have on X is left truncated by the patient's age at entry into the hospital and right censored by the end of the study.

(a) Plot the number at risk, Y_i, as a function of age.

(b) Estimate the conditional survival function for a psychiatric patient who has survived to age 30 without entering a psychiatric hospital.

4.9 Hoel and Walburg (1972) report results of an experiment to study the effects of radiation on life lengths of mice. Mice were given a dose of 300 rads of radiation at 5–6 weeks of age and followed to death. At death each mouse was necropsied to determine if the cause of death was thymic lymphoma, reticulum cell sarcoma, or another cause. The ages of the mice at death are shown below:

Cause of Death	*Age at Death (Days)*
Thymic lymphoma	158, 192, 193, 194, 195, 202, 212, 215, 229, 230, 237, 240, 244, 247, 259, 300, 301, 337, 415, 444, 485, 496, 529, 537, 624, 707, 800
Reticulum cell sarcoma	430, 590, 606, 638, 655, 679, 691, 693, 696, 747, 752, 760, 778, 821, 986
Other causes	136, 246, 255, 376, 421, 565, 616, 617, 652, 655, 658, 660, 662, 675, 681, 734, 736, 737, 757, 769, 777, 801, 807, 825, 855, 857, 864, 868, 870, 873, 882, 895, 910, 934, 942, 1,015, 1,019

(a) For each of the three competing risks estimate the cumulative incidence function at 200, 300, . . . , 1,000 days by considering the two other risks as a single competing risk.

(b) Show that the sum of the three cumulative incidence functions found in part a is equal to the Kaplan-Meier estimate of the overall survival function for this set of data.

(c) Repeat part a using the complement of the marginal Kaplan-Meier estimates. What are the quantities estimating and how different from the results found in part a are these estimates?

(d) Compute the conditional probability function for thymic lymphoma at 500 and 800 days. What are the quantities estimating?

4.10 Using the data reported in section 1.3 for the AML low risk and AML high risk groups, find the following quantities for the two competing risks of relapse and death:

(a) The estimated cumulative incidence at one year.

(b) The standard errors of the two estimates in part a.

(c) The estimated conditional probabilities of relapse and of death in remission.

(d) The standard errors of the probabilities found in part c.

(e) Graphically express the development of relapse and death in remission for these two disease groups.

5 Estimation of Basic Quantities for Other Sampling Schemes

5.1 Introduction

In Chapter 4, we examined techniques for estimating the survival function for right-censored data in sections 4.2-4.5 and for left-truncated data in section 4.6. In this chapter, we discuss how to estimate the survival function for other sampling schemes, namely, left, double, and interval censoring, right-truncation, and grouped data. Each sampling scheme provides different information about the survival function and requires a different technique for estimation.

In section 5.2, we examine estimating for three censoring schemes. In the first scheme, left censoring, censored individuals provide information indicating only that the event has occurred prior to entry into the study. Double-censored samples include some individuals that are left-censored and some individuals that are right-censored. In both situations, some individuals with exact event times are observed. The last censoring scheme considered in this section is interval censoring, where individual event times are known to occur only within an interval.

In section 5.3, we present an estimator of the survival function for right-truncated data. Such samples arise when one samples individuals from event records and, retrospectively determines the time to event.

In section 5.4, we consider estimation techniques for grouped data. In elementary statistics books, the relative frequency histogram is often used to describe such data. In survival analysis, however, the complicating feature of censoring renders this simple technique ineffective because we will not know exactly how many events would have occurred in each interval had all subjects been observed long enough for them to have experienced the event. The life table methodology extends these elementary techniques to censored data.

Grouped survival data arises in two different situations. In the first, discussed in section 5.4, we follow a large group of individuals with a common starting time. The data consists of only the number who die or are lost within various time intervals. The basic survival quantities are estimated using a *cohort* (sometimes called a generation) life table.

In the second, a different sampling scheme is considered. Here a cross-sectional sample of the number of events and number at risk at different ages in various time intervals are recorded. In this instance, the cohort life table, which is based on longitudinal data, is not appropriate, and the basic survival quantities are estimated by the *current* life table. We refer the reader to Chiang (1984) for details of constructing this type of life table.

5.2 Estimation of the Survival Function for Left, Double, and Interval Censoring

In this section we shall present analogues of the Product-Limit estimator of the survival function for left-, double-, and interval-censored data. As discussed in section 3.3, left-censoring occurs when some individuals have experienced the event of interest prior to the start of the period of observation, while interval censoring occurs when all that is known is that the event of interest occurs between two known times. Double censoring occurs when both left censoring and right censoring are present. In addition some exact event times are observed. Each censoring scheme requires a distinct construction of the survival function.

For left censoring for some individuals, all we know is that they have experienced the event of interest prior to their observed study time, while for others their exact event time is known. This type of censoring is handled quite easily by reversing the time scale. That is, instead of measuring time from the origin we fix a large time τ and define new times by τ minus the original times. The data set based on these reverse

times is now right-censored and the estimators in sections 4.2–4.4 can be applied directly. Note that the Product-Limit estimator in this case is estimating $P[\tau - X > t] = P[X < \tau - t]$. Examples of this procedure are found in Ware and Demets (1976).

Examples of pure left censoring are rare. More common are samples which include both left and right censoring. In this case a modified Product-Limit estimator has been suggested by Turnbull (1974). This estimator, which has no closed form, is based on an iterative procedure which extends the notion of a self-consistent estimator discussed in Theoretical Note 3 of section 4.2. To construct this estimator we assume that there is a grid of time points $0 = t_0 < t_1 < t_2 < \cdots < t_m$ at which subjects are observed. Let d_i be the number of deaths at time t_i (note here the t_i's are not event times, so d_i may be zero for some points). Let r_i be the number of individuals right-censored at time t_i (i.e., the number of individuals withdrawn from the study without experiencing the event at t_i), and let c_i be the number of left-censored observations at t_i (i.e., the number for which the only information is that they experienced the event prior to t_i). The only information the left-censored observations at t_i give us is that the event of interest has occurred at some $t_j \leq t_i$. The self-consistent estimator estimates the probability that this event occurred at each possible t_j less than t_i based on an initial estimate of the survival function. Using this estimate, we compute an expected number of deaths at t_j, which is then used to update the estimate of the survival function and the procedure is repeated until the estimated survival function stabilizes. The algorithm is as follows:

Step 0: Produce an initial estimate of the survival function at each t_j, $S_0(t_j)$. Note any legitimate estimate will work. Turnbull's suggestion is to use the Product-Limit estimate obtained by ignoring the left-censored observations.

Step (K + 1)1: Using the current estimate of S, estimate $p_{ij} = P[t_{j-1} < X \leq t_i \mid X \leq t_i]$ by $\frac{S_K(t_{j-1}) - S_K(t_j)}{1 - S_K(t_i)}$, for $j \leq i$.

Step (K + 1)2: Using the results of the previous step, estimate the number of events at time t_i by $\hat{d}_i = d_i + \sum_{l=j}^{m} c_l p_{ij}$.

Step (K + 1)3: Compute the usual Product-Limit estimator (4.2.1) based on the estimated right-censored data with $\hat{d}_i$ events and r_i right-censored observations at t_i, ignoring the left-censored data. If this estimate, $S_{K+1}(t)$, is close to $S_K(t)$ for all t_i, stop the procedure; if not, go to step 1.

