2

Descriptive approaches of survival analysis

In this book, inferences of model parameters on survival processes are the main focus. Before proceeding with statistical procedures for such inferences, I would like to portray some descriptive approaches first. In survival analysis, such descriptive methods and techniques are used to summarize main features of raw survival data. Other than numbers, tables, graphs, and some other simple statistics, these approaches include significance tests on group differences in the survival and the hazard functions. Though generally viewed as simplistic ways for recapitulating survival data, the descriptive approaches are sometimes applied for deriving conclusions in biomedical research. In clinical trials, for example, results directly from descriptive approaches are commonly used to generate analytic results, particularly since in those studies the sample size is often too small to consider a large number of parameters and the effects on survival are partially accounted for in the process of randomization.

Many of the descriptive approaches in survival analysis are counting methods on the survival function, from which other lifetime indicators, such as the cumulative hazard function, can be easily computed. This chapter introduces some popular methods in this area used by statisticians, demographers, and other quantitative methodologists in summarizing survival data. Because they rely completely on empirical data without making assumptions on the form of the probability distribution, these approaches are also referred to as *nonparametric methods*. In particular, I start with the description of the Kaplan–Meier and the Nelson–Aalen estimators, the two related and well-known nonparametric methods used in analyzing the survival probability and the cumulative hazard function. Then I provide a brief introduction of the life table method, initially developed and used by demographers and epidemiologists. Next, I describe a variety of testing techniques for comparing two or more group-wise survival functions. Lastly, a summary of these descriptive approaches is provided.

Survival Analysis: Models and Applications, First Edition. Xian Liu.

2.1 The Kaplan–Meier (product-limit) and Nelson–Aalen estimators

The Kaplan–Meier estimator (1958), also known as the *product-limit method*, provides a simple but effective scheme that calculates a single lifetime indicator, the survival function $S(t)$. As time t is divided into a series of intervals according to observed event or censored times, the Kaplan–Meier survival estimates are calculated by the product of a series of interval-specific conditional probabilities. This method is designed in such a simple way that the censored survival function can be easily computed by hand. The Nelson–Aalen estimator (1972), as an alternative to the Kaplan–Meier approach, calculates the cumulative hazard function using the same rationale. The application of both estimators is based on the assumption that the occurrence of censoring is independent of actual survival times.

2.1.1 Kaplan–Meier estimating procedures with or without censoring

I start the description of the Kaplan–Meier estimator with a simple example in the absence of censoring. Suppose there are ten older women who have died of breast cancer during a seven-year observation period. Using 'month' as the time scale, their survival times are ordered by rank according to the number of months elapsed from time 0 to the occurrence of the event:

Survival times in months: 5, 17, 24, 32, 40, 46, 47, 50, 59, and 74.

Given this series of survival times, the survival rate at time 0 is 1 because all ten older women are alive at the beginning of the observation. The proportion surviving to month 5 is 0.9 as nine out of those ten older women survive beyond month 5. Likewise, by the end of month 17, two women have died in total, so the proportion surviving at this time is 0.8. At month 24 and month 32, the proportions of survival are, respectively, 0.7 and 0.6. The survival rates at the following observed survival times can also be computed easily from the ratio of the number of women still alive at each survival time over value 10. While the above calculation is straightforward, the proportion surviving beyond a given month can be calculated in a different perspective. For example, the survival rate at month 17 can be estimated by the proportion of survivors at month 5 (0.9) multiplied by the proportion surviving between month 5 and month 17. At month 24, the survival probability can be calculated by the survival rate at month 17 times the proportion surviving between month 17 and month 24. Placing the previous step into the calculation, the survival rate at month 24 can be further expressed as the product of three interval-specific conditional proportions of survival. Similarly, the survival rate at month 32 can be estimated as the product of four interval-specific conditional probabilities of survival, the survival rate at month 40 the product of five conditional surviving proportions, and so forth. The survival rate at each survival time thus computed is called the Kaplan–Meier estimate. The logic involved in this estimator is that for an older woman with breast cancer who survives beyond month 40, she must survive through months 5, 17, 24, and 32 first, so that her chance of survival throughout 40 months is composed of a series of interval-specific survival rates.

Below, I take the liberty to summarize the above procedure up to month 48 (four years) in the format of a typical Kaplan–Meier table. Table 2.1 shows the procedure that the survival rate at each month is simply the product of a series of interval-specific conditional probabilities of survival. In the absence of censoring, the survival rate can be more easily obtained

Table 2.1 Kaplan–Meier survival estimates for older women with breast cancer.

Month	Calculating steps	Survival rate
0	10/10 = 1.0	1.0
5	$1.0 \times (9/10) = 0.9$	0.9
17	$1.0 \times (9/10) \times (8/9) = 0.9$	0.8
24	$1.0 \times (9/10) \times (8/9) \times (7/8) = 0.7$	0.7
32	$1.0 \times (9/10) \times (8/9) \times (7/8) \times (6/7) = 0.6$	0.6
40	$1.0 \times (9/10) \times (8/9) \times (7/8) \times (6/7) \times (5/6) = 0.5$	0.5
46	$1.0 \times (9/10) \times (8/9) \times (7/8) \times (6/7) \times (5/6) \times (4/5) = 0.4$	0.4
47	$1.0 \times (9/10) \times (8/9) \times (7/8) \times (6/7) \times (5/6) \times (4/5) \times (3/4) = 0.3$	0.3

Table 2.2 Kaplan–Meier survival estimates with right censoring.

Month	Calculating steps	Survival rate
0	12/12 = 1.0	1.00
5	$1.0 \times (11/12) = 0.92$	0.92
17	$1.0 \times (11/12) \times (10/11) = 0.83$	0.83
20^+	$1.0 \times (11/12) \times (10/11) \times (9/9) = 0.83$	0.83
24	$1.0 \times (11/12) \times (10/11) \times (9/9) \times (8/9) = 0.74$	0.74
32	$1.0 \times (11/12) \times (10/11) \times (9/9) \times (8/9) \times (7/8) = 0.65$	0.65
35^+	$0.65 \times (6/6) = 0.65$	0.65
40	$0.65 \times (6/6) \times (5/6) = 0.54$	0.54
46	$0.65 \times (6/6) \times (5/6) \times (4/5) = 0.43$	0.43
47	$0.65 \times (6/6) \times (5/6) \times (4/5) \times (3/4) = 0.32$	0.32

from the number of survivors at each observed survival time over the total number of survivors at the beginning of observation. When censoring exists, however, this ratio function does not apply very well because the survival status for a censored case is unknown. Removal of censored cases from the calculation leads to erroneous results due to the neglect of survival times for those lost to observation in the middle of a survival interval. In such situations, the Kaplan–Meier estimator has the capability to account for some types of censored data, particularly right censoring. This property is a major appeal of the Kaplan–Meier estimator, given frequent occurrences of censoring in longitudinal surveys and clinical trials.

To demonstrate how the Kaplan–Meier estimator handles right censoring, I extend the previous example by adding two older women with breast cancer who entered the study at the beginning of the investigation but are then lost to observation, one at month 20 and one at month 35. Now, the total number of older women increases to 12, with their survival times given by

Survival times in months: 5, 17, 20^+, 24, 32, 35^+, 40, 46, 47, 50, 59, and 74,

where the two censored cases are designated by the sign +. Given these survival and censored times, a revised Kaplan–Meier table is displayed in Table 2.2, which presents the proportions

surviving at ten survival times after adding two censored cases. At time 5, the survival rate now increases to 0.92 (11/12) because the two censored patients are known to be alive at month 5 and are thus counted as survivors. Likewise, $S(17)$ is elevated to 0.83 after taking additional survivors into consideration. At month 20, one survivor is lost to observation, but from month 17 to month 20, none of the other survivors is deceased so $S(20)$ is still 0.83. At this time point, nine patients remain exposed to the risk of death. At month 24, another patient is deceased; therefore, with nine survivors at month 20, the proportion of survival at month 24 is computed as $(11/12) \times (10/11) \times (9/9) \times (8/9) = 0.74$. Similarly, the proportion surviving to month 32 is $(11/12) \times (10/11) \times (9/9) \times (8/9) \times (7/8) = 0.65$. At month 35, another patient is lost to observation and hence there are five patients left exposed to the risk of death and the survival rate at this time is still 0.65. Given the number of survivors at month 35, the proportions surviving to months 40, 46, and 47 are, respectively, 0.54, 0.43, and 0.32, computed by following the Kaplan–Meier estimating procedure. It is worth noting that at each observed survival time the value of the survival rate is higher than the corresponding rate without including censored cases because the survival times for the two censored patients are considered in the counting procedure.

Using the survival estimates computed in Table 2.2, I generate a plot to highlight that the Kaplan–Meier survival function follows a declining step process, using the following SAS program.

SAS Program 2.1:

```
data Kaplan_Meier;
   input Months Status@@;
datalines;

5  1 17 1 20 0 24 1 32 1 35 0 40 1 46 1 47 1 50 1
59 1 74 1
;

ods html;
ods graphics on;

proc lifetest data = Kaplan_Meier;
  time Months*Status(0);
run;

ods graphics off;
ods html close;
```

In SAS Program 2.1, I create a temporary dataset titled 'Kaplan_Meier,' containing only two variables: Months (survival time in months) and Status (0 = censored and 1 = not censored). Then the data of survival times and the censoring indicator STATUS are entered in order. The PROC LIFETEST procedure is used to plot the Kaplan–Meier survival function. For this simple analysis without involving other covariates, only the TIME statement is required in the PROC LIFETEST statement. For the graphic display, the ODS GRAPHICS statement is specified.

SAS Program 2.1 generates a plot of the Kaplan–Meier estimates, given in Figure 2.1, which displays the plot of the Kaplan–Meier survival estimates against the time scale,

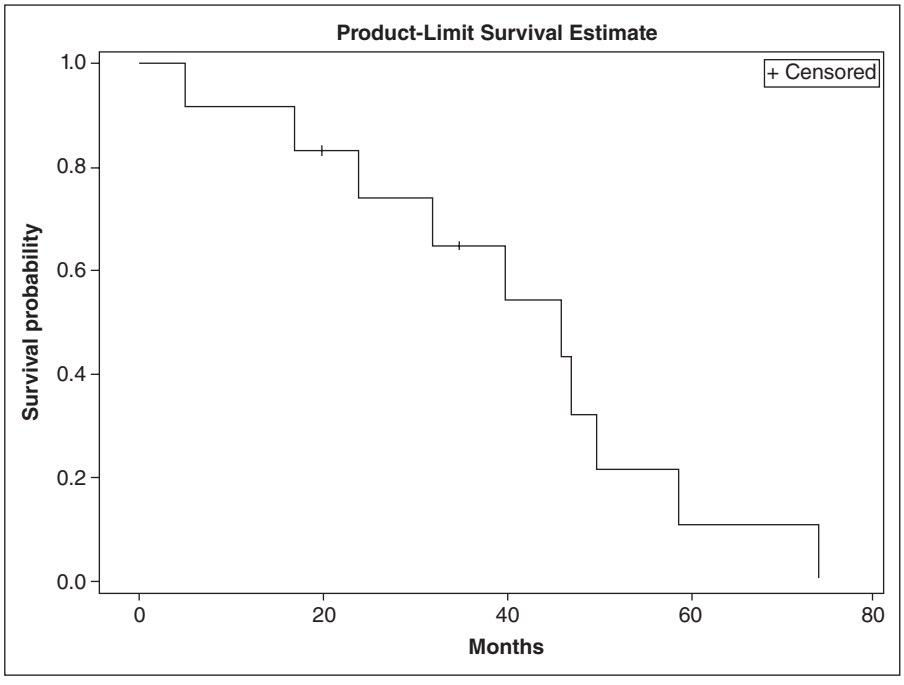

Figure 2.1 Plot of Kaplan–Meier estimates.

Months. Clearly, the Kaplan–Meier (product-limit) survival estimates follow a step function at survival times. When a death occurs, the survival curve drops vertically to a lower level of survival and then remains constant over time until another death occurs. In the plot, the symbol '+' indicates the horizontal location in months where right censoring occurs. Notice that in a standard Kaplan–Meier plot, the survival function is termed the 'survival probability,' rather than the 'survival rate' or the 'proportion surviving,' as pertaining to the large-sample approximation theory. For analytic convenience, the terms 'survival probability,' 'survival rate,' and 'proportion of survival' are interchangeable in the following text.

2.1.2 Formulation of the Kaplan–Meier and Nelson–Aalen estimators

The example given in Subsection 2.1.1 uses a very small sample of individuals, with each having a unique survival time, censored or not censored. If a large sample is used for calculating the Kaplan–Meier estimates, there may be too many lifetimes to be ranked, arranged, and computed. Additionally, in large-scale longitudinal surveys, some individuals may share the same survival times and some others may be lost to observation at exactly the same time points. In survival analysis, this shared survival time is referred to as the *tied observation time*, with observations sharing the same survival time, called the *tied cases*. In the presence of observation ties, the Kaplan–Meier estimator needs to be formulated to fit into various situations.

For a sample of n individuals, there are potentially n survival times until all experience a particular event (e.g., death). These survival times can be ordered by rank as

$$t_1 \le t_2 \le t_3 \le t_4 \le \cdots \le t_n,$$

where t_i represents the time at which individual i experiences a particular event or right censoring ($i = 1, 2, \ldots, n$). Because at time t_i individuals with actual survival or censored times smaller than t_i have already exited, there is a specific number of survivors who remain exposed to the risk of event at t_i, denoted by n_i, where $n_1 \ge n_2 \ge \cdots \ge n_n$. As t_n here is the lifetime for the last survivor in the rank list, $n_n = 1$. If there are no observation ties, the total number of survival times is equal to the number of observations and, accordingly, survival times are ordered by

$$t_1 < t_2 < t_3 < t_4 < \cdots < t_n.$$

With the existence of tied cases at some of the survival times, however, the total number of recorded times is smaller than n.

Let d_i be the number of events at time t_i ($d_i = 1$ if there are no tied cases at t_i); then the Kaplan–Meier estimator for the probability of survival at time t is

$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}, \tag{2.1}$$

where $\hat{S}(t)$ is the Kaplan–Meier estimate for the probability of survival at time t. As d_i can be any number including 1, Equation (2.1) takes into consideration the existence of tied observations. Although censoring is not particularly specified in this equation, its presence does not affect the validity of this formulation. If t_i is a censored survival time, for example, d_i is 0 from t_{i-1} to t_i and $n_i = n_{i-1} - 1$, so that the conditional probability of survival between t_{i-1} and t_i is 1. If t_i indicates a tied censored time with c_i being the number of censored observations tied at t_i, then $n_i = n_{i-1} - c_i$ and the conditional probability of survival between t_{i-1} and t_i is still 1.

