- The confidence intervals are asymptotically exact for $0 < S(t) < 1$.
- The confidence intervals do not degenerate to confidence intervals of width zero for $n(t) = 0$ or $n(t) = n$ as was the case with the Wald confidence interval.

This concludes the discussion concerning finding point and interval estimators for $S(t)$ from a complete data set of lifetimes. We now introduce techniques for estimating $S(t)$ from a right-censored data set.

Survivor Function Estimation for Randomly Right-Censored Data Sets

The general case in which there are both ties and right-censored data values is now considered. Some new notation must be established in order to derive the nonparametric estimator for $S(t)$. As before, assume that n items are on test. Let $y_1 < y_2 < \cdots < y_k$ denote the k distinct observed failure times, and let d_j denote the number of observed failures at time y_j, for $j = 1, 2, \ldots, k$. Let $n_j = n(y_j)$ denote the number of items on test just prior to time y_j, for $j = 1, 2, \ldots, k$, and it is customary to include any values that are right censored at y_j in this count.

The search for a survivor function estimator begins by assuming that the data arose from a discrete distribution with mass values $y_1 < y_2 < \cdots < y_k$. For a discrete distribution, $h(y_j)$ is a conditional probability with interpretation $h(y_j) = P[T = y_j \,|\, T \geq y_j]$ for $j = 1, 2, \ldots, k$. As shown in Appendix F, the survivor function can be written in terms of the hazard function at the mass values as

$$S(t) = \prod_{j \,|\, y_j \leq t} \big[1 - h(y_j)\big] \qquad\qquad t \geq 0.$$

Thus, a reasonable estimator for $S(t)$ is $\prod_{j \,|\, y_j < t} \big[1 - \hat{h}(y_j)\big]$, which reduces the problem of estimating the survivor function to that of estimating the hazard function at each mass value. An appropriate element in the likelihood function at mass value y_j for a randomly right-censored data set is

$$h(y_j)^{d_j} \big[1 - h(y_j)\big]^{n_j - d_j}$$

for $j = 1, 2, \ldots, k$. The above expression is correct because d_j is the number of failures at y_j, $h(y_j)$ is the conditional probability of failure at y_j, $n_j - d_j$ is the number of items on test not failing at y_j, and $1 - h(y_j)$ is the probability of failing after time y_j conditioned on survival to time y_j. Thus, the likelihood function for $h(y_1), h(y_2), \ldots, h(y_k)$ is

$$L\big(h(y_1), h(y_2), \ldots, h(y_k)\big) = \prod_{j=1}^{k} h(y_j)^{d_j} \, \big[1 - h(y_j)\big]^{n_j - d_j}$$

and the log likelihood function is

$$\ln L\big(h(y_1), h(y_2), \ldots, h(y_k)\big) = \sum_{j=1}^{k} \Big\{ d_j \ln h(y_j) + (n_j - d_j) \ln \big[1 - h(y_j)\big] \Big\}.$$

The ith element of the score vector is

$$\frac{\partial \ln L\big(h(y_1), h(y_2), \ldots, h(y_k)\big)}{\partial h(y_i)} = \frac{d_i}{h(y_i)} - \frac{n_i - d_i}{1 - h(y_i)}$$

for $i = 1, 2, \ldots, k$. Equating this element of the score vector to zero and solving for $h(y_i)$ yields the maximum likelihood estimate

$$\hat{h}(y_i) = \frac{d_i}{n_i}$$

for $i = 1, 2, \ldots, k$. This estimate for $\hat{h}(y_i)$ is sensible because d_i of the n_i items on test at time y_i fail, so the ratio of d_i to n_i is an appropriate estimate of the conditional probability of failure at time y_i. This derivation may strike a familiar chord because at each time y_i, estimating $h(y_i)$ with d_i divided by n_i is equivalent to estimating the probability of success (that is, failing at time y_i) for each of the n_i items on test. Thus, this derivation is equivalent to finding the maximum likelihood estimators for the probability of success for k binomial random variables.

