# Analysis of Survival Data

<div style="text-align: right">**11**</div>

Chapter Outline

This chapter introduces methods for analyzing data collected from a longitudinal study in which a group of subjects are followed for a defined time period or until some specified event occurs. We frequently encounter such data in the health field — for example, newly diagnosed cancer patients in a registry were followed annually until they died. Another example consists of smokers who completed a smoking cessation program and were then contacted every three months to find out whether or not they had relapsed. The focus in these studies is the length of time from a meaningful starting point until the time at which either some well-defined event happens, such as death or relapse to a certain condition, or the study ends. The data from such studies are called *survival data*. We have previously encountered survival data in our consideration of the life table in Chapter 4. In this chapter, we will consider a special type of life table: the follow-up life table.

We first discuss the collection and organization of the data. This discussion is followed by the presentation of two related methods for analyzing survival data. The life-table method is used for larger data sets, and the product-limit method is generally used for smaller data sets. We also show how the CMH test statistic from Chapter 10 can be used for comparing two survival distributions.

## 11.1 Data Collection in Follow-up Studies

Perhaps an example best illustrates the nature of the data required for a survival analysis.

---

**Example 11.1**

The California Tumor Registry (1963) identified a total of 2711 females with ovarian cancer initially diagnosed between 1942 and 1956 in 37 hospitals in California. The follow-up system of the Central Registry was designed to identify deaths through the statewide vital registration system and to facilitate the follow-up activities of the

participating hospital registries. The Central Registry received yearly follow-up information on each case. The registry program served not only to furnish the information essential for statistical study of cancer cases, but also to stimulate periodic medical checkups of the cancer patients. Based on the data accumulated in the Central Registry up to 1957, the researchers were able to analyze ovarian cancer patients who had been followed for up to 17 years.

In this data set, patients were observed for different lengths of time and not all of the patients had died by 1957. In addition, others could not be contacted — that is, they were lost to follow-up. Despite the different lengths of observation and the incomplete observations, it is possible to analyze the survival experience of these patients. An appropriate survival analysis is not restricted to those who had died, but it incorporates all the patients who entered the study. It is essential to include all those who entered the study because the exclusion of any patient from the analysis could introduce a selection bias as well as reducing the sample size.

The survival time cannot be calculated for those patients who were still alive at the closing date of the study or for those patients whose survival status was unknown. For these incomplete observations, the survival time is said to be *censored*. Those patients who were still alive at the closing date are known as *withdrawn alive*, and those patients whose status could not be assessed (because, for example, they moved away or refused to participate) are known as *lost-to-follow-up*.

To include the censored observations in the analysis, we calculate a censored survival time from the date of diagnosis to (1) the closing date of the study for those withdrawn alive and (2) the last known date of observation for the lost-to-follow-up. This allows the number of years from the date of diagnosis to the date of death or to the termination date to be calculated for each patient in the study.

By tabulating the uncensored and censored survival times of all 2711 female ovarian cancer patients by one-year intervals, we obtain the data shown in Table 11.1. Within the first year after diagnosis, 1421 of 2711 patients had died and 68 were lost-to-follow-up. There were no patients in the category withdrawn alive since every patient was followed for at least one year. The last column of the table can be created by adding the total column entries from the bottom. This reverse cumulative total indicates the number of patients alive at the beginning of each interval. The entry in the first row of this column is the total number of patients in the study. The other entries in this last column can also be found by subtracting the sum of the number of deaths, lost-to-follow-up, and withdrawn alive from the number of persons who started the previous interval. For example, the second entry in this column is 1222, which is determined by subtracting the sum of 1421, 68, and 0 from 2711, the number of subjects who began the previous interval.

The essential data items required for a survival analysis include $d_i$, the number of deaths; $l_i$, the number of patients lost-to-follow-up; $w_i$, the number of patients withdrawn alive; and $n_i$, the number of patients alive at the beginning of the $i$th interval. These data, presented in Table 11.1, are analyzed by the life-table method presented in the next section.