Sometimes, specification of censored times is necessary in formalizing the survival function due to reasons of generalization. In such situations, a status indicator for a survival or a censored time, denoted by δ_{ti}, can be created where $\delta_{ti} = 0$ if t_i is a censored survival time and $\delta_{ti} = 1$ if t_i is an actual survival time. Let $\tilde{d}_i = d_i$ if $\delta_{ti} = 1$ and $\tilde{d}_i = c_i$ if $\delta_{ti} = 0$ (or $\tilde{d}_i = \min(d_i, c_i)$). Then the Kaplan–Meier estimator can be written by

$$\hat{S}(t) = \prod_{t_i < t} \left(\frac{n_i - \tilde{d}_i}{n_i} \right)^{\delta_{ti}}. \tag{2.2}$$

Clearly, when t_i is a censored survival time, $\delta_{ti} = 0$ and $\left[\left(n_i - \tilde{d}_i \right) / n_i \right] = 1$; hence the proportion surviving between t_{i-1} and t_i is 1. In contrast, when t_i is an actual survival time, $\delta_{ti} = 1$ and the term $\left[\left(n_i - \tilde{d}_i \right) / n_i \right]$ is smaller than 1, thus indicating an actual proportion of survival in the interval (t_{i-1}, t_i).

Given Equations (2.1) and (2.2), several scenarios are suggested to deal with various situations. First, if the number of events is small, the data can be arranged in the order of time without grouping and the number of censored cases in the intervening time intervals is also counted, as applied in Subsection 2.1.1. Second, if the above procedure is considered to be time consuming because of a relatively large sample size but the number of censored

survival times is relatively small, some of the successive intervals only containing censored times can be combined. Lastly, when the number of events is large, some selected division points need to be specified with events and censored cases counted in corresponding intervals. For such data, a summary table may be created to display the selected time intervals, the number of events, the number of censored cases, and the probability of survival at each of the selected intervals.

From the Kaplan–Meier estimator, other lifetime functions can be readily derived given the intimate associations among various indicators described in Chapter 1. A popular application, for example, is to convert the survival function to the cumulative hazard function. Given the equation $H(t) = -\log S(t)$, the cumulative hazard function can be expressed in terms of the Kaplan–Meier survival estimate $\hat{S}(t_i)$, given by

$$\hat{H}(t) = -\log\left[\prod_{t_i<t}\left(\frac{n_i - \tilde{d}_i}{n_i}\right)^{\delta_{ti}}\right]. \tag{2.3}$$

For small x, $\log(1 + x) \approx x$. Therefore, Equation (2.3) can be further expanded:

$$\hat{H}(t) = -\sum_{t_i\le t}\delta_{t_i}\log\left(1-\frac{\tilde{d}_i}{n_i}\right) \approx -\sum_{t_i\le t}\delta_{t_i}\left(-\frac{\tilde{d}_i}{n_i}\right) = \sum_{t_i\le t}\left(\frac{\tilde{d}_i\delta_{t_i}}{n_i}\right). \tag{2.4}$$

If censored times are not particularly specified in the formulation of the cumulative hazard function, Equation (2.4) becomes

$$\hat{H}(t) \approx \sum_{t_i\le t}\left(\frac{d_i}{n_i}\right) = \sum_{t_i\le t}[1-\hat{s}(t_i)]. \tag{2.5}$$

Equation (2.5) is the celebrated *Nelson–Aalen estimator*, initially proposed by Nelson (1972) and later mathematically formalized and justified by Aalen (1978). Clearly, the Nelson–Aalen estimator, developed independently, is an approximation to the Kaplan–Meier estimate transformation on $H(t)$. As mentioned in Chapter 1, at a single survival time t_i, the ratio of the number of events over the number of those exposed to the risk of an event is identical to the hazard rate at t_i, and between two survival times in a step function, the hazard rates all take value 0. Consequently, the Nelson–Aalen estimator has a solid theoretical base (Aalen, 1978; Andersen *et al*., 1993; Fleming and Harrington, 1991). In Chapter 6, I will provide a simple proof for its validity using the martingale central limit theorem.

The cumulative hazard function is frequently used to test the parametric form of the hazard function, given its intimate associations with some of the parametric distributions, as will be described in some of the later chapters. While both approaches are widely used, the Nelson–Aalen estimator on the cumulative hazard function is generally considered to behave better than the Kaplan–Meier estimator for data of small samples.

Given the mathematical association among lifetime indicators, the survival function can be expressed in terms of the Nelson–Aalen estimator, given by

$$\hat{S}(t) = \exp\left[-\hat{H}(t)\right] \approx \exp\left(-\sum_{t_i\le t}\frac{d_i}{n_i}\right), \tag{2.6}$$

or

$$\hat{S}(t) \approx \exp\left[-\sum_{t_i \le t}\left(\frac{\tilde{d}_i \delta_{ti}}{n_i}\right)\right]. \tag{2.7}$$

In the SAS software package, the Nelson–Aalen estimates of both survival probabilities and the cumulative hazard function can be made. Using the prior example of Table 2.2, here I reformat SAS Program 2.1 by adding the option 'NELSON' (or 'AALEN') in the PROC LIFETEST statement, requesting SAS to apply the Nelson–Aalen estimator, instead of the Kaplan–Meier method, for computing the probability of survival and the cumulative hazard rate at each survival time. The revised SAS program creates a temporary SAS data file named Nelson_Aalen, as displayed below.

SAS Program 2.2:

```
.....

data Nelson_Aalen;
  input Months Status@@;
datalines;

5  1 17 1 20 0 24 1 32 1 35 0 40 1 46 1 47 1 50 1
59 1 74 1
;

ods html;
ods graphics on;

proc lifetest data = Nelson_Aalen NELSON;
  time Months*Status(0);
run;

ods graphics off;
ods html close;
```

SAS Program 2.2 generates a plot of the Nelson–Aalen survival estimates that is almost identical to Figure 2.1. From the output data derived from SAS Program 2.2, not presented here, the Nelson–Aalen estimates of survival probabilities are analogous to those reported in Table 2.2.

2.1.3 Variance and standard error of the survival function

When a population sample is large enough, the Kaplan–Meier estimator approximates the mean of the survival probability, asymptotically normally distributed. Given this property, the variance of this survival estimate, from the theoretical standpoint, can be well specified for assessing the dispersion of the survival probability. Nevertheless, estimation of the variance for $\hat{S}(t)$, denoted by $\hat{V}\left[\hat{S}(t)\right]$, cannot easily be formulated without further inference. The difficulty here is the ambiguity resulting from the fact that the variance estimated from a sample does not depend on all limits of an observation. In survival data, the greatest observed lifetime, denoted t^*, is often a censored time, so that $\hat{S}\left(t^*\right) > 0$ and $n_{t^*+1} = 0$.

Therefore, $\hat{S}(t)$ is undefined for $t > t^*$. When the probability of $t > t^*$ for a population is sizable, a nonparametric estimate of the variance for $S(t)$ is not highly informative. In such situations, some approximation approaches need to be applied.

A number of techniques have been developed to derive an unbiased estimator for the variance of the Kaplan–Meier survival function (Kalbfleisch and Prentice, 2002; Peto *et al*., 1977). Originally, Kaplan and Meier (1958) used the well-known Greenwood (1926) formula to yield an estimate of the variance for $\hat{S}(t)$, and this formula remains a popular estimator in the analysis of the nonparametric survival function. Researchers prefer to use the *delta method* for the derivation of this approximation, as presented in most textbooks on survival analysis. In brief, the delta method is an approximation method involving transformations of random variables. Specifically, let X be a random variable distributed as $N(\mu, \sigma^2)$ with p.d.f. $f(x)$ and $g(X)$ be a single-valued and measurable function of X. If $g(X)$ is differentiable, its integral is the expected value of $g(X)$. Although the variance of $g(X)$ is often not directly obtainable, a linear approximation of $g(X)$ in the neighborhood of μ, through some expansion series, leads to the approximation that $V[g(X)] \approx [g'(\mu)]^2\sigma^2$. Appendix A provides a detailed description of this approximation method.

The derivation of the Greenwood formula on the Kaplan–Meier survival function takes several transformation steps. It starts with the estimation for the variance of $\log \hat{S}(t)$. Taking the log values of both sides of Equation (2.1) yields

$$\log\left[\hat{S}(t)\right] = \sum_{t_i<t} \log\left(\frac{n_i - d_i}{n_i}\right) = \sum_{t_i<t} \log\left[\hat{s}(t_i)\right], \tag{2.8}$$

where $\hat{s}(t_i)$ is the conditional probability of survival in interval (t_{i-1}, t_i). As $\hat{s}(t_i)$ can be expressed as an estimate of a proportion, its variance is

$$\hat{V}\left[\hat{s}(t_i)\right] = \frac{\hat{s}(t_i)\left[1 - \hat{s}(t_i)\right]}{n_i}. \tag{2.9}$$

Using the delta method, the variance of $\log \hat{s}(t_i)$ can be approximated from the variance of $\hat{s}(t_i)$. As the derivative of $\log(X)$ is $(1/X)$, the variance of $\log \hat{s}(t_i)$ is approximated by

$$\begin{aligned} \hat{V}\left[\log \hat{s}(t_i)\right] &\approx \left[\frac{1}{\hat{s}(t_i)}\right]^2 \left\{\frac{\hat{s}(t_i)\left[1 - \hat{s}(t_i)\right]}{n_i}\right\} \\ &\approx \frac{1 - \hat{s}(t_i)}{\hat{s}(t_i) n_i}. \end{aligned} \tag{2.10}$$

For analytic convenience, both numerator and denominator are multiplied by a common term n_i, which leads to

$$\hat{V}\left[\log \hat{s}(t_i)\right] \approx \frac{n_i\left[1 - \hat{s}(t_i)\right]}{\hat{s}(t_i)(n_i)^2}. \tag{2.11}$$

As $\hat{s}(t_i) \times n_i = (n_i - d_i)$, Equation (2.11) can be written by

$$\hat{V}\left[\log \hat{s}(t_i)\right] \approx \frac{d_i}{(n_i - d_i)n_i}. \tag{2.12}$$

Given Equation (2.12), the variance of $\log \hat{S}(t)$ can be obtained by summing up variances of all $\log \hat{s}(t_i)$ values, where $t_i \le t$, given by

$$\hat{V}\left[\log \hat{S}(t_i)\right] \approx \sum_{t_i \le t} \frac{d_i}{(n_i - d_i)n_i}. \tag{2.13}$$

Lastly, the variance of the survival probability $\hat{S}(t)$ can be approximated by performing a retransformation procedure using the delta method. Because the derivative of $\exp(X)$ is still $\exp(X)$, $\exp[\log(x)]$ is X and hence the final equation is

$$\hat{V}\left[\hat{S}(t_i)\right] \approx \left[\hat{S}(t)\right]^2 \sum_{t_i \le t} \frac{d_i}{(n_i - d_i)n_i}. \tag{2.14}$$

Equation (2.14) is the famous Greenwood formula, widely used in the descriptive analysis of survival data. The square root of this equation yields an estimate of the standard error for $\hat{S}(t)$. Given the Greenwood formula, once the point estimate of $S(t)$ is obtained, an approximate of its variance can be readily calculated using empirical data of d_i and n_i from a Kaplan–Meier table. Kalbfleisch and Prentice (2002) consider it valid to use a normal approximation of the distribution of $\hat{S}(t)$ with mean $S(t)$ and the variance estimate if censoring is not sizable and the sample size is large. Such validity holds whether the event time T is discrete or continuous or mixed with discrete and continuous components.

With the same rationale, the variance of the cumulative hazard rate at t can be estimated by $V\left[-\log \hat{S}(t)\right]$, using the same procedure that derives the Greenwood formula. After some algebra, the variance of the Nelson–Aalen estimator can be written by

$$\hat{V}\left[H(t)\right] \approx \sum_{t_i \le t} \frac{d_i}{n_i^2}. \tag{2.15}$$

In SAS, the PROC LIFETEST procedure not only calculates the Kaplan–Meier survival estimates but it also computes corresponding standard errors using the square root of the Greenwood formula. SAS Program 2.1, used to generate a plot demonstrating the Kaplan–Meier estimates, also derives the probability of survival and its standard error at each survival time, as summarized in Table 2.3. The interested reader might want to practice whether his or her hand calculation agrees with the estimates reported in the table.

2.1.4 Confidence intervals and confidence bands of the survival function

Conventionally, given the estimate of the variance and a significance level α, the confidence interval of an asymptotically normally distributed estimate can be readily computed. For example, for a continuous random variable X distributed as $N(\mu, \sigma^2)$, its 95 % confidence interval with $\alpha = 0.05$ is simply given by $\bar{X} \pm 1.96\left[\hat{V}(X)\right]^{1/2}$. A serious defect, however, arises from using this conventional procedure for estimating the variance of the survival

Table 2.3 Kaplan–Meier estimates and standard errors from SAS Program 3.1.

Time (months)	Probability of survival	Survival standard error
0	1.0000	0.0000
5	0.9167	0.0798
17	0.8333	0.1076
20+	0.8333	0.1076
24	0.7407	0.1295
32	0.6481	0.1426
35+	0.6481	0.1426
40	0.5401	0.1544
46	0.4321	0.1568
47	0.3241	0.1503
50	0.2160	0.1335
59	0.1080	0.1014
74	0.0000	—

function. Whereas a standard continuous variable has range $(-\infty, \infty)$, the probability of survival ranges between 0 and 1. Given this restriction, using the standard equation can yield the value of $\hat{S}(t)$ out of its bounds, thus yielding impossible estimates. Given this concern, the confidence interval of $\hat{S}(t)$ needs to be estimated by some transformation approaches, from which the range of $\hat{S}(t)$ can be restricted.

There are a number of popular transformation functions that can be applied to derive a confidence interval of $\hat{S}(t)$ with range (0, 1). These transformation approaches include the Wilson score method, arcsine–square root transformation, logit transformation, and log–log transformation. The Wilson score method, developed by Edwin B. Wilson (1927), improves the normal approximation interval by imputing a new asymptotic variance based on a new parameter instead of the proportion itself. For the 95 % confidence, the Wilson score method derives a nearly identical interval to that derived from the normal approximation. The arcsine–square root transformation, a widely used method to compute point-wise confidence limits for the survival function, takes the arcsine of the square root of a number with the range (−1, 1). This transformation derives the variance and a confidence interval stabilizing for situations with no censoring. The logit transformation, also widely used, yields the confidence interval for the logit of $\hat{S}(t)$ first and then this logit-based interval is retransformed to the confidence limits of $\hat{S}(t)$. The reader familiar with generalized linear modeling might remember that the probability linked to a logit function takes a value between 0 and 1.

In survival analysis, the most popular transformation function for the confidence interval of $\hat{S}(t)$ is perhaps the so-called *log–log transformation*, developed by Kalbfleisch and Prentice (2002). The logic of this transformation method is that the asymptotic normal distribution of $\hat{S}(t)$ should be first transformed to a continuous function with unrestricted bounds. Specifically, they propose first to estimate the variance of the following transformed function:

$$\hat{v}(t) = \log\left[-\log \hat{S}(t)\right]. \tag{2.16}$$

Equation (2.16) specifies a well-known transformation function in survival analysis, referred to as the *log–log survival function*. As discussed in Chapter 1, $-\log S(t)$ is simply the cumulative hazard rate at time t; therefore the log–log survival function is actually the log transformation of the cumulative hazard function with a range from minus infinity to plus infinity. Given this property, the log–log survival function is also called the *log function of the cumulative hazard*.