EXAMPLE 5.1 To illustrate Turnbull's algorithm, consider the data in section 1.17 on the time at which California high school boys first smoked marijuana. Here left censoring occurs when boys respond that they have used

TABLE 5.1
Initial Estimate of the Survival Function Formed by Ignoring the Left-Censored Observations

i	Age t_i	Number Left-Censored c_i	Number of Events d_i	Number Right-Censored r_i	$Y_i = \sum_{j=i}^{m} d_j + r_j$	$S_o(t_i)$
0	0					1.000
1	10	0	4	0	179	0.978
2	11	0	12	0	175	0.911
3	12	0	19	2	163	0.804
4	13	1	24	15	142	0.669
5	14	2	20	24	103	0.539
6	15	3	13	18	59	0.420
7	16	2	3	14	28	0.375
8	17	3	1	6	11	0.341
9	18	1	0	0	4	0.341
10	>18	0	4	0	4	0.000
Total		12	100	79	0	

marijuana but can not recall the age of first use, while right-censored observations occur when boys have never used marijuana. Table 5.1 shows the data and the initial Product-Limit estimator, S_o, obtained by ignoring the left-censored observations.

In step 1, we estimate the p_{ij}'s. Note we only need estimates for those i with $c_i > 0$ for the computations in step 2. For the left-censored observation at t_4 we have

$$p_{41} = \frac{1.000 - 0.978}{1 - 0.669} = 0.067; \quad p_{42} = \frac{0.978 - 0.911}{1 - 0.669} = 0.202;$$
$$p_{43} = \frac{0.911 - 0.804}{1 - 0.669} = 0.320; \quad p_{44} = \frac{0.804 - 0.669}{1 - 0.669} = 0.410.$$

Similar computations yield the values for p_{ij} in Table 5.2.

Using these values, we have $\hat{d}_1 = 4 + 0.067 \times 1 + 0.048 \times 2 + 0.039 \times 3 + 0.036 \times 2 + 0.034 \times 3 + 0.034 \times 1 = 4.487$, $\hat{d}_2 = 13.461$, $\hat{d}_3 = 21.313$, $\hat{d}_4 = 26.963$, $\hat{d}_5 = 22.437$, $\hat{d}_6 = 14.714$, $\hat{d}_7 = 3.417$, $\hat{d}_8 = 1.206$, $\hat{d}_9 = 0$, and $\hat{d}_{10} = 4$. These values are then used in Table 5.3 to compute the updated estimate of the survival function, $S_1(t)$.

Then using these estimates of the survival function the p_{ij}'s are re-computed, the $\hat{d}$'s are re-estimated, and the second step estimator $S_2(t_i)$ is computed. This estimate is found to be within 0.001 of S_1 for all t_i, so

TABLE 5.2
Values of p_{ij} in Step 1

i/j	4	5	6	7	8	9
1	0.067	0.048	0.039	0.036	0.034	0.034
2	0.202	0.145	0.116	0.107	0.102	0.102
3	0.320	0.230	0.183	0.170	0.161	0.161
4	0.410	0.295	0.234	0.218	0.206	0.206
5		0.281	0.224	0.208	0.197	0.197
6			0.205	0.190	0.180	0.180
7				0.072	0.068	0.068
8					0.052	0.052
9						0.000

TABLE 5.3
First Step of the Self-Consistency Algorithm

t_i	$\hat{d}$	r_i	Y_i	$S_1(t_i)$
0				1.000
10	4.487	0	191.000	0.977
11	13.461	0	186.513	0.906
12	21.313	2	173.052	0.794
13	26.963	15	149.739	0.651
14	22.437	24	107.775	0.516
15	14.714	18	61.338	0.392
16	3.417	14	28.624	0.345
17	1.207	6	11.207	0.308
18	0.000	0	4.000	0.308
>18	4.000	0	4.000	0.000

the iterative process stops. The final estimate of the survival function, to three decimal places, is given by $S_1(t)$ in Table 5.3.

In some applications the data may be interval-censored. Here the only information we have for each individual is that their event time falls in an interval $(L_i, R_i]$, $i = 1, \ldots, n$, but the exact time is unknown. An estimate of the survival function can be found by a modification of above iterative procedure as proposed by Turnbull (1976). Let $0 = \tau_0 < \tau_1 < \cdots < \tau_m$ be a grid of time points which includes all the points L_i, R_i for $i = 1, \ldots, n$. For the ith observation, define a weight α_{ij} to be 1 if the interval $(\tau_{j-1}, \tau_j]$ is contained in the interval $(L_i, R_i]$, and 0 otherwise. Note that α_{ij} is an indicator of whether the event which

occurs in the interval $(L_i, R_i]$ could have occurred at τ_j. An initial guess at $S(\tau_j)$ is made. The algorithm is as follows:

Step 1: Compute the probability of an event's occurring at time τ_j, $p_j = S(\tau_{j-1}) - S(\tau_j)$, $j = 1, \ldots, m$.

Step 2: Estimate the number of events which occurred at τ_i by

$$d_i = \sum_{i=1}^{n} \frac{\alpha_{ij} p_j}{\sum_k \alpha_{ik} p_k}.$$

Note the denominator is the total probability assigned to possible event times in the interval $(L_i, R_i]$.

Step 3: Compute the estimated number at risk at time τ_i by $Y_i = \sum_{k=j}^{m} d_k$.

Step 4: Compute the updated Product-Limit estimator using the pseudo data found in steps 2 and 3. If the updated estimate of S is close to the old version of S for all τ_i's, stop the iterative process, otherwise repeat steps 1–3 using the updated estimate of S.

EXAMPLE 5.2

To illustrate the estimation procedure for interval-censored data consider the data on time to cosmetic deterioration for early breast cancer patients presented in section 1.18.

Consider first the 46 individuals given radiation therapy only. The end points of the intervals for the individuals form the τ_i's as listed in Table 5.4. An initial estimate is found by distributing the mass of 1/46 for the ith individual equally to each possible value of τ contained in $(L_i, R_i]$. For example, the individual whose event time is in the interval $(0, 7]$ contributes a value of $(1/46)(1/4)$ to the probability of the event's occurring at 4, 5, 6, and 7 months, respectively. Using this initial approximation in step 1, we can compute the p_j's. Here, for example, we have $p_1 = 1. - 0.979 = 0.021$, $p_2 = 0.979 - 0.955 = 0.024$, $p_3 = 0.0214$, etc. The estimated number of deaths as shown in Table 5.4 is then computed. As an example, at $\tau = 4$ we have $d_1 = 0.021/(0.021 + 0.024 + 0.021 + 0.029 + 0.031) + 0.021/(0.021 + 0.024 + 0.021 + 0.029) + 0.021/(0.0201 + 0.024) = 0.842$. These estimates are then used to compute the estimated number at risk at each τ_i.

Using the estimated number of deaths and number at risk we compute the updated estimate of the survival function, as shown in Table 5.4. This revised estimate is then used to re-estimate the number of deaths, and the process continues until the maximum change in the estimate is less than 10^{-7}. This requires, in this case, 305 iterations of the process. The final estimate is shown in the second half of Table 5.4.