**Table 11.1    Survival times for ovarian cancer patients initially diagnosed 1942–1956, followed to 1957.**

| Years after Diagnosis | Death $d_i$ | Censored | | | Number Entering Interval $n_i$ |
| | | Lost $l_i$ | Withdrawn $w_i$ | Total | |
|---|---|---|---|---|---|
| 0–1 | 1,421 | 68 | 0 | 1,489 | 2,711 |
| 1–2 | 335 | 19 | 37 | 391 | 1,222 |
| 2–3 | 132 | 17 | 84 | 233 | 831 |
| 3–4 | 64 | 10 | 47 | 121 | 598 |
| 4–5 | 44 | 12 | 48 | 104 | 477 |
| 5–6 | 20 | 12 | 39 | 71 | 373 |
| 6–7 | 19 | 10 | 35 | 64 | 302 |
| 7–8 | 14 | 14 | 19 | 47 | 238 |
| 8–9 | 7 | 10 | 25 | 42 | 191 |
| 9–10 | 7 | 9 | 19 | 35 | 149 |
| 10–11 | 5 | 4 | 14 | 23 | 114 |
| 11–12 | 5 | 4 | 17 | 26 | 91 |
| 12–13 | 1 | 4 | 11 | 16 | 65 |
| 13–14 | 3 | 1 | 15 | 19 | 49 |
| 14–15 | 1 | 0 | 13 | 14 | 30 |
| 15–16 | 0 | 0 | 7 | 7 | 16 |
| 16–17 | 0 | 0 | 9 | 9 | 9 |

*Source:* California Tumor Registry, 1963

## 11.2    The Life-Table Method

In Chapter 4, the population life table was introduced to illustrate the idea of probability and its connection to life expectancy. The estimated life expectancy is generally used as a descriptive statistic. To use the life-table technique as an analytical tool, we shall combine ideas from Chapter 5 on probability distributions with the life-table analysis framework.

In survival analysis, our focus is on the length of survival. Let $X$ be a continuous random variable representing survival time. Consider a new function, the *survival function*, defined in symbols as

$$S(x) = \Pr(X > x).$$

This function is the probability that a subject survives beyond time $x$. Since $F(x)$, the cdf, is defined as

$$F(x) = \Pr(X \le x)$$

the survival function is one minus the cdf — that is,

$$S(x) = 1 - F(x).$$

It is more convenient to work with $S(x)$ rather than $F(x)$, since we usually talk about survival being greater than some value rather than being less than a value.

The idea of a survival function is contained in the population life table presented in Chapter 4. It is represented by the $l_x$ column, the number of survivors at the beginning of each age interval. Specifically, $S(x)$ in the population life table is $l_x/l_0$. Recall that the $l_x$ column starts with $l_0$, usually set at 100,000, and all subsequent $l_x$ values are derived

by multiplying the conditional probability of surviving in an age interval by the number of those who have survived all previous age intervals.

To analyze the data in Table 11.1 by the life-table method, we shall estimate the survival distribution in the same manner. The results of these calculations are shown in Table 11.2. The first two columns (the time interval and the number of deaths) are transferred from Table 11.1. The other columns show the results of the life-table analysis.

**Table 11.2   Estimates of probabilities and standard errors for ovarian cancer patients.**

| (1) Years after Diagnosis | (2) Deaths $d_i$ | (3) Exposed to Risk $n_i'$ | (4) Dying $q_i$ | (5) Surviving $(1 - q_i)$ | (6) Cumulative Probability Surviving $P_i$ | (7) Standard Error $SE(P_i)$ |
|---|---|---|---|---|---|---|
| | | | Conditional Probability | | | |
| 0–1 | 1,421 | 2,677.0 | 0.531 | 0.469 | 1.000 | 0.0000 |
| 1–2 | 335 | 1,194.0 | 0.281 | 0.719 | 0.469 | 0.0096 |
| 2–3 | 132 | 780.5 | 0.169 | 0.831 | 0.338 | 0.0092 |
| 3–4 | 64 | 569.5 | 0.112 | 0.888 | 0.280 | 0.0089 |
| 4–5 | 44 | 447.0 | 0.098 | 0.902 | 0.249 | 0.0087 |
| 5–6 | 20 | 347.5 | 0.058 | 0.942 | 0.224 | 0.0086 |
| 6–7 | 19 | 279.5 | 0.068 | 0.932 | 0.212 | 0.0086 |
| 7–8 | 14 | 221.5 | 0.063 | 0.937 | 0.197 | 0.0086 |
| 8–9 | 7 | 173.5 | 0.040 | 0.960 | 0.185 | 0.0087 |
| 9–10 | 7 | 135.0 | 0.052 | 0.948 | 0.177 | 0.0088 |
| 10–11 | 5 | 105.0 | 0.048 | 0.952 | 0.168 | 0.0090 |
| 11–12 | 5 | 80.5 | 0.062 | 0.938 | 0.160 | 0.0093 |
| 12–13 | 1 | 57.5 | 0.017 | 0.983 | 0.150 | 0.0097 |
| 13–14 | 3 | 41.0 | 0.073 | 0.927 | 0.147 | 0.0099 |
| 14–15 | 1 | 23.5 | 0.043 | 0.957 | 0.137 | 0.0109 |
| 15–16 | 0 | 12.5 | 0.000 | 1.000 | 0.131 | 0.0119 |
| 16–17 | 0 | 4.5 | 0.000 | 1.000 | 0.131 | 0.0119 |