Applying the delta method with respect to the Greenwood formula, the variance of Equation (2.16) can be derived:

$$\hat{V}\left[\hat{v}(t)\right] \approx \frac{1}{\left[\log \hat{S}(t)\right]^2} \sum_{t_i \le t} \frac{d_i}{(n_i - d_i)n_i} \approx \frac{\hat{V}\left[\hat{S}(t)\right]}{\left[\hat{S}(t)\log \hat{S}(t)\right]}. \tag{2.17}$$

Given Equation (2.17), the confidence interval for the log–log survival function is given by

$$\log\left[-\log \hat{S}(t)\right] \pm z_{1-\alpha/2}\sqrt{\frac{\hat{V}\left[\hat{S}(t)\right]}{\left[\hat{S}(t)\log \hat{S}(t)\right]}}, \tag{2.18}$$

where $z_{1-\alpha/2}$ is the z-score for the upper $\alpha/2$ percentile of the standard normal distribution. As the log–log survival function is the log transformation of $H(t)$, the confidence interval for the survival function can be estimated by taking the exponential form of Equation (2.18) twice, which leads to

$$\left[\hat{S}(t)\right]^{1/\tilde{\theta}} \le \left[\hat{S}(t)\right] \le \left[\hat{S}(t)\right]^{\tilde{\theta}}, \tag{2.19}$$

where

$$\tilde{\theta} = \exp\left\{z_{1-\alpha/2}\sqrt{\frac{\hat{V}\left[\hat{S}(t)\right]}{\left[\hat{S}(t)\log \hat{S}(t)\right]}}\right\}. \tag{2.20}$$

Equations (2.19) and (2.20) yield a confidence interval of $\hat{S}(t)$ ranging between 0 and 1. This retransformed confidence interval is considered to provide the correct coverage probability for a $(1 - \alpha)$ interval, valid even for very small samples with heavy censoring (Borgan and Liestøl, 1990). Given the nonlinearity after retransformation, however, this confidence interval is not symmetric about $\hat{S}(t)$.

As limits of an independent Kaplan–Meier survival estimate, the confidence interval of $\hat{S}(t)$ *is not associated with the entire lifetime process; rather, it only covers the true value of* $S(t)$ with probability $(1 - \alpha)$ at a single time point. Therefore, the confidence interval of the survival rate is not considered highly informative in survival analysis. For a lifetime process, it is essential statistically to combine a series of confidence intervals for the survival function, thereby constituting a $(1 - \alpha)$ confidence region for the entire survival curve. In statistics, such a confidence region is referred to as the *confidence bands*. Though closely connected with each other, the confidence bands differ conceptually and mathematically from the point-wise confidence intervals. While a point-wise confidence interval only attaches to the survival estimate at an individual time point, the confidence bands cover the entire survival curve simultaneously, with each point-wise confidence band behaving as an

integral element of the whole confidence region. For this reason, confidence bands are also called *simultaneous confidence bands* or *simultaneous confidence intervals*. One of the important features in confidence bands is that if each confidence interval individually has probability $(1 - \alpha)$, the simultaneous coverage probability is generally less than $(1 - \alpha)$. Therefore, it is unacceptable to derive confidence bands by connecting the endpoints of all point-wise confidence intervals.

The derivation of confidence bands involves complex mathematical justifications and inferences. In particular, a point-wise confidence band, denoted by $\hat{S}(t) \pm \tilde{w}(t)$ with coverage probability $(1 - \alpha)$, satisfies the following condition for each value of t:

$$\Pr\left\{\left[\hat{S}(t) - \tilde{w}(t)\right] \le \hat{S}(t) \le \left[\hat{S}(t) + \tilde{w}(t)\right]\right\} = 1 - \alpha, \tag{2.21}$$

where $\tilde{w}$ is the confidence width determined by α. The lower and upper limits, $\left[\hat{S}(t) - \tilde{w}(t)\right]$ and $\left[\hat{S}(t) + \tilde{w}(t)\right]$, are sometimes denoted by $\breve{L}$ and $\breve{U}$, respectively.

There are two popular methods for calculating the confidence bands with respect to the Kaplan–Meier survival estimates. The first method is proposed by Hall and Wellner (1980), referred to as the *Hall–Wellner band*. The second approach, developed by Nair (1984), is called the *equal precision band*. For both methods, the confidence bands are constructed using the confidence coefficients taken from special distributions. These coefficients are provided in some textbooks and academic works (e.g., Klein and Moeschberger, 2003). Both bands can be computed by most statistical software packages including SAS (in PROC LIFETEST). Therefore, I do not include those coefficients tables in this book.

The mathematical inferences and the estimation procedures for both the Hall–Wellner and the equal precision bands are complex; for details of mathematical justifications and inferences, the interested reader is referred to Borgan and Liestøl (1990), Hall and Wellner (1980), Klein and Moeschberger (2003), and Nair (1984). Operationally, the calculation can be performed by taking the following four steps:

Step 1. Pick two time points, t_L and t_U. If the Hall–Wellner band is used, $t_L = 0$ and t_U is the event time just less than the largest observed event time; if the equal precision band is applied, t_L is just larger than the first observed event time and t_U is the same as above.

Step 2. Calculate the values of two confidence band coefficients, denoted $\acute{a}_L$ and $\acute{a}_U$, respectively, given by

$$\acute{a}_L = \frac{n\hat{V}\left[\hat{S}(t_L)\right]}{1 + n\hat{V}\left[\hat{S}(t_L)\right]} \tag{2.22}$$

and

$$\acute{a}_U = \frac{n\hat{V}\left[\hat{S}(t_U)\right]}{1 + n\hat{V}\left[\hat{S}(t_U)\right]}. \tag{2.23}$$

Step 3. Using values of $\acute{a}_L$ and $\acute{a}_U$, find the third coefficient, denoted either by $\acute{\kappa}_\alpha(\acute{a}_L, \acute{a}_U)$ for the Hall–Wellner band or by $\acute{c}_\alpha(\acute{a}_L, \acute{a}_U)$ for the equal precision band, from the tables for confidence coefficients of $100(1 - \alpha)$ confidence bands.

Step 4. There are several transformation forms for the confidence bands, as for confidence intervals. With respect to the popular log–log transformation, the confidence band can be expressed as

$$\left[\hat{S}(t)\right]^{1/\tilde{\theta}} \le \left[\hat{S}(t)\right] \le \left[\hat{S}(t)\right]^{\tilde{\theta}}, \tag{2.24}$$

where

$$\tilde{\theta} = \exp\left\{\frac{k_\alpha(\acute{a}_L, \acute{a}_U)\left[1+n\hat{V}\left(\hat{S}(t)\right)\right]}{\sqrt{n}\log\left[\hat{S}(t)\right]}\right\} \quad \text{for Hall}-\text{Wellner band,} \tag{2.25}$$

$$\tilde{\theta} = \exp\left\{\frac{c_\alpha(\acute{a}_L, \acute{a}_U)\sqrt{\hat{V}\left[\hat{S}(t)\right]}}{\log\left[\hat{S}(t)\right]}\right\} \quad \text{for equal precision band.} \tag{2.26}$$

In Equations (2.25) and (2.26), $\acute{\kappa}_\alpha$ or $\acute{c}_\alpha$ is the upper α fractile of the least observed upper bound. In practice, only the intervals for values of t greater than the first observed event time and smaller than the greatest observed event time need to be computed.

The two types of bands differ in several perspectives. The Hall–Wellner confidence bands are not proportional to the point-wise confidence intervals, as their derivation uses some ad hoc formulas. The equal precision method, on the other hand, yields proportional results to point-wise confidence intervals; specifically, the approach applies identical formulas to calculate confidence intervals and confidence bands, and the only difference between the two estimators is that the z-score is used to calculate confidence intervals and a different coefficient to derive confidence bands. Overall, the Hall–Wellner method provides better results than the equal precision when applying the log–log transformation, whereas the equal precision bands are more suitable when the arcsine–square root transformation is used.

To illustrate differences between confidence intervals and confidence bands for the Kaplan–Meier survival estimates, the data of older Americans described in Chapter 1 are used. In particular, I want to examine the two-year survival function among those diagnosed with cancer at baseline. In the AHEAD data, there are 39 survival times observed by the end of the 24th month, 19 actual events (those who died of cancer) and 20 censored cases. The following SAS program is constructed to estimate the survival function, confidence intervals, and confidence bands using lifetime data of those 39 individuals.

SAS Program 2.3:

```
options ls=80 ps=56 nodate number pageno=1 center;
ods select all;
ods trace off;
ods listing;
title1;
run;

data Kaplan_Meier;
  input Months Status @@;
datalines;
```

```
1 0 2 0 3 1 3 1 3 1 4 1 4 1 5 0 5 0 5 0 6 1 6 1 7 1 8 1
8 1 9 1 9 1 9 1 10 0 12 0 12 0 12 0 12 1 13 1 14 1 15 1
19 0 20 0 21 0 21 1 22 0 22 0 22 1 22 1 22 0 23 0 23 0
23 0 23 0
;

ods html;
ods graphics on;

ods select SurvivalPlot;
proc lifetest data=Kaplan_Meier OUTSURV=out1 Confband=HW Conftype=loglog
plots=survival(CB=HW);
  time Months * Status(0);
run;

proc print data=out1;
     title2 "OUTSURV data set";
     title3 "CONFBAND=all, CONFTYPE=Loglog";
run;

ods graphics off;
ods trace off;
```

In SAS Program 2.3, the input data of 39 survival times and their censoring statuses are saved in the SAS temporary dataset 'Kaplan_Meier.' The two variables, 'Months' and 'Status,' are defined previously in SAS Program 2.1. In the PROC LIFETEST statement, I ask SAS to create a temporary output dataset 'out1' containing the survival estimates and then to plot them with graphics using the ODS. To obtain the Hall–Wellner bands in the out = out1 dataset, I specify the CONFBAND = HW option (for a brief illustration, only the Hall–Wellner bands are used in this example; if both the Hall–Wellner and equal precision bands are needed, the CONFBAND = all option needs to specified). The CONFTYPE = LOGLOG option is specified to apply the log–log transformation for calculating the confidence intervals and confidence bands. Lastly, I request SAS to display a specific plot of the Hall–Wellner bands for the survival function. Other commands have been described previously.

SAS Program 2.3 derives Table 2.4, where the second column is the survival time and the third column displays whether a given survival time is an actual or a censored survival time, with 1 = censored. While the fourth column presents the Kaplan–Meier survival probabilities, the fifth and sixth columns are the lower and upper limits of the confidence interval for each survival estimate given $\alpha = 0.05$. Similarly, the last two columns demonstrate the lower and upper limits of the confidence band derived from the Hall–Wellner method. As evidenced in this table, the confidence band of each survival probability has much wider confidence limits than the corresponding confidence interval. The latter, associated independently with a single time point, seriously underestimates the true variability of the survival function. Therefore, in survival analysis, especially in clinical and epidemiological settings, the confidence bands should be regularly used for demonstrating a true confidence range of the survival function.

SAS Program 2.3 also yields a plot of the Hall–Wellner confidence bands on the Kaplan–Meier estimates, shown in Figure 2.2. The plot displays the confidence region of the survival

Table 2.4 Kaplan–Meier estimates, confidence intervals, and confidence bands.

Obs	Months	OUTSURV data aet				HW_LCL	HW_UCL
		CONFBAND=all, CONFTYPE=Loglog					
		CENSOR	SURVIVAL	SDF_LCL	SDF_UCL		
1	0	—	1.00000	1.00000	1.00000	—	—
2	1	1	1.00000	—	—	—	—
3	2	1	1.00000	—	—	—	—
4	3	0	0.91892	0.76931	0.97311	0.24868	0.99488
5	4	0	0.86486	0.70534	0.94140	0.43793	0.97480
6	5	1	0.86486	—	—	—	—
7	5	1	0.86486	—	—	—	—
8	5	1	0.86486	—	—	—	—
9	6	0	0.80522	0.63366	0.90225	0.46431	0.94066
10	7	0	0.77540	0.59982	0.88108	0.45706	0.92067
11	8	0	0.71575	0.53506	0.83624	0.42621	0.87709
12	9	0	0.62628	0.44371	0.76377	0.36158	0.80633
13	10	1	0.62628	—	—	—	—
14	12	0	0.59497	0.41254	0.73749	0.33582	0.78107
15	12	1	0.59497	—	—	—	—
16	12	1	0.59497	—	—	—	—
17	12	1	0.59497	—	—	—	—
18	13	0	0.55778	0.37391	0.70720	0.30109	0.75282
19	14	0	0.52060	0.33699	0.67587	0.26638	0.72461
20	15	0	0.48341	0.30159	0.64353	0.23199	0.69654
21	19	1	0.48341	—	—	—	—
22	20	1	0.48341	—	—	—	—
23	21	0	0.43946	0.25877	0.60648	0.18760	0.66767
24	21	1	0.43946	—	—	—	—
25	22	0	0.34181	0.17709	0.52175	0.09299	0.61559
26	22	1	—	—	—	—	—
27	22	1	—	—	—	—	—
28	22	1	—	—	—	—	—
29	23	1	—	—	—	—	—
30	23	1	—	—	—	—	—
31	23	1	—	—	—	—	—
32	23	1	—	—	—	—	—

probabilities within a two-year observation period. Notice that the Hall–Wellner confidence bands exclude the initial observed survival times because, according to Borgan and Liestøl (1990), anomalous values of the lower confidence band are often detected at the start of the survival process when the log–log or arcsine transformations are used. Consequently, displaying the Hall–Wellner bands in those local regions is not informative.

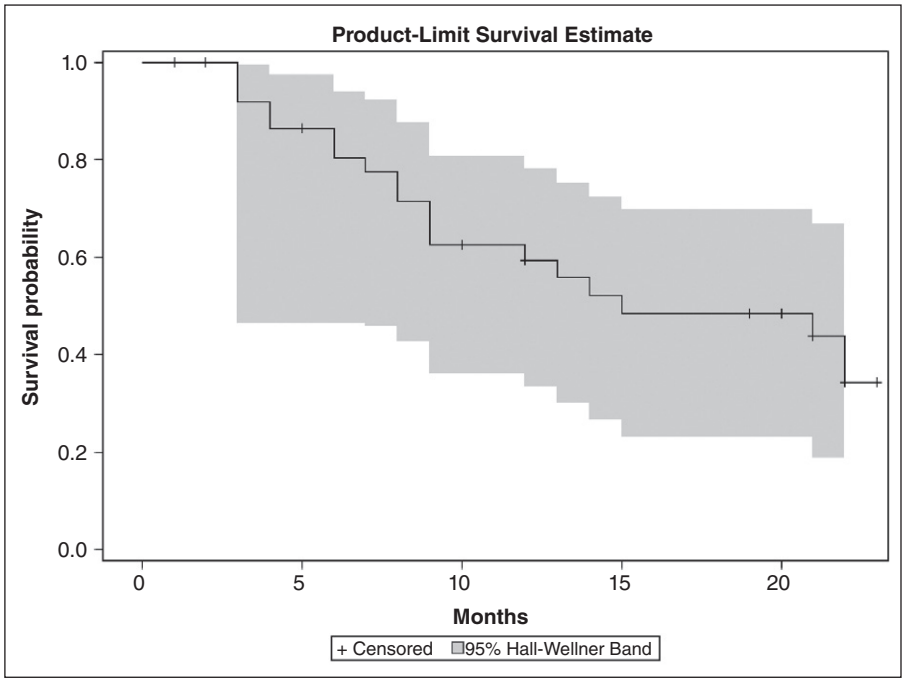

Figure 2.2 Hall–Wellner confidence bands for the probability of survival.

