Spring 5-14-2016

Confidence Intervals For The Survival Function

Luke H. Yang
Southern Illinois University Carbondale, hzyang2@gmail.com

CONFIDENCE INTERVALS FOR THE SURVIVAL FUNCTION

by

Luke Yang

B.S., Maryville University, 2013

A Research Paper
Submitted in Partial Fulfillment of the Requirements for the
Master of Science

Department of Mathematics
in the Graduate School
Southern Illinois University Carbondale
May, 2016

RESEARCH PAPER APPROVAL

CONFIDENCE INTERVALS FOR THE SURVIVAL FUNCTION

by

Luke Yang

A Research Paper Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Master of Science

in the field of Mathematics

Approved by:

David J. Olive

S. Yaser Samadi

Kwangho Choiy

Graduate School
Southern Illinois University Carbondale
April 7, 2016

AN ABSTRACT OF THE RESEARCH PAPER OF

LUKE YANG, for the Master of Mathematics degree in MATHEMATICS, presented on APRIL 07, 2016, at Southern Illinois University Carbondale.

TITLE: CONFIDENCE INTERVALS FOR THE SURVIVAL FUNCTION

MAJOR PROFESSOR: Dr. David J. Olive

This manuscript, taken from Olive(2010, ch. 16), suggests confidence intervals for the survival function as estimated by the Kaplan Meier estimator and the empirical survival function.

ACKNOWLEDGMENTS

I would like to thank Dr. David Olive for his invaluable assistance and insights leading to the writing of this paper. My sincere thanks also goes to Dr. S. Yaser Samadi and Dr. Kwangho Choiy who are my graduate committee members for their patience and understanding during the two years of effort that went into the production of this paper.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

In the analysis of "time to event" data, there are n individuals and the time until an event is recorded for each individual. Typical events are failure of a product or death of a person or reoccurrence of cancer after surgery, but other events such as first use of cigarettes or the time that baboons come down from trees (early in the morning) can also be modeled. The data is typically right skewed and censored data is often present.

Censoring occurs because of time and cost constraints. A product such as light bulbs may be tested for 1000 hours. Perhaps 30% fail in that time but the remaining 70% are still working. These are censored: they give partial information on the lifetime of the bulbs because it is known that about 70% last longer than 1000 hours. Handling censoring and time dependent covariates is what makes the analysis of time to event data different from other fields of statistics.

Reliability analysis is used in *engineering* to study the lifetime (time until failure) of manufactured products while survival analysis is used in *actuarial sciences*, *statistics* and *biostatistics* to study the lifetime (time until death) of humans, often after contracting a deadly disease. In the *social sciences*, the study of the time until the occurrence of an event is called the analysis of event time data or event history analysis. In *economics*, the study is called duration analysis or transition analysis. Hence reliability data = failure time data = lifetime data = survival data = event time data.

For univariate survival analysis, there is a response but no predictors. Let $\log(t) = \ln(t) = \log_e(t)$, and $\exp(t) = e^t$.

One of the difficulties with survival analysis is that the response Y = survival time is usually not observed, instead the censored response is observed. In this thesis the data will be right censored, and "right" will often be omitted. In the following definition, note that both $T \geq 0$ and $Y \geq 0$ are nonnegative.

Definition 1. Let $Y \geq 0$ be the time until an event occurs. Then Y is called the survival time. The survival time is censored if the event of interest has not been observed. Let Y_i be the ith survival time. Let Z_i be the time the ith observation (possibly an individual or machine) leaves the study for any reason other than the event of interest. Then Z_i is the time until the ith observation is censored. Then the right censored survival time T_i of the ith observation is $T_i = \min(Y_i, Z_i)$. Let $\delta_i = 0$ if T_i is (right) censored ($T_i = Z_i$) and let $\delta_i = 1$ if T_i is not censored ($T_i = Y_i$). Then the univariate survival analysis data is $(T_1, \delta_1), (T_2, \delta_2), ..., (T_n, \delta_n)$. Alternatively, the data is $T_1, T_2^*, T_3, ..., T_{n-1}^*, T_n$ where the * means that the case was (right) censored. Sometimes the asterisk * is replaced by a plus +, and Y_i, y_i or t_i can replace T_i. In this manuscript we will assume that the censoring mechanism is independent of the time to event: Y_i and Z_i are independent.

For example, in a study breast cancer patients who receive a lumpectomy, suppose the researchers want to keep track of 100 patients for five years after receiving a lumpectomy (tumor removal). The response is time until death after a lumpectomy. Patients who are lost to the study (move or eventually refuse to cooperate) and patients who are still alive after the study are censored. Perhaps 15% die, 5% move away and so leave the study and 80% are still alive after 5 years. Then 85% of the cases are (right) censored. The actual study may take two years to recruit patients, follow each patient for 5 years, but end 5 years after the end of the two year recruitment period. So patients enter the study at different times, but the censored response is the time until death or censoring from the time the patient entered the study.

Definition 2. i) The cumulative distribution function (cdf) of Y is $F(t) = P(Y \leq t)$. Since $Y \geq 0$, $F(0) = 0, F(\infty) = 1$, and $F(t)$ is nondecreasing.

ii) The probability density function (pdf) of Y is $f(t) = F'(t)$.

iii) The survival function of Y is $S(t) = P(Y > t) = 1 - F(t)$. $S(0) = 1, S(\infty) = 0$ and $S(t)$ is nonincreasing.

CHAPTER 2

THE EMPIRICAL ESTIMATOR

Notation: Let the indicator variable $I_A(Y_i) = 1$ if $Y_i \in A$ and $I_A(Y_i) = 0$ otherwise. Often write $I_{(t,\infty)}(Y_i)$ as $I(Y_i > t)$.

Definition 3. If none of the survival times are censored, then the empirical survival function $\hat{S}_E(t) =$ (number of individual with survival times $> t$)/(number of individuals) $= a/n$. So

$$\hat{S}_E(t) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i > t) = \hat{p}_t =$$

sample proportion of lifetimes $> t$.