The first task is to estimate the conditional probability of dying for each interval of observation. When there is no censoring in an interval, the estimate of the probability of dying in the interval is simply the ratio of the number who died during the interval to the number alive at the beginning of the interval. However, it is not appropriate to use this ratio as the estimator of the probability of dying if censoring occurred in the interval. The use of this denominator, the number alive at the beginning of the interval, means that those who were lost-to-follow-up or withdrawn alive during the interval are treated as if they survived the entire interval. Thus, using this ratio when there is censoring likely results in an underestimate of the probability of dying in the interval.

The problem with the censored individuals is that we do not know their actual length of survival during the interval. We do know that it is extremely unlikely that they all survived the entire interval. The assumption used most often in practice (although there are other more reasonable assumptions) is that the censored individuals survived to the midpoint of the interval. Under this assumption, we can calculate $q_i$, an estimator of the conditional probability of dying during the $i$th interval, as follows:

$$q_i = \frac{d_i}{n_i - \dfrac{l_i + w_i}{2}} = \frac{d_i}{n_i}.$$

The denominator in the above equation is the effective number of subjects exposed to the risk of dying during the interval, denoted by $n'_i$. Table 11.2 shows the estimated effective number of patients exposed to the risk of dying in column 3 and the estimate of the conditional probability of dying in column 4. The use of $n'_i$ implies that those patients who were lost or withdrawn were subjected to one half the risk of dying during the interval.

The estimator of the conditional probability of survival in the $i$th interval is one minus the estimator of the probability of dying, that is, $1 - q_i$. The result of this subtraction is shown in column 5.

Next, we calculate $P_i$, the sample estimator of the probability of surviving until the beginning of the $i$th interval. The set of the $P_i$ are used to estimate the survival distribution $S(x)$. By definition, $P_1 = 1$, and the estimators of the other survival probabilities are calculated in the following manner:

$$P_2 = (1 - q_1), \ P_3 = (1 - q_2)\,(1 - q_1), \ldots$$

and in general

$$P_i = (1 - q_{i-1})(1 - q_{i-2}) \ldots (1 - q_1) = (1 - q_{i-1}) \ P_{i-1}.$$

The results of these products are shown in column 6 of Table 11.2. From column 6, we see that the estimate of the one-year survival probability for ovarian cancer patients in California who were diagnosed during the 1942–1956 period was 0.47 and the estimate of the five-year survival probability was 0.22. More recent statistics estimate the five-year survival probability for ovarian cancer to be 0.39 for white females and 0.38 for black females in the 1981–1986 period (National Cancer Institute 1990), suggesting some improvement in cancer treatment. However, this improvement may be due more to the early detection of ovarian cancer in recent years. Cancer-related statistics, including estimates of survival rates, are routinely provided by the National Cancer Institute's SEER (Surveillance, Epidemiology, and End Results) program, which includes many population-based cancer registries throughout the United States.

Besides knowing the point estimate of a population survival probability, we also wish to have a confidence interval for the survival probability. We shall assume that, in large samples, an estimated cumulative survival probability approximately follows a normal distribution. The variance of the estimated cumulative survival probability is estimated by

$$\hat{V}ar(P_i) = P_i^2 \sum_{j=1}^{i-1} \frac{q_i}{n'_j(1-q_j)}.$$

The estimated standard errors (the square root of the estimated variance) of the $P_i$ are shown in column 7 of Table 11.2.