2.2 Life table methods

Other than the Kaplan–Meier and Nelson–Aalen estimators, a popular nonparametric approach in survival analysis is the life table method, which also has some capacity to handle right censoring. Demographers and epidemiologists have a long history of using life tables for analyzing survival data. A typical life table generates the probability of survival at a particular year of age, the life expectancy at birth, and the life expectancy remaining at an exact age, based on the hypothesis of a synthetic cohort. Separate life tables can be created for comparing mortality rates and life expectancies among individuals of different demographic and socioeconomic characteristics. Though originally designed to analyze mortality and survival, the life table method has been extended to calculate actuaries, disability-free life expectancy, geographical mobility, marital status patterns, occupational careers, and multidimensional transitions in health care (Crimmins, Hayward, and Saito, 1996; Hayward and Grady, 1990; Liu *et al*., 1995, 1997; Rogers, 1975; Rogers, Rogers, and Belanger, 1990; Schoen, 1988; Schoen and Land, 1979; Sullivan, 1971). In a survival analysis, Gehan (1969) provides specifications of the continuous hazard function and the probability density function from discrete estimates of a life table, thereby advancing the life table method to the analysis of event times and survival processes.

In this section, I first describe Gehan's basic formulations of several continuous lifetime functions based on the theory of large-sample approximations. Next, I briefly introduce some of the advanced life table techniques, such as the multistate life table and the

multidimensional Markov processes for events with frequent turnovers. Lastly, an empirical example is provided to display how to construct a life table using empirical survival data.

2.2.1 Life table indicators

Like the Kaplan–Meier estimator, a life table is created to calculate the probability of survival and its changes. In demography and epidemiology, however, the survival function in a life table is generally viewed as a function of age, rather than of time, largely due to the fact that survival data used in these fields often come from the age-specific death rates in a specific period. For analytic convenience, the probability of survival is usually multiplied by 100 000 in a conventional life table, yielding a new indicator, denoted by l_a, where the letter 'a' represents an exact age ($a = 0, 1, \ldots, \omega$), referred to as the *number of survivors on a radix*. As l_a is simply $S(a) \times 100\,000$, in a conventional life table $l_0 = 100\,000$ and $l_\omega = 0$. Using the vital registration statistics or survey data with respect to a specific period, a conventional life table is often constructed by assuming a synthetic birth cohort following current age-specific death rates (Keyfitz, 1985; Siegel and Swanson, 2004). Consequently, survival processes described in a conventional life table do not reflect actual trajectories of survival dynamics because they are derived from current mortality schedules across many population birth cohorts.

In survival analysis, the conventional life table method is extended by counting the actual number of survivors, of censored observations, and of events, from longitudinal data of event histories. Accordingly, observational time, rather than age, is used to define intervals to highlight the dynamic nature of a lifetime event. In many aspects, this modified life table method bears a tremendous resemblance to the Kaplan–Meier and Nelson–Aalen estimators, as will be seen below. To be in line with the focus of this book, therefore, many of the following specifications are based on Gehan (1969) with some minor notational modifications.

Let the survival data be grouped into $J + 1$ intervals, denoted by (t_{j-1}, t_j), where $j = 1, 2, \ldots, J + 1$, with the unit of interval j defined as $\tilde{b}_j = t_j - t_{j-1}$. By definition, $t_0 = 0$ and $t_{J+1} = \infty$, and accordingly $\tilde{b}_0 = 0$, $\tilde{b}_{J+1} = \infty$. Also, n_{j-1} is the number of individuals entering the interval (t_{j-1}, t_j) and d_j is the number of events occurring in interval j. In the absence of censoring, the difference between n_{j-1} and n_j yields the number of events in (t_{j-1}, t_j) and the ratio of the number of events over n_{j-1} generates the probability of experiencing a particular event in the interval.

In the presence of censoring, the *effective sample size* in (t_{j-1}, t_j), denoted by $\check{n}_j$, is defined and used for calculating the life table measures. Suppose c_j, the number of censored cases, falls in (t_{j-1}, t_j). The effective sample size at the start of the interval is conventionally given by

$$\check{n}_j = n_j - \frac{c_j}{2}, \tag{2.27}$$

where $(c_j/2)$ is the adjustment that only half of c_j, assumed to be evenly distributed in (t_{j-1}, t_j), should be counted in the total number of individuals exposed to the risk of the event. This adjustment is one of the key features in the life table method that counts survival times of censored cases in the estimation procedure. Accordingly, the conditional probability of experiencing the event, denoted by q_j, is estimated by

$$\hat{q}_j = \frac{d_j}{\check{n}_j} = \frac{d_j}{n_j - \frac{c_j}{2}}, \tag{2.28}$$

where the denominator represents the unbiased amount of exposure to the risk, given the assumption of a uniform distribution of censored observations.

Given the association that $\hat{s}_j = 1 - \hat{q}_j$, the conditional probability of survival in (t_{j-1}, t_j) is

$$\hat{s}_j = 1 - \frac{d_j}{n_j - \frac{c_j}{2}}. \tag{2.29}$$

This estimate of the conditional probability is another main feature of the life table method. In the Kaplan–Meier estimator, censored survival times are counted only across intervals, as shown in Table 2.3; in the life table method, nevertheless, censored times are taken into account both across and within intervals.

The variance for the conditional probability of experiencing the event in (t_{j-1}, t_j) can be estimated by the conventional approach, given by

$$V(\hat{q}_j) = \frac{\hat{q}_j \hat{s}_j}{\check{n}_j} = \frac{2\hat{q}_j \hat{s}_j}{2n_j - c_j}. \tag{2.30}$$

The square root of Equation (2.30) gives rise to an estimate of the standard error for $\hat{q}_j$. Notice that if n_j is small, Equation (2.30) can cause strong inconsistencies in the estimate, thereby affecting the quality of this estimator.

Given the conditional probability of the event, the survival function at the end of (t_{j-1}, t_j) is estimated as

$$\hat{S}(t_j) = \hat{S}(t_{j-1})\hat{s}_{j-1} = \prod_{j'=1}^{j} (1 - \hat{q}_{j'}). \tag{2.31}$$

As Equation (2.31) demonstrates, the life table estimator of the survival function is actually an approximation of the Kaplan–Meier method, expressed as the product of a series of interval-specific conditional probabilities of survival. When the sample size is large, the two approaches are asymptotically equivalent.

The estimation of the variance for the survival function also bears some resemblance to the estimator described in Section 2.1, given by

$$V\left[\hat{S}(t_j)\right] = \left[\hat{S}(t_j)\right]^2 \sum_{j'=1}^{j} \frac{\hat{q}_{j'}}{\hat{s}_{j'}\left(n_{j'} - \frac{c_{j'}}{2}\right)}. \tag{2.32}$$

The square root of Equation (2.32) yields an estimate for the standard error of the survival function at t_j. The reader might want to compare Equation (2.32) with the Greenwood formula (Equation (2.14)). In fact, in the absence of censoring, the variance estimate of the

survival function in the life table method approximates the estimate derived from Equation (2.14).

According to Gehan (1969), the probability density function at the midpoint of (t_{j-1}, t_j), denoted by t_{mj}, can be approximated straightforwardly:

$$\hat{f}(t_{mj}) = \frac{\left[\hat{S}(t_{j-1}) - \hat{S}(t_j)\right]}{\tilde{b}_j} = \frac{\hat{S}(t_{j-1})\hat{q}_j}{\tilde{b}_j}, \tag{2.33}$$

with the variance estimator

$$\hat{V}\left[\hat{f}(t_{mj})\right] \cong \left[\hat{f}(t_{mj})\right]^2 \sum_{j'=1}^{j} \left[\frac{\hat{q}_{j'}}{\left(n_{j'} - \frac{c_{j'}}{2}\right)\hat{s}_{j'}} + \frac{\hat{s}_{j'}}{\left(n_{j'} - \frac{c_{j'}}{2}\right)\hat{q}_{j'}} \right]. \tag{2.34}$$

Obviously, the above p.d.f. estimate at the midpoint of a time interval is computed as the total probability of experiencing a particular event for the entire interval divided by the interval width $\tilde{b}_j$. The validity of Equations (2.33) and (2.34) is based on the assumption that events occurring within a time interval are evenly or linearly distributed.

Likewise, given this hypothesis, the hazard function at t_{mj} can also be estimated in the same fashion:

$$\hat{h}(t_{mj}) = \frac{2\hat{q}_j}{\tilde{b}_j(1+\hat{s}_j)} = \frac{d_j}{\tilde{b}_j\left[n_j - \frac{1}{2}(d_j + c_j)\right]}. \tag{2.35}$$

Notice that in Equation (2.35), only half of d_j, the number of events occurring in (t_{j-1}, t_j), are counted in the number of individuals exposed to the risk. As mentioned in Chapter 1, if $\tilde{b}_j$ represents a considerably wide unit of a time interval, the continuous $S(t)$ is a decreasing function within the interval, so not all n_{j-1} individuals are at risk in the entire interval. If events within the interval occur uniformly or linearly, counting half of d_j in the denominator yields a reasonable estimate of the hazard rate at the midpoint of the interval. This estimate, however, is an average, not necessarily reflecting an instantaneous rate, especially when events occur irregularly or nonlinearly. As demographers term it, Equation (2.35) essentially provides an estimate of the average hazard function (Siegel and Swanson, 2004).

Gehan (1969) also specifies an estimate for the variance of the estimated hazard rate at t_{mj}. With some notational modifications, it can be written as

$$\hat{V}\left[\hat{h}(t_{mj})\right] \cong \left[\hat{h}(t_{mj})\right]^2 \frac{1 - \left[b_j\hat{h}(t_{mj})/2\right]^2}{\left(n_j - \frac{c_j}{2}\right)\hat{q}_j}. \tag{2.36}$$

The above measures can be used when constructing a life table with the time interval as the basic unit. I would like to emphasize that the above formulas are mostly based on the theory of large-sample approximations, so the life table method is applicable only when the sample size is large enough for every time interval. If the sample size for a given interval

is small (less than 20, say), the Kaplan–Meier or Nelson–Aalen estimators are preferable for an efficient nonparametric analysis of survival data. Given this restriction, the life table method is not recommended for a descriptive analysis of survival data obtained from clinical trials.

2.2.2 Multistate life tables

In survival analysis, an individual's event histories are sometimes linked to more than one single event process. At a given time, individuals may be exposed to the risks of several related events, thereby making survival processes attach to a set of competing risks. Examples of such phenomena include transitions in multiple modes of health (Crimmins, Hayward, and Saito, 1996; Land, Guralnik, and Blazer, 1994; Liu *et al*., 1995), labor force participation (Hayward and Grady, 1990), and multidimensional transitions in health care (Liang *et al*., 1996; Liu *et al*., 1997). A number of demographers and statisticians have developed a series of statistical models for the description and analysis of such multidimensional processes in a life table format, generally referred to as the *multistate life table*. These models are generally associated with one or more states of origin (the state at the beginning of observation) and more than one state of destination (the state at the end of observation), which, combined, constitute a finite space for a set of stochastic and multidimensional survival processes.

To describe a multistate life table more effectively, here I provide a flow chart about transitions in functional status to aid in the interpretation of the mathematical specifications given below. Suppose that at the beginning of a time interval individuals are divided into two groups according to function status, 'functional independence' and 'functional dependence.' As observed at the origin of time, this functional status is referred to as the *state of origin*, denoted by $\tilde{i} = 1, 2$. At the end of the observation, there are three possible outcomes in terms of that individual's functional status: 'functionally independent,' 'functionally dependent,' and 'dead,' referred to as the *state of destination* and denoted by $\tilde{j} = 1, 2, 3$, respectively. Between the two functional states, 'functional independence' and 'functional dependence,' a transition can occur from either direction within the time interval, and therefore they are called the *transient states*. The third status at destination, 'dead,' is a permanently ending state, and conventionally it is called the *absorbing state*. The multidimensional transitions between these states are displayed in Figure 2.3, where $P_{\tilde{i}\tilde{j}}$ indicates a transition process from the origin state $\tilde{i}$ to the destination state $\tilde{j}$. Given the assumption

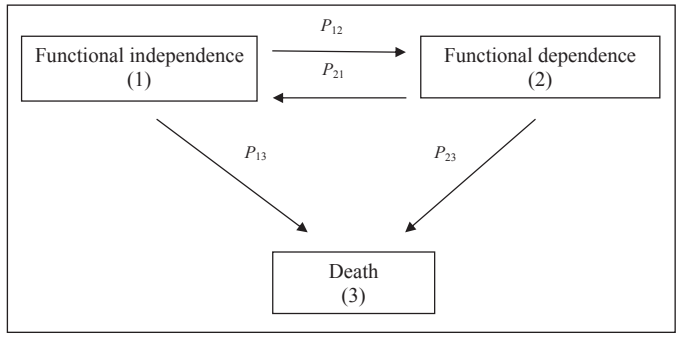

Figure 2.3 Three-state multistate life table model on transitions in functional status.

that only one transition is permitted within a specific time interval, four transition processes can be identified: from functional independence to functional dependence, from functional independence to death, from functional dependence to functional independence, and from functional dependence to death. As defined, individuals can move in and out of the transient states, as shown by the arrowed lines in Figure 2.3.

Like a conventional life table, the multistate life table estimates the probability of survival at a particular year of age, the life expectancy at birth, and the life expectancy remaining at an exact age. Each of these life table measures, however, needs to be calculated for each state of origin. Descriptively, the multistate life table is generally defined as a time-inhomogeneous and continuous-time Markov process model with finite space Ω. The state space Ω of the stochastic process has $\mathcal{K}+1$ states, where $\mathcal{K}$ is a positive integer greater than 1 (in terms of the example about transitions in functional status, $\mathcal{K}$ is 2). The $(\mathcal{K}+1)$ th state is the absorbing state. As indicated by Figure 2.3, two-way transitions are allowed between the transient states; that is, while a functionally independent person at the beginning can become functionally dependent within the time interval, an individual is permitted to recover from functional dependence within the period.