Assume $Y_1, ..., Y_n$ are iid with $Y_i \geq 0$. Fix $t > 0$. Then $I(Y_i > t)$ are iid binomial(1,$p = P(Y_i > t)$). So $n\hat{S}_E(t) \sim$ binomial(n,$p = P(Y_i > t)$). Hence $E[n\hat{S}_E(t)] = nP(Y > t)$ and $V[n\hat{S}_E(t)] = nS(t)F(t)$. Thus $E[\hat{S}_E(t)] = S(t)$ and $V[\hat{S}_E(t)] = S(t)F(t)/n = [S(t)(1 - S(t))]/n \leq 0.25/n$. Thus $SD[\hat{S}_E(t)] = \sqrt{V[\hat{S}_E(t)]} \leq 0.5/\sqrt{n}$. So need $n \approx 100$ for $SD[\hat{S}_E(t)] < 0.05$.

Let $t_{(1)} \leq t_{(2)} \leq \cdots \leq t_{(n)}$ be the observed ordered survival times (= lifetimes = death times). Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times. Let $d_i =$ number of deaths at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are no ties. If $m < n$ and some $d_i \geq 2$, then there are ties.

Then $\hat{S}_E(t)$ is a step function with $\hat{S}_E(0) = 1$ and $\hat{S}_E(t) = \hat{S}_E(t_{i-1})$ for $t_{i-1} \leq t < t_i$. Note that $\sum_{i=1}^{m} d_i = n$. The table below is useful for computing and plotting $\hat{S}_E(t)$ given the $t_{(i)}$ or given the t_i and d_i. Let $a_0 = n$ and $a_i = \sum_{k=1}^{n} I(T_i > t_i) =$ # of cases $t_{(j)} > t_i$ for $i = 1, ..., m$. Then $\hat{S}_E(t_i) = a_i/n = \sum_{k=1}^{n} I(T_i > t_i)/n = \hat{S}_E(t_{i-1}) - \dfrac{d_i}{n}$.

Let $\hat{S}(t)$ be the estimated survival function. Let $t(p)$ be the pth percentile of Y: $P(Y \leq t(p)) = F(t(p)) = p$ so $1 - p = S(t(p)) = P(Y > t(p))$. Then $\hat{t}(p)$, the estimated time when 100 p % have died, can be estimated from a graph of $\hat{S}(t)$ with "over" and

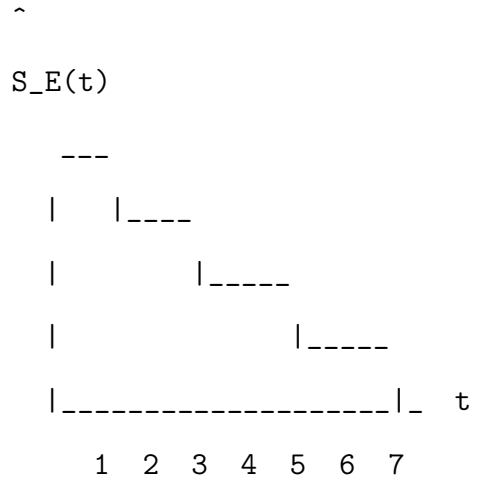
Table 2.1. Method for Computing Empirical Estimator

t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
$t_0 = 0$		$\hat{S}_E(0) = 1 = \frac{n}{n} = \frac{a_0}{n}$
t_1	d_1	$\hat{S}_E(t_1) = \hat{S}_E(t_0) - \frac{d_1}{n} = \frac{a_0 - d_1}{n} = \frac{a_1}{n}$
t_2	d_2	$\hat{S}_E(t_2) = \hat{S}_E(t_1) - \frac{d_2}{n} = \frac{a_1 - d_2}{n} = \frac{a_2}{n}$
$\vdots$	$\vdots$	$\vdots$
t_j	d_j	$\hat{S}_E(t_j) = \hat{S}_E(t_{j-1}) - \frac{d_j}{n} = \frac{a_{j-1} - d_j}{n} = \frac{a_j}{n}$
$\vdots$	$\vdots$	$\vdots$
t_{m-1}	d_{m-1}	$\hat{S}_E(t_{m-1}) = \hat{S}_E(t_{m-2}) - \frac{d_{m-1}}{n} = \frac{a_{m-2} - d_{m-1}}{n} = \frac{a_{m-1}}{n}$
t_m	d_m	$\hat{S}_E(t_m) = 0 = \hat{S}_E(t_{m-1}) - \frac{d_m}{n} = \frac{a_{m-1} - d_m}{n} = \frac{a_m}{n}$

"down" lines. a) Find $1-p$ on the vertical axis and draw a horizontal "over" line to $\hat{S}(t)$. Draw a vertical "down" line until it intersects the horizontal axis at $\hat{t}(p)$. Usually want $p = 0.5$ but sometimes $p = 0.25$ and $p = 0.75$ are used.

Example 1. Smith (2002, p. 68) gives steroid induced remission times for leukemia patients. The $t_{(j)}$, t_i and d_i are given in the following table. The a_i and $\hat{S}_E(t)$ needed to be computed. Note that $a_i = \#$ of cases with $t_{(j)} > t_i$.

The 2nd column $t_{(j)}$ gives the 21 ordered survival times. The 3rd column t_i gives the distinct ordered survival times. Often just the number is given, so $t_1 = 1$ would be replaced by 1. The 4th column d_i tells how many events (remissions) occurred at time t_i and the last column computes $\hat{S}_E(t_i)$. A good check is that the 1st column entry divided by n is equal to $a_i/n = \hat{S}_E(t_i) =$ last column entry. A graph of the estimated survival function would be a step function with times 0, 1, ..., 23 on the horizontal axis and $\hat{S}_E(t)$ on the vertical axis. A convention is to draw vertical lines at the jumps (at the t_i). So the step function would be 1 on (0,1), 19/21 on (1,2), ..., 1/21 on (22,23) and 0 for $t > 23$. The vertical lines connecting the steps are at $t = 1, 2, ..., 23$.