Given these estimated standard errors plus the assumption of the approximate normality of the estimated survival probabilities, we can calculate confidence intervals for the survival probabilities. The approximate $(1 - \alpha)*100$ percent confidence interval for a survival probability is given by

$$P_i \pm (z_{1-\alpha/2}) \ s.e.(P_i)].$$

For example, an approximate 95 percent confidence interval for the five-year survival probability is

$$0.224 - 1.96\ (0.0086) \text{ to } 0.224 + 1.96\ (0.0086)$$

or from 0.207 to 0.241.

Although this procedure is adequate in most cases, there are other more complicated approaches to constructing a confidence interval for $P_i$ that cause the actual confidence level to agree more closely with the nominal confidence level, especially for small samples (Thomas and Grunkemeier 1975).

It is also possible to calculate the confidence interval for the difference between two survival probabilities from different study groups — for example, the five-year survival probability of ovarian cancer for white females and black females — by using the procedure discussed in Chapter 7.

Let us further explore the estimated survival distribution by creating Figure 11.1, the plot of the cumulative survival probabilities against the years after diagnosis. Although we have values of $P_i$ for only the integer values of $t$, we have connected the points to show the shape of the survival distribution. It starts with survival probability of 1 at time 0 and drops quickly as time progresses, indicating a very high early mortality for ovarian cancer patients. Note that the survival curve does not descend all the way to zero, since some women survive more than 17 years.

The rapid decrease in the estimated survival curve suggests that the mean and the median survival times will be short. To verify this, let us estimate the mean and the median survival times from the estimated survival distribution. Since some of the women survive longer than the 17 years of the study, this complicates the estimation of the population mean survival time. Instead of estimating the population mean, we shall therefore estimate the mean restricted to the time frame of 17 years, the length of the study. This restricted value will thus underestimate the true unrestricted mean. If no patient survived longer than the time frame of the study, the following procedure provides an estimate of the unrestricted mean.
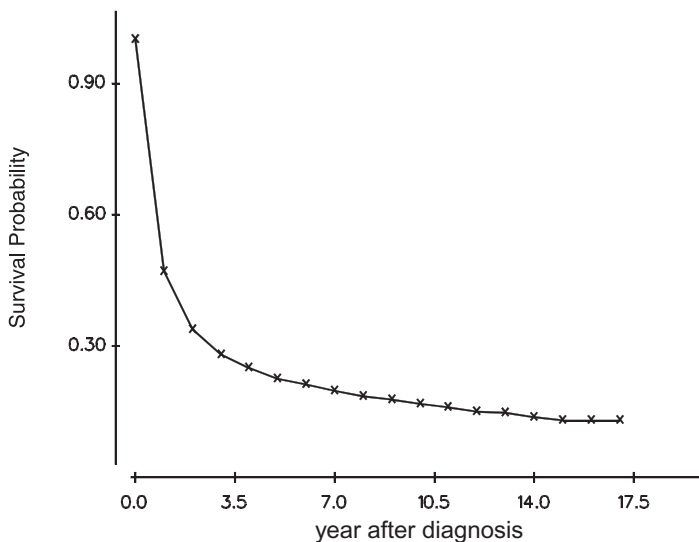


**Figure 11.1**   Estimated survival distribution of ovarian cancer patients.

The sample mean, restricted to the 17-year time frame, is found by summing the number of years (or other unit of time) survived during each time interval and dividing this sum by the sample size. However, the process of determining the number of years survived in an interval is complicated by the deaths, losses, and withdrawals that occurred during the interval. Instead of directly attempting to calculate the years survived, we shall use the following method to deal with this complication.

We calculate the sample mean by forming a weighted average of the years provided by each interval. The weight used with each interval is the cumulative survival probability associated with the interval. This approach deals with the complications mentioned above, since the probability takes the deaths, losses, and withdrawals into account. Since there are two cumulative survival probabilities associated with each interval — the probability at the beginning, $P_i$, and the probability at the end, $P_{i+1}$ — we use their average. Thus, the formula for the restricted sample mean is

$$\bar{x}_r = \sum_{i=1}^{k} a_i \left( \frac{P_i + P_{i+1}}{2} \right)$$

where $k$ is the number of intervals and $a_i$ is the width of the $i$th interval.

