

B.5 Censoring and truncation, Kaplan-Meier estimator

1. (a) See lecture notes.
 - (b) There is right censoring: The depression may not have recurred at the time that the study ended, or the patient died or dropped out. There is left truncation: The first episode of depression made the patients eligible for the study, but not immediately. Thus, the event of interest — the recurrence of depression — could already have happened before the patient was enrolled in the study.
 - (c) This study design involves right truncation: The entire study population has already experienced the event of interest (AIDS diagnosis). Any individual whose incubation period extended beyond the truncation time would not have appeared in the study.
2. Suppose $t_1 < t_2 < \dots < t_k$ are times at which events occur. The Kaplan-Meier estimator for the survival function is

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right),$$

where n_i is the number at risk at time t_i and d_i is the number of events at t_i . If there is no censoring, then $n_{j+1} = n_j - d_j = n_0 - \sum_{i=1}^j d_i$. Thus

$$\begin{aligned} \hat{S}(t) &= \prod_{t_i \leq t} \left(\frac{n_i - d_i}{n_i}\right) \\ &= \prod_{t_i \leq t} \left(\frac{n_{i+1}}{n_i}\right) \\ &= \frac{n_{j+1}}{n_0} \text{ where } j = \max\{i : t_i \leq t\} \\ &= 1 - n_0^{-1} \sum_{i: t_i \leq t} d_i \\ &= 1 - \hat{F}(t). \end{aligned}$$

Sketches are given in Figure B.1. The Kaplan-Meier estimate for the survival to age 4 is 0.667. Greenwood's formula estimates the variance of this estimate as

$$\hat{S}(4)^2 \sum_{t_i \leq 4} \frac{d_i}{n_i n_{i+1}} = 0.44 \left(\frac{2}{21 \cdot 19} + \frac{2}{19 \cdot 17} + \frac{1}{17 \cdot 16} + \frac{2}{16 \cdot 14} \right) = 0.0106.$$

The standard error is then about $\sqrt{0.0106} = 0.103$. We then approximate a 95% confidence interval as $0.667 \pm 1.96 \cdot 0.103 = (.465, .869)$. Alternatively, we can find the 95% confidence interval for $\log \hat{S}(4)$ to be $-0.405 \pm 1.96 \cdot \sqrt{.0238}$, which transforms into the CI $(.493, .903)$ for $\hat{S}(4)$.

Why are these not the same? It depends on whether we think $\hat{S}(4)$ is normal or $\log \hat{S}(t)$ is normal. In the limit as the number of samples goes to ∞ both are, but with only 21 samples the two approaches disagree.

3. The log likelihood is

$$\ell(p) = \log \binom{n}{x} + x \log p + (n - x) \log(1 - p).$$

This has solution $0 = \ell'(\hat{p}) = x/\hat{p} - (n - x)/(1 - \hat{p})$, implying $\hat{p} = x/n$. We know that the variance of a binomial random variable is $np(1 - p)$. Substituting \hat{p} for p yields the estimate

$$\text{Var}(\hat{p}) = \text{Var}(x/n) = n^{-2} \text{Var}(x) = n^{-1} p(1 - p) = n^{-1} \frac{x}{n} \frac{n - x}{n} = \frac{x(n - x)}{n^3}.$$