Example 2. If $d_i = 1, 1, 1, 1$ and if $t_i = 1, 3, 5, 7$, then $a_1 = 3$, $a_2 = 2$ and $a_3 = 1$. Hence $\hat{S}_E(1) = 0.75, \hat{S}_E(3) = 0.5, \hat{S}_E(5) = 0.25$, and $\hat{S}_E(7) = 0$, and the estimated survival function is graphed above.

Table 2.2. Example for Computing Empirical Estimator

a_i	$t_{(j)}$	t_i	d_i	$\hat{S}_E(t_i) = \hat{S}_E(t_{i-1}) - \frac{d_i}{n}$
21		$t_0 = 0$		$\hat{S}_E(0) = 1 = 21/21$
	1			
19	1	$t_1 = 1$	2	$\hat{S}_E(1) = (21-2)/21 = 19/21$
	2			
17	2	$t_2 = 2$	2	$\hat{S}_E(2) = (19-2)/21 = 17/21$
16	3	$t_3 = 3$	1	$\hat{S}_E(3) = (17-1)/21 = 16/21$
	4			
14	4	$t_4 = 4$	2	$\hat{S}_E(4) = (16-2)/21 = 14/21$
	5			
12	5	$t_5 = 5$	2	$\hat{S}_E(5) = (14-2)/21 = 12/21$
	8			
	8			
	8			
8	8	$t_6 = 8$	4	$\hat{S}_E(8) = (12-4)/21 = 8/21$
	11			
6	11	$t_7 = 11$	2	$\hat{S}_E(11) = (8-2)/21 = 6/21$
	12			
4	12	$t_8 = 12$	2	$\hat{S}_E(12) = (6-2)/21 = 4/21$
3	15	$t_9 = 15$	1	$\hat{S}_E(15) = (4-1)/21 = 3/21$
2	17	$t_{10} = 17$	1	$\hat{S}_E(17) = (3-1)/21 = 2/21$
1	22	$t_{11} = 22$	1	$\hat{S}_E(22) = (2-1)/21 = 1/21$
0	23	$t_{12} = 23$	1	$\hat{S}_E(23) = (1-1)/21 = 0$

Let $t_1 \leq t < t_m$. Then the classical large sample 95% CI for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\hat{S}_E(t_c) \pm 1.96\sqrt{\frac{\hat{S}_E(t_c)[1-\hat{S}_E(t_c)]}{n}} = \hat{S}_E(t_c) \pm 1.96SE[\hat{S}_E(t_c)].$$

Let $0 < t$ and let

$$\tilde{p}_{t_c} = \frac{n\hat{S}_E(t_c) + 2}{n+4}.$$

Then the Agresti and Coull (1998) plus four 95% CI for $S(t_c)$ based on $\hat{S}_E(t)$ is

$$\tilde{p}_{t_c} \pm 1.96\sqrt{\frac{\tilde{p}_{t_c}[1-\tilde{p}_{t_c}]}{n+4}} = \tilde{p}_{t_c} \pm 1.96SE[\tilde{p}_{t_c}].$$

The 95% large sample CI $\hat{S}_E(t_c) \pm 1.96SE[\tilde{p}_{t_c}]$ is also interesting. Alternative confidence intervals for a binomial parameter p could also be used. See Olive (2014, pp. 268-269, 285-286) and Agresti and Coull (1998) for references.

Example 3. Let $n = 21$ and $\hat{S}_E(12) = 4/21$.

a) Find the 95% classical CI for $\hat{S}_E(12)$.

b) Find the 95% plus four CI for $\hat{S}_E(12)$.

Solution: a)

$$\frac{4}{21} \pm 1.96\sqrt{\frac{\frac{4}{21}(1-\frac{4}{21})}{21}} = \frac{4}{21} \pm 0.16795 = (0.0225, 0.3584).$$

b)

$$\tilde{p}_{12} = \frac{21\frac{4}{21} + 2}{21+4} = \frac{6}{25}.$$

So the 95% CI is

$$\frac{6}{25} \pm 1.96\sqrt{\frac{\frac{6}{25}(1-\frac{6}{25})}{25}} = \frac{6}{25} \pm 0.16742 = (0.0726, 0.4074).$$

Note that the CIs are not very short since $n = 21$ is small.

CHAPTER 3

THE KAPLAN MEIER ESTIMATOR

Let $[0,\infty) = I_1 \cup I_2 \cup \cdots \cup I_m = [t_0,t_1) \cup [t_1,t_2) \cdots \cup [t_{m-1},t_m)$ where $t_0 = 0$ and $t_m = \infty$. It is possible that the 1st interval will have left endpoint > 0 ($t_0 > 0$) and the last interval will have finite right endpoint ($t_m < \infty$).

The Kaplan Meier estimator is used to estimate $S_Y(t) = P(Y > t)$ when there is censoring. Let $p_j = \mathrm{P}(\text{surviving through } I_j | \text{ alive at the start of } I_j) = P(Y > t_j | Y > t_{j-1}) = \dfrac{P(Y > t_j, Y > t_{j-1})}{P(Y > t_{j-1})} = \dfrac{S(t_j)}{S(t_{j-1})}$. Now $p_1 = S(t_1)/S(t_0) = S(t_1)$ since $S(0) = S(t_0) = 1$. Writing $S(t_k)$ as a telescoping product gives

$$S(t_k) = S(t_1)\frac{S(t_2)}{S(t_1)}\frac{S(t_3)}{S(t_2)}\cdots\frac{S(t_{k-1})}{S(t_{k-2})}\frac{S(t_k)}{S(t_{k-1})} = p_1 p_2 \cdots p_k = \prod_{j=1}^{k} p_j.$$

Let $\hat{p}_j = 1-$ (number dying in I_j)/(number with potential to die in I_j). Then $\hat{p}_j = 1 - d_j/n_j$ is the estimate of p_j used by the Kaplan Meier estimator.