On the state space Ω, I define a stochastic process, $[\ddot{y}(t): t \geq 0]$, as seen by transition probabilities from the state of origin to the state of destination. The transition probabilities between the $\mathcal{K}+1$ states of $\ddot{y}$, assumed to be continuous, are given by

$$\tilde{\pi}_{\tilde{i}\tilde{j}}[t, \Delta t) = \Pr\left[\ddot{y}(t+\Delta t) = \tilde{j} \middle| \ddot{y}(t) = \tilde{i}\right], \tag{2.37}$$

where $\tilde{\pi}_{\tilde{i}\tilde{j}}[t, \Delta t)$ represents the probability that an individual in state $\tilde{i}$ at time t will be in state $\tilde{j}$ at time $(t + \Delta t)$. The corresponding transition force, statistically referred to as the gross flow hazard rate, is

$$h_{\tilde{i}\tilde{j}}(t) = \lim_{\Delta t \to 0} \frac{\tilde{\pi}_{\tilde{i}\tilde{j}}[t, \Delta t)}{\Delta t}, \quad \text{for } \tilde{i} \neq \tilde{j}, \tag{2.38}$$

where $h_{\tilde{i}\tilde{j}}(t)$ is the force of decrement for a transition from state $\tilde{i}$ to state $\tilde{j}$ at time t. As defined in Chapter 1, it is nonnegative but not necessarily smaller than 1.

If persistence in state $\tilde{i}$ from time t to time $(t + \Delta t)$ can be viewed as a special type of transition, its transition probability can be written by

$$\tilde{\pi}_{\tilde{i}\tilde{i}}[t, \Delta t) = 1 - \sum_{\tilde{j}=1}^{\mathcal{K}+1} \tilde{\pi}_{\tilde{i}\tilde{j}}[t, \Delta t), \quad \text{for } \tilde{j} \neq \tilde{i}. \tag{2.39}$$

The derivation of Equation (2.39) is based on the constraint that a set of transition probabilities must sum up to 1.

Likewise $h_{\tilde{i}\tilde{i}}(t)$, the force of persistence in $\tilde{i}$ at time t, is

$$h_{\tilde{i}\tilde{i}}(t) = -\lim_{\Delta t \to 0} \left[\frac{1 - \sum_{\tilde{j}=1}^{\mathcal{K}+1} \tilde{\pi}_{\tilde{i}\tilde{j}}[t, \Delta t)}{\Delta t} \right] = -\sum_{\tilde{j}=1}^{\mathcal{K}+1} h_{\tilde{i}\tilde{j}}(t), \quad \text{for } \tilde{j} \neq \tilde{i}, \tag{2.40}$$

where $h_{\tilde{i}\tilde{i}}(t)$, as a counterforce, is always nonpositive, referred to as the *force of retention* (Schoen, 1988). The transition probabilities and the forces of transition can be conveniently arranged into two $(\mathcal{K}+1)$ by $(\mathcal{K}+1)$ stochastic matrices, defined as $\Pi(t, \Delta t)$ and $\boldsymbol{h}(t)$, respectively. By definition, each row of the Π matrix sums to 1 and each row of the $\boldsymbol{h}$ matrix sums to 0.

Given $S(0) = 1$, the initial distribution of survival probability in state $\tilde{i}$, denoted by $S_{\tilde{i}}(0)$, is mathematically defined by

$$S_{\tilde{i}}(0) = \Pr\left[\ddot{y}(0) = \tilde{i}\right], \quad \text{for } \tilde{i} \in \boldsymbol{\Omega}, \tag{2.41}$$

with the range (0, 1). Within the context of transitions in functional status, for example, $S_{\tilde{i}}(0)$ indicates the probability distribution of individuals in the two states of origin at time 0 given that $S_1(0) + S_2(0) = 1$. For younger populations, it is likely that no one is functionally dependent, so $S_1(0) = 1$ and $S_2(0) = 0$. Similarly, for individuals beyond a certain old age, everyone may be functionally disabled, so we have $S_1(0) = 0$ and $S_2(0) = 1$. When constructing a multistate life table, this distribution can be obtained from empirical data.

Given $S_i(0)$, the survival function at state $\tilde{i}$ can be defined as follows

$$S_{\tilde{i}}(t) = \Pr\left[\ddot{y}(t) = \tilde{i}\right] = \sum_{\tilde{k} \in \Omega} S_{\tilde{k}}(0)\tilde{\pi}_{\tilde{k}\tilde{i}}(0, t), \tag{2.42}$$

Demographers call $S_{\tilde{i}}(0)$ the radix of a multistate life table, as the sequence of $S_{\tilde{i}}(t)$ is the survival function corresponding to the Markov chain. As usually applied, $\mathrm{S}_{\tilde{i}}(t)$ can be multiplied by a value, such as 10^6, for analytic convenience (Hoem and Jensen, 1982), thereby generating a new life table indicator, termed $l_{\tilde{i}}(t)$. In the literature of multistate life table modeling, $l_{\tilde{i}}(t)$ is referred to as the stationary population corresponding to the Markov process.

The gross flows of the stationary population are specified as the function

$$l_{\tilde{i}\tilde{j}}[t, \Delta t) = l_{\tilde{i}}(t)\tilde{\pi}_{\tilde{i}\tilde{j}}[t, \Delta t), \tag{2.43}$$

where $l_{\tilde{i}\tilde{j}}[t, \Delta t)$ represents the number of individuals in state $\tilde{i}$ at time t who are in state $\tilde{j}$ at time $(t + \Delta t)$ with respect to the stationary population. Accordingly, a $(\mathcal{K}+1)$ by $(\mathcal{K}+1)$ matrix $\boldsymbol{I}(t, \Delta t)$ can be created containing elements $l_{\tilde{i}\tilde{j}}(t, \Delta t)$, referred to as the *matrix of gross flows*.

Within the context of event times, the expected life in state $\tilde{j}$ between time t_{j-1} and time t_j spent by those in state $\tilde{i}$ at time t_{j-1} can be written as

$$\tilde{L}_{\tilde{i}\tilde{j}}(t_{j-1}, t_j) = \int_0^1 l_{\tilde{i}\tilde{j}}(t_{j-1}, u)\,\mathrm{d}u, \tag{2.44}$$

where the unit $(t_j - t_{j-1})$ represents a discrete interval with width $\tilde{b}_j$, often set at 1 year in demographic and epidemiologic studies. $\tilde{L}_{\tilde{i}\tilde{j}}(\cdot)$ is also called the sojourn time, representing the total person-years lived. These person-years lived at the level of gross flows can be aggregated to the level of net flows, defined by

$$\tilde{L}_{\tilde{i}}[t_{j-1}, t_j) = \sum_{\tilde{j}=1}^{\tilde{k}} \tilde{L}_{\tilde{j}\tilde{i}}[t_{j-1}, t_j) = \int_0^1 l_{\tilde{i}}[t_{j-1} + u)\,\mathrm{d}u, \tag{2.45}$$

where $\tilde{L}_{\tilde{i}}(t_{j-1}, t_j)$ is the person-years lived in state $\tilde{i}$ between t_{j-1} and t_j without the constraint of being in state $\tilde{i}$ at time t_{j-1}.

The above equations summarize basic specifications of a traditional multistate life table. Researchers have developed a variety of estimating algorithms to formalize these functions (Land and Schoen, 1982; Liu *et al*., 1995; Namboodiri and Suchindran, 1987; Rogers, 1975; Schoen, 1988; Schoen and Land, 1979). Because these accounting procedures are based on varying assumptions on patterns of transitions within a discrete interval (usually one year), there are distinct differences in the results derived from those approaches as well as a general distinction between the underlying stochastic processes of a given event and accounting procedures (Hoem and Jensen, 1982). While some traditional procedures only permit a single transition within a one-year period, several refined methods relax the single-transition assumption, thereby taking into account the return of those who have left a given state at an earlier stage (Schoen, 1988).

These methods, however, may still lead to substantial bias for lifetime events with rapid turnovers because those who have returned to the state of origin may leave there again soon. The pattern of health care use, for example, typifies such rapid processes, given the frequent and intense turnovers of hospitalization and institutionalization over time. Other dynamic processes that occur rapidly over time include adolescent dating behavior, the employment experiences of marginal workers, mental disorders like depression, and the like. Indeed, the traditional accounting procedures are not capable of handling these frequent events, resulting in a characterization of life-cycle experiences at variance with the true set of the stochastic processes.

Theoretically, difficulties in estimating more intense and rapidly unfolding processes can be resolved by using shorter time intervals. If the interval unit is sufficiently short to suit the circumstances of a rapid process, it may be reasonable to assume that the rate of transition from one state to another is constant across all subintervals, thereby retaining the standard ways of estimating interval-specific transition rates. Such a strategy, however, is usually not realistic, given the scarcity of empirical data on frequent transitions. Additionally, the use of survey data from multiple random samples, which is often the case in constructing a multistate life table, would be highly restrictive, given an insufficient sample size for each much shortened subinterval. Hence, it is necessary to adapt the conventional estimation procedures to characterize accurately the phenomena that occur intensely and rapidly over time.

For example, if use of a wide time interval is unavoidable in constructing a multistate life table, the generation of transition probabilities needs to be based on the principle of the Chapman–Kolmogorov relation (Cox and Miller, 1978), given by

$$\tilde{\pi}_{\tilde{i}\tilde{j}}^{(\check{n})} = \sum_{\tilde{j}=1}^{\tilde{k}} \tilde{\pi}_{\tilde{i}\tilde{j}}^{(\check{n}-1)} \tilde{\pi}_{\tilde{j}\tilde{i}}^{(l)}, \quad \check{n} = (1, 2, \ldots), \tag{2.46}$$

where $\tilde{\pi}_{\tilde{i}\tilde{j}}^{(\check{n})}$ is defined as the probability of being in state $\tilde{j}$ at time $\check{n}$ for those who are at state $\tilde{i}$ at time 0, termed the *n-step transition probability*. Similarly, $\tilde{\pi}_{\tilde{i}\tilde{j}}^{(\check{n}-1)}$ is an $(\check{n}-1)$-step transition probability and $\tilde{\pi}_{\tilde{i}\tilde{j}}^{(l)}$ is a one-step transition probability. This Markov process equation indicates that, in the presence of repeated transitions within a time interval, the $\check{n}$-step transition probability is virtually the outcome of a series of one-step transition probabilities within an interval represented by $\check{n}$ steps. Other multistate life table indicators need to be adapted as well, according to repeated flows of transitions.

The mathematical algorithms of a multistate life table for events with rapid processes are complex, so I do not describe the detailed procedures further in this text. The interested reader is referred to Liu *et al*. (1997).

2.2.3 Illustration: Life table estimates for older Americans

In this subsection, I provide an empirical example to demonstrate how to construct a life table using the SAS code. Empirical data come from a random sample of older Americans diagnosed with lung cancer. Event time is measured as the number of months elapsed from the time of diagnosis to the time of death or the time of censoring. Given the information of actual survival and censored times, I first create a dataset containing the number of events and the number of censored cases in each month for a total of 12 months. Below is the SAS program for this study.

SAS Program 2.4:

```
title 'survival of older persons diagnosed with lung cancer';

data Life_Table;
   keep Freq Months Censored;
   retain Months -.5;
   input fail withdraw @@;
   Months + 1;
   Censored = 0;
   Freq = fail;
   output;
   Censored = 1;
   Freq = withdraw;
   output;
   datalines;

16 0 22 5 19 3 20 4 27 2 20 11 21 32 38 59 40 61
24 33 18 19 23 36
;
.....
```

In this program, I create three variables for the construction of the life table: Months (defined earlier), Censored (1 = censored, 0 = not censored), and Freq (the frequency variable). From the above program, two types of observations are created for each time interval, one indicating the event observations and the other the censored observations. As displayed, input data are the frequencies of events and censored cases in each month.

The next step is to specify the ODS GRAPHICS ON statement for generating graphics of the survival and hazard curves and invoke the PROC LIFETEST again to calculate various life table estimates. As a result, the remainder of SAS Program 2.4 is given below.

SAS Program 2.4 (continued):

```
.....
ods graphics on;

proc lifetest data = Life_Table method = lt intervals = (0 to 11 by 1)
            plots = (s, ls, lls, h, p);
```

```
  time Months * Censored(1);
  freq Freq;
run;

ods graphics off;
```

In SAS Program 2.4 (continued), I ask SAS to compute the life table survival estimates by specifying METHOD = LT. The INTERVALS = (0 to 11 by 1) option specifies that estimates for 12 intervals (0 to 11) are computed. The PLOTS = (s, ls, lls, h, p) option requests SAS to display graphs of the life table survival function estimate, negative log of the estimate (the cumulative hazard function), log of the negative log of the estimate (the log–log survival function), estimated density function at the midpoint of each interval, and estimated hazard function.

As a result of SAS Program 2.4, SAS constructs a life table containing the requested survival function estimates and the standard errors (each function is described in Subsection 2.2.1) and five requested lifetime graphs. While the life table thus produced includes a large amount of data, I only display a portion of the estimates in this text, as summarized in Table 2.5. The table shows that among the older persons diagnosed with lung cancer, the survival probability throughout the first month is 0.97 with a standard error of 0.01. As time progresses, the probability of survival declines initially and then accelerates in later stages of the observation period. The five-month probability of survival is about 0.86 ($SE = 0.01$). In the 10th month, the survival rate is slightly higher than 0.5, suggesting that about half of those older persons are expected to survive throughout 10 months. By the end of the 12th month, the chance of survival is only 0.34 with a standard error of 0.03. This twelve-month survival rate indicates that for older persons diagnosed with lung cancer, only about 35 % are expected to survive beyond a one-year period.

SAS Program 2.4 also produces many other survival estimates, such as the conditional probability of failure, the median residual lifetime, the probability density function, the

Table 2.5 Life table survival estimates for patients with lung cancer.

Interval (t_{i-1}, t_i)	Number failed	Number censored	Effective sample size	Survival function	Standard error of survival
0–1	16	0	553.0	1.0000	0.0000
1–2	22	5	534.5	0.9711	0.0071
2–3	19	3	508.5	0.9311	0.0108
3–4	20	4	486.0	0.8963	0.0130
4–5	27	2	463.0	0.8594	0.0149
5–6	20	11	429.5	0.8093	0.0168
6–7	21	32	388.0	0.7716	0.0180
7–8	38	59	321.5	0.7299	0.0192
8–9	40	61	223.5	0.6436	0.0214
9–10	24	33	136.5	0.5284	0.0241
10–11	18	19	86.5	0.4355	0.0263
11–12	23	36	41.0	0.3449	0.0282

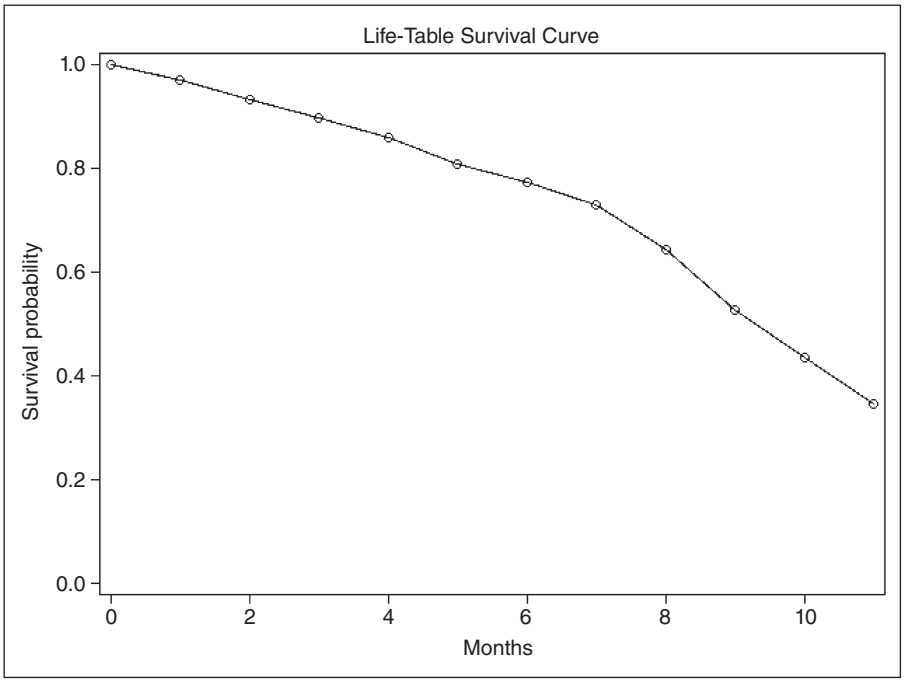

Figure 2.4 Life-table survival function for older persons with lung cancer.

hazard rate, and the standard error of each of these life table measures. The interested reader might want to rerun SAS Program 2.4 for viewing more results.