Figure 5.1 shows the estimated survival functions for the radiotherapy only and combination radiotherapy and chemotherapy groups. The

TABLE 5.4
Calculations for Estimating the Survival Function Based on Interval-Censored Data

τ	*Initial* $S(t)$	*Estimated Number of Deaths* d	*Estimated Number at Risk* Y	*Updated* $S(t)$	*Change*
0	1.000	0.000	46.000	1.000	0.000
4	0.979	0.842	46.000	0.982	−0.002
5	0.955	1.151	45.158	0.957	−0.002
6	0.934	0.852	44.007	0.938	−0.005
7	0.905	1.475	43.156	0.906	−0.001
8	0.874	1.742	41.680	0.868	0.006
10	0.848	1.286	39.938	0.840	0.008
11	0.829	0.709	38.653	0.825	0.004
12	0.807	1.171	37.944	0.799	0.008
14	0.789	0.854	36.773	0.781	0.008
15	0.775	0.531	35.919	0.769	0.006
16	0.767	0.162	35.388	0.766	0.001
17	0.762	0.063	35.226	0.764	−0.002
18	0.748	0.528	35.163	0.753	−0.005
19	0.732	0.589	34.635	0.740	−0.009
22	0.713	0.775	34.045	0.723	−0.011
24	0.692	0.860	33.270	0.705	−0.012
25	0.669	1.050	32.410	0.682	−0.012
26	0.652	0.505	31.360	0.671	−0.019
27	0.637	0.346	30.856	0.663	−0.026
32	0.615	0.817	30.510	0.646	−0.031
33	0.590	0.928	29.693	0.625	−0.035
34	0.564	1.056	28.765	0.602	−0.039
35	0.542	0.606	27.709	0.589	−0.047
36	0.523	0.437	27.103	0.580	−0.057
37	0.488	1.142	26.666	0.555	−0.066
38	0.439	1.997	25.524	0.512	−0.073
40	0.385	2.295	23.527	0.462	−0.077
44	0.328	2.358	21.233	0.410	−0.082
45	0.284	1.329	18.874	0.381	−0.097
46	0.229	1.850	17.545	0.341	−0.112
48	0.000	15.695	15.695	0.000	0.000

Interval	*Survival Probability*
0–4	1.000
5–6	0.954
7	0.920
8–11	0.832
12–24	0.761
25–33	0.668
34–38	0.586
40–48	0.467
≥48	0.000

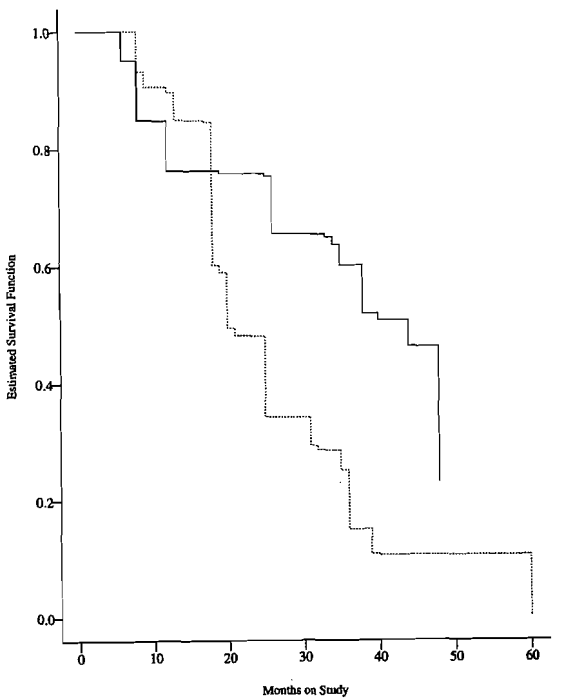

Figure 5.1 *Estimated probability of no evidence of breast retraction based on interval-censored data. Radiation only group* (———). *Radiation plus chemotherapy group* (——)

figure seems to indicate a longer time to cosmetic deterioration for patients given only radiotherapy.

Practical Notes

1. An example of the left-censored Product-Limit estimator is found in Ware and DeMets (1976). A key assumption needed for these calculations is that the death and censoring times are independent.
2. For the case of combined left and right censoring, Turnbull (1974) shows that an estimator of the variance-covariance matrix of $\hat{S}(t)$ is given by the matrix $\mathbf{V} = (V_{ij})$, where $V_{ij} = \text{Cov}[\hat{S}(t_i), \hat{S}(t_j)]$, constructed as follows. Let

$$A_{ii} = \frac{d_i}{[\hat{S}(t_{i-1}) - \hat{S}(t_i)]^2} + \frac{d_{i+1}}{[\hat{S}(t_i) - \hat{S}(t_{i+1})]^2} + \frac{r_i}{\hat{S}(t_i)^2} + \frac{c_i}{(1 - \hat{S}(t_i))^2}, \; i = 1, 2, \ldots, m-1;$$

$$A_{mm} = \frac{d_m}{[\hat{S}(t_{m-1}) - \hat{S}(t_m)]^2} + \frac{r_m}{\hat{S}(t_m)^2} + \frac{c_m}{[1 - \hat{S}(t_m)]^2};$$

$$A_{i+1,i} = A_{i,i+1} = -\frac{d_{i+1}}{[\hat{S}(t_i) - \hat{S}(t_{i+1})]^2}, \; i = 1, 2, \ldots, m-1; \text{and}$$

$$A_{ij} = 0 \text{ for } |i - j| \geq 2.$$

Define the matrix $\mathbf{J}$ to be the symmetric matrix given by

$$\mathbf{J} = \begin{pmatrix} b_1 & q_1 & 0 & 0 & \ldots & 0 & 0 \\ q_1 & b_2 & q_2 & 0 & \ldots & 0 & 0 \\ 0 & q_2 & b_3 & q_3 & \ldots & 0 & 0 \\ 0 & 0 & q_3 & b_4 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & 0 & 0 \\ 0 & 0 & 0 & 0 & \ldots & b_{m-1} & q_{m-1} \\ 0 & 0 & 0 & 0 & \ldots & q_{m-1} & b_m \end{pmatrix}$$

where $b_i = A_{ii}$ and $q_i = A_{i,i+1}$. The inverse of this matrix is the estimated covariance matrix $\mathbf{V}$.