Now suppose the data is censored but the event and censoring times are known. Let Y_i = time to event for ith person. Let $Y_i^* = T_i = \min(Y_i, Z_i)$ where Y_i and Z_i are independent and Z_i is the censoring time for the ith person (the time the ith person is lost to the study for any reason other than the time to event under study). The censored data is, for example, $y_1, y_2+, y_3, ..., y_{n-1}, y_n+$ where y_i means the time was uncensored and y_i+ means the time was censored. A status variable will be 1 if the time was uncensored and 0 if censored.

Let $\delta_i = I(Y_i \le Z_i)$ so $\delta_i = 1$ if T_i is uncensored and $\delta_i = 0$ if T_i is censored. Let $t_{(1)} \le t_{(2)} \le \cdots \le t_{(n)}$ be the observed ordered survival times. Let $\gamma_j = 1$ if $t_{(j)}$ is uncensored and 0, otherwise. Let $t_0 = 0$ and let $0 < t_1 < t_2 < \cdots < t_m$ be the distinct survival times corresponding to the $t_{(j)}$ with $\gamma_j = 1$. Let d_i = number of events (deaths) at time t_i. If $m = n$ and $d_i = 1$ for $i = 1, ..., n$ then there are no ties. If $m < n$ and some $d_i \ge 2$, then there are ties. Let $n_i = \sum_{j=1}^{n} I(t_{(j)} \ge t_i) = \#$ at risk at $t_i = \#$ alive and not

yet censored just before t_i.

Definition 4. The Kaplan Meier estimator = product limit estimator of $S_Y(t_i) = P(Y > t_i)$ is $\hat{S}_K(0) = 1$ and

$$\hat{S}_K(t_i) = \prod_{k=1}^{i}(1 - \frac{d_k}{n_k}) = \hat{S}_K(t_{i-1})(1 - \frac{d_i}{n_i}).$$

$\hat{S}_K(t)$ is a step function with $\hat{S}_K(t) = \hat{S}_K(t_{i-1})$ for $t_{i-1} \leq t < t_i$ and $i = 1, ..., m$. If $t_{(n)}$ is uncensored then $t_m = t_{(n)}$ and $\hat{S}_K(t) = 0$ for $t > t_m$. If $t_{(n)}$ is censored, then $\hat{S}_K(t) = \hat{S}_K(t_m)$ for $t_m \leq t \leq t_{(n)}$, but $\hat{S}_K(t)$ is undefined for $t > t_{(n)}$.

The table below is useful for computing and plotting $\hat{S}_k(t_i)$ given the $t_{(j)}$ and γ_j or given the t_i, n_i and d_i. Let $n_0 = n$. If f_{i-1} = number of events (deaths) and number censored in time interval $[t_{i-1}, t_i)$, then $n_i = n_{i-1} - f_{i-1}$ = number of $t_{(j)} \geq t_i$.

Example 4. Modifying Smith (2002, p. 113) slightly, suppose that the ordered censored survival times in days until repair of $n = 13$ street lights is 36, 38, 38, 38+, 78 112, 112, 114+, 162+, 189, 198, 237, 489+.

In general, a 95% CI for $S_Y(t_i)$ is $\hat{S}(t_i) \pm 1.96\ SE[\hat{S}(t_i)]$. If the lower endpoint of the CI is negative, round it up to 0. If the upper endpoint of the CI is greater than 1, round it down to 1. Do not use impossible values of $S_Y(t)$.

R plots the KM survival estimator along with the pointwise 95% CIs for $S_Y(t)$. If we guess a distribution for Y, say $Y \sim W$, with a formula for $S_W(t)$, then the guessed $S_W(t_i)$ can be added to the plot. If roughly 95% of the $S_W(t_i)$ fall within the bands, then $Y \sim W$ may be reasonable. For example, if $W \sim EXP(1)$, use $S_W(t) = \exp(-t)$. If $W \sim EXP(\lambda)$, then $S_W(t) = \exp(-\lambda t)$. Recall that $E(W) = 1/\lambda$.

If $\lim_{t\to\infty} tS_Y(t) \to 0$, then $E(Y) = \int_0^\infty t f_Y(t)dt = \int_0^\infty S_Y(t)dt$. Hence an estimate of the mean $\hat{E}(Y)$ can be obtained from the area under $\hat{S}(t)$.

Greenwood's formula is

$$SE[\hat{S}_K(t_j)] = \hat{S}_K(t_j)\sqrt{\sum_{i=1}^{j}\frac{d_j}{n_j(n_j - d_j)}}$$

Table 3.1. Method for Computing Kaplan Meier Estimator

t_i	n_i	d_i	$\hat{S}_K(t)$
$t_0 = 0$			$\hat{S}_K(0) = 1$
t_1	n_1	d_1	$\hat{S}_K(t_1) = \hat{S}_K(t_0)[1 - \frac{d_1}{n_1}]$
t_2	n_2	d_2	$\hat{S}_K(t_2) = \hat{S}_K(t_1)[1 - \frac{d_2}{n_2}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_j	n_j	d_j	$\hat{S}_K(t_j) = \hat{S}_K(t_{j-1})[1 - \frac{d_j}{n_j}]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_{m-1}	n_{m-1}	d_{m-1}	$\hat{S}_K(t_{m-1}) = \hat{S}_K(t_{m-2})[1 - \frac{d_{m-1}}{n_{m-1}}]$
t_m	n_m	d_m	$\hat{S}_K(t_m) = 0 = \hat{S}_K(t_{m-1})[1 - \frac{d_m}{n_m}]$

Table 3.2. Example for Computing Kaplan Meier Estimator

f_j	$t_{(j)}$	γ_j	t_i	n_i	d_i	$\hat{S}(t)$
						$\hat{S}(0) = 1$
1	36	1	36	13	1	$\hat{S}(36) = 0.9231$
3	38	1	38	12	2	$\hat{S}(38) = 0.7692$
	38	1				
	38	0				
1	78	1	78	9	1	$\hat{S}(78) = 0.6837$
4	112	1	112	8	2	$\hat{S}(112) = 0.5128$
	112	1				
	114	0				
	162	0				
1	189	1	189	4	1	$\hat{S}(189) = 0.3846$
1	198	1	198	3	1	$\hat{S}(198) = 0.2564$
1	237	1	237	2	1	$\hat{S}(36) = 0.1282$
	489	0				

where $j = 1, ..., m-1$.

