

Lecture 12

Confidence intervals and left truncation

We need to find confidence intervals (pointwise) for the estimators of $S(t)$ at each time point. We differentiate the log-likelihood and use likelihood theory,

$$l = \sum_i d_i \log h_i + \sum_i (n_i - d_i) \log(1 - h_i),$$

differentiated twice to find the Hessian matrix $\left\{ \frac{\partial^2 l}{\partial h_i \partial h_j} \right\}$.

Note that since l is a sum of functions of each individual hazard the Hessian must be diagonal.

The estimators $\{\widehat{h}_1, \widehat{h}_2, \dots, \widehat{h}_n\}$ are asymptotically unbiased and are asymptotically jointly normally distributed with approximate variance I^{-1} , where the information matrix is given by

$$I = \mathbf{E} \left(- \left\{ \frac{\partial^2 l}{\partial h_i \partial h_j} \right\} \right).$$

Since the Hessian is diagonal, the covariances are all asymptotically zero, and coupled with asymptotic normality, this ensures that all pairs $\widehat{h}_i, \widehat{h}_j$ are asymptotically independent.

$$-\frac{\partial^2 l}{\partial h_i^2} = \frac{d_i}{h_i^2} + \frac{n_i - d_i}{(1 - h_i)^2}$$

We use the observed information J and so replace h_i in the above by its estimator $\widehat{h}_i = \frac{d_i}{n_i}$. Hence we have

$$\mathbf{var} \widehat{h}_i \approx \frac{d_i(n_i - d_i)}{n_i^3}.$$

12.1 Greenwood's formula

12.1.1 Reminder of the δ method

If the random variation of Y around μ is small (for example if μ is the mean of Y and $\mathbf{var} Y$ has order $\frac{1}{n}$), we use:

$$g(Y) \approx g(\mu) + (Y - \mu)g'(\mu) + \frac{1}{2}(Y - \mu)^2 g''(\mu) + \dots$$

Taking expectations

$$\begin{aligned}\mathbf{E}(g(Y)) &= g(\mu) + O\left(\frac{1}{n}\right) \\ \mathbf{var}(g(\mathbf{Y})) &= \mathbf{g}'(\mu)^2 \mathbf{var}Y + o\left(\frac{1}{n}\right)\end{aligned}$$

12.1.2 Derivation of Greenwood's formula for $\mathbf{var}(\widehat{S}(t))$

$$\log \widehat{S}(t) = \sum_{t_i \leq t} \log(1 - \widehat{h}_i)$$

But

$$\mathbf{var} \widehat{h}_i \approx \frac{d_i(n_i - d_i)}{n_i^3} \quad \text{and} \quad \widehat{h}_i \xrightarrow{P} h_i$$

so that, given $g(h_i) = \log(1 - h_i)$,

$$g'(h_i) = \frac{-1}{(1 - h_i)}$$

we have

$$\begin{aligned}\mathbf{var} \log(1 - \widehat{h}_i) &\approx \frac{1}{(1 - h_i)^2} \mathbf{var} \widehat{h}_i \\ &\approx \frac{1}{\left(1 - \frac{d_i}{n_i}\right)^2} \frac{d_i(n_i - d_i)}{n_i^3} \\ &= \frac{d_i}{n_i(n_i - d_i)}\end{aligned}$$

Since $\widehat{h}_i, \widehat{h}_j$ are asymptotically independent we can put all this together to get

$$\mathbf{var} \log(\widehat{S}(t)) = \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \quad (1)$$

Let $Y = \log \widehat{S}$ and note that we need $\mathbf{var}(e^Y) \approx (e^Y)^2 \mathbf{var}Y$, again using the delta-method. Finally we have *Greenwood's formula*

$$\mathbf{var}(\widehat{S}(t)) \approx \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}. \quad (2)$$

Applying this to the same sort of argument to the Nelson-Aalen estimator and its extension to the survival function we also see

$$\mathbf{var} \widehat{H}(t) \approx \sum_{t_i \leq t} \frac{d_i(n_i - d_i)}{n_i^3}$$

and

$$\begin{aligned}\mathbf{var}\tilde{S}(t) &= \mathbf{var}\left(\exp(-\hat{H}(t))\right) \\ &\approx (e^{-H})^2 \sum_{t_i \leq t} \frac{d_i(n_i - d_i)}{n_i^3} \\ &\approx (\tilde{S}(t))^2 \sum_{t_i \leq t} \frac{d_i(n_i - d_i)}{n_i^3}\end{aligned}$$

Clearly these estimates are only reasonable if each n_i is sufficiently large, since they rely heavily on asymptotic calculations.

