CHAPTER 2

CONFIDENCE INTERVALS FOR VALUES OF SURVIVOR FUNCTION

2.1 Introduction

The survivor function, $S(t)$, is defined to be the probability that the survival time T of an individual is at least t, that is

$$S(t) = P(T \geq t). \qquad (2.1.1)$$

Suppose that $\hat{S}(t)$ denotes the Kaplan-Meier estimate of the survivor function $S(t)$. The classical approximate confidence interval for the survivor function $S(t)$ is given by

$$\hat{S}(t) \pm \text{a suitable multiple of the standard error of } \hat{S}(t), \qquad (2.1.2)$$

and the expression for the standard error of the Kaplan-Meier estimate can be obtained by using Greenwood's formula (see Greenwood (1926)).

In this chapter, an alternative confidence interval for the survivor function $S(t)$ is obtained by using the first four moments of the Kaplan-Meier estimate. A simulation study is carried out to compare the confidence interval based on Greenwood's formula and the proposed alternative confidence interval. The simulation results show that the alternative confidence interval tends to perform better both in coverage probability and expected length.

2.2 Standard Error of the Survivor Function

In general, suppose that there are n individuals with observed survival times $t_1, t_2, t_3, ..., t_n$. Some of these observations may be right-censored, and there may also be more than one individual with the same observed survival time. We therefore assume that there are *r* death times amongst the individuals, where $r \le n$. After arranging these death times in an ascending order, the *j*-th is denoted by $t_{(j)}$, for *j* = 1, 2, …, *r*, and so the *r* ordered death times are $t_1 < t_2 < \ldots < t_r$. Next let n_j be the number of individuals who are alive just before $t_{(j)}$, and d_j be the number of death at $t_{(j)}$. Assume that the deaths of the individuals in the sample occur independently of one another. Hence the estimated probability of survival from $t_{(j)}$ to t_{j+1} is given by

$$\hat{P}_j = (n_j - d_j)/n_j . \qquad (2.2.1)$$

Then the Kaplan-Meier estimate, $\hat{S}(t)$, of the survivor function at any time *t* satisfying $t_k \le t < t_{k+1}$ for some $1 \le k \le r$ is given by the following product of the $\hat{P}_j$:

$$\hat{S}(t) = \prod_{j=1}^{k} \hat{P}_j = \prod_{j=1}^{k} \frac{n_j - d_j}{n_j} . \qquad (2.2.2)$$

To find the expected value and variance of the Kaplan-Meier estimate, two important assumptions are required. First, the random variable $n_j \hat{P}_j$ has a binomial distribution with parameters n_j and P_j where P_j is the true probability of survival from $t_{(j)}$ to t_{j+1}. Second, the $\hat{P}_j$ are mutually independent. By using the transformation

$$\ln \hat{S}(t) = \sum_{j=1}^{k} \ln \hat{P}_j , \qquad (2.2.3)$$

and the Taylor series approximation to the variance of a function of a random variable, it can be shown that the standard error of $\hat{S}(t)$ is given approximately by

$$s.e.\{\hat{S}(t)\} \approx \hat{S}(t) \left\{ \sum_{j=1}^{k} \frac{d_j}{n_j (n_j - d_j)} \right\}^{\frac{1}{2}} . \qquad (2.2.4)$$

Equation (2.2.4) is known as Greenwood’s formula.

An approximate 100(1 - α)% confidence interval based on the Greenwood's formula for $S(t)$ is then given by

$$\left\{S(t): \hat{S}(t) - z_{\alpha/2}\, s.e.\left(\hat{S}(t)\right) \le S(t) \le \hat{S}(t) + z_{\alpha/2}\, s.e.\left(\hat{S}(t)\right)\right\}, \qquad (2.2.5)$$

where $z_{\alpha/2}$ is the (1-$\alpha/2$)-th quantile of the standard normal distribution.

Another estimate of the survivor function $S(t)$ is given by

$$\bar{S}(t;t^*,p^*) = \prod_{j:t_{(j)}\le t} \frac{n_j + \bar{\lambda} - d_j}{n_j + \bar{\lambda}}, \qquad 0 \le t \le t^* \qquad (2.2.6)$$

where $\bar{\lambda}$ is the unique solution to $\bar{S}(t^*;t^*,p^*) = p^*$. This estimate would be relevant under the null hypothesis $H_0 : S(t^*) = p^*$ (see Thomas and Grunkemeier (1975), Barber and Jennison (1999)). In the process of using Equation (2.2.6) to construct confidence intervals for the survivor function $S(t)$, again frequently only the first two moments of the estimate of the survivor function are used.