Using the data in Example 5.1, we find that

$$\mathbf{J} = \begin{pmatrix} 9941.9 & -2380.5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -2380.5 & 3895.1 & -1514.7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1514.7 & 2691.5 & -1173.7 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1173.7 & 2314.6 & -1097.4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1097.4 & 2041.5 & -845.5 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -845.5 & 2328.8 & -1358.1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1358.1 & 2210.8 & -730.5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -730.5 & 800 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 44.3 & -42.2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -42.2 & 42.2 \end{pmatrix}$$

and the matrix $\mathbf{V}$ is

$$\begin{pmatrix} 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0001 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ 0.0001 & 0.0005 & 0.0004 & 0.0003 & 0.0003 & 0.0002 & 0.0002 & 0.0002 & 0.0000 & 0.0000 \\ 0.0001 & 0.0004 & 0.0009 & 0.0008 & 0.0006 & 0.0004 & 0.0004 & 0.0004 & 0.0000 & 0.0000 \\ 0.0001 & 0.0003 & 0.0008 & 0.0013 & 0.0010 & 0.0007 & 0.0007 & 0.0006 & 0.0000 & 0.0000 \\ 0.0001 & 0.0003 & 0.0006 & 0.0010 & 0.0015 & 0.0011 & 0.0010 & 0.0009 & 0.0000 & 0.0000 \\ 0.0000 & 0.0002 & 0.0004 & 0.0007 & 0.0011 & 0.0017 & 0.0015 & 0.0014 & 0.0000 & 0.0000 \\ 0.0000 & 0.0002 & 0.0004 & 0.0007 & 0.0010 & 0.0015 & 0.0020 & 0.0018 & 0.0000 & 0.0000 \\ 0.0000 & 0.0002 & 0.0004 & 0.0006 & 0.0009 & 0.0014 & 0.0018 & 0.0029 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.3263 & 0.3187 \\ 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.3187 & 0.3345 \end{pmatrix}$$

Thus, the estimated survival function and its standard error are obtained from the square root of the diagonal elements of the matrix **V**. In this example,

Age	$\hat{S}(t)$	*Standard Error*
0.000	1.000	0.000
10.000	0.977	0.011
11.000	0.906	0.022
12.000	0.794	0.031
13.000	0.651	0.036
14.000	0.516	0.039
15.000	0.392	0.041
16.000	0.345	0.044
17.000	0.308	0.054
18.000	0.308	0.571
> 18	0.000	0.578

3. Standard errors for the estimator of the survival function based on interval-censored data are found in Turnbull (1976) or Finkelstein and Wolfe (1985). Finkelstein (1986) and Finkelstein and Wolfe (1985) provide algorithms for adjusting these estimates for possible covariate effects.

Theoretical Notes

1. For left-censored data, Gomez et al. (1992) discuss the derivation of the left-censored Kaplan–Meier estimator and a "Nelson–Aalen" estimator of the cumulative backward hazard function defined by $G(t) = \int_t^\infty \frac{f(x)}{F(x)}\,dx$. These derivations are similar to those discussed in the notes after section 4.2. Further derivations are found in Andersen et al. (1993), using a counting process approach.
2. The estimator of the survival function, based on Turnbull's algorithms for combined left and right censoring or for interval censoring, are generalized maximum likelihood estimators. They can be derived by a self-consistency argument or by using a modified EM algorithm. For both types of censoring, standard counting process techniques have yet to be employed for deriving results.

5.3 Estimation of the Survival Function for Right-Truncated Data

For right-truncated data, only individuals for which the event has occurred by a given date are included in the study. Right truncation arises commonly in the study of infectious diseases. Let T_i denote the chronological time at which the ith individual is infected and X_i the time between infection and the onset of disease. Sampling consists of observing (T_i, X_i) for patients over the period (0 to τ). Note that only patients who have the disease prior to τ are included in the study. Estimation for this type of data proceeds by reversing the time axis. Let $R_i = \tau - X_i$. The R_i's are now left truncated in that only individuals with values of $T_i \leq R_i$ are included in the sample. Using the method discussed in section 4.6 for left-truncated data, the Product-Limit estimator of $Pr[R > t \mid R \geq 0]$ can be constructed. In the original time scale, this is an estimator of $Pr[X < \tau - t \mid X \leq \tau)]$. Example 5.3 shows that this procedure is useful in estimating the induction time for AIDS.

EXAMPLE 5.3 To illustrate the analysis of right-truncated data, consider the data on the induction time for 37 children with transfusion-related AIDS, described in section 1.19. The data for each child consists of the time of infection T_i (in quarter of years from April 1, 1978) and the waiting time to induction X_i. The data was based on an eight year observational window, so $\tau = 8$ years.

Table 5.5 shows the calculations needed to construct the estimate of the waiting time to infection distribution. Here $R_i = 8 - X_i$. The column headed d_i is the number of individuals with the given value of R_i or, in the original time scale, the number with an induction time of X_i. The number at risk column, Y_i, is the number of individuals with a value of R between X_i and R_i or, in the original time scale, the number of individuals with induction times no greater than X_i and infection times no greater than $8 - X_i$. For example, when $X_i = 1.0$ ($R_i = 7.0$) in the original time scale, there are 19 individuals with induction times greater than 1 and one individual with an infection time greater than 7, so $Y_i = 37 - 19 - 1 = 17$. The final column of Table 5.5 is the Product-Limit estimator for R_i based on d_i and Y_i. This is an estimate of the probability that the waiting time to AIDS is less than x, given X is less than 8 years, $G(t) = Pr[X < x \mid X \leq 8]$. Figure 5.2 shows the estimated distribution of the induction time for AIDS for the 37 children and 258 adults.

TABLE 5.5
Estimation of the Distribution of the Induction Time to AIDS Based on Right-Truncated Data

T_i	X_i	R_i	d_i	Y_i	$\hat{P}r[X < x_i \mid X \leq 8]$
5.25	0.25	7.75			
7.25	0.25	7.75	2	2	0.0000
5.00	0.50	7.50			
5.50	0.50	7.50			
6.00	0.50	7.50			
6.25	0.50	7.50			
6.75	0.50	7.50	5	7	0.0243
3.50	0.75	7.25			
3.75	0.75	7.25			
5.00	0.75	7.25			
6.50	0.75	7.25			
6.75	0.75	7.25			
7.00	0.75	7.25	6	13	0.0850
2.75	1.00	7.00			
3.75	1.00	7.00			
4.00	1.00	7.00			
4.75	1.00	7.00			
5.50	1.00	7.00	5	17	0.1579
6.00	1.25	6.75			
6.25	1.25	6.75	2	18	0.2237
5.00	1.50	6.50			
5.25	1.50	6.50			
5.50	1.50	6.50	3	19	0.2516
3.00	1.75	6.25			
4.25	1.75	6.25			
5.75	1.75	6.25	3	21	0.2988
1.50	2.25	5.75			
4.75	2.25	5.75	2	19	0.3486
5.00	2.50	5.50			
5.25	2.50	5.50	2	20	0.3896
3.75	2.75	5.25	1	18	0.4329
2.25	3.00	5.00			
3.75	3.00	5.00	2	17	0.4584
4.50	3.25	4.75	1	14	0.5195
3.75	3.50	4.50	1	13	0.5594
3.75	4.25	3.75	1	11	0.6061
1.00	5.50	2.50	1	3	0.6667

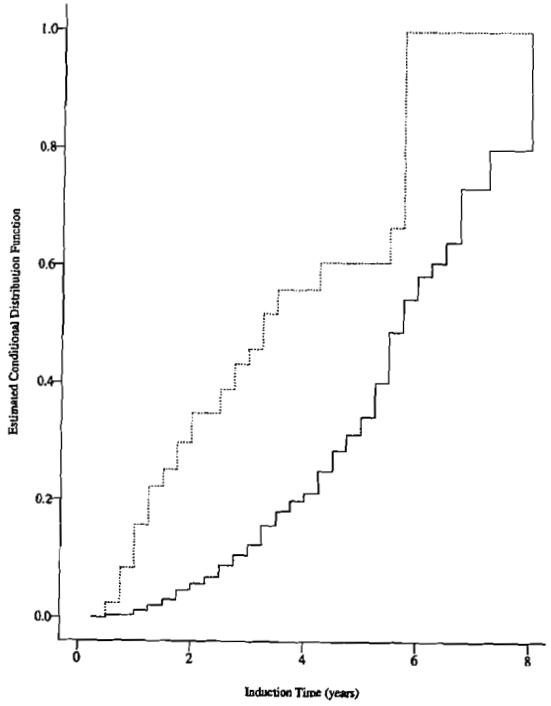

Figure 5.2 *Estimated conditional distribution of the induction time for AIDS for the 258 adults* (———) *and 37 children* (——)

Practical Note

1. For right-truncated data, standard errors of the survival estimator function follow directly by using Greenwood's formula. Lagakos et al. (1988) discuss techniques for comparing two samples based on right-truncated data. Gross and Huber–Carol (1992) discuss regression models for right-truncated data.