Using this particular estimate for the hazard function at y_i, the survivor function estimate becomes

$$\hat{S}(t) = \prod_{j|y_j \le t} \left[1 - \hat{h}(y_j)\right] = \prod_{j|y_j \le t} \left[1 - \frac{d_j}{n_j}\right],$$

for $t \ge 0$, commonly known as the Kaplan–Meier or product–limit estimate. When the largest data value recorded corresponds to a failure, the Kaplan–Meier product–limit estimator drops to zero; when the largest data value recorded corresponds to a right-censored observation, a common convention is to cut off the Kaplan–Meier product–limit estimator at the current positive value of $\hat{S}(t)$. The original journal article by American mathematician Edward Kaplan and American statistician Paul Meier in 1958 that established the Kaplan–Meier product–limit estimator is one of the most cited papers in the statistics literature. The following example illustrates the process of calculating the Kaplan–Meier product–limit estimate.

Example 10.3 Use the Kaplan–Meier product–limit estimator to calculate a point estimate of the probability that a remission time in the treatment group in the 6–MP clinical trial described in Example 8.3 exceeds 14 weeks. In other words, estimate $S(14)$ using the Kaplan–Meier product–limit estimator.

The data set contains $n = 21$ patients on test, $r = 9$ observed failures (leukemia relapses), and $k = 7$ distinct observed failure times. The data values, in weeks, are

$$\begin{array}{cccccccccccc} 6 & 6 & 6 & 6^* & 7 & 9^* & 10 & 10^* & 11^* & 13 & 16 \\ 17^* & 19^* & 20^* & 22 & 23 & 25^* & 32^* & 32^* & 34^* & 35^*. \end{array}$$

Table 10.2 gives the values of y_j, d_j, n_j, and $1 - d_j/n_j$ for $j = 1, 2, \ldots, 7$. Assuming a random right-censoring scheme, the Kaplan–Meier product–limit survivor function

j	y_j	d_j	n_j	$1 - \frac{d_j}{n_j}$
1	6	3	21	$1 - \frac{3}{21}$
2	7	1	17	$1 - \frac{1}{17}$
3	10	1	15	$1 - \frac{1}{15}$
4	13	1	12	$1 - \frac{1}{12}$
5	16	1	11	$1 - \frac{1}{11}$
6	22	1	7	$1 - \frac{1}{7}$
7	23	1	6	$1 - \frac{1}{6}$

Table 10.2: Product–limit calculations for 6–MP treatment case.

estimate at $t = 14$ weeks is

$$\begin{aligned}\hat{S}(14) &= \prod_{j|y_j \le 14} \left[1 - \frac{d_j}{n_j}\right] \\ &= \left[1 - \frac{3}{21}\right]\left[1 - \frac{1}{17}\right]\left[1 - \frac{1}{15}\right]\left[1 - \frac{1}{12}\right] \\ &= \frac{176}{255} \\ &= 0.69.\end{aligned}$$

The Kaplan–Meier product–limit survivor function estimate for all t values is plotted in Figure 10.4. Downward steps occur at the $k = 7$ observed failure times. Some software packages place a vertical hash mark on the survivor function estimate to highlight censored values that occur between observed failure times; these occur at times 9, 11, 17, 19, 20, 25, 32, and 34 in Figure 10.4. The effect of censored observations in the survivor function estimate is a larger downward step at the next subsequent observed failure time. If there is a tie between an observed failure time and censoring time (as there is at time 6 in this example) the standard convention of including the censored value(s) in the risk set when computing the number of items at risk means that there will be a larger downward step in the survivor function estimate following the tied value. Since the last observed data value, 35*, corresponds to a right-censored observation, the survivor function estimate is truncated at time 35 and is assumed to be undefined for $t > 35$.