We may hope to get better confidence interval if we make use of the third and fourth moments of $\hat{S}(t)$ (or $\bar{S}(t;t^*,p^*)$) as well. In Section 2.3, the first four moments of $\hat{S}(t)$ are used to construct a confidence interval for the survivor function $S(t)$. Section 2.4 gives some numerical results to demonstrate the advantage of using higher moments in constructing a confidence interval for $S(t)$.

2.3 Approximate Confidence Interval Based on Higher Order Moments

The random variable $Y_j = n_j \hat{P}_j$ has a binomial distribution with parameters n_j and P_j. Thus the first four moments Y_j can be found. From the first four moments of Y_j, we can obtain the first four moments of $\hat{P}_j$. The results are as follows:

$$E\;\hat{P}_j\; = P_j, \tag{2.3.1}$$

$$E\;\hat{P}_j^{\,2}\; = \frac{1}{n_j}\left[P_j(1-P_j) + n_j P_j^{\,2}\right], \tag{2.3.2}$$

$$E\;\hat{P}_j^{\,3}\; = \frac{1}{n_j^{\;2}}\left[3P_j(1-P_j) + 3n_j P_j^{\,2} - 2P_j + P_j^{\,3}\;\; n_j - 1\;\;\; n_j - 2\;\;\right], \tag{2.3.3}$$

$$E\;\hat{P}_j^{\,4}\; = \frac{1}{n_j^{\;3}}\left[7P_j(1-P_j) + 7n_j P_j^{\,2} - 6P_j + 6P_j^{\,3}\;\; n_j - 1\;\;\; n_j - 2\right.$$

$$\left. + P_j^{\,4}\;\; n_j - 1\;\;\; n_j - 2\;\;\; n_j - 3\;\;\right]. \tag{2.3.4}$$

As the random variable $\hat{P}_1, \hat{P}_2, ..., \hat{P}_k$ are independent, the first four moments of

$$w = \hat{S}\;\;t\;\; = \prod_{j=1}^{k} \hat{P}_j\,, \tag{2.3.5}$$

will be given by

$$E(w^l) = \prod_{j=1}^{k} E(\hat{P}_j^{\;l})\,, \qquad\qquad 1 \le l \le 4. \tag{2.3.6}$$

Let m_l represent the *l*-th central moment of $\hat{S}(t)$, that is

$$m_l = E\left(\left[\hat{S}(t) - E(\hat{S}(t))\right]^l\right), \quad l = 2, 3, 4. \tag{2.3.7}$$

Furthermore let

$$\bar{m}_3 = m_3 \big/ m_2^{\;3/2} \tag{2.3.8}$$

and $$\bar{m}_4 = m_4 \big/ m_2^{\;2} \tag{2.3.9}$$

be respectively the measures of skewness and kurtosis of the distribution of $\hat{S}\;\;t\;\;$.

Consider the random variable ε_i given by

$$\varepsilon_i = \begin{cases} \lambda_1 e_i + \lambda_2 (e_i^{\,2} - \dfrac{(1+\lambda_3)}{2}), & e_i \ge 0 \\ \lambda_1 e_i + \lambda_2 (\lambda_3 e_i^{\,2} - \dfrac{(1+\lambda_3)}{2}), & e_i < 0 \end{cases}, \quad i = 1, 2, ..., n\,, \tag{2.3.10}$$

where $e_i \sim N(0,1)$ and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)^T$ is a vector of constants (see Pooi (2003)). The random variable ε_i is said to have a quadratic-normal distribution with parameters 0 and $\boldsymbol{\lambda}$, ($\varepsilon_i \sim \mathrm{QN}(0, \boldsymbol{\lambda})$).

As noted in Pooi (2003), the constants λ_i should be such that ε_i given by Equation (2.3.10) is a one-to-one function when $|e_i| < Z_q$ ($q > 0$ is a small value and Z_q is the $(1-q)100\%$ point of the standard normal distribution). By examining the extreme values of ε_i, it can be shown that ε_i is a one-to-one function of e_i for $|e_i| < Z_q$ provided that $-\lambda_1/(2\lambda_2\lambda_3) < -Z_q$ when $-\lambda_1/(2\lambda_2\lambda_3) < 0$, and $-\lambda_1/(2\lambda_2) > Z_q$ when $-\lambda_1/(2\lambda_2) \geq 0$.

The function ε_i in Equation (2.3.10) can be approximated by

$$\varepsilon_i = \lambda_1 e_i - \frac{1}{2}\lambda_2(1+\lambda_3) + \frac{2}{7}\lambda_2(1-\lambda_3)e_i + \frac{1}{2}\lambda_2(1+\lambda_3)e_i^2$$

$$+\frac{41}{180}\lambda_2(1-\lambda_3)e_i^3 - \frac{1}{72}\lambda_2(1-\lambda_3)e_i^5 + \frac{1}{2520}\lambda_2(1-\lambda_3)e_i^7 \qquad (2.3.11)$$

for $|e_i| < 4$.