12.2 Left truncation

Left truncation is easily dealt with in the context of nonparametric survival estimation. Suppose the invented data set comes from the following hidden process: There is an event time, and an independent censoring time, and, in addition, a truncation time, which is the time when that individual becomes available to be studied. For example, suppose this were a nursing home population, and the time being studied is the number of years after age 80 when the patient first shows signs of dementia. The censoring time might be the time when the person dies or moves away, or when the study ends. The study population consists of those who have entered the nursing home free of dementia. The truncation time would be the age at which the individual moves into the nursing home.

Table 12.1: Invented data illustrating left truncation. Event times after the censoring time may be purely nominal, since they may not have occurred at all; these are marked with *. The row *Observation* shows what has actually been observed. When the event time comes before the truncation time the individual is not included in the study; this is marked by a \circ .

Patient ID	5	2	9	0	1	3	7	6	4	8
Event time	2	5	5	*	7	*	12	*	*	*
Censoring time	10	8	7	8	11	7	14	14	14	14
Truncation time	-2	3	6	0	1	0	6	6	-5	1
Observation	2	5	\circ	8+	7	7+	12	14+	14+	14+

Table 12.2: Computations of survival estimates for invented data set of Table 12.1.

t_i	d_i	n_i	\hat{h}_i	$\hat{S}(t_i)$	$\tilde{S}(t_i)$
2	1	6	0.17	0.83	0.85
5	1	6	0.17	0.69	0.72
7	1	7	0.14	0.58	0.62
12	1	4	0.25	0.45	0.48

We give a version of these data in Table 12.1. Note that patient number 9 was truncated at time 6 (i.e., entered the nursing home at age 86) but her event was at time 5 (i.e., she had already suffered from dementia since age 85), hence was not included in the study. In table 12.2 we give the computations for the Kaplan-Meier estimate of the survival function. The computations are exactly the same as those of section 11.4.4, except for one important change: The number at risk n_i is not simply the number $n - \sum_{t_i < t} d_i - \sum_{t_i < t} k_i$ of individuals who have not yet had their event or censoring time. Rather, an individual is at risk at time t if her event time and censoring time are both $\geq t$, and if the truncation time is $\leq t$. (As usual, we assume that individuals who have their event or are censored in a given year, were at risk during that year. We are similarly assuming that those who entered the study at age x are at risk during that year.) At the start of our invented study there are only 6 individuals at risk, so the estimated hazard for the event at age 2 becomes $1/6$.

In the most common cases of truncation we need do nothing at all, other than be careful in interpreting the results. For instance, suppose we were simply studying the age after 80 at which individuals develop dementia by a longitudinal design, where 100 healthy individuals 80 years old are recruited and followed for a period of time. Those who are already impaired at age 80 are truncated. All this means is that we have to understand (as we surely would) that the results are conditional on the individual not suffering from dementia until age 80.

We can compute variances for the Kaplan-Meier and Nelson-Aalen estimators using Greenwood's formula exactly as before, only taking care to use the reinterpreted number at risk. The one problem that arises is that individuals may enter into the study slowly, yielding a small number at risk, and hence very wide error bounds, which of course will carry through to the end.

12.3 Example: The AML study

In the 1970s it was known that individuals who had gone into remission after chemotherapy for acute lymphatic leukemia would benefit — by longer remission times — from a course of continuing “maintenance” chemotherapy. A study [EEH⁺77] pointed out that “Despite a lack of conclusive evidence, it has been assumed that maintenance chemotherapy is useful in the management of acute myelogenous leukemia (AML).” The study set out to test this assumption, comparing the duration of remission between an experimental group that received the additional chemotherapy, and a control group that did not. (This analysis is based on the discussion in [MGM01].)

The data are from a preliminary analysis of the data, before completion of the study. The duration of complete remission in weeks was given for each patient (11 maintained, 12 non-maintained controls); those who were still in remission at the time of the analysis are censored observations. The data are given in Table 12.3. They are included in the `survival` package of R, under the name `aml`.

