7, 13]. However, in general, investigation of the cause of failure is expensive and requires time, and hence sometimes the cause of failure is not observed, even if the failure time is observed [5, 9, 10]. This paper considers such incomplete data where some of the failure times are observed without observing the cause of failure. In order to analyse the incomplete data, it is assumed that these failure times are from the same population as the complete data; that is, the population remains unchanged irrespective of the knowledge of the cause of failure. Under these assumptions, parametric and nonparametric estimation methods for life characteristics of a failure mode from the incomplete data are proposed. In the parametric case, under the assumption of exponentiality, the uniformly minimum variance unbiased estimator of the failure rate of a failure mode and its variance are derived. In the nonparametric case, under an additional assumption, estimators of the reliability function based on the Kaplan-Meier estimator [7] are derived and the properties of these estimators are investigated by Monte Carlo simulation.

2. MODEL

Notation

X_i	lifetime of system i
X_{ji}	lifetime of mode j for system i; $j = 1, 2$
$F(t)$	Cdf of X_i
$F_j(t)$	Cdf of X_{ji}; $j = 1, 2$
$\bar{\psi}(t)$	$1 - \psi(t)$; ψ is any function
δ_i	indicator variable denoting the cause of failure for system i
$I[\bullet]$	indicator function of event $[\bullet]$

Other, standard notation is given in "Information for Readers & Authors" at rear of each issue.

Assumptions

1. (X_{1i}, X_{2i}) $i = 1, 2, \ldots, n$ are n s-independent identically distributed (iid) pairs of random variables.
2. X_{1i} and X_{2i} are s-independent for all i; hence $\bar{F}(t) = \bar{F}_1(t)\,\bar{F}_2(t)$.
3. $\Pr\{X_{1i} = X_{2i}\} = 0$.
4. The observable quantities are:

$$X_i \equiv \min(X_{1i}, X_{2i}),\ i = 1, 2, \ldots, n$$

$$\delta_i \equiv j, \text{ if } X_i = X_{ji};\ j = 1, 2;\ i = 1, 2, \ldots, m\ (m \le n).$$

That is to say, the first m observations consist of the system failure time and cause of failure, while remaining observations include only the failure time (see figure 1).

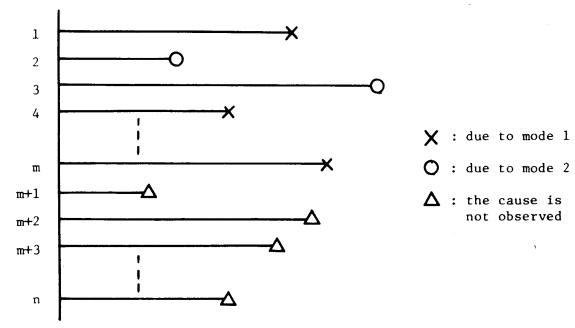

Fig. 1. The incomplete data.

3. PARAMETRIC CASE

The likelihood function for observed data x_i ($i = 1, 2, \ldots, n$) and δ_i ($i = 1, 2, \ldots, m$) is:

$$L = \prod_{i=1}^{m} [dF_1(x_i)\bar{F}_2(x_i)]^{I[\delta_i=1]}[dF_2(x_i)\bar{F}_1(x_i)]^{I[\delta_i=2]} \prod_{i=m+1}^{n} dF(x_i). \tag{3.1}$$

Hence, when the parameter vector of $F_j(t)$ denoted by θ_j, the maximum likelihood estimator of $\theta \equiv (\theta_1, \theta_2)$ is $\hat{\theta}$, which is the solution to $\partial/\partial\theta(\log L) = 0$. However, generally speaking, the solution cannot be obtained in closed form, and hence numerical methods have to be employed. One exception is the case when $F_1(t)$ and $F_2(t)$ are both exponential distributions. This case is explained in detail.