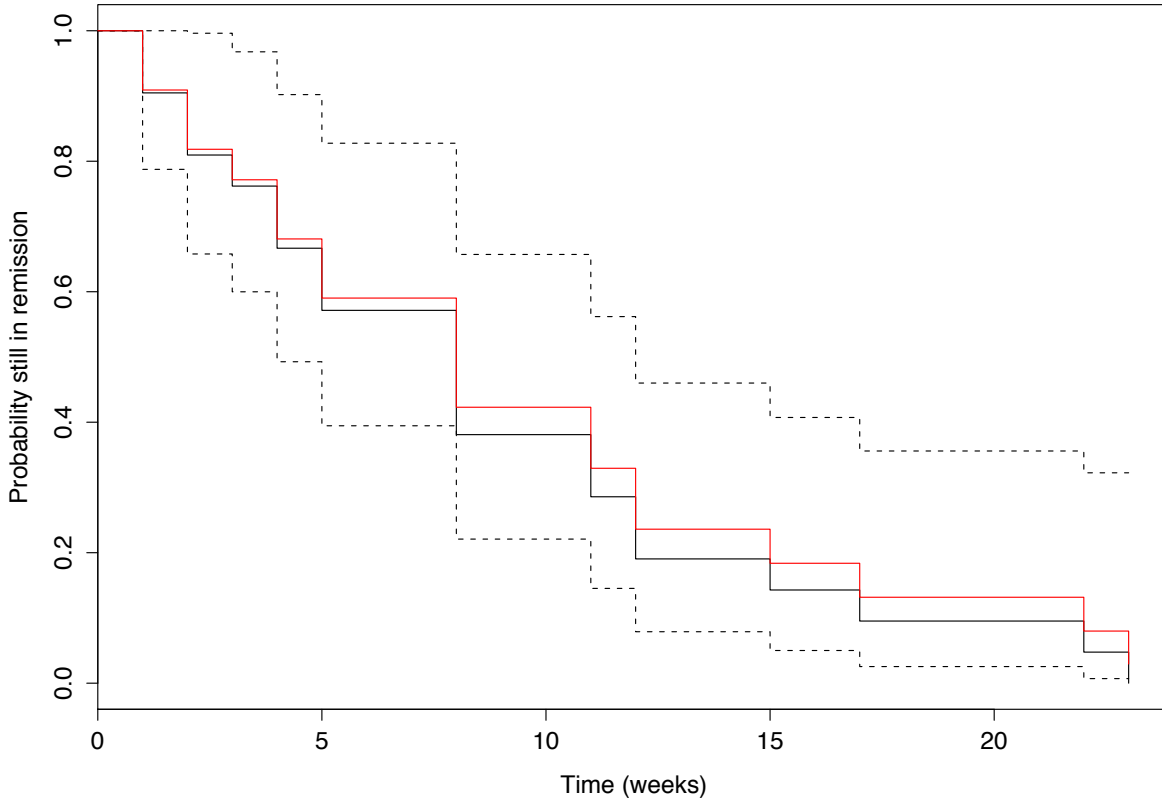


Figure B.1: Estimated survival (remission) curves for leukaemia patients. Black is Kaplan-Meier, red is Nelson-Aalen. Dashed lines show approximate 95% confidence intervals for Kaplan-Meier estimate, using Greenwood's formula.

If all the censoring occurs at $t = 0$ then the number of individuals at risk of dying in $(0, t)$ is actually $n(t) + d(t)$. Thus alive at time t is binomial with parameters $n = n(0) = n(t) + d(t)$ and $p = S(t)$. The MLE for p is thus

$$\hat{S}(t) = \hat{p} = \frac{n(t)}{n(t) + d(t)} = \frac{n(t)}{n(0)}.$$

(If the censoring all happens at time 0, then the number at risk at time $0+$ will be the same as the sum of the number who die up to time t , and the number still at risk at time t .) The variance estimate is

$$\frac{d(t)n(t)}{n(0)^3} = n(t)^{-1} \frac{d(t)}{n(0)} \frac{n(t)}{n(0)} \frac{n(t)}{n(0)} = n(t)^{-1} (1 - \hat{S}(t)) \hat{S}(t)^2.$$

Greenwood’s estimate in the case of no censoring is

$$\begin{aligned}
 \text{Var } \hat{S}(t) &\approx \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \\
 &= \hat{S}(t)^2 \sum_{t_i \leq t} \frac{n_{i+1} - n_i}{n_i(n_{i+1})} \\
 &= \hat{S}(t)^2 \sum_{t_i \leq t} \left(\frac{1}{n_{i+1}} - \frac{1}{n_i} \right) \\
 &= \hat{S}(t)^2 \left(\frac{1}{n_j} - \frac{1}{n_0} \right) \\
 &= \hat{S}(t)^2 \frac{d(t)}{n(t)n(0)} \\
 &= n(t)^{-1} \hat{S}(t)^2 (1 - \hat{S}(t))
 \end{aligned}$$

as before.

4. (a) Right censoring and left truncation.
- (b) If individuals who enter at age x are considered immediately available to count at risk at age x , and those who die at age x are also at risk.

Age	65	66	67	68	69	70	71	72	73	74	75
# at risk	3	9	11	13	14	17	14	12	12	8	4

We are planning to use the actuarial estimator — so we count those who are censored or died as having had half a year at risk, and count those who entered at a given age as having half a year at risk in that year, we get the following counts:

Age	65	66	67	68	69	70	71	72	73	74	75
# at risk	1.5	6.0	9.5	9.5	11.5	13.0	11.5	10.5	9.0	6.0	3.5

- (c) Again, counting whole years at risk for those who enter, die, or are right-censored, we have

Age	n_i	d_i	h_i	$\hat{S}(t_i)$	$\tilde{S}(t_i)$
70	17	4	0.235	0.765	0.790
72	12	1	0.083	0.701	0.727
73	12	3	0.250	0.526	0.566
74	8	4	0.500	0.263	0.343
75	4	1	0.250	0.197	0.268

The actuarial estimate gives us

Age	n_i	d_i	h_i	$\hat{S}(t_i)$	$\tilde{S}(t_i)$
70	13.0	4	0.308	0.692	0.735
72	10.5	1	0.095	0.626	0.668
73	9.0	3	0.333	0.418	0.479
74	6.0	4	0.667	0.139	0.246
75	3.5	1	0.286	0.099	0.185

- (d) We use the whole-year method, rather than the actuarial estimate. Our central estimate for the probability of surviving from age 70 to age 75 is $\hat{S}(74) = 0.343$. Using Greenwood’s

estimate, we estimate the variance of $\log \hat{S}(74)$ to be

$$\begin{aligned}\sum_{t_i \leq 74} \frac{d_i}{n_i(n_i - d_i)} &= \frac{4}{17 \cdot 13} + \frac{1}{12 \cdot 11} + \frac{3}{12 \cdot 9} + \frac{4}{8 \cdot 4} \\ &= 0.178,\end{aligned}$$

so the standard error is $\sqrt{0.178} = 0.422$. Thus an approximate 95% confidence interval for $S(74)$ is

$$(0.343e^{-0.422 \cdot 1.96}, 0.343e^{0.422 \cdot 1.96}) = (0.150, 0.784).$$