The first thing we do is to estimate the survival curves. The summary data and computations are given in Table 12.4. The Kaplan-Meier survival curves are shown in Figure 12.1. In Table 12.5 we show the computations for confidence intervals just for the Kaplan-Meier curve of the maintenance group. The confidence intervals are based on the logarithm of survival, using (1) directly. That is, the bounds on the confidence interval are

$$\exp \left\{ \log \hat{S}(t) \pm z \sqrt{\sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)}} \right\},$$

Table 12.3: Times of complete remission for preliminary analysis of AML data, in weeks. Censored observations denoted by +.

maintained	9 13 13+ 18 23 28+ 31 34 45+ 48 161+
non-maintained	5 5 8 8 12 16+ 23 27 30 33 43 45

where z is the appropriate quantile of the normal distribution. Note that the approximation cannot be assumed to be very good in this case, since the number of individuals at risk is too small for the asymptotics to be reliable. We show the confidence intervals in Figure 12.2.

Table 12.4: Computations for the Kaplan-Meier and Nelson-Aalen survival curve estimates of the AML data.

t_i	Maintenance						Non-Maintenance (control)					
	n_i	d_i	\hat{h}_i	$\hat{S}(t_i)$	\hat{H}_i	$\tilde{S}(t_i)$	n_i	d_i	\hat{h}_i	$\hat{S}(t_i)$	\hat{H}_i	$\tilde{S}(t_i)$
5	11	0	0.00	1.00	0.00	1.00	12	2	0.17	0.83	0.17	0.85
8	11	0	0.00	1.00	0.00	1.00	10	2	0.20	0.67	0.37	0.69
9	11	1	0.09	0.91	0.09	0.91	8	0	0.00	0.67	0.37	0.69
12	10	0	0.00	0.91	0.09	0.91	8	1	0.12	0.58	0.49	0.61
13	10	1	0.10	0.82	0.19	0.83	7	0	0.00	0.58	0.49	0.61
18	8	1	0.12	0.72	0.32	0.73	6	0	0.00	0.58	0.49	0.61
23	7	1	0.14	0.61	0.46	0.63	6	1	0.17	0.49	0.66	0.52
27	6	0	0.00	0.61	0.46	0.63	5	1	0.20	0.39	0.86	0.42
30	5	0	0.00	0.61	0.46	0.63	4	1	0.25	0.29	1.11	0.33
31	5	1	0.20	0.49	0.66	0.52	3	0	0.00	0.29	1.11	0.33
33	4	0	0.00	0.49	0.66	0.52	3	1	0.33	0.19	1.44	0.24
34	4	1	0.25	0.37	0.91	0.40	2	0	0.00	0.19	1.44	0.24
43	3	0	0.00	0.37	0.91	0.40	2	1	0.50	0.10	1.94	0.14
45	3	0	0.00	0.37	0.91	0.40	1	1	1.00	0.00	2.94	0.05
48	2	1	0.50	0.18	1.41	0.24	0	0				

Important: The estimate of the variance is more generally reliable than the assumption of, particularly for small numbers of events. Thus, the first line in Table 12.5 indicates that the estimate of $\log \hat{S}(9)$ is associated with a variance of 0.009. The error in this estimate is on the order of n^{-3} , so it's potentially about 10%. On the other hand, the number of events observed has binomial distribution, with parameters around $(11, 0.909)$, so it's very far from a normal distribution. We could improve our confidence interval by using the Poisson confidence intervals worked out in Problem Sheet 3, question 2, or binomial confidence interval. We will not go into the details in this course.