As mentioned above, five graphs are produced from SAS Program 2.4. Here, I select for display the graphs of the survival function and the cumulative hazard function, the two lifetime functions I consider to be most important for the description of survival processes for a population. First, Figure 2.4 displays the plot of the life table survival function estimate. It is interesting to note that in the first seven months or so, the survival function declines linearly; nevertheless, from that time point forward the probability of survival drops more sharply, highlighting the increased mortality acceleration in later months of the observation interval.

The next plot, Figure 2.5, displays the negative log of the survival estimates, the cumulative hazard function, and an accelerated cumulative hazard function over time. As will be discussed in Chapter 3, if the plot of the negative log of the survival function versus survival time approximates a straight line, the hazard function is constant, thereby highlighting an exponential distribution of survival times. If it is not, as shown by the above curve, the hazard function does not tend to be constant within this twelve-month period.

2.3 Group comparison of survival functions

The Kaplan–Meier estimator can be applied to compare survival functions by adding certain stratification factors. The stratification factors often selected include treatments in clinical

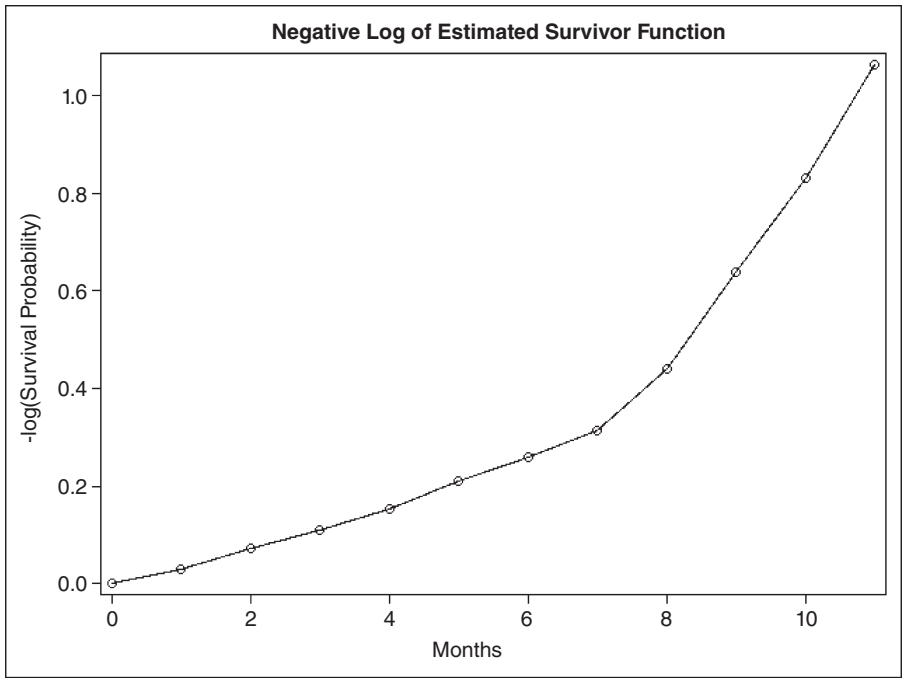

Figure 2.5 Negative log of the estimated survival function (the cumulative Hazard function) for older persons diagnosed with lung cancer.

trials and such sociodemographic variables as age group, gender, ethnicity, and marital status. In comparing the survival functions between two or more groups, an observed difference can be either the outcome of an actual disparity or a reflection of the sampling error. Therefore, it is essential to perform significance tests for determining whether an observed difference is true.

Significance tests on survival curves of different population groups generally begin with a null hypothesis, denoted by H_0, assuming no statistically significant difference. A significance level α, the predetermined critical value in a probability distribution, is regularly used to help determine whether to accept or reject the H_0 hypothesis. In particular, if the *p*-value of an observed difference is less than or equal to α, the difference is considered statistically significant, thereby resulting in the rejection of the underlying null hypothesis. If the *p*-value is greater than α, the null hypothesis is probably true and a group difference in the survival function may occur merely by chance.

Statisticians use various probability functions for hypothesis testing, the normal and the chi-squared distributions being the most widely applied. In survival analysis, there are a number of methods that can be used to test group differences in the survival function statistically. These techniques include, but are not limited to, the Mantel–Haenszel logrank test (Mantel and Haenszel, 1959), the Peto and Peto logrank test (Peto and Peto, 1972), the Gehan generalized Wilcoxon rank sum test (Gehan, 1967), the Peto and Peto and Prentice generalized Wilcoxon test (Peto and Peto, 1972; Prentice, 1978), and the Tarone and Ware modified

Wilcoxon test (Tarone and Ware, 1977). In this section, I describe these methods with empirical illustrations.

2.3.1 Logrank test for survival curves of two groups

I start the description of the logrank test by assuming two separate groups in a population of interest, termed G_1 and G_2, respectively. Each group is described by a different survival function, denoted by $S_1(t)$ and $S_2(t)$. As specified in Subsection 2.1.2, a sample of n observed survival times is ranked as $t_1 \le t_2 \le t_3 \le t_4 \le \cdots \le t_n$, among whom some may be tied. At each specific survival time t_i ($i = 1, \ldots, n'$), there are d_i individuals who experience a particular event of interest, among whom d_{1i} are those affiliated with G_1 and d_{2i} with G_2. If there are no tied cases, $d_i = 1$, then either d_{1i} or d_{2i} would take the value 0 and the other takes the value 1, and $n' = n$. If there are tied observations, however, the number of observed survival times n' is smaller than the number of individuals n. The number of survivors exposed to the risk of the event just before t_i, denoted by n_i, is also divided into n_{1i} for G_1 and n_{2i} for G_2. Therefore, $n_i = n_{1i} + n_{2i}$ and $d_i = d_{1i} + d_{2i}$. This classification can be recapitulated by a (2×2) contingency table displaying the number of events and the number of nonevents at t_i, as classified by G_1 and G_2 (Table 2.6). Using this table, I first set up the underlying null hypothesis that G_1 and G_2 have the identical survival function, written by H_0: $S_1(t) = S_2(t)$. If this hypothesis of no association holds, the marginal totals should all be fixed and, consequently, d_{1i} can be viewed as a random variable with parameters n_i, n_{1i}, and d_i following a specific probability distribution called *hypergeometric distribution* (Peto and Peto, 1972). A detailed description of the hypergeometric distribution is provided in Chapter 3 (Section 3.7).

Briefly, the hypergeometric probability of having d_{1i} in n_{1i}, given the fixed values of n_i, n_{1i}, and d_i, can be written by

$$\Pr\left(\tilde{Y}_{1i} = d_{1i}\right) = \frac{\binom{d_i}{d_{1i}}\binom{n_i - d_i}{n_{1i} - d_{1i}}}{\binom{n_i}{n_{1i}}}, \tag{2.47}$$

where $\tilde{Y}_{1i}$ is the random variable for d_{1i}. Specification of each binomial coefficient in Equation (2.47) is described in Section 3.7.

The hypergeometric random variable d_{1i}, given n_i, n_{1i}, and d_i, is well defined, with the expected value

Table 2.6 Number of events and of nonevents at t_i in two groups.

Group	Event		Total
	Yes	No	
G_1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
G_2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
Total	d_i	$n_i - d_i$	n_i

$$\mathrm{E}(d_{1i}; n_i, n_{1i}, d_i) = \frac{d_i n_{1i}}{n_i} \tag{2.48}$$

and variance

$$V(d_{1i}; n_i, n_{1i}, d_i) = \frac{d_i(n_i - d_i)n_{1i}n_{2i}}{n_i^2(n_i - 1)}. \tag{2.49}$$

As will be further discussed in Chapter 3, the interpretation of Equation (2.48) is straightforward: if the variable d_{1i} is random, its expected value is simply the proportion of n_i selected to n_{1i}, then multiplied by the total number of events d_i. In other words, if the null hypothesis holds, d_i should be proportionally allocated into G_1 and G_2. Therefore, from the variability of the observed d_{1i}, the null hypothesis on the association between survival and group can be statistically tested given a value of α.

Mantel and Haenszel (1959) propose to sum the differences between d_{1i} and $\mathrm{E}(d_{1i})$ for all observed survival times, given by

$$\check{D} = \sum_{i=1}^{n'} [d_{1i} - \mathrm{E}(d_{1i})], \tag{2.50}$$

where $\check{D}$ is the sum of differences between the observed and the expected values of d_{1i} over all the observed survival times. Likewise, the variance of $\check{D}$ is the sum of the variances of d_{1i} over the total number of survival times:

$$V(\check{D}) = \sum_{i=1}^{n'} V(d_{1i}). \tag{2.51}$$

As $\check{D}$ tends to be normally distributed with increasing sample size, its standardized form has mean zero and variance 1, written by

$$\frac{\check{D}}{\sqrt{\mathrm{var}(\check{D})}} \sim N(0,1).$$

Therefore, a z-score can be derived for testing the independence of survival and group, with the test statistic defined as

$$z = \frac{\sum_{i=1}^{n'} [d_{1i} - \mathrm{E}(d_{1i})]}{\sqrt{\sum_{i=1}^{n'} \mathrm{var}(d_{1i})}}. \tag{2.52}$$

Given the standard procedure for the z-test, significance testing on the null hypothesis can be performed: if the z-score is smaller than z_α, the null hypothesis that H_0: $S_1(t) = S_2(t)$ should be accepted with the conclusion that survival and group are independent. If $z \le z_\alpha$, on the other hand, the H_0 hypothesis should be rejected, implying that G_1 and G_2 are probably subject to two different survival processes.

When the total number of observed events is large, a more plausible statistic for testing the difference in two group-specific survival functions is to convert the standard normal to the chi-square distribution. Statistically, the square of a standard normal random variable has a chi-squared distribution, so a robust and efficient test statistic based on the chi-square distribution, generally denoted by Q, is

$$Q_{\text{logrank}} = \frac{\sum_{i=1}^{n'} [d_{1i} - \mathrm{E}(d_{11})]^2}{\sum_{i=1}^{n'} \mathrm{var}(d_{1i})} \sim \chi^2(1), \tag{2.53}$$

where $\chi^2(1)$ indicates the chi-square distribution with one degree of freedom for two groups. If the Q score is greater than or equal to the score associated with α, the p-value is lower than or equal to α and thus the null hypothesis is to be rejected. If the p-value is lower than α, then the difference in the two survival functions is probably due to sampling error, so H_0 is accepted. Generally, the p-values generated from both z-score and Q-score are identical when the sample size is large.

Peto and Peto (1972) and Prentice (1978) mathematically formalize the logrank test developed by Mantel and Haenszel (1959). This formalization includes the derivation of logrank scores for survival data with right censoring, starting with the specification of a real-valued random variable $\tilde{Y}$ with c.d.f. $F(t)$ and survival curve $S(t) = 1 - F(t)$. Suppose that $S_k(t_i, \theta_i)$ is a survival function parameterized by θ_i, where $\theta_i = \theta_{1i}$ for G_1 and $\theta_i = \theta_{2i}$ for G_2, and $k = 1, 2$. Here, the test on the independence of survival and group can be performed in terms of the parameter θ with the null hypothesis H_0: $\theta_{1i} = \theta_{2i} = \theta_i$. The parameter θ_i can be linked to a parametric distribution and estimated from a specific parametric likelihood function, denoted by $\hat{\theta}_i$. By such specifications, the logrank test can be expressed in terms of ordered residuals relative to a parametric distribution (Andersen *et al*., 1982; Prentice, 1978).

Given a continuous survival function, the hazard rate for $G_k(k = 1, 2)$ at t_i, given $\hat{\theta}_i$, is given by

$$\hat{h}_k\left(t_i, \hat{\theta}_i\right) = -\frac{\mathrm{d}\log \hat{S}_k\left(t_i, \hat{\theta}_i\right)}{\mathrm{d}t_i}. \tag{2.54}$$

Therefore, the cumulative hazard rate for $G_k(k = 1, 2)$ at t_i, given $\hat{\theta}_i$, is

$$\hat{H}_k\left(t_i, \hat{\theta}_i\right) = -\log \hat{S}_k\left(t_i, \hat{\theta}_i\right), \tag{2.55}$$

where the function S_k is known and continuous.

In the presence of censoring and without ties, let $\upsilon_{1i} = 0$ if d_{1i} is right censored; otherwise $\upsilon_{1i} = 1$. Assuming θ_0 is a fixed form of θ_i for all t values, the fixed overall parameter θ_0, under the null hypothesis, can be estimated by

$$\hat{\theta} = \frac{\sum d_{1i}}{\sum -\log S(\upsilon_{1i})}. \tag{2.56}$$

As $S(\upsilon_{1i})$ is relative to the occurrence of a single event, the denominator in Equation (2.56) is the expected number of total events from t_0 to survival time t_i in G_1. Alternatively, Equation (2.56) can be mathematically expressed as

$$\hat{\theta} = \frac{\sum_{i=1}^{n_i'} \frac{d_{1i}}{n_{1i}}}{\int_0^{n_i'} h_1(u)\,\mathrm{d}u}. \tag{2.57}$$

Therefore, the estimate of θ_0 can be expressed as the Nelson–Aalen estimate over the cumulative hazard function from a continuous distribution.

Given the null hypothesis and letting

$$e_{1i} = -\theta_0 \log S(\upsilon_{1i}), \tag{2.58}$$

Peto and Peto (1972) specify the following equations:

$$\mathrm{E}(e_{1i}) = \mathrm{E}\left[-\log S^{\theta_0}(\upsilon_{1i})\right] = \mathrm{E}(d_{1i}) = \mathrm{var}(d_{1i} - e_{1i}), \tag{2.59}$$

which is regardless of the fixed censoring point of $\tilde{Y}_{1i}$.