This formula has an interesting geometrical interpretation: It provides an approximation to the area under the estimated survival curve. For example, consider a curve with three intervals (Figure 11.2). We are using rectangles to estimate the area under the curve. As we can see, some of the area under the curve is not included in the rectangles. However, this area is approximately offset by the areas included in the rectangles that are not under the curve. The formula for the area of a rectangle is the height multiplied by the width. In this case, the width is one unit or, in general, $a_i$ units, and the height is taken to be the average of the points at the beginning and end of the interval, that is, $(P_i + P_{i+1})/2$. Hence, another way of interpreting the mean is that it is the area under the survival curve. We approximate this area by calculating the area of the rectangles that can be superimposed on the survival curve.

When the intervals are all of the same width — for example, a — then the formula can be simplified to
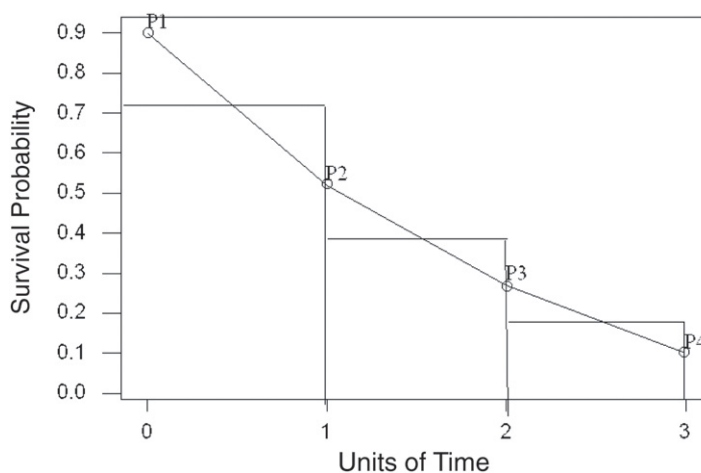


**Figure 11.2** Survival curve with rectangles superimposed.

$$\overline{x}_r = a_i \left( \sum_{i=1}^{k+1} P_i - \frac{P_1 + P_{k+1}}{2} \right).$$

Because the intervals are all of width one in this example, the sample mean is simply the sum of the entries in column 6 of Table 11.2 minus one half of the first and last entries in the column. This is

$$(1.000 + 0.469 + 0.338 + \ldots + 0.131) - 0.5(1.000 + 0.131)$$

which equals 3.92 years. This restricted mean survival time appears to be larger than what the first-year survival probability might suggest. As we saw in Chapter 3, the mean can be affected by a few large observations, and that is the case here. The sample mean reflects the presence of a few long-term survivors. Let us now calculate the median length of survival.

The median survival time is estimated in the following manner. First, we read down the list of estimated cumulative survival probabilities, column 6 in Table 11.2, until we find the interval for which $P_i$ is greater than or equal to 0.5 and $P_{i+1}$ is less than 0.5. In Table 11.2, this is the first interval, since $P_1$ is greater than 0.5 and $P_2$ is less than 0.5. Thus, we know that the estimated median survival time is between 0 and 1 year. Since 47 percent of the patients survived the first year, we suspect that the estimated median survival time is much closer to one year than to zero years. To find a more precise value, we shall use linear interpolation.

In using linear interpolation, we are assuming that the deaths occurred at a constant rate throughout the interval. This is the same assumption we made when we connected the survival probabilities in Figure 11.1. In using linear interpolation, we know that to reach the median, we only require a portion of the interval, not the entire interval. The portion that we need is simply the ratio of the difference of $P_i$ and 0.5 to the length of the interval. In symbols, this is

$$(P_i - 0.5) / (P_i - P_{i+1}).$$

We multiply this ratio by the width of the interval and add that to the survival time at the beginning of the interval. Replacing these words by symbols, the formula is

$$Sample\ median = x_i + a_i \left( \frac{P_i - 0.5}{P_i - P_{i+1}} \right)$$

where $x_i$ is the survival time at beginning of the interval and $a_i$ is the width of the interval. In this example, the sample median survival time is

$$Sample\ median = 0 + 1 \left( \frac{1 - 0.5}{1 - 0.469} \right) = 0.94.$$

The sample median survival time of about one year is much shorter than the estimated restricted mean survival time. As we just mentioned, the mean survival time is affected by a small number of long-term survivors. This is why the median is more often used with survival data.