Theoretical Note

1. For right-truncated data, as for left censoring, the reversal of time allows direct estimation of the cumulative backward hazard function. Keiding and Gill (1990) discuss the large-sample-size properties of the estimated survival function for this type of data, using a counting process approach.

5.4 Estimation of Survival in the Cohort Life Table

A "cohort" is a group of individuals who have some common origin from which the event time will be calculated. They are followed over time and their event time or censoring time is recorded to fall in one of $k+1$ adjacent, nonoverlapping intervals, $(a_{j-1}, a_j]$, $j = 1, \ldots, k+1$. A traditional cohort life table presents the actual mortality experience of the cohort from the birth of each individual to the death of the last surviving member of the cohort. Censoring may occur because some individuals may migrate out of the study area, drop out of observation, or die unrecorded.

The cohort life table has applications in assessing survival in animal or human populations. The event need not be death. Other human studies may have, as an end point, the first evidence of a particular disease or symptom, divorce, conception, cessation of smoking, or weaning of breast-fed newborns, to name a few.

The basic construction of the cohort life table is described below:

1. The first column gives the adjacent and nonoverlapping fixed intervals, $I_j = (a_{j-1}, a_j]$, $j = 1, \ldots, k+1$, with $a_0 = 0$ and $a_{k+1} = \infty$. Event and censoring times will fall into one and only one of these intervals. The lower limit is in the interval and the upper limit is the start of the next interval.

2. The second column gives the number of subjects Y_j', entering the jth interval who have not experienced the event.

3. The third column gives the number of individuals W_j lost to follow-up or withdrawn alive, for whatever reason, in the jth interval. As for the product limit estimator, the censoring times must be independent of the event times.

4. The fourth column gives an estimate of the number of individuals Y_j at risk of experiencing the event in the jth interval, assuming that censoring times are uniformly distributed over the interval $Y_j = Y_j' - W_j/2$.

5. The fifth column reports the number of individuals d_j who experienced the event in the jth interval.

6. The sixth column gives the estimated survival function at the start of the jth interval $\hat{S}(a_{j-1})$. For the first interval, $\hat{S}(a_0) = 1$. Analogous to the product-limit estimator for successive intervals (see 4.2.1),

$$\hat{S}(a_j) = \hat{S}(a_{j-1})[1 - d_j/Y_j]. \tag{5.4.1}$$

$$= \prod_{i=1}^{j}(1 - d_i/Y_i)$$

7. The seventh column gives the estimated probability density function $\hat{f}(a_{mj})$ at the midpoint of the jth interval, $a_{mj} = (a_j + a_{j-1})/2$. This quantity is defined as the probability of having the event in the jth interval per unit time, i.e.,

$$\hat{f}(a_{mj}) = [\hat{S}(a_{j-1}) - \hat{S}(a_j)]/(a_j - a_{j-1}) \tag{5.4.2}$$

8. The eighth column gives the estimated hazard rate, $\hat{h}(a_{mj})$ at the midpoint of the jth interval, a_{mj}. Based on (2.3.2), this quantity is defined in the usual way as

$$\begin{aligned}\hat{h}(a_{mj}) &= \hat{f}(a_{mj})/\hat{S}(a_{mj}) \\ &= \hat{f}(a_{mj})/\{\hat{S}(a_j) + [\hat{S}(a_{j-1}) - \hat{S}(a_j)]/2\} \\ &= \frac{2\hat{f}(a_{mj})}{[\hat{S}(a_j) + \hat{S}(a_{j-1})]}\end{aligned} \tag{5.4.3}$$

Note that $\hat{S}(a_{mj})$ is based on a linear approximation between the estimate of S at the endpoints of the interval.

It may also be calculated as the number of events per person-units, i.e.,

$$\hat{h}(a_{mj}) = d_j/[(a_j - a_{j-1})(Y_j - d_j/2)]. \tag{5.4.4}$$

Because the last interval is theoretically infinite, no estimate of the hazard or probability density function (and, of course, their standard errors) may be obtained for this interval.

Other useful quantities in subsequent calculations are the estimated conditional probability of experiencing the event in the jth interval, $\hat{q}_j = d_j/Y_j$, and the conditional probability of surviving through the jth interval, $\hat{p}_j = 1 - \hat{q}_j = 1 - d_j/Y_j$. Specifically, we could write (5.4.1) as

$$\hat{S}(a_j) = \hat{S}(a_{j-1})\hat{p}_j.$$

Note, also, that (5.4.2) and (5.4.3) could be written as

$$\hat{f}(a_{mj}) = \hat{S}(a_{j-1})\hat{q}_j/(a_j - a_{j-1}) \text{ and}$$

$$\hat{h}(a_{mj}) = 2\hat{q}_j/[(a_j - a_{j-1})(1 + \hat{p}_j)],$$

respectively.

9. The ninth column gives the estimated standard deviation of survival at the beginning of the jth interval (see Greenwood, 1926) which is

approximately equal to

$$\hat{S}(a_{j-1})\sqrt{\sum_{i=1}^{j-1}\frac{\hat{q}_i}{Y_i\hat{p}_i}} = \hat{S}(a_{j-1})\sqrt{\sum_{i=1}^{j-1}\frac{d_i}{Y_i(Y_i-d_i)}}, \tag{5.4.5}$$

for $j = 2, \ldots, k+1$, and, of course, the estimated standard deviation of the constant $\hat{S}(a_o) = 1$ is 0. Note that this estimated standard error is identical to the standard error obtained for the product limit estimator in (4.2.2).