The probability density function (p.d.f.) of ε_i is given by

$$f(\varepsilon_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}e_i^2} \left| \frac{de_i}{d\varepsilon_i} \right|$$

Figures 2.3.1 - 2.3.3 show the p.d.f. of ε_i for a number of different values of $\boldsymbol{\lambda}$. The figures show that skewed distributions and narrow waist distributions can be generated by suitable choices of the value of $\boldsymbol{\lambda}$.

Now, suppose that $\hat{\boldsymbol{\lambda}}$ is the value of $\boldsymbol{\lambda}$ such that

$$E\ \varepsilon_i^{l}\ = \hat{m}_l, \qquad l = 2, 3, 4. \qquad (2.3.12)$$

where $\hat{m}_l$ is the value of m_l evaluated at $P_i = \hat{P}_i$, $i = 1, 2, \ldots, k$.

Then an approximate $100(1-\alpha)\%$ confidence interval for $S(t)$ is given by

$$L \le S(t) \le U, \tag{2.3.13}$$

where

$$L = \hat{m}_1 + \hat{\lambda}_1 \;\; -z_{\alpha/2} \;\; + \hat{\lambda}_2 \left(\hat{\lambda}_3 \;\; -z_{\alpha/2} \;\; {}^2 - \frac{1+\hat{\lambda}_3}{2} \right), \tag{2.3.14}$$

and

$$U = \hat{m}_1 + \hat{\lambda}_1 \;\; z_{\alpha/2} \;\; + \hat{\lambda}_2 \left(\;\; z_{\alpha/2} \;\; {}^2 - \frac{1+\hat{\lambda}_3}{2} \right). \tag{2.3.15}$$

2.4 Numerical Results

Consider the case when the survival times $t_1,\ t_2,\ t_3,\ ...,\ t_n$ are generated from a Weibull distribution with parameters λ and γ (i.e. $T \sim$ Weibull (λ,γ)). The generation of the Weibull survival times may be achieved by using the following procedure:

(1) Generate a random number R from the uniform distribution U(0, 1).

(2) (a) Generate a non-censored t_i using

$$t_i = \left[-\frac{1}{\lambda} \ln(1-R) \right]^{\frac{1}{\gamma}}$$

(b) Generate a right-censored t_i using the following steps:

(i) Generate a random number U from the uniform distribution U(0, $Q_{0.9999}$) where $Q_{0.9999}$ is the 0.9999-quantile of the Weibull (λ,γ) distribution.

(ii) Generate a random number W from the Weibull (λ,γ) distribution.

(iii) If U < W, then U is a generated right-censored t_i. If U $\geq$ W, then go to step (i).

Suppose a total of *N* values of ($t_1\ t_2\ t_3\ ...\ t_n$) are generated. For a generated value of ($t_1\ t_2\ t_3\ ...\ t_n$) and a given value of $t = I \times 0.1 \times$ median of *T*, where *I* = 1, 2, …, 20, we first find all the t_i which are less than t. Suppose these t_i when arranged in an ascending order are

$$t_{(1)}, t_{(2)}, t_{(3)}, \ldots, t_{(k)}.$$

From these t_i, we compute the Type I confidence interval based on the Greenwood's formula for $\hat{S}(t)$ (see Section 2.2) and the Type II confidence interval based on the first four moments of $\hat{S}(t)$ (see Section 2.3). The coverage probability of a given type of confidence interval is estimated by the proportion of the confidence intervals (among the *N* confidence intervals) which cover the actual value of $S(t)$ given by

$$S(t) = \exp(-\lambda t^{\gamma}) . \tag{2.4.1}$$

The expected length of a given type of confidence interval is estimated by the average value of the lengths of the *N* computed confidence intervals.

The results for the coverage probabilities and expected lengths of the two types of confidence intervals are given in Table 2.4.1. The results show that the coverage probability of the Type II confidence interval tends to be closer to the target value 0.95 than that of the Type I confidence interval. Furthermore the expected length of the Type II confidence interval is very often much shorter than that of the Type I confidence interval.

2.5 Concluding Remarks

After finding the value of $\bar{\lambda}$ (see Equation (2.2.6)) we may set $P_j = (n_j + \bar{\lambda} - d_j)/(n_j + \bar{\lambda})$ for *j* satisfying $t_{(j)} \le t^*$ and use Equations (2.3.1) – (2.3.4) to find the first four moments of $\hat{S}(t^*)$. Thus it would be possible to extend the present work further to the case when we find confidence interval via testing of the null hypothesis $H_0 : S(t^*) = p^*$.