Exponential Case

Let λ_j be the failure rate for $F_j(t)$. Under the exponentiality assumption, the likelihood function is:

$$L = \lambda_1^r \lambda_2^{m-r} (\lambda_1 + \lambda_2)^{n-m} \exp\left[-(\lambda_1 + \lambda_2) \sum_{i=1}^{n} x_i\right], \tag{3.2}$$

$r \equiv \sum_{i=1}^{m} I[\delta_i = 1]$ denotes the number of systems that failed due to mode 1. Thus, by solving:

$$\frac{\partial}{\partial\lambda_1}\log L = r\frac{1}{\lambda_1} + (n-m)\frac{1}{\lambda_1 + \lambda_2} - \sum_{i=1}^{n} x_i = 0 \tag{3.3a}$$

$$\frac{\partial}{\partial\lambda_2}\log L = (m-r)\frac{1}{\lambda_2} + (n-m)\frac{1}{\lambda_1 + \lambda_2} - \sum_{i=1}^{n} x_i = 0, \tag{3.3b}$$

the maximum likelihood estimator of λ_1 is:

$$\hat{\lambda}_1 = \frac{r}{m}\frac{n}{\sum_{i=1}^{n} x_i}. \tag{3.4}$$

Under the above conditions, $\left(\sum_{i=1}^{n} X_i, r = \sum_{i=1}^{m} I[\delta_i = 1]\right)$ is a s-sufficient and complete set of statistics for (λ_1, λ_2) [8, p 132]. $\sum_{i=1}^{n} X_i$ follows the gamma distribution with a scale parameter $1/(\lambda_1 + \lambda_2)$ and a shape parameter n, and r follows the binomial distribution with a parameter $\lambda_1/(\lambda_1 + \lambda_2)$ and an index m. Moreover, $\sum_{i=1}^{n} X_i$ and r are s-independent. Hence, the uniformly minimum variance unbiased estimator of λ_1 is:

$$\lambda_1^* = \frac{r}{m}\frac{n-1}{\sum_{i=1}^{n} X_i}.$$

and its variance is:

$$\mathrm{Var}\{\lambda_1^*\} = \frac{m\lambda_1^2 + (n-1)\lambda_1\lambda_2}{m(n-2)}.$$

It is of interest to see how the variance of the estimator can be reduced by using the data excluding the cause of failure as compared to not using failure-cause data at all. Figure 2 indicates how the variance of λ_1^*, which is represented as a function of $(n\text{-}m)$, decreased as $(n\text{-}m)$ increases when $m = 10$.

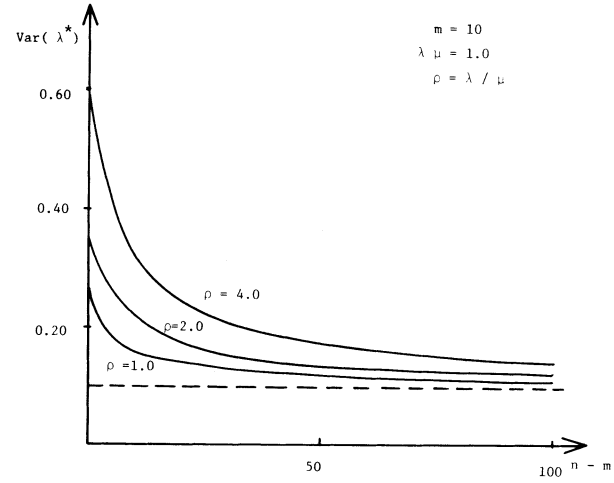

Fig. 2. Variance of λ^*.

4. NONPARAMETRIC CASE

The purpose of this section is to obtain the nonparametric estimators for the reliability function $\bar{F}_1(t)$ from the incomplete data. It is assumed that there exist ties in the field lifetimes data, since they are measured in time units of either months, weeks, days, etc. Therefore, there is a possibility of two or more failures in different systems being observed at a particular point of time. These failures might be caused by similar or different failure modes. But assumption 3 above rules out the possibility of more than one component failure occurring in a particular system at a given moment of time.

If $n = m$, ie, all the data consist of a failure time and a cause of failure, the Kaplan-Meir estimator [7] is an estimator of $\bar{F}_1(t)$. This estimator is nonparametric maximum likelihood and has strong s-consistency [7, 13]. If $n > m$, the problem is to derive a way of using X_i ($i = m + 1, \ldots, n$) for the estimation of $\bar{F}_1(t)$. Let $p_1(t)$ be the conditional probability of the cause of failure due to mode 1 given that the failure time of system X is t:

$$p_1(t) \equiv \Pr\{X = X_1 | X = t\}. \tag{4.1}$$

Then the likelihood function (3.1) can be rewritten in terms of $F(t)$ and $p_1(t)$:

$$L = \prod_{i=1}^{m} p_1(x_i)^{I[\delta_i=1]}[1 - p_1(x_i)]^{I[\delta_i=2]} \prod_{i=1}^{n} dF(x_i). \tag{4.2}$$