10. The tenth column shows the estimated standard deviation of the probability density function at the midpoint of the jth interval which is approximately equal to

$$\left[\frac{\hat{S}(a_{j-1})\hat{q}_j}{(a_j-a_{j-1})}\sqrt{\sum_{i=1}^{j-1}[\hat{q}_i/(Y_i\hat{p}_i)]+[\hat{p}_j/(Y_j\hat{q}_j)]}\right] \tag{5.4.6}$$

11. The last column gives the estimated standard deviation of the hazard function at the midpoint of the jth interval which is approximately equal to

$$\left\{\frac{1-[\hat{h}(a_{mj})(a_j-a_{j-1})/2]^2}{Y_jq_j}\right\}^{1/2}\cdot\hat{h}(a_{mj}) \tag{5.4.7}$$

As noted in Chapter 2, the mean would be computed as in formula (2.4.2) with $S(x)$ replaced by $\hat{S}(x)$. There is some ambiguity regarding the mean lifetime because $\hat{S}(x)$ may be defined in several ways, as explained in Chapter 4, when the largest observation is a censored observation. For this reason, the median lifetime is often used. The median survival time may be determined by using relationship (2.4.4). For life tables, one first determines the interval where $\hat{S}(a_j) \leq 0.5$ and $\hat{S}(a_{j-1}) \geq 0.5$. Then, the median survival time can be estimated by linear interpolation as follows:

$$\begin{aligned}\hat{x}_{0.5} &= a_{j-1} + [\hat{S}(a_{j-1}) - 0.5](a_j - a_{j-1})/[\hat{S}(a_{j-1}) - \hat{S}(a_j)] \qquad (5.4.8)\\ &= a_{j-1} + [\hat{S}(a_{j-1}) - 0.5]/\hat{f}(a_{mj})\end{aligned}$$

Because we are often interested in the amount of life remaining after a particular time, the mean residual lifetime and the median residual lifetime are descriptive statistics that will estimate this quantity. For reasons stated above, the median residual lifetime at time x is often the preferable quantity. If the mean residual lifetime can be estimated without ambiguity, then, formula (2.4.1) with $S(x)$ replaced by $\hat{S}(x)$ is used. If the proportion of individuals surviving at time a_{i-1} is $S(a_{i-1})$, then the median residual lifetime is the amount of time that needs to be added to a_{i-1} so that $S(a_{i-1})/2 = S(a_{i-1} + \text{mdrl}(a_{i-1}))$, i.e., the $\text{mdrl}(a_{i-1})$ is the increment of time at which half of those alive at time a_{i-1} are expected to survive beyond. Suppose the jth interval contains the survival probability $S(a_{i-1} + \text{mdrl}(a_{i-1}))$, then an estimate of $\text{mdrl}(a_{i-1})$, determined in a similar fashion as (5.4.8) is given by

$$\widehat{\text{mdrl}}(a_{i-1}) = \tag{5.4.9}$$
$$(a_{j-1} - a_{i-1}) + [\hat{S}(a_{j-1}) - \hat{S}(a_{i-1})/2](a_j - a_{j-1})/[\hat{S}(a_{j-1}) - \hat{S}(a_j)]$$

Hence the median residual lifetime at time 0 will, in fact, be the median lifetime of the distribution.

The variance of this estimate is approximately

$$\widehat{\text{Var}}[\widehat{\text{mdrl}}(a_{i-1})] = \frac{[\hat{S}(a_{i-1})]^2}{4Y_i[\hat{f}(a_{mj})]^2} \tag{5.4.10}$$

Some major statistical packages will provide the median residual lifetime and its standard error at the beginning of each interval.

EXAMPLE 5.4

Consider The National Labor Survey of Youth (NLSY) data set discussed in section 1.14. Beginning in 1983, females in the survey were asked about any pregnancies that have occurred since they were last interviewed (pregnancies before 1983 were also documented). Questions regarding breast feeding are included in the questionnaire.

This data set consists of the information from 927 first-born children to mothers who chose to breast feed their child and who have complete information for all the variables of interest. The universe was restricted to children born after 1978 and whose gestation was between 20 and 45 weeks. The year of birth restriction was included in an attempt to eliminate recall problems.

The response variable in the data set is the duration of breast feeding in weeks, followed by an indicator if the breast feeding is completed (i.e., the infant is weaned).

The quantities described above are shown in Table 5.6 for this data set. Because none of the mothers claimed to wean their child before one week, the first interval will be from birth to two weeks. As always, when data are grouped, the selection of the intervals is a major decision. Generally, guidelines used in selecting intervals for frequency histograms apply, namely, the number of intervals should be reasonable, there should be enough observations within each interval to adequately represent that interval, and the intervals should be chosen to reflect the nature of the data. For example, in this data set, it is of interest to examine early weaners in smaller intervals and later weaners in broader intervals. This principle is also true in most population mortality studies where one wishes to study infant mortality in smaller intervals and later mortality may be studied in broader intervals.

TABLE 5.6
Life Table for Weaning Example

Week weaned [lower, upper)	*Number of infants not weaned entering interval*	*Number lost to follow-up or withdrawn without being weaned*	*Number exposed to weaning*	*Number weaned*	*Est. Cum. proportion not weaned at beginning of interval*	*Est. p.d.f. at middle of interval*	*Est. hazard at middle of interval*	*Est. stand. dev. of survival at beginning of interval*	*Est. stand. dev of. p.d.f. at middle of interval*	*Est. stand. dev of. hazard at middle of interval*
0– 2	927	2	926	77	1.0000	0.0416	0.0434	0	0.0045	0.0049
2– 3	848	3	846.5	71	0.9168	0.0769	0.0875	0.0091	0.0088	0.0104
3– 5	774	6	771	119	0.8399	0.0648	0.0836	0.0121	0.0055	0.0076
5– 7	649	9	644.5	75	0.7103	0.0413	0.0618	0.0149	0.0046	0.0071
7–11	565	7	561.5	109	0.6276	0.0305	0.0537	0.0160	0.0027	0.0051
11–17	449	5	446.5	148	0.5058	0.0279	0.0662	0.0166	0.0021	0.0053
17–25	296	3	294.5	107	0.3381	0.0154	0.0555	0.0158	0.0014	0.0052
25–37	186	0	186	74	0.2153	0.0071	0.0414	0.0138	0.0008	0.0047
37–53	112	0	112	85	0.1296	0.0061	0.0764	0.0114	0.0006	0.0066
53–	27	0	27	27	0.0313			0.0059		

An interesting feature of these data is that the hazard rate for weaning is high initially (many mothers stop breastfeeding between 1 and 5 weeks), levels off between 5 and 37 weeks, and begins to rise after 37 weeks as can be seen in Figure 5.3.

The median weaning time for all mothers starting to breast-feed is determined from (5.4.8) to be 11.21 weeks (with a standard error of 0.5678 weeks) and the median residual weaning time at 25 weeks is 15.40 weeks (with a standard error of 1.294 weeks).

Practical Notes

1. Summarizing the assumptions made in the life table methodology, we have seen that i) censored event times (including loss or withdrawal) are assumed to be independent of the time those individuals would have realized the event had they been observed until the event occurred, ii) the censoring times and death times are assumed to be uniformly distributed within each interval, (hence $Y_j' - W_j/2$ is taken to be the number exposed (or at risk) in the jth interval (see the number of people at risk in column 4 and the calculation of the number of person-units in the denominator of eq. (5.4.4), and iii) the hazard rate is assumed constant within intervals.
2. Individuals lost to follow-up are lost to observation if they move, fail to return for treatment, or, for some other reason, their survival status becomes unknown in the jth interval. On the other hand, individuals withdrawn alive are those known to be alive at the closing date of the study. Such observations typically arise in cohort studies or clinical trials. One assumption, as stated in the preceding note, is that the survival experience after the date of last contact of those lost to follow-up and withdrawn alive is similar to that of the individuals who remain under observation. Cutler and Ederer (1958) point out that the survival experience of lost individuals may be better than, the same as, or worse than individuals continuing under observation. Thus, every attempt should be made to trace such individuals and to minimize the number of individuals lost.