The Agresti and Coull (1998) plus four 95% CI adds two successes (deaths) and two failures (survives) to the data set from a binomial distribution, and then computes the classical binomial 95% CI from the modified data set. For $t \in [t_1, t_m]$, Olive (2010, problem 16.45) modifies this procedure by adding two artificial deaths just before time t_1 and two artificial censored observations after the largest death time t_m. Then the classical 95% CI for the Kaplan Meier estimator is computed from the modified data set.

Hence

$$\tilde{S}_K(t_i) = \left(1 - \frac{1}{n+4}\right)\left(1 - \frac{1}{n+3}\right)\prod_{k=1}^{i}\left(1 - \frac{d_k}{n_k+2}\right)$$

for $i = 1, ..., m$ where the first two terms are due to the two artificial deaths at the just before t_1 and $n_k + 2$ is used in the product due to the two artificial cases censored at time t_m. Also $[SE(\tilde{S}_K(t_i)]^2 =$

$$[\tilde{S}_K(t_i)]^2\left(\sum_{k=1}^{i}\frac{d_k}{(n_k+2)(n_k+2-d_k)} + \frac{1}{(n+4)(n+4-1)} + \frac{1}{(n+3)(n+3-1)}\right)$$

for $i = 1, ..., m-1$.

If the CI is initially (L,U), then the CI $(\max(0, L), \min(1, U))$ is used. In addition to the classical Kaplan Meier CI, there is a log CI that uses $log(\hat{S})$ and a log–log CI that uses $\log(-\log(\hat{S}))$ that are easy to compute with software.

CHAPTER 4

EXAMPLES AND SIMULATIONS

Simulations were done in *R*. See R Core Team (2015). The function *kmsim2* simulates the classical, log, log–log, and plus four CIs for the Kaplan Meier estimator and is in the collection of *R* functions *regpack* available from (http://lagrange.math.siu.edu/Olive/regpack.txt).

The program *kmsim2* computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the n $t_{(j)}$. This is done for runs=5000 data sets and the program computes the proportion of times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The average scaled CI lengths (the average of $\sqrt{n}$ CI length) are also computed. The ccov is the proportion for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The three classical CIs are not made if the last case is censored so NA is given. The plus four CI seems to be good at $t_{(1)}$ and $t_{(n)}$. With 5000 runs, coverage between 0.94 and 0.96 would not give much evidence that the coverage is different from the nominal covarage of 0.95.

```
> library(survival)
> kmsim2(n=10,runs=5000)
$ccov
 [1] 0.8808 0.9648 0.9740 0.9748 0.9644 0.9536 0.9368 0.9088 0.8400     NA

$lcov
 [1] 0.8730 0.9490 0.9570 0.9652 0.9664 0.9646 0.9750 0.9762 0.9826     NA

$llcov
```

```
 [1] 0.7768 0.8954 0.9144 0.9222 0.9210 0.9216 0.9234 0.9230 0.9214     NA

$p4cov
 [1] 0.9964 0.9114 0.9108 0.9148 0.9184 0.9194 0.9326 0.9414 0.9554 0.9738

$clen
 [1] 0.8170504 1.3276870 1.7097334 1.8942508 1.9756001 1.9786097 1.9024568
 [8] 1.5967784 1.0986384       NaN

$llen
 [1] 0.7657591 1.2264927 1.5981921 1.9133880 2.0764107 2.1498071 2.1682851
 [8] 2.1503575 2.2076806        NA

$lllen
 [1] 1.463784 1.682308 1.776004 1.825388 1.831936 1.790259 1.692386 1.525528
 [9] 1.265297       NA

$p4len
 [1] 1.325905 1.473112 1.569981 1.632562 1.665454 1.668856 1.641050 1.577583
 [9] 1.470264 1.189196
```

The above output is for $n = 10$ with 5000 runs. The tables below summarizes the CI coverages and scaled lengths for t_1, t_3, t_{n-2}, and t_{n-1} for various values on n. The figures and tables are explained further in the conclusions chapter. The sample size n is the last number on the horizontal axis for a figure.

Table 4.1. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
10	t_1	cov	0.8808	0.8730	0.7768	0.9964
		len	0.8171	0.7658	1.4638	1.3259
10	t_3	cov	0.9740	0.9570	0.9144	0.9108
		len	1.7097	1.5982	1.7760	1.5700
10	t_{n-2}	cov	0.9088	0.9762	0.9230	0.9414
		len	1.5968	2.1504	1.5255	1.5776
10	t_{n-1}	cov	0.8400	0.9826	0.9214	0.9554
		len	1.0986	2.2077	1.2653	1.4703

Table 4.2. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
20	t_1	cov	0.8762	0.8734	0.7768	0.9974
		len	0.5896	0.5708	1.2077	1.1241
20	t_3	cov	0.9588	0.9486	0.9180	0.9302
		len	1.2810	1.2254	1.5181	1.4217
20	t_{n-2}	cov	0.8850	0.9740	0.9368	0.9608
		len	1.2486	1.7692	1.3159	1.4844
20	t_{n-1}	cov	0.8246	0.9776	0.9360	0.9708
		len	0.8423	1.7547	1.0856	1.3348

Table 4.3. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
50	t_1	cov	0.8788	0.8770	0.7764	0.9982
		len	0.3783	0.3734	0.8413	0.8062
50	t_3	cov	0.9510	0.9458	0.9206	0.9558
		len	0.8303	0.8157	1.0705	1.0646
50	t_{n-2}	cov	0.8832	0.9716	0.9486	0.9722
		len	0.8771	1.2793	1.0062	1.1901
50	t_{n-1}	cov	0.8220	0.9810	0.9530	0.9804
		len	0.5865	1.2472	0.8375	1.0419