Hence the maximum likelihood estimator of $p_1(t)$ can be obtained only at the observed time points $x_1, x_2, \ldots, x_m$:

$$\hat{p}_1(x_i) = \frac{\sum_{j=1}^{m} I[x_j = x_i, \delta_j = 1]}{\sum_{j=1}^{m} I[x_j = x_i]}. \tag{4.3}$$

The derived maximum likelihood estimator of $F(t)$ is an ordinary empirical distribution. In the continuous case, $F_1(t)$ is functionally related to $p_1(t)$ and $F(t)$ by:

$$\bar{F}_1(t) = \exp\left\{ - \int_0^t p_1(s) \frac{dF(s)}{1 - F(s)} \right\}, \tag{4.4a}$$

and in the discrete case, by:

$$\bar{F}_1(t) = \exp\left\{ \sum_{\substack{s \le t \\ s: \text{ jump point}}} p_1(s) \log \frac{1 - F(s)}{1 - F(s - 0)} \right\}, \tag{4.4b}$$

where $F(s - 0) \equiv \lim_{\epsilon \to 0} F(s - \epsilon)$.

Thus, X_i $(i = m + 1, \ldots, n)$ cannot contribute to the estimation of $F_1(t)$ in the nonparametric case where there is no information about the cause of failure corresponding to the failure time. Thus an additional assumption is needed, viz, both $F_1(t)$ and $F_2(t)$ have continuous and smooth failure rate functions $\lambda_1(t)$ and $\lambda_2(t)$ respectively. This assumption seems quite reasonable even in the nonparametric case since failure times can be thought of as continuous variables. Under the above assumption, several iterative estimation methods are proposed.

Estimation Methods

The estimation methods proposed here consist of two steps and apply the concept of the self-consistent estimator proposed by Efron [6] and the EM algorithm developed by Dempster et al. [4]. Suppose that the values of x_i $(i = 1, 2, \ldots, n)$ are arranged in ascending order. In the presence of ties, let $t_1 < t_2 < \ldots < t_{n'} (n' \le n)$ denote the n' distinct system failure times that are obtained by this arrangement. For failure times satisfying values of t_i, let d_{ji} be the number of failures due to mode j, $j = 1, 2$, and d_{0i} be the number of failures where the cause of failure is not observed. Define $d_i \equiv d_{1i} + d_{2i} + d_{0i}$, and $n_i \equiv \sum_{j=i}^{n'} d_j$.

Step A: Estimation of $\bar{F}_1(t)$ and $\bar{F}_2(t)$, given an estimate of $p_1(t)$

Suppose that $p_1^*(t_i)$ which denote estimates of $p_1(t_i)$ have been obtained for all i. One idea is that for each i the pseudo-data which contain $d'_{ji} = d_{ji} + p_1^*(t_i)\, d_{0i}$ failures due to mode j are generated based on $p_1^*(t_i)$. Based on this idea, which is similar to the one by Turnbull [4], as an extension of the Kaplan-Meier estimator, a pair of estimators can be obtained:

$$\hat{\bar{F}}_j(t) = \prod_{t_i \le t} \left(\frac{n_i - d'_{ji}}{n_i} \right), j = 1, 2. \tag{4.5}$$

However, it is unreasonable for the product $\hat{\bar{F}}_1(t)\hat{\bar{F}}_2(t)$ to disagree with the empirical Cdf, $\bar{F}(t)$. Another idea is that it is possible to obtain an estimator of $F_1(t)$ by substituting the estimators of $p_1(s)$ and $F(s)$ for the original values of $p_1(s)$ and $F(t)$ in (4.4b). Based on this idea, another pair of estimators is:

$$\tilde{\bar{F}}_j(t) = \prod_{t_i \le t} \left(\frac{n_i - d_i}{n_i} \right)^{d'_{ji}/d_i}. \tag{4.6}$$

Step B: Estimation of $p_1(t)$, given estimates of $\bar{F}_j(t)$

Once $\bar{F}_j^*(t)$, which denotes the estimate of $\bar{F}_j(t)$, has been obtained, we can construct the estimators of $p_1(t)$ on the assumption that the failure rate functions $\lambda_1(t)$ and $\lambda_2(t)$ are continuous and smooth, based on the equation $p_1(t) = \lambda_1(t)/[\lambda_1(t) + \lambda_2(t)]$. When we intend to estimate $p_1(t)$ for t_i where $d_i = d_{0i} > 0$, it is important to consider what portion of $t_1 < t_2 < \ldots < t_{n'}$ has to be used. If we use preceding failure times t_k's $(k < i)$ and successive failure times $t_{k'}$'s $(k' > i)$ jointly, iterative calculation using a computer is necessary until the estimates of $p_1(t)$ and $\bar{F}_j(t)$ converge. However, if we use only preceding failure times, such calculation is no longer necessary. For example, in the case of $d_{ji} = 0$, a simple estimator based on each immediately preceding failure time for mode j is:

$$\hat{p}_1(t_i) = \frac{\hat{\lambda}_1(t_i)}{\hat{\lambda}_1(t_i) + \hat{\lambda}_2(t_i)} \tag{4.7}$$

$$\hat{\lambda}_j(t_i) \equiv \frac{\bar{F}_j^*(t_{-ji} - 0) - \bar{F}_j^*(t_{-ji})}{\bar{F}_j^*(t_{-ji})\,(t_i - t_{-ji})}, j = 1, 2$$

where t_{-1i} is a maximum value of t_k such that $k < i$ and $d_{1k} > 0$; t_{-2i} is a maximum value of t_k such that $k < i$ and $d_{2k} > 0$; $\hat{\lambda}_j(t_i) \equiv 0$ if the corresponding values of t_{-ji} do not exist; $\hat{p}_1(t_i)$ is assigned a value of 1/2 when $\hat{\lambda}_j(t_i) = 0$. Whereas, if $d_{1i} > 0$ or $d_{2i} > 0$, $\hat{p}_1(t_i)$ is given by (4.3). By using (4.7), estimates of $p_1(t_i)$ and $\bar{F}_j(t_i)$ can be calculated in the order $t_1 < t_2 < \ldots < t_{n'}$ and these values once determined never change. For comparison, consider another simple estimator which uses both immediately preceding and successive failure times in the case of $d_{ji} = 0$:

$$\tilde{p}_1(t_i) = \frac{\tilde{\lambda}_1(t_i)}{\tilde{\lambda}_1(t_i) + \tilde{\lambda}_2(t_i)}, \tag{4.8}$$

$$\tilde{\lambda}_j(t_i) \equiv \frac{\bar{F}_j^*(t_{-ji} - 0) - \bar{F}_1^*(t_{+ji} - 0)}{\bar{F}_j^*(t_i)\,(t_{+ji} - t_{-ji})}, j = 1, 2$$

where the values of t_{-ji} are the same as given before, and t_{+1i} is a minimum value of t_k such that $k > i$ and $d_{1k} > 0$, and t_{+2i} is a minimum value of t_k such that $k > i$ and $d_{2k} > 0$. In this case, if one of the values among t_{-1i} and t_{+1i} or t_{-2i} and t_{+2i} does not exist, we use only the existing one and substitute it in (4.7) or in an equation similar to (4.7). If $d_{1i} > 0$ or $d_{2i} > 0$, then $\tilde{p}_1(t_i)$ is given by (4.3).

Simulation Study

The biases and the mean square errors of the following five estimators for $F_1(t)$ are investigated by Monte Carlo simulation.

A: estimator obtained from the combination of (4.5) and (4.7)
B: estimator obtained from the combination of (4.6) and (4.7)
C: estimator obtained from the combination of (4.5) and (4.8)
D: estimator obtained from the combination of (4.6) and (4.8)
E: Kaplan-Meier estimator based on (X_i, δ_i) $i = 1, 2, \ldots, m$

Figures 3 and 4 indicate estimates of the biases and the mean square errors respectively obtained from 3000 random numbers for $n = 20$, $m = 10$, and $F_1(t)$ and $F_2(t)$ are both exponential distributions with scale parameter 1. The following findings were obtained.

1. Estimators obtained in A and C are larger than those obtained in B and D respectively.
2. D has the smallest mean square error, and B is not so inferior to D.
3. A, B, C, D are all superior to E in terms of mean square errors.

Under other conditions, similar results were obtained where D is the best estimator among these five estimators. However, iterative numeration is necessary for the calculation of D and therefore B is more convenient.

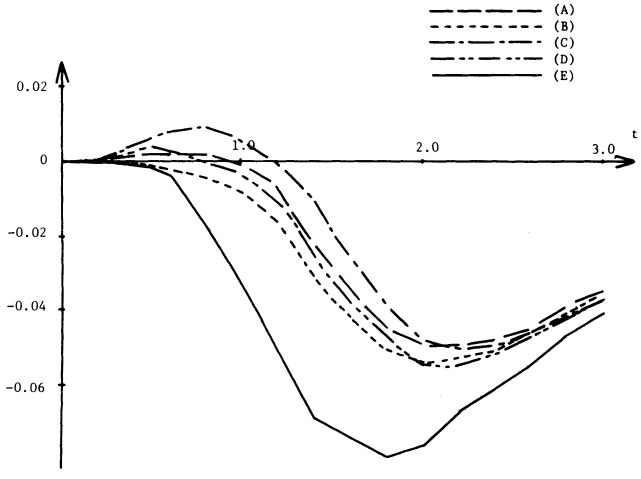

Fig. 3. The biases of estimators for $\bar{F}_1(t)$.