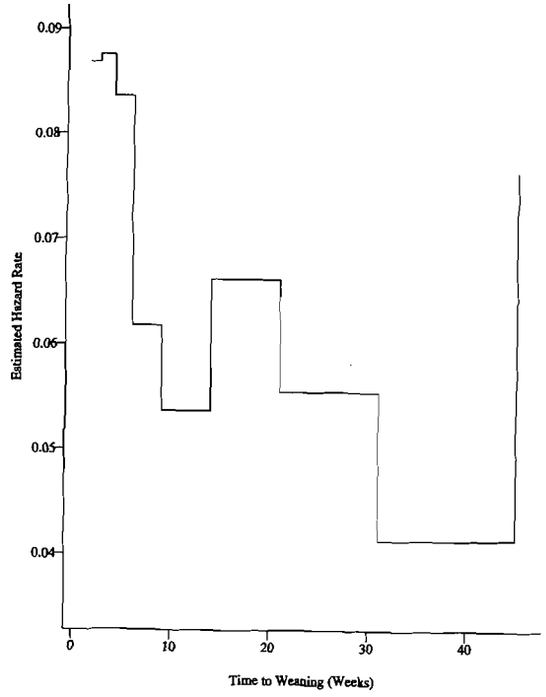

Figure 5.3 *Life table estimate of the hazard rate of the time to infant weaning*

3. SAS and SPSS have routines which reproduce the cohort life table.

Theoretical Notes

1. An alternative estimator of the hazard function is given by Sacher (1956) assuming that the hazard rate is constant within each interval

but is allowed to vary between intervals. This estimator is given by $\hat{h}(a_{mj}) = (-\ln \hat{p}_j)/(a_j - a_{j-1})$, which, Gehan and Siddiqui (1973) show, is slightly more biased than (5.4.3).

2. If the lengths of the grouping intervals approach 0, then, the life table estimates of the survival function are equivalent to the Kaplan–Meier estimate (Thompson, 1977). This limiting process provides a framework to link life table estimates with those using exact lifetimes, even in the presence of covariates.

5.5 Exercises

5.1 A study of 190 first-year medical students asked the question, How old were you when you first smoked a cigarette? Responses were either the exact ages at which they started smoking, that they never smoked, or that they currently smoke but cannot remember when they started. The data is summarized below. Using this sample, estimate the survival function to within 0.001.

Age (t)	*Number Who Started Smoking at Age t*	*Number of Age t Who Smoke Now but Do Not Know the Age They Started*	*Number of Age t Who Do Not Smoke*
14	2	0	0
15	3	0	0
16	10	0	0
17	13	0	0
18	5	0	0
19	3	0	1
20	2	4	13
21	1	6	44
22	2	8	39
23	1	2	19
24	0	0	3
25	0	0	4
26	1	0	4
Total	43	20	127

5.2 A study involving 100 veterinarians was performed to estimate the time until their first needlestick injury. They completed a survey which asked, How many years after graduation from veterinarian school did you experience your first needlestick injury? Many of them could remember or determine from their records the month and year their first injury occurred, but others could only say that it happened before a certain time. Others had had no needlestick injury yet at the time of the survey. The data below reflects these times after graduation.

Time (t) After Graduation (in Months)	*Number Who Had Needlestick Injury at Time t*	*Number Who Had Needlestick Injury Prior to Time t*	*Number Who Never Had Needlestick Injury at Time t*
2	3	0	0
4	2	0	0
8	1	0	0
10	2	1	0
12	4	2	0
15	6	2	1
20	3	4	1
24	3	3	2
28	2	3	3
34	1	4	5
41	0	2	3
62	0	3	4
69	0	2	6
75	0	1	6
79	0	2	3
86	0	3	7
Total	27	32	41

Estimate the survival (injury-free) function to an accuracy of three decimal places.

5.3 Eighteen elderly individuals who had entered a nursing home in the past five years were asked when they experienced their first fall (post-admittance). Some of the individuals only indicated that it occurred within a certain time period (in months), whereas others said they had never had a fall. The data (in months post-admittance) is as follows:

Falls occurred in (6–12], (48–60], (24–36], (12–24], (18–24], (9–12], (36–42], (12–36]

Times since admittance for individuals who never had a fall: 23, 41, 13, 25, 59, 39, 22, 18, 49, 38.

Estimate the survival function of the time from admittance to first fall to within three decimal places.

5.4 Twenty women who had a lumpectomy as a primary treatment for breast cancer were followed periodically for the detection of a metastasis. When a metastasis was detected it was only known that the time of the clinical appearance of the metastasis was between the times of the last two visits to the physician, so the data is interval-censored. Suppose the data is as follows:

Times in months between which a metastasis could be detected: (12, 18], (20, 24], (10, 13], (14, 15], (25, 33], (33, 44], (18, 22], (19, 25], (13, 22], (11, 15].

Times last seen for patients disease free at the end of study:
25, 27, 33, 36, 30, 29, 35, 44, 44, 44.

Estimate the survival time for the distribution of the time from surgery to first clinical evidence of a metastasis.

5.5 A study was performed to estimate the distribution of incubation times of individuals known to have a sexually transmitted disease (STD). Twenty-five patients with a confirmed diagnosis of STD at a clinic were identified on June 1, 1996. All subjects had been sexually active with a partner who also had a confirmed diagnosis of a STD at some point after January 1, 1993 (hence $\tau = 42$ months). For each subject the date of the first encounter was recorded as well as the time in months from that first encounter to the clinical confirmation of the STD diagnosis. Based on this right-truncated sample, compute an estimate of the probability that

Date of First Encounter	*Months From 1/93 to Encounter*	*Time (in months) until STD Diagnosed in Clinic*
2/93	2	30
4/93	4	27
7/93	7	25
2/94	14	19
8/94	20	18
6/94	18	17
8/93	8	16
1/94	13	16
5/94	17	15
2/95	26	15
8/94	20	15
3/94	15	13
11/94	23	13
5/93	5	12
4/94	16	11
3/94	15	9
11/93	11	8
6/93	6	8
9/95	33	8
4/93	4	7
8/93	8	6
11/95	35	6
10/93	10	6
12/95	36	4
1/95	25	4

the infection period is less than x months conditional on the infection period's being less than 42 months.

Estimate the distribution of infection-free time (survival).

5.6 Using the data on 258 adults with AIDS reported in section 1.19, estimate the probability that the waiting time to AIDS is less than x, given the waiting time is less than eight years.

5.7 The following data is based on a cohort of 1,571 men in the Framingham Heart Study who were disease free at age 40 and followed for a period of 40 years. (See Klein, Keiding, and Kreiner (1995) for a detailed description of the cohort.) Of interest is the distribution of the time to development or coronary heart disease (CHD). The following life table data is available to estimate this distribution.

Age Interval	*Number of CHD Events*	*Number Lost to Follow-Up*
45–50	17	29
50–55	36	60
55–60	62	83
60–65	76	441
65–70	50	439
70–75	9	262
75–80	0	7

Construct a cohort life table for this data.

5.8 Individuals seen at a large city sexually transmitted disease (STD) clinic are considered at high risk for acquiring HIV. The following data is recorded on 100 high-risk individuals who are infected with some STD, but did not have HIV, when they were seen at the clinic in 1980. Their records were checked at subsequent visits to determine the time that HIV was first detected.