The median can also be obtained from the plot of the estimated survival curve shown in Figure 11.1. We move up the vertical axis until we reach the survival probability value of 0.5. We then draw a line parallel to the time axis and mark where it intersects the
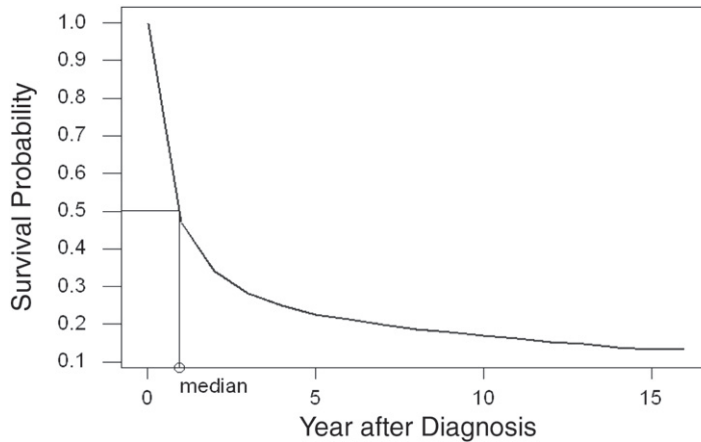
**Figure 11.3**  Using the estimated survival curve to find the median.

survival curve. We next draw a line, parallel to the vertical axis, from the intersection point to the time line. The sample median survival time is the value where the line intersects the time axis. Figure 11.3 shows the estimated survival curve plot with these lines used to find the sample median drawn in the plot as well. The accuracy of the estimate of the median is limited by the scales used in plotting the survival curve. In Figure 11.3, the precision of the estimate is likely not to be high because of the scales used. It appears that the sample estimate of the median is approximately one year.

Another statistic often used in survival analysis is the *hazard rate*, which is also known as the life-table mortality rate, force of mortality, and instantaneous failure rate. It is used to measure the proneness to failure during a very short time interval. It is analogous to an age-specific death rate or interval-specific failure rate. It is the proportion of subjects dying or failing in an interval per unit of time. The hazard rate is usually estimated by the following formula:

$$h_i = \frac{d_i}{a_i(n_i' - d_i/2)} = \frac{2q_i}{a_i(2 - q_i)}.$$

The denominator of this formula uses the number of survivors — again assuming that death is occurring at a constant rate throughout the interval — at the midpoint of the interval. When the interval is very short, it makes little difference whether the number of survivors at the beginning or at the midpoint of the interval is used in the denominator. The sample hazard rates are calculated and shown in Table 11.3 for the first 10 years of follow-up. The estimate of the first year hazard or mortality rate is quite high with 723 deaths per 1000 patients. The hazard is concentrated in the first five years after diagnosis and stabilizes at a low level after five years of survival. The variance of the sample hazard rate is estimated by

**Table 11.3  Estimates of hazard rates and their standard errors.**

| Year | Hazard Rate | Standard Error | Year | Hazard Rate | Standard Error |
|---|---|---|---|---|---|
| 0–1 | 0.723 | 0.0179 | 5–6 | 0.059 | 0.0132 |
| 1–2 | 0.326 | 0.0176 | 6–7 | 0.070 | 0.0161 |
| 2–3 | 0.185 | 0.0160 | 7–8 | 0.065 | 0.0174 |
| 3–4 | 0.119 | 0.0149 | 8–9 | 0.041 | 0.0156 |
| 4–5 | 0.104 | 0.0156 | 9–10 | 0.053 | 0.0201 |

$$\hat{V}ar(h_{i}) = h_{i}^{2}\left(\frac{1-(h_{i}a_{i}/2)^{2}}{n_{i}'q_{i}}\right).$$

The estimated standard errors (the square root of the estimated variance) of the sample hazard rates are calculated and shown in Table 11.3. If we assume that the sample hazard rates are asymptotically normally distributed, these sample standard errors can be used to calculate confidence intervals for the population hazard rates. For example, the 95 percent confidence interval for the first year hazard or mortality rate ranges from

$$0.723 - 1.96(0.0179) \text{ to } 0.723 + 1.96(0.0179)$$

or from 0.688 to 0.758.