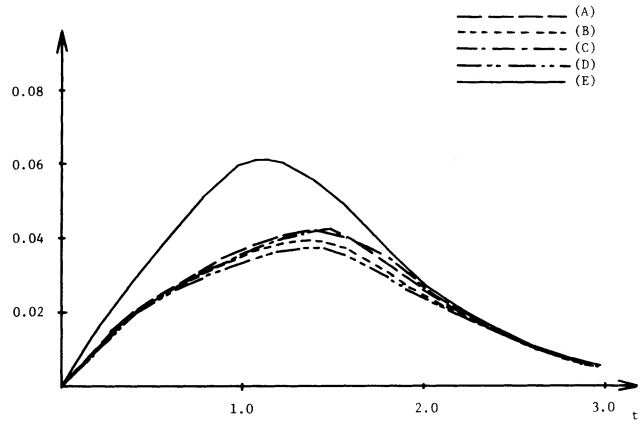

Fig. 4. The mean square errors of estimators for $\bar{F}_1(t)$.

REFERENCES

[1] J. Berkson, L, Elveback, "Competing exponential risks with particular inference to the study of smoking lung cancer," *J. Amer. Statist. Assoc.*, vol 55, 1960, pp 415-428.
[2] D. R. Cox, "The analysis of exponentially distributed life times with two types of failures," *J. Roy. Statist. Soc.* B, vol 21, 1959, pp 411-421.
[3] H. A. David, M. L. Moeschberger. *The Theory of Competing Risks,* Griffin, 1978.
[4] A. P. Dempster, N. M. Laird, D. B. Rubin, "The maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc.* B, vol 39, 1977, pp 1-22.
[5] G. E. Dinse, "Nonparametric estimation for partially-incomplete time and types of failure data," *Biometrics,* vol 38, 1982, pp 417-431.
[6] B. Efron, "The two sample problem with censored data," *Proc. Fifth Berkeley Symp. Math. Statist. Prob.*, 1967, pp 831-853.
[7] E. L. Kaplan, P. Meier, "Nonparametric estimation from incomplete observation," *J. Amer. Statist. Assoc.*, vol 53, 1958, pp 457-481.
[8] E. L. Lehmann. *Testing Statistical Hypotheses,* John White, 1959.
[9] M. Miyakawa, "Statistical analysis of incomplete data in competing risks model," *J. Japanese Soc. for Quality Control,* vol 12, 1982, pp 23-29 (in Japanese).
[10] M. Miyakawa, H. Yamamoto, "Nonparametric estimation of reliability function in competing risks model," *Abstract in Conf. Japanese Soc. for Quality Control,* 1983, pp 49-52 (in Japanese).
[11] T. Miyamura, "Statistical analysis of reliability data in randomly censored life testing," *J. Opr. Res. Soc. Japan,* vol 23, 1980, pp 191-204.
[12] M. L. Moeschberger, H. A. David, "Life tests under competing cause of failure and the theory of competing risks," *Biometrics,* vol 27, 1971, pp 909-933.
[13] A. P. Peterson Jr., "Expressing the Kaplan-Meier estimator as a function of empirical survival functions," *J. Amer. Statist. Assoc.*, vol 72, 1977, pp 854-858.
[14] B. W. Turnbull, "Nonparametric estimation of a survivorship function with doubly censored data," *J. Amer. Statist. Assoc.*, vol 69, 1974, pp 169-173.

AUTHOR

Masami Miyakawa; Department of Management Science and Engineering; Faculty of Engineering; Tokyo Institute of Technology; Tokyo 152 JAPAN.

M. Miyakawa was born in Chiba, Japan on 1957 February 1. He received his B.Eng. and M.Eng. degrees from Tokyo Institute of Technology in 1979 and 1981 respectively. He is an Assistant Professor in the Department of Management Science and Engineering at Tokyo Institute of Technology. His areas of interest include statistical analysis of life data with engineering and biomedical areas.

Manuscript TR84-077 received 1983 July 21; revised 1983 November 21.

★★★