Let $O_{ki} = d_{ki}$ and $E_{ki} = e_{ki}$ ($i = 1, \ldots, n'$; $k = 1, 2$) in a traditional fashion. Under the null hypothesis, $\mathrm{E}(O_{ki} - E_{ki}) = 0$ and $\mathrm{var}(O_{ki} - E_{ki}) = \mathrm{E}(E_{ki})$. As a result, a test statistic can be expressed by

$$Q_{\text{logrank}} \approx \sum_k \sum_i \frac{(O_{ki} - E_{ki})^2}{\mathrm{var}(O_{ki} - E_{ki})} \sim \chi^2(1). \tag{2.60}$$

Equation (2.60) agrees with Equation (2.53) using a different expression. According to Equation (2.59), Equation (2.60) further reduces to

$$Q_{\text{logrank}} \approx \sum_k \sum_i \frac{(O_{ki} - E_{ki})^2}{\hat{E}_{ki}} \sim \chi^2(1). \tag{2.61}$$

Equation (2.61) indicates that the Mantel–Haenszel statistic is basically an approximate to the familiar Pearson chi-square test for equality of two groups. The term 'logrank test' actually comes from Peto and Peto's inference, in which the method uses the log transformation of the survival function to test a series of ranked survival times. In the absence of censoring, as Andersen *et al*. comment (Andersen et al., 1993, p. 349), the logrank test generates test scores approximately linearly related to the log rank of the observations ordered from the largest to the smallest (the so-called Savage test).

2.3.2 The Wilcoxon rank sum test on survival curves of two groups

The original Wilcoxon two-sample rank sum test (Wilcoxon, 1945) is perhaps the most popular nonparametric testing technique for two population groups. Traditionally, this test has been widely used as an alternative to the paired *t*-test when the assumption of normality cannot be satisfied. In particular, the Wilcoxon method is applied to the ordinal or continuous

response variables with the null hypothesis that the distribution of a given variable is the same for two population groups. The test is based on the calculation of a statistic, generally called U, whose distribution under the null hypothesis is known.

Gehan (1967) and Breslow (1970) extend the Wilcoxon rank sum test to the context of survival analysis. Before describing the extended approach handling right censoring, I review the Wilcoxon rank sum test for uncensored data for familiarizing the reader with this type of test. Suppose n_1 and n_2 individuals are allocated randomly into G_1 and G_2, with $n_1 + n_2 = n$. Then the observations are rank ordered by group:

$$\left(n_{11}, n_{12}, \ldots, n_{1n_1}\right) \text{and} \left(n_{21}, n_{22}, \ldots, n_{2n_2}\right).$$

All the observations are then arranged into a single ranked series, regardless of which group they belong to, given by

$$\left\{n_{(1)} < n_{(2)} < \cdots < n_{(n_1+n_2)}\right\}.$$

The ranks for observations in G_1 are added up:

$$R_1 = \sum_{i'=1}^{n_1} R_{1i'}, \tag{2.62}$$

where $R_{1i'}$ is the rank of $n_{1i'}$ in $n_{(i)}$, where $i' = 1, \ldots, n_1$). R_2 can be obtained by adding ranks of $n_{2i''}$ in $n_{(i)}$, where $i'' = 1, \ldots, n_2$.

Given that the ranked observations are survival times that follow cumulative distribution functions $F_1(n_1)$ and $F_2(n_2)$, I specify the null hypothesis that H_0: $F_1(t) = F_2(t)$. The Wilcoxon rank sum test is based on the calculation of the statistic U, whose distribution under the null hypothesis is known, written by

$$\frac{R_1 - E(R_1)}{\sqrt{\text{var}(R_1)}} \sim N(0,1), \tag{2.63}$$

where

$$\text{E}(R_1) = \frac{n_2(n_1 + n_2 + 1)}{2} \tag{2.64}$$

and

$$\text{var}(R_1) = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}. \tag{2.65}$$

The U score can be tested according to the distribution specified in Equation (2.63). Specifically, for testing the null hypothesis the following counting process is specified:

$$U(n_{1i'}, n_{2i''}) = U_{i'i''} = \begin{cases} +1 & \text{if } n_{1i'} > n_{2i''}, \\ 0 & \text{if } n_{1i'} = n_{2i''}, \text{ where } i' = 1, \ldots, n_1 \text{ and } i'' = 1, \ldots, n_2\ . \\ -1 & \text{if } n_{1i'} < n_{2i''}, \end{cases} \tag{2.66}$$

The total U score is defined by

$$U = \sum_{i',i''} U_{i'i''}. \tag{2.67}$$

As defined by Equation (2.67), U is the cumulative number of observations in G_2 whose rank is definitely less than $n_{1i'}$ minus the number of observations in G_2 whose rank is definitely greater than $n_{1i'}$. Therefore, if the null hypothesis is true, the value of the U-score should be zero.

From the above equations, a close relationship between U and R_1 can be identified, given by

$$R_1 = \frac{n_1(n_1 + n_2 + 1) + U}{2}. \tag{2.68}$$

As a result, U can be expressed in terms of R_1:

$$U = 2R_1 - n_1(n_1 + n_2 + 1). \tag{2.69}$$

Gehan's approach (1965) adapts Equation (2.67) to survival processes in the presence of right censoring. Let $t_{i'}$ and $t_{i''}$ be actual survival times and $t_{i'}^+$ and $t_{i''}^+$ censored times for G_1 and G_2, respectively. Then, Gehan redefines the $U_{i'i''}$ score as

$$U_{i'i''}^{\text{Gehan}} = \begin{cases} +1 & \text{if } t_{i'} > t_{i''} \text{ or } t_{i'}^+ \geq t_{i''}, \\ 0 & \text{if } t_{i'} = t_{i''} \text{ or } \left(t_{i'}^+, t_{i''}^+\right) \text{ or } t_{i'}^+ < t_{i''} \text{ or } t_{i''}^+ < t_{i'}, \\ -1 & \text{if } t_{i'} < t_{i''} \text{ or } t_{i'} \leq t_{i''}^+. \end{cases} \tag{2.70}$$

In Gehan's approach, whether a censored observation is counted as 1, 0, or −1 depends on the timing of censoring as compared to survival times of the other group. If censoring occurs to a member of G_1 but the censored time is greater than or equal to the actual survival time for a given member of G_2, the rank of the actual survival time for the member of G_1 is greater than the actual survival time for the member of G_2. Consequently, the score should be 1. In contrast, if the actual survival time for the member of G_1 is less than or equal to the censored time for the member of G_2, the rank of the survival time for the member of G_1 is definitely less than the actual survival time for the member of G_2. Then the value −1 should be assigned. If censoring occurs to an individual in one group before an event takes place for a matched member in the other group, a comparison of their ranks is difficult, so the score is simply 0.

A summary statistic for observation i' ($i' = 1, \ldots, n_1$), denoted by $W_{i'}$, is defined by

$$W_{i'} = \sum_{i''=1}^{n_2} U_{i'i''}^{\text{Gehan}}. \tag{2.71}$$

Equation (2.71) indicates that for an individual of G_1 with event time i', $W_{i'}$ is the number of observations in G_2 whose lifetimes are definitely less than $t_{i'}$ minus the number whose survival times are definitely greater than $t_{i'}$, taking into account the occurrence of right censoring. If survival is truly independent of group, survival times should be randomly distributed; then the expected value of $W_{i'}$ is 0. Given this rationale, the Gehan statistic, denoted by W, is

$$W = \sum_{i'=1}^{n_1} W_{i'} = \sum_{i'} \sum_{i''} U_{i'i''}^{\text{Gehan}}, \tag{2.72}$$

where the sum is over all n_1-versus-n_2 comparisons.

If the null hypothesis holds, the statistic W has the properties

$$\mathrm{E}(W) = 0, \tag{2.73}$$

$$\mathrm{var}(W) = \frac{n_1 n_2}{(n_1 + n_2)(n_1 + n_2 - 1)} \sum_{i'=1}^{n_1} W_{i'}^2. \tag{2.74}$$

Finally, the null hypothesis that $S_1(t) = S_2(t)$ can be tested by the Q test, given by

$$Q_{\text{Wilcoxon}} = \frac{[W - E(W)]^2}{\mathrm{var}(W)} \sim \chi^2(1). \tag{2.75}$$

Therefore, given the value of α, the equality of survival curves between G_1 and G_2 can be statistically tested by the p-value of the above chi-square distributed statistic. This method is the so-called *generalized Wilcoxon test* in survival analysis. From Equation (2.75), the reader might notice its striking similarity to the logrank test.

As the number of individuals exposed to the risk of a particular event decreases with the rank of survival times, the generalized Wilcoxon rank sum test implicitly uses the number of exposures just before a survival time as weight for the derivation of the statistic. This implied feature is the major difference between the generalized Wilcoxon test and the logrank test. As more weights are given to differences in the survival function at smaller t values, Tarone and Ware (1977) suggest the use of the square root of n_i as the weight to perform the generalized Wilcoxon test, so that a more balanced Q statistic can be derived.

Peto and Peto (1972) and Prentice (1978) advocate a different scoring method in the presence of censoring, using the Kaplan–Meier survival estimates. Here, the status indicator is used for a survival or a censored time, defined as $\delta_{ti'}$, where $\delta_{ti'} = 0$ if $t_{i'}$ is a censored survival time and $\delta_{ti'} = 1$ if $t_{i'}$ is an actual survival time. Then the $W_{i'}$ is redefined as

$$W_{i'} = \begin{cases} \hat{S}(t_{i'+1}) + \hat{S}(t_{i'-1}) - 1 & \text{if } \delta_{ti'} = 1, \\ \hat{S}(t_{i'+1}) - 1 & \text{if } \delta_{ti'} = 0. \end{cases} \tag{2.76}$$

Accordingly, the overall test score, W, is redefined as the sum of the scores generated by performing Equation (2.76) for G_1. In the presence of tied cases, this test score is reformulated by

$$W = \sum_{i'=1}^{n_1} W_{i'} d_{i'}. \tag{2.77}$$

The Peto and Peto (1972) and Prentice (1978) generalization of the Wilcoxon test is considered to be preferable to Gehan's approach because their generalized scores are consistent for an exact observation in the presence of right censoring. In Gehan's method, scores

at particular survival times vary according to the pattern of censoring imposed on the observations, so that differences in the pattern of censoring between G_1 and G_2 can somewhat affect the quality of a test (Andersen *et al*., 1993; Peto and Peto, 1972).

In general, compared to the logrank test, the generalized Wilcoxon rank sum test is sensitive to early differences between two survival curves, given the assignment of weight to each observation. Tarone and Ware (1977) argue that these two popular approaches can be unified by a general counting system in which they differ only in the choice of weight. This characterization enables statisticians to develop generalized procedures with choices of weight, as described in Subsection 2.3.3.

2.3.3 Comparison of survival functions for more than two groups

In empirical analyses, researchers often need to compare survival processes among more than two population groups. In biomedical studies, for example, there are frequently more than two treatments or testing groups in clinical trials. In survey data analysis, scientists are often interested in whether or not the occurrence of a particular event differs among several socioeconomic and demographic groups for gaining important information with policy implications. Given survival data allocated into more than two population groups, the theoretical question is whether any of those population subgroups differs from any others in the survival function.

Technically, tests on survival curves of more than two groups are simply the extension of the two-sample perspectives described above. Suppose there are K different groups, where $K > 2$, denoted by $G_1, G_2, \ldots, G_K$, respectively. In the presence of right censoring, the survival data from group k ($k = 1, \ldots, K$) is

$$(t_{k1}, \delta_{k1}), (t_{k2}, \delta_{k2}), \ldots, (t_{kn_k}, \delta_{kn_k}),$$

where δ_{ki} is the censoring status indicator defined earlier. In this survival data structure, at each survival time t_i ($i = 1, \ldots, n'$) the sample n_i is allocated into K groups, given by n_{1i}, $n_{2i}, \ldots, n_{Ki}$. Likewise, d_i is divided into $d_{1i}, d_{2i}, \ldots, d_{Ki}$, where d_{ki} is defined as the number of individuals in G_k who experience a particular event of interest at survival time t_i. Therefore, $n_i = n_{1i} + n_{2i} + \cdots + n_{Ki}$ and $d_i = d_{1i} + d_{2i} + \cdots + d_{Ki}$. If there are no tied observations, one of the d_{ki} takes the value 1 and the others take the value 0. This classification can be illustrated by a $(K \times 2)$ contingency table displaying the number of events and the number of nonevents at t_i, as classified by K groups (Table 2.7). Using this table, the null hypothesis to be tested is that all K groups are subject to an identical survival function, written by H_0: $S_1(t) = S_2(t) = \cdots = S_K(t)$. This null hypothesis is to be rejected if one of the $S_k(t)$ values deviates significantly from any of the others. A $\chi^2(K-1)$ test can be performed by comparing the observed and the expected values of events, by extending the methods described in Subsections 2.3.1 and 2.3.2.

With K groups involved in a comparison, a matrix expression of mathematical equations is more convenient. Let $\boldsymbol{O}_i = [d_{1i}, \ldots, d_{(K-1)i}]$ be a vector for the observed number of events in group 1 to group $(K-1)$ at event time t_i. Given a series of numbers of exposure, denoted $n_{1i}, n_{2i}, \ldots, n_{Ki}$, the distribution of counts in $\boldsymbol{O}_i$ is assumed to follow a multivariate hypergeometric function, conditional on both the row and the column totals (a detailed description of the hypergeometric distribution is provided in Chapter 3). This multivariate hypergeometric distribution, under the null hypothesis, is associated with a mean vector

Table 2.7 Number of events and of exposures at t_i in K groups.

Group	Event		Total
	Yes	No	
G_1	d_{1i}	$n_{1i} - d_{1i}$	n_{1i}
G_2	d_{2i}	$n_{2i} - d_{2i}$	n_{2i}
—	—	—	—
G_K	d_{Ki}	$n_{Ki} - d_{Ki}$	n_{Ki}
Total	d_i	$n_i - d_i$	n_i

$$\mathbf{E}_i = \left[\frac{d_i n_{1i}}{n_i}, \frac{d_i n_{2i}}{n_i}, \frac{d_i n_{(K-1)i}}{n_i} \right]' \tag{2.78}$$

and a variance–covariance matrix

$$\boldsymbol{V}_i = \begin{bmatrix} v_{11i} & v_{12i} & \cdots & v_{1(K-1)i} \\ & v_{22i} & \cdots & v_{2(K-1)i} \\ & & \ddots & \\ & & & v_{(K-1)(K-1)i} \end{bmatrix}. \tag{2.79}$$

In the standard logrank test, the kth diagonal element in $\boldsymbol{V}_i$ is defined as

$$v_{kki} = \frac{n_{ki}(n_i - n_{ki}) d_i (n_i - d_i)}{n^2 (n_i - 1)} \tag{2.80}$$

and the $k\imath$th off-diagonal element is

$$v_{k\imath i} = \frac{n_{ki} n_{\imath i} d_i (n_i - d_i)}{n^2 (n_i - 1)}, \tag{2.81}$$

where $k \neq \imath$.