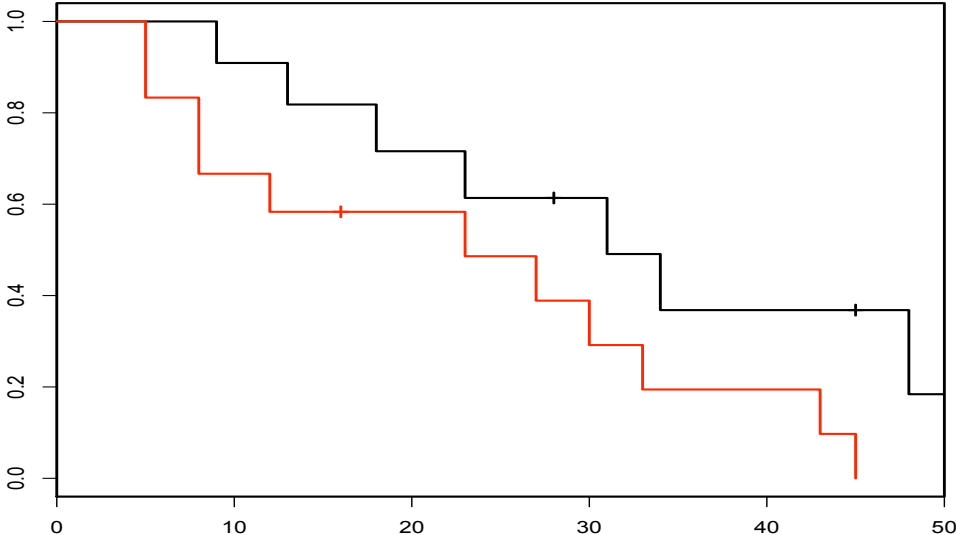


Figure 12.1: Kaplan-Meier estimates of survival in maintenance (black) and non-maintenance groups in the AML study.

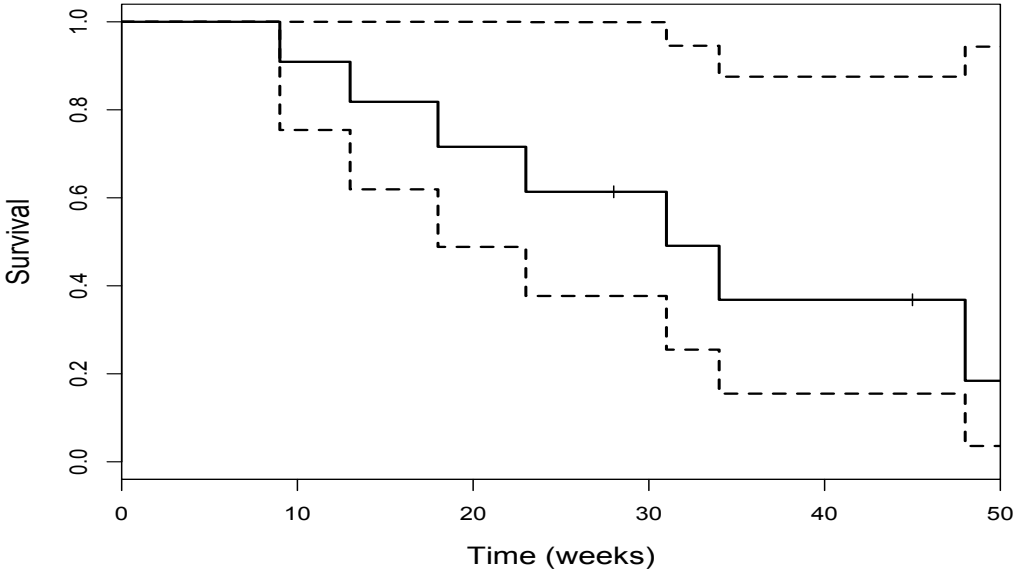


Figure 12.2: Greenwood's estimate of 95% confidence intervals for survival in maintenance group of the AML study.

Table 12.5: Computations for Greenwood's estimate of the standard error of the Kaplan-Meier survival curve from the maintenance population in the AML data. "lower" and "upper" are bounds for 95% confidence intervals, based on the log-normal distribution.

t_i	n_i	d_i	$\frac{d_i}{n_i(n_i-d_i)}$	$\text{Var}(\log \hat{S}(t_i))$	lower	upper
9	11	1	0.009	0.009	0.754	1.000
13	10	1	0.011	0.020	0.619	1.000
18	8	1	0.018	0.038	0.488	1.000
23	7	1	0.024	0.062	0.377	0.999
31	5	1	0.050	0.112	0.255	0.946
34	4	1	0.083	0.195	0.155	0.875
48	2	1	0.500	0.695	0.036	0.944

12.4 Actuarial estimator

The **actuarial estimator** is a further estimator for $S(t)$. It is given as

$$S^*(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i - \frac{1}{2}c_i} \right)$$

The intervals between consecutive failure times are usually of constant length, and it is generally used by actuaries and demographers following a cohort from birth to death. Age will normally be the time variable and hence the unit of time is 1 year.