Table 4.4. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
100	t_1	cov	0.8806	0.8802	0.7806	0.9996
		len	0.2688	0.2670	0.6145	0.5964
100	t_3	cov	0.9534	0.9512	0.9258	0.9638
		len	0.5905	0.5853	0.7835	0.7988
100	t_{n-2}	cov	0.8660	0.9720	0.9522	0.9770
		len	0.6676	0.9820	0.7981	0.9499
100	t_{n-1}	cov	0.8158	0.9722	0.9504	0.9818
		len	0.4441	0.9528	0.6706	0.8231

Table 4.5. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
200	t_1	cov	0.8740	0.8736	0.7836	0.9980
		len	0.1897	0.1891	0.4397	0.4313
200	t_3	cov	0.9536	0.9524	0.9246	0.9718
		len	0.4191	0.4173	0.5636	0.5826
200	t_{n-2}	cov	0.8692	0.9674	0.9482	0.9760
		len	0.5049	0.7456	0.6220	0.7361
200	t_{n-1}	cov	0.8090	0.9812	0.9598	0.9828
		len	0.3349	0.7216	0.5277	0.6342

Table 4.6. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
400	t_1	cov	0.8748	0.8744	0.7784	0.9990
		len	0.1342	0.1340	0.3133	0.3085
400	t_3	cov	0.9474	0.9466	0.9202	0.9680
		len	0.2973	0.2967	0.4023	0.4187
400	t_{n-2}	cov	0.8668	0.9712	0.9572	0.9772
		len	0.3789	0.5623	0.4785	0.5611
400	t_{n-1}	cov	0.8076	0.9766	0.9600	0.9836
		len	0.2518	0.5426	0.4096	0.4822

Table 4.7. Simulated CI Coverages and Scaled Lengths

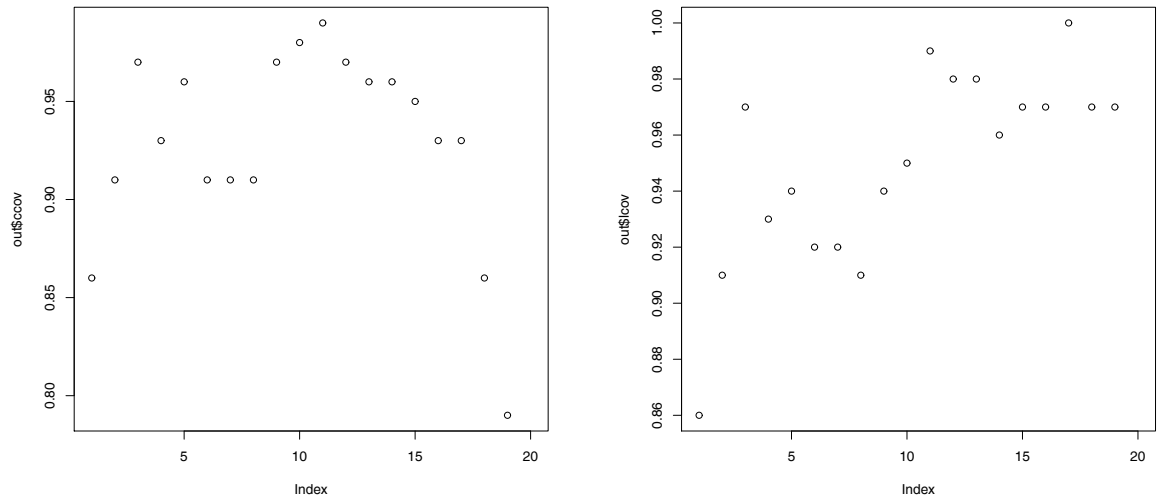
n	t_i	cov/len	clas	log	loglog	plus4
600	t_1	cov	0.8872	0.8872	0.7782	0.9988
		len	0.1104	0.1103	0.2583	0.2534
600	t_3	cov	0.9498	0.9496	0.9238	0.9748
		len	0.2428	0.2425	0.3294	0.3436
600	t_{n-2}	cov	0.8622	0.9720	0.9586	0.9798
		len	0.3227	0.4783	0.4114	0.4787
600	t_{n-1}	cov	0.8160	0.9840	0.9622	0.9820
		len	0.2149	0.4622	0.3547	0.4116

Table 4.8. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
800	t_1	cov	0.8816	0.8814	0.7722	0.9988
		len	0.0959	0.0958	0.2247	0.2200
800	t_3	cov	0.9436	0.9422	0.9152	0.9712
		len	0.2097	0.2095	0.2851	0.2979
800	t_{n-2}	cov	0.8708	0.9670	0.9582	0.9774
		len	0.2865	0.4248	0.3677	0.4263
800	t_{n-1}	cov	NA	NA	NA	0.9836
		len	NaN	NA	NA	0.3664

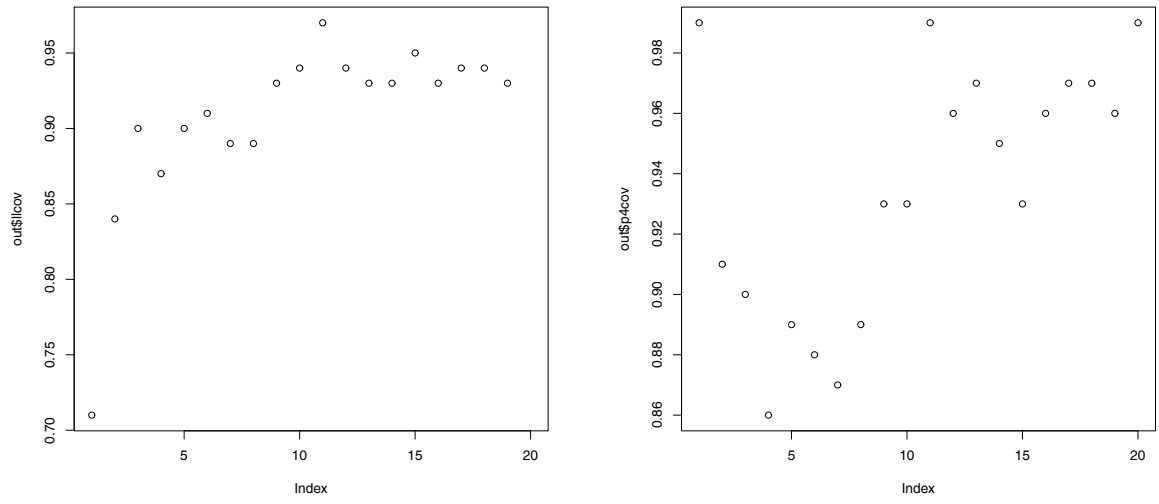
Table 4.9. Simulated CI Coverages and Scaled Lengths