Year Intervals	*Number of HIV-Positive*	*Number Lost to Follow-Up*
0–2	2	3
2–4	1	2
4–6	4	8
6–8	3	10
8–10	2	18
10–12	2	21
12–14	3	21

Construct a cohort life table for the incidence of HIV.

5.9 An investigator, performing an animal study on mammary carcinogenesis risk, wants to describe the distribution of times (in days) until the

onset of the first palpable tumor for rats fed a control diet. Mammary tumors were induced by a single oral dose (5 mg dissolved in 1.0 ml. com oil) of 7,12-dimethylbenz(a)anthracene (DMBA) administered by intragastric intubation when the animals were seven weeks old. Starting six weeks after DMBA administration, each rat was examined once daily and the time (in days) until the onset of the first palpable tumor was recorded. Three rats had a palpable tumor when the first examination was made at day 62. The remaining times when the first palpable tumor was detected are below.

Times (in days) when the first palpable tumor was detected:
46, 49, 54, 61, 62, 64, 68, 120, 150, 160.

Estimate the survival time for the distribution of the time from DBMA administration until the first palpable evidence of a tumor occurred.

5.10 Wagner and Altmann (1973) report data from a study conducted in the Amboseli Reserve in Kenya on the time of the day at which members of a baboon troop descend from the trees in which they sleep. The time is defined as the time at which half of the troop has descended and begun that day's foraging. On some days the observers arrived at the site early enough to observe at what time this event occurred, whereas on other days they arrived after this median descent time, so that day's observation was left censored at their arrival time. That data is in the following tables. By reversing the time scale to be the number of minutes from midnight (2400 hours), estimate the distribution of the time to descent for a randomly selected troop of baboons.

Observed Time of Day When Half of the Troop Descended from the Trees

Day	Date	Descent Time	Day	Descent Date	Time	Descent Day	Date	Time
1	25/11/63	0656	20	12/7/64	0827	39	10/6/64	0859
2	29/10/63	0659	21	30/6/64	0828	40	11/3/64	0900
3	5/11/63	0720	22	5/5/64	0831	41	23/7/64	0904
4	12/2/64	0721	23	12/5/64	0832	42	27/2/64	0905
5	29/3/64	0743	24	25/4/64	0832	43	31/3/64	0905
6	14/2/64	0747	25	26/3/64	0833	44	10/4/64	0907
7	18/2/64	0750	26	18/3/64	0836	45	22/4/64	0908
8	1 /4/64	0751	27	15/3/64	0840	46	7/3/64	0910
9	8/2/64	0754	28	6/3/64	0842	47	29/2/64	0910
10	26/5/64	0758	29	11/5/64	0844	48	13/5/64	0915
11	19/2/64	0805	30	5/6/64	0844	49	20/4/64	0920
12	7/6/64	0808	31	17/7/64	0845	50	27/4/64	0930
13	22/6/64	0810	32	12/6/64	0846	51	28/4/64	0930
14	24/5/64	0811	33	28/2/64	0848	52	23/4/64	0932
15	21/2/64	0815	34	14/5/64	0850	53	4/3/64	0935
16	13/2/64	0815	35	7/7/64	0855	54	6/5/64	0935
17	11/6/64	0820	36	6/7/64	0858	55	26/6/64	0945
18	21/6/64	0820	37	2/7/64	0858	56	25/3/64	0948
19	13/3/64	0825	38	17/3/64	0859	57	8/7/64	0952
						58	21/4/64	1027

Observer Arrival Time on Days Where the Descent Time Was Not Observed

Day	Date	Arrival Time	Day	Date	Arrival Time	Day	Date	Arrival Time
1	1/12/63	0705	32	13/10/63	0840	63	2/5/64	1012
2	6/11/63	0710	33	4/7/64	0845	64	1/3/64	1018
3	24/10/63	0715	34	3/5/64	0850	65	17/10/63	1020
4	26/11/63	0720	35	25/5/64	0851	66	23/10/63	1020
5	18/10/63	0720	36	24/11/63	0853	67	25/7/64	1020
6	7/5/64	0730	37	15/7/64	0855	68	13/7/64	1031
7	7/11/63	0740	38	16/2/64	0856	69	8/6/64	1050
8	23/11/63	0750	39	10/3/64	0857	70	9/3/64	1050
9	28/11/63	0750	40	28/7/64	0858	71	26/4/64	1100
10	27/11/63	0753	41	18/6/64	0858	72	14/10/63	1205
11	28/5/64	0755	42	20/2/64	0858	73	18/11/63	1245
12	5/7/64	0757	43	2/8/64	0859	74	2/3/64	1250
13	28/3/64	0800	44	27/5/64	0900	75	8/5/64	1405
14	23/3/64	0805	45	28/10/64	0905	76	1/7/64	1407
15	26/10/63	0805	46	15/5/64	0907	77	12/10/63	1500
16	11/7/64	0805	47	10/5/64	0908	78	31/7/64	1531
17	27/7/64	0807	48	27/6/64	0915	79	6/10/63	1535
18	9/6/64	0810	49	11/10/63	0915	80	19/6/64	1556
19	24/6/64	0812	50	17/2/64	0920	81	29/6/64	1603
20	16/ 10/63	0812	51	22/10/63	0920	82	9/5/64	1605
21	25/2/64	0813	52	10/7/64	0925	83	9/10/63	1625
22	6/6/64	0814	53	14/7/64	0926	84	8/3/64	1625
23	22/11/63	0815	54	11/4/64	0931	85	11/2/64	1653
24	10/10/63	0815	55	23/5/64	0933	86	30/5/64	1705
25	2/11/63	0815	56	30/7/64	0943	87	5/3/64	1708
26	23/6/64	0817	57	18/7/64	0945	88	26/2/64	1722
27	24/4/64	0823	58	29/7/64	0946	89	4/5/64	1728
28	3/7/64	0830	59	16/7/64	0950	90	12/3/64	1730
29	29/4/64	0831	60	22/7/64	0955	91	25/10/63	1730
30	4/8/63	0838	61	15/10/63	0955	92	29/11/63	1750
31	7/10/63	0840	62	19/10/63	1005	93	22/2/64	1801
						94	22/3/64	1829

6 Topics in Univariate Estimation

6.1 Introduction

In Chapter 4, we presented two techniques for providing summary curves which tell us about the survival experience of a cohort of individuals. These two estimators were the Kaplan–Meier estimator, which provides an estimate of the survival function, and the Nelson–Aalan estimator, which provides an estimate of the cumulative hazard rate. These statistics are readily available in many statistical packages.

Although these two statistics provide an investigator with important information about the eventual death time of an individual, they provide only limited information about the mechanism of the process under study, as summarized by the hazard rate. The slope of the Nelson–Aalan estimator provides a crude estimate of the hazard rate, but this estimate is often hard to interpret. In section 6.2, we discuss how these crude estimates of the hazard rate can be smoothed to provide a better estimator of the hazard rate by using a kernel-smoothing technique.

In some applications of survival analysis, an investigator has available very precise information about the mortality rates in a historical control or standard population. It is of interest to compare the hazard rates in the sample group to the known hazard rates in the reference population to determine how the mortality experience of the experimental subjects differs. The "excess" mortality in the experimental group can