These life table calculations can be performed by the computer (see **Program Note 11.1** on the website).

## 11.3   The Product-Limit Method

When we analyze a smaller data set — for example, a sample size less than 100 — the life-table method may not work very well because the grouping of survival times becomes problematic. Instead we use a method that is based on the actual survival time for each subject rather than grouping the subjects into intervals. The product-limit method, also known as the Kaplan-Meier method (Kaplan and Meier 1958), is used to estimate the cumulative survival probability from a small data set, without relying on groupings of survival times. As we can see following, the basic principles and computational procedures involved in the product-limit method are similar to the life-table method.

We start with an example.

---

**Example 11.2**

Suppose that 14 alcohol-dependent patients went through an intensive detoxification treatment for four years from 1990 to 1993 at a small clinic. There was a follow-up contact every month to check on their drinking status. The data shown in Table 11.4

**Table 11.4   Status of 14 alcohol-dependent patients discharged from a clinic.**

| Patient Number | Date of Discharge | Date of Termination | Follow-up Status | Gender |
|---|---|---|---|---|
| 1 | 9001 | 9312 | 2 Still sober (withdrawn) | 1 Female |
| 2 | 9003 | 9009 | 1 Relapsed | 1 Female |
| 3 | 9005 | 9209 | 1 Relapsed | 2 Male |
| 4 | 9009 | 9111 | 2 Lost-to-follow-up | 2 Male |
| 5 | 9011 | 9306 | 1 Relapsed | 1 Female |
| 6 | 9102 | 9312 | 2 Still sober (withdrawn) | 1 Female |
| 7 | 9104 | 9211 | 1 Relapsed | 1 Female |
| 8 | 9108 | 9304 | 1 Relapsed | 1 Female |
| 9 | 9110 | 9202 | 1 Relapsed | 2 Male |
| 10 | 9203 | 9308 | 2 Lost-to-follow-up | 2 Male |
| 11 | 9207 | 9311 | 1 Relapsed | 2 Male |
| 12 | 9212 | 9310 | 1 Relapsed | 1 Female |
| 13 | 9303 | 9312 | 2 Still sober (withdrawn) | 2 Male |
| 14 | 9304 | 9310 | 1 Relapsed | 2 Male |

were abstracted from the clinic patient records. The date of discharge and the date of termination are shown in year and month (9001 indicates 1990, January). The follow-up status is coded 2 if censored (withdrawn or lost-to-follow-up) and 1 if relapsed to drinking. Gender is coded 1 for females and 2 for males. The purpose of our study is to analyze the length of alcohol-free time among these 14 patients.

The first step of analysis is to calculate the survival time, $x$, in months for all subjects, censored and uncensored, and arrange them in order from the smallest to the largest with the censoring status indicated. If an uncensored subject and a censored subject have survival times of the same length, the uncensored one precedes the corresponding censored observation. For the data shown in Table 11.4, the ordered list of alcohol-free times in months, with the censored observations marked by asterisks, is as follows:

$$4, 6, 6, 9*, 10, 14*, 16, 17*, 19, 20, 28, 31, 34*, 47*.$$

The second step is to create a worksheet like that shown in Table 11.5. In Table 11.5, the column headings refer to death and survival. For this problem, death is equated with relapse and survival is remaining alcohol free. The first three columns in the worksheet are created according to the following procedures.

1. List the uncensored alcohol-free times in order. These are 4, 6, 10, 16, 19, 20, 28, and 31. We shall refer to these times as $x_1, x_2, \ldots, x_8$, respectively.
2. Count the number of relapses at each $x_i$. There is one relapse at each time unless there are ties. The numbers are 1, 2, 1, 1, 1, 1, 1, and 1.
3. Count the number of subjects who are at risk of relapse at $x_i$. For example, when the survival time is 10 months, three people have already relapsed and one person was withdrawn. Thus, there are only 10 persons at risk of relapse at 10 months. The list of these numbers is 14, 13, 10, 8, 6, 5, 4, and 3.