n	t_i	cov/len	clas	log	loglog	plus4
1000	t_1	cov	0.8732	0.8732	0.7726	0.9986
		len	0.0849	0.0848	0.1989	0.1965
1000	t_3	cov	0.9460	0.9460	0.9220	0.9734
		len	0.1873	0.1871	0.2551	0.2667
1000	t_{n-2}	cov	0.8682	0.9700	0.9576	0.9744
		len	0.2616	0.3882	0.3375	0.3898
1000	t_{n-1}	cov	0.8056	0.9808	0.9654	0.9852
		len	0.1740	0.3748	0.2926	0.3347



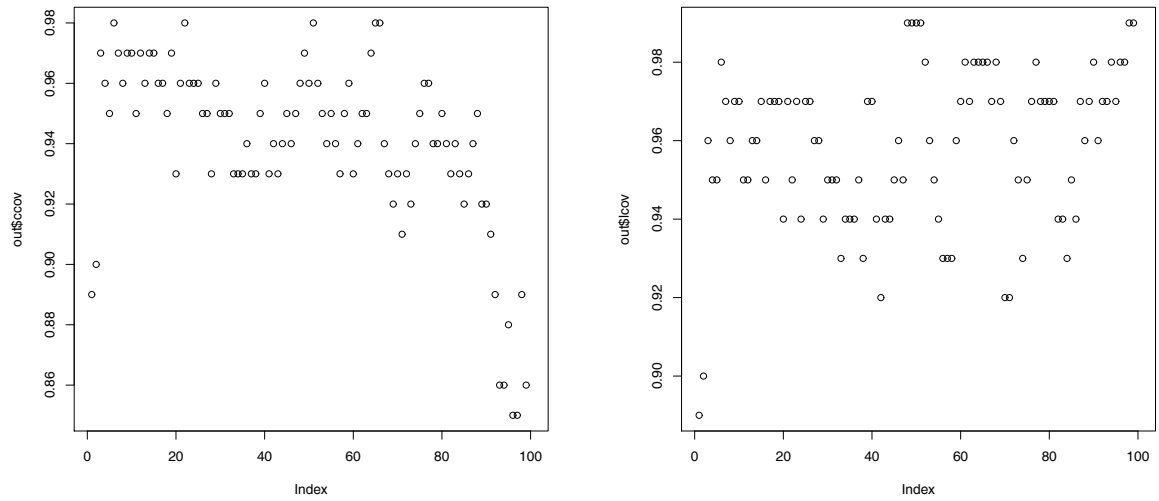
(a) (b)

Figure 4.1. CI coverages: (a) classic (b) log



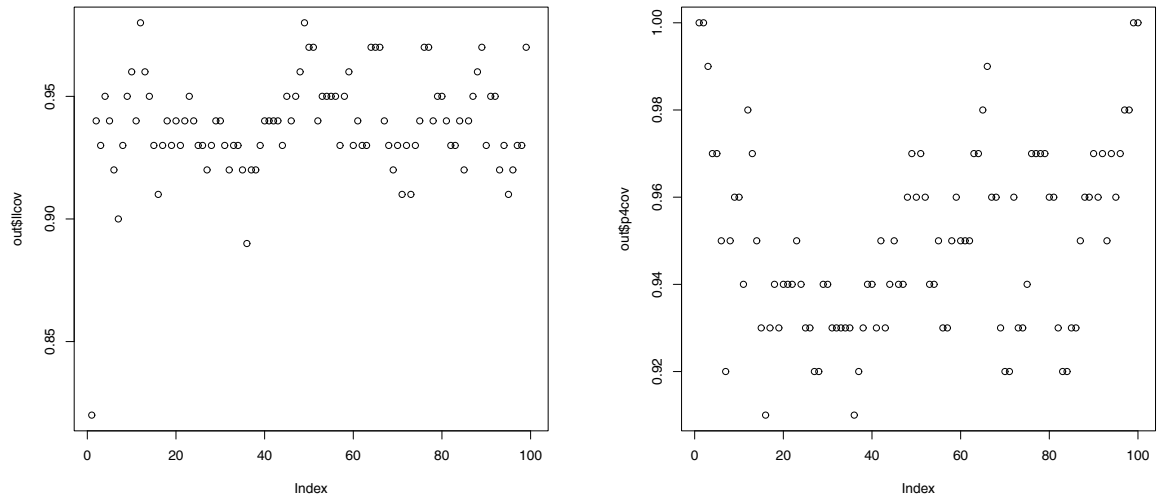
(a) (b)