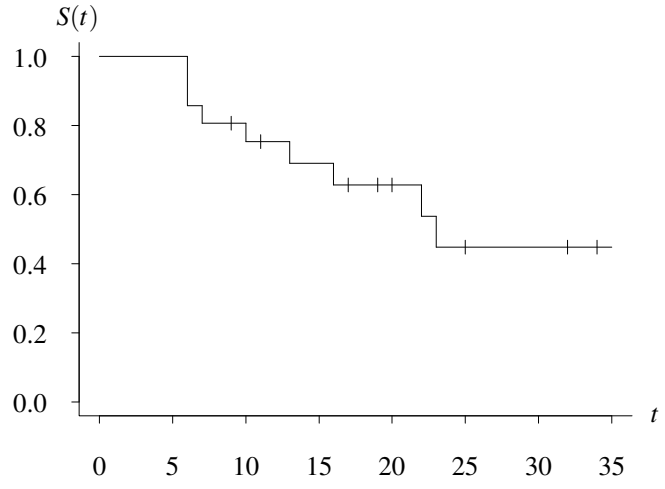

Figure 10.4: Product–limit survivor function estimate for the 6–MP treatment group.

There is a second and perhaps more intuitive way of deriving the Kaplan–Meier product–limit estimator, often referred to as the "redistribute-to-the-right" algorithm. This technique begins by defining an initial probability mass function that apportions equal probability to each of the n data values. In subsequent passes through the data, this probability mass function estimate is modified as

the probability is redistributed to the right, with special treatment given to right-censored observations. The algorithm is illustrated next on the 6–MP treatment group data set from Example 8.3.

Example 10.4 Implement the redistribute-to-the-right algorithm for calculating the Kaplan–Meier product–limit estimate of the survivor function for the remission time in the treatment group in the 6–MP clinical trial from Example 8.3.

For the $n = 21$ patients in the treatment group for the 6–MP experiment, each failure or censoring time is initially assigned a mass value of $1/n$ as follows:

6	6	6	6^*	7	9^*	10	10^*	11^*	13	...
$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	...

If there were no censored observations, the fractions would be the appropriate estimators for the probability mass function values. This probability mass function corresponds to the empirical survivor function described earlier in this section. Combining the three tied observed failures at $t = 6$ yields

6	6^*	7	9^*	10	10^*	11^*	13	...
$\frac{1}{7}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	$\frac{1}{21}$	...

As indicated earlier, there are mass values in the Kaplan–Meier product–limit estimator only at observed failure times. Since the random censoring model is assumed, the mass associated with the individual whose remission time is right censored at 6 weeks can be split evenly among each of the 17 subsequent failure/censoring times:

6	6^*	7	9^*	10	10^*	11^*	13	...
$\frac{1}{7}$	0	$\frac{6}{119}$	$\frac{6}{119}$	$\frac{6}{119}$	$\frac{6}{119}$	$\frac{6}{119}$	$\frac{6}{119}$	...

because $\frac{1}{21} + \frac{1}{17} \cdot \frac{1}{21} = \frac{6}{119}$. The probability mass function estimates at $t = 6$ and $t = 7$ have now been determined. The mass value $\frac{6}{119}$ associated with the right-censored observation at time 9 can be allocated among the 15 subsequent failure/censoring times as

6	6^*	7	9^*	10	10^*	11^*	13	...
$\frac{1}{7}$	0	$\frac{6}{119}$	0	$\frac{32}{595}$	$\frac{32}{595}$	$\frac{32}{595}$	$\frac{32}{595}$	...

because $\frac{6}{119} + \frac{1}{15} \cdot \frac{6}{119} = \frac{96}{1785} = \frac{32}{595}$. After allocating the mass at 10^* to the subsequent 13 data values and the mass at 11^* to the subsequent 12 data values, the estimator becomes

6	6^*	7	9^*	10	10^*	11^*	13	...
$\frac{1}{7}$	0	$\frac{6}{119}$	0	$\frac{32}{595}$	0	0	$\frac{16}{255}$	...

When this process is continued through all the data values, the resulting probability mass function defined on the observed failure times corresponds to the Kaplan–Meier product–limit estimator. To check this for one specific time value, the survivor function estimate at time 14 is

$$\hat{S}(14) = 1 - \frac{1}{7} - \frac{6}{119} - \frac{32}{595} - \frac{16}{255} = \frac{176}{255} = 0.69,$$

which matches the result from the previous example.