With $(K - 1)$ degrees of freedom, the Q score for the logrank test of the $(K \times 1)$ table can be written as

$$Q_{\text{logrank}} = (\boldsymbol{O} - \mathbf{E})' \boldsymbol{V}^{-1} (\boldsymbol{O} - \mathbf{E}) \sim \chi^2 (K - 1), \tag{2.82}$$

where

$$\boldsymbol{O} = \sum_{i=1}^{n'} \boldsymbol{O}_i, \quad \mathbf{E} = \sum_{i=1}^{n'} \mathbf{E}_i, \quad \text{and} \quad \boldsymbol{V} = \sum_{i=1}^{n'} \boldsymbol{V}_i.$$

It is recognizable that the procedures of comparing survival curves for more than two groups, as formulated by Equations (2.78) through (2.82), are simply the expansion of the logrank test on two groups described in Subsection 2.3.1. The logrank test, however, does not consider group weight in calculating the test statistic. This procedure may not cause

serious bias when comparing two population samples, given relatively similar numbers of individuals. In comparing more than two survival curves, however, this problem can be serious because the number of individuals is usually unevenly allocated. Using weights can reduce sensitivity to early or late departures in testing the relationship among multiple samples (Klein and Moeschberger, 2003). As a result, some other techniques in this area, like the generalized Wilcoxon sum rank test and its modified version, may be used as alternatives.

As indicated earlier, some scientists (Klein and Moeschberger, 2003; Tarone and Ware, 1977) suggest that the logrank test and the modified Wilcoxon rank sum statistic can be specified in a way that they differ only in the choice of weight. This clarification results in the standardization of various techniques for comparing survival data of different population groups. Specifically, this unification is conducted by defining a weight function $w_k(t)$, given the property that $w_k(t) = 0$ whenever n_{ki} is 0.

Let $t_1 \le t_2 \le t_3 \le t_4 \le \cdot \cdot \cdot \le t_n$ be the distinct survival times in a combined sample and $w(t_i)$ be a positive weight function at event time t_i. Then the rank test statistic for comparing survival functions of K groups have the form of a K-dimensioned vector $\ddot{\boldsymbol{\upsilon}} = (\ddot{\upsilon}_1, \ddot{\upsilon}_2, \ldots, \ddot{\upsilon}_K)'$ with $\ddot{\upsilon}_k \ (k = 1, 2, \ldots, K)$,given by

$$\ddot{\upsilon}_k = \sum_{i=1}^{n'} w(t_i)\left(d_{ki} - \frac{n_{ki}d_i}{n_i}\right). \tag{2.83}$$

The estimated variance–covariance matrix of $\ddot{\boldsymbol{\upsilon}}$, $\boldsymbol{V} = (V_{k\iota})$, is defined as

$$V_{k\iota} = \sum_{i=1}^{n'} w^2(t_i)\left[\frac{(n_i n_u \delta_{k\iota} - n_{ki}n_{ui})\, d_i(n_i - d_i)}{n^2(n_i - 1)}\right], \tag{2.84}$$

where $\delta_{k\iota}$ is 1 if $k = \iota$ and 0 otherwise. The term $\ddot{\upsilon}_k \ (k = 1, 2, \ldots, K)$ is a weighted sum of observed minus expected numbers of events under the null hypothesis of identical survival curves. Given the standardization, the overall test statistic, denoted by Q_{standard}, is generally written as a multivariate sandwich equation $\ddot{\boldsymbol{\upsilon}}'\boldsymbol{V}^{-1}\ddot{\boldsymbol{\upsilon}}$. To spread it out with consideration of weights, the equation is

$$Q_{\text{standard}} = \left[\sum_{i=1}^{n'} \boldsymbol{w}_i(\boldsymbol{O}_i - \mathbf{E}_i)\right]' [\boldsymbol{w}_i \boldsymbol{V}_i \boldsymbol{w}_i]^{-1} \left[\sum_{i=1}^{n'} \boldsymbol{w}_i(\boldsymbol{O}_i - \mathbf{E}_i)\right] \sim \chi^2(K-1) \tag{2.85}$$

where $\boldsymbol{w}_{\text{i}} = \text{diag}(w_{\text{i}})$ for a $(K - 1)$ by $(K - 1)$ diagonal matrix.

Equation (2.85) is used as a generalized formula for testing survival curves of K groups, in which several models of this sort are reflected with differences only in the choice of weight. As a standardized expression, Equation (2.85) can also be applied to a two-group comparison as a special case with $K = 2$.

Given a unified specification, differences in various methods can be evaluated by examining the specification of weight for each testing method. The logrank test is based on the comparison between the observed and the expected numbers of events at each event time regardless of the distribution of n, so that $\boldsymbol{w}_{\text{i}}$ is an identity matrix over all event times for this test. Gehan's Wilcoxon rank sum test takes into account the distribution of sample size across all t_i values and, therefore, in this test $\boldsymbol{w}_{\text{i}} = \text{diag}(n_{\text{i}})$. Tarone and Ware (1977) suggest, as mentioned previously, that more weight should be assigned to later differences between the observed and the expected numbers of events, so that, in this method, $\boldsymbol{w}_{\text{i}} = \text{diag}(n_{\text{i}})^{1/2}$. In

the Peto–Peto Wilcoxon test, on the other hand, the Kaplan–Meier estimate for the pooled sample, $\hat{S}(t)$, is used as the weight. Based on the survival function of a pooled sample, instead of survival times and censoring distributions, the Peto–Peto weight specification is believed to have the capacity of generating more reliable results when censoring patterns differ over individual samples (Prentice and Marek, 1979).

Prentice (1978) and Andersen *et al*. (1982) propose a method to modify the weight used by the Peto–Peto method. They contend that in the Peto–Peto test, the Kaplan–Meier estimator for $S(t)$ should be replaced by the following estimator:

$$\bar{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i + 1}\right). \tag{2.86}$$

According to Prentice (1978) and Andersen *et al*. (1982), Equation (2.86) provides a consistent estimator of $S(t)$ under mild conditions on censoring. The test with such a weight is called the *modified Peto–Peto test*. When n is large, however, the modified Peto–Peto test generates very close or even identical results to those derived from the test originally proposed by Peto and Peto (1972).

There are some other tests on survival curves of different population groups dealing with other lifetime situations, such as the stratified and logrank tests for trend. For those additional testing techniques, the interested reader can refer to Collett (2003), Lawless (2003), and Klein and Moeschberger (2003).

2.3.4 Illustration: Comparison of survival curves between married and unmarried persons

To display how to perform various statistical tests on differences in the survival function between two or more population groups, I provide an empirical example about marital status and the probability of survival among older Americans. In particular, I want to estimate the five-year survival function among currently married and currently not married persons separately, and then assess whether or not the two survival curves differ significantly. In the survival data of older Americans, there is a variable called 'Married,' with 1 = 'currently married' and 0 = else, used for stratifying the survival data. The null hypothesis is that the survival curve for the individuals who are 'currently married' does not differ distinctively from the survival curve among the 'currently not married,' written by H_0: $(S_1(t) = S_2(t)$. I propose to use the above-mentioned five methods to test this hypothesis.

The SAS program for performing the five tests largely assembles SAS Program 2.1, with the addition of the variable 'Married' in the PROC LIFETEST statement as the stratification factor, shown below.

SAS Program 2.5:

```
.............

proc format;
  value Rx 1 = 'Married' 0 = 'Not married';

.........
```

```
ODS graphics on;

proc lifetest data = new plots = survival(atrisk = 0 to 60 by 10);
  time Months * Status(0);
  strata married  / test = all;
run;

ODS graphics off;
```

Compared to SAS Program 2.1, this program adds the stratification factor 'Married,' a PROC FORMAT statement that specifies values of that factor, and a STRATA option in the PROC LIFETEST statement. The PLOTS = survival option requests SAS to plot the survival curves and the 'ATRISK =' option specifies the time points at which the numbers exposed to the risk of death are displayed. In the STRATA statement, the TEST = all option specifies that all the nonparametric test scores are calculated.

SAS Program 2.5 yields a large quantity of output data. Therefore, I select to display the basic descriptive information and the results of the five tests in the following table.

SAS Program Output 2.1:

The LIFETEST Procedure

Summary of the Number of Censored and Uncensored Values

Stratum	married	Total	Failed	Censored	Percent Censored
1	Married	1090	181	909	83.39
2	Not married	892	218	674	75.56
Total		1982	399	1583	79.87

Rank Statistics

married	Log-Rank	Wilcoxon	Tarone	Peto	Modified Peto	Fleming
Married	-43.348	-76900	-1820	-39.368	-39.345	-39.556
Not married	43.348	76900	1820	39.368	39.345	39.556

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	19.2336	1	<.0001
Wilcoxon	19.1857	1	<.0001
Tarone	19.1951	1	<.0001
Peto	19.6290	1	<.0001
Modified Peto	19.6285	1	<.0001

The first part of SAS Program Output 2.1 presents the distribution of the total sample, of the number of events, and of the number of censored observations, classified by marital status. Of 1982 individuals, 1090 are currently married and 892 currently not married. There are 399 persons who are deceased by the end of the 60th month, among whom 181 are currently married and 218 are currently not married. Additionally, 1583 older persons are right censored: 909 'currently married' and 674 'currently not married.'

The second section of the output table displays the test scores, varying over different methods as anticipated. The final section demonstrates the test results on the five methods described previously. Those tests generate very close chi-square statistics assuming a $\chi^2(1)$ distribution, especially those generated from the Peto–Peto and the modified Peto–Peto tests. With a negligible difference (19.6290 versus 19.6285), the modified Peto–Peto test is not shown to improve the original estimate significantly in this example. Each of the test scores attaches to a *p*-value, all smaller than 0.0001. Given the consistency of the test scores and the corresponding *p*-values, it can be concluded that the null hypothesis should be rejected; that is, the survival curve for currently married persons differs significantly from the survival curve among their unmarried counterparts.

In SAS Program 2.5, the PLOT = SURVIIVAL option requests SAS to produce two survival curves, one for currently married and one for currently not married. The graph in Figure 2.6 is the resulting plot, where the survival curves for the two groups, currently married and currently not married, are displayed, along with the number of older persons

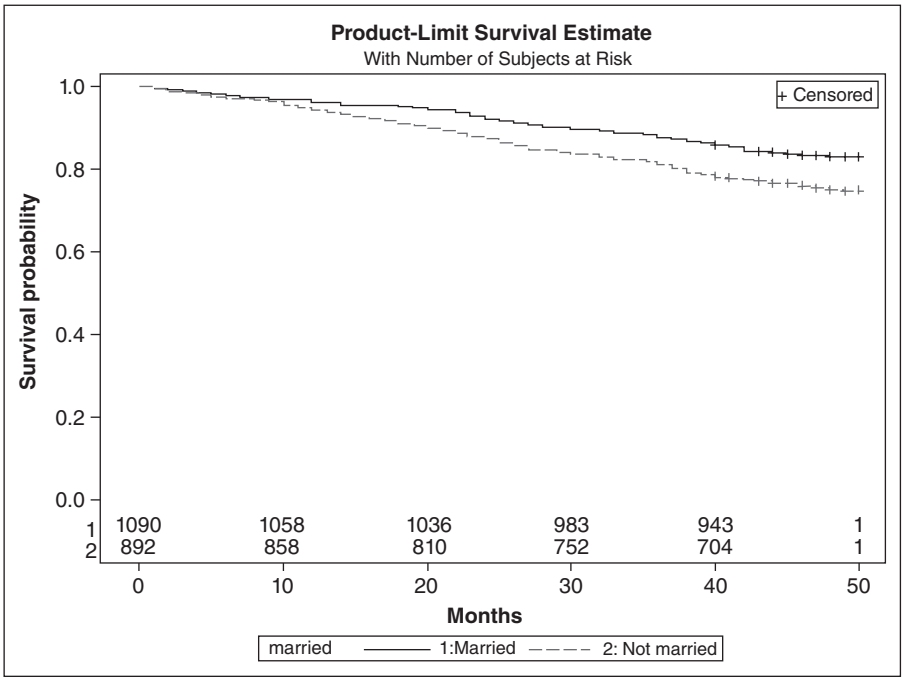

Figure 2.6 Plot of survival curves for currently married and currently not married persons.

exposed to the risk of death at month 0, 10, 20, 30, 40, and 50. The dominance of the survival curve among currently married older persons is obvious as the probability of survival declines at a much slower pace than does the curve among the 'currently not married.' The two survival curves start to separate after month 0, then the separation widens consistently over time. According to the results of the five tests, the separation between the two survival curves is statistically significant, demonstrating that older Americans who are currently married are expected to live longer than those currently not married.

2.4 Summary

This chapter describes some widely used descriptive approaches in survival analysis. Both the Kaplan–Meier and Nelson–Aalen estimators are described extensively, including the derivation of variances, confidence intervals, and confidence bands for the survival function. As most lifetime indicators are intimately associated, the Kaplan–Meier and Nelson–Aalen estimators can be applied to estimate other lifetime measures, given relevant formulas described in Chapter 1. Additionally, I provide a brief introduction about the life table method. While a life table accounts for survival times of censored observations both across and within fixed intervals, in many aspects the life table estimates approximate those generated from the Kaplan–Meier and the Nelson–Aalen approaches. One distinct advantage of using the Kaplan–Meier and Nelson–Aalen estimators over the life table method is the flexibility of using data with a small sample size, which is perhaps the reason why the life table method is rarely applied in biomedical studies.

In this chapter, I also describe several popular methods for testing survival functions of two or more population groups statistically, including the logrank test, Gehan's generalized Wilcoxon rank sum test, the Peto–Peto logrank and Wilcoxon methods, and the Tarone and Ware modified test. All these methods can be articulated by a unified formula in which they differ only in the choice of weight. This characterization facilitates the development of computer software procedures that program all those methods within an integrated estimating process, as exemplified by the programmed options in SAS PROC LIFETEST. While all these techniques often generate very close and even identical testing results, the logrank and Gehan's generalized Wilcoxon rank sum tests are the most widely used approaches to compare two or more group-specific survival curves.

As a stratification factor can be causally associated with some other explanatory factors, there are potentially conflicting possibilities for a bivariate relationship between a single covariate and survival outcomes. Methodologically, one of the potential possibilities is a spurious association, defined as a mathematical relationship in which two factors have no actual causal connection but look correlated due to the existence of one or more 'lurking' or confounding variables. A good example is found in a study of the relationship between education and smoking behavior among older Taiwanese (Liu, Hermalin, and Chuang, 1998). In the traditional Chinese culture, the majority of older Taiwanese women are illiterates, but, at the same time, they are less likely to smoke cigarettes and more likely to survive than older men. Consequently, a descriptive analysis displays a strong positive association between education and smoking cigarettes among older Taiwanese, a phenomenon contrary to what is expected. This observed association is obviously spurious because most lowly educated Taiwanese women do not smoke, thereby confounding the actual association between education and smoking behavior. After the lurking variable 'gender' is controlled,

the association between education and smoking behavior among older Taiwanese becomes significantly negative. Therefore, caution must apply when performing the descriptive approaches described in this chapter. These methods only have the capability to provide tentative results in survival analysis. There are some exceptions in this regard, including some of the clinical trials in which the potential confounding effects are taken into account in the process of randomization.