Figure 4.2. CI coverages: (a) log-log (b) plus4



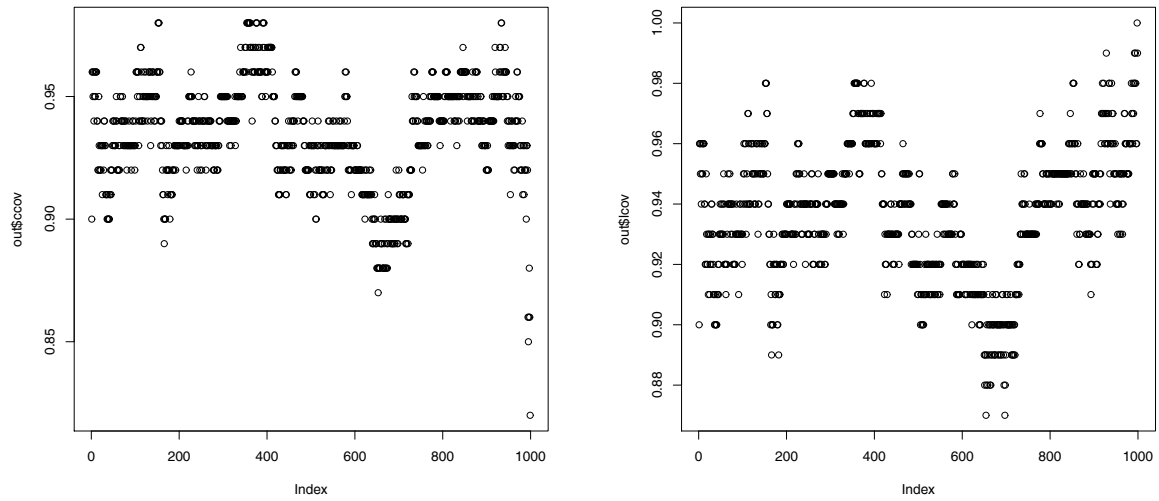
(a) (b)

Figure 4.3. CI coverages: (a) classic (b) log



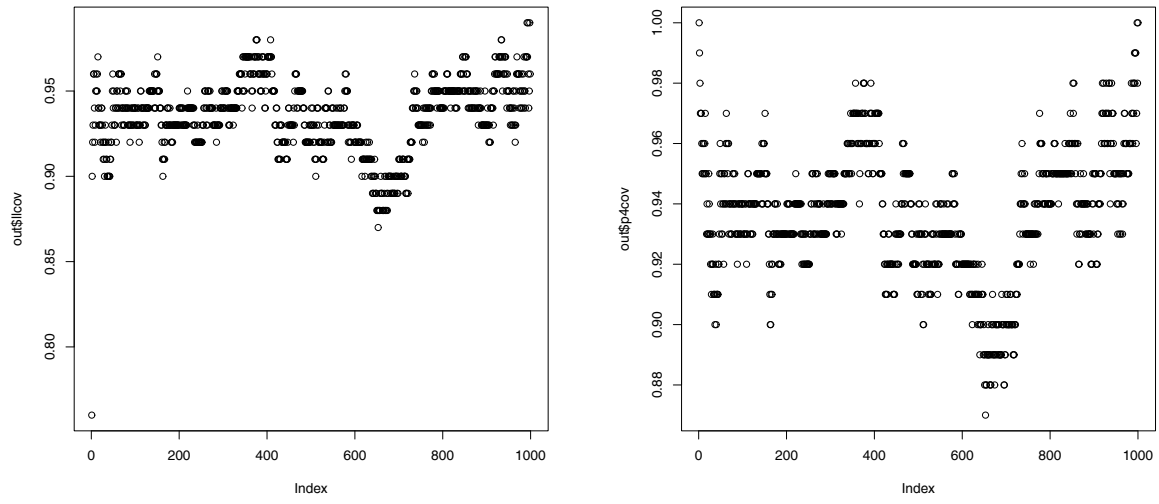
(a) (b)

Figure 4.4. CI coverages: (a) log-log (b) plus4



(a) (b)

Figure 4.5. CI coverages: (a) classic (b) log



(a) (b)

Figure 4.6. CI coverages: (a) log-log (b) plus4

CHAPTER 5

CONCLUSIONS

Table 5.1. Best Method

n	10	20	50	100	200
t_1	p4	p4	p4	p4	p4
t_3	log	log	clas	log	log
t_{n-2}	p4	p4	llog	llog	llog
t_{n-1}	p4	llog	llog	llog	llog
n	400	600	800	1000	conclusion
t_1	p4	p4	p4	p4	p4
t_3	log	clas	clas	clas/log	log
t_{n-2}	llog	llog	llog	llog	llog
t_{n-1}	llog	llog	p4	llog	llog

Table 5.2. Ranges on Figures

n	clas	log	llog	p4	conclusion
20	0.80 - 1.00	0.86 - 1.00	0.70 - 1.00	0.86 - 1.00	log,p4
100	0.84 - 1.00	0.89 - 1.00	0.80 - 1.00	0.90 - 1.00	p4
1000	0.80 - 1.00	0.86 - 1.00	0.75 - 1.00	0.86 - 1.00	log,p4

From the tables, the best CIs are plus4 for t_1, log for t_3, and loglog for t_{n-2} and t_{n-1}. From the figures, the best CIs are log and plus4 if n=20, plus4 if n=100, and log and plus4 if n=1,000. Examine the ranges of the vertical axis of the figures. These ranges are summarized in table 4.2.

REFERENCES

[1] Agresti, A., and Coull, B.A., "Approximate is Better than Exact for Interval Estimation of Binomial Parameters," *The American Statistician*, 52 (1998), 119–126.

[2] Olive, D.J., *"Multiple Linear and 1D Regression Models"* online course notes, see (http://lagrange.math.siu.edu/Olive/regbk.htm) (2010).

[3] Olive, D.J., *"Statistical Theory and Inference"* Springer, New York, NY (2014).

[4] R Core Team, "R: a Language and Environment for Statistical Computing," "R Foundation for Statistical Computing," Vienna, Austria,(www.R-project.org) (2015).

[5] Smith, P.J., *"Analysis of Failure and Survival Data,"* "Chapman and Hall/CRC," Boca Raton, FL, (2002).

VITA

Graduate School
Southern Illinois University

Luke Yang

hzyang2@gmail.com

Maryville University in Saint Louis
Bachelor of Science, Mathematics, May 2013

Major Professor: Dr. David J. Olive