The fourth and fifth columns, estimates of the conditional probability of survival ($1 - q_x$) and the cumulative probability of survival ($P_x$) are calculated next, followed by the calculation of estimated standard error of $P_x$, shown in column 6. The estimator of the conditional probability of relapse is the number of relapses divided by the number at risk, that is, $q_x = d_x/n_x$. The estimator of the conditional probability of survival is

**Table 11.5   Kaplan-Meier estimates of survival probabilities with standard errors.**

| (1) Survival Time $x_i$ | (2) Number of Deaths $d_x$ | (3) Number at Risk $n_x$ | (4) Conditional Probability of Survival $(1 - q_x)$ | (5) Cumulative Probability of Survival $P_x$ | (6) Standard Error $SE(P_x)$ | (7) Approx. Standard Error $SE(P_x)$ |
|---|---|---|---|---|---|---|
| 0 | 0 | 14 | 1.000 | 1.000 | — | — |
| 4 | 1 | 14 | 0.929 | 0.929 | 0.069 | 0.066 |
| 6 | 2 | 13 | 0.846 | 0.786 | 0.110 | 0.101 |
| 10 | 1 | 10 | 0.900 | 0.707 | 0.124 | 0.121 |
| 16 | 1 | 8 | 0.875 | 0.619 | 0.136 | 0.135 |
| 19 | 1 | 6 | 0.833 | 0.516 | 0.148 | 0.146 |
| 20 | 1 | 5 | 0.800 | 0.412 | 0.150 | 0.141 |
| 28 | 1 | 4 | 0.750 | 0.309 | 0.144 | 0.128 |
| 31 | 1 | 3 | 0.667 | 0.206 | 0.127 | 0.106 |

$$1 - q_x = 1 - \frac{d_x}{n_x} = \frac{n_x - d_x}{n_x}.$$

For example, $1 - q_4 = (1 - 1/14) = 0.929$ and $1 - q_6 = (1 - 2/13) = 0.846$. The estimator of the cumulative probability of survival is found from the estimators of the conditional probabilities of survival in the same way as in the life-table method — that is,

$$P_x = \prod_{t \leq x}(1 - q_t) = \prod_{t \leq x} \frac{n_t - d_t}{n_t}.$$

The product symbol, $\Pi$, means that we multiply each term in the expression by one another for the indicated values of $t$. For example,

$$P_6 = \prod_{t \leq 6}(1 - q_t) = (1 - q_4)(1 - q_6) = (0.929)(0.846) = 0.786.$$

We could have included $1 - q_0$ in the product, but since $q_0$ is defined to be zero, its inclusion would not have changed the product.

As we have just seen, the censored observations have not been excluded from the analysis. They played a role in the determination of the number at risk at each time of relapse. If the censored observations were totally excluded from the analysis, the estimate of the conditional survival probabilities for the uncensored observations would be different.

The variance of $P_x$ is estimated by

$$\hat{Var}(P_x) = P_x^2 \left( \sum_{t \leq x} \frac{d_t}{(n_t - d_t)n_t} \right) \cong P_x^2 \left( \frac{1 - P_x}{n_x} \right).$$

The approximation shown in the preceding equation is much simpler to calculate, and it works reasonably well in most situations (Peto et al. 1977). Taking the square root of the variance, we obtain the estimated standard errors of the $P_x$ that are shown in column 6. The approximate standard errors are shown in column 7. The approximate estimate of the standard error of $P_4$ is 0.066, compared to the value of 0.069 obtained from the use of the first expression for the sample variance.

Figure 11.4 graphically displays the estimated survival distribution shown in the fifth column of Table 11.5. The plot includes a survival probability of 1 at time 0. The plot of the survival probabilities is referred to as a step function, since it looks like a stair step. It has this appearance because the probability of survival stays the same over a time period — this causes the horizontal lines — and then drops whenever there is another relapse — the vertical lines. However, long horizontal lines, showing no change in survival probability for a long period of time, should not be interpreted as a period with no risk, for these may occur because of a small number of subjects under observation during those time periods.

We can estimate the mean survival time from the survival distribution. Again, just as in the life-table method, if the largest survival time is a censored time, we are really estimating a restricted mean. If the largest survival time is uncensored, then the survival probability will decrease to zero, and we will be estimating the unrestricted mean. As in the life table, the mean survival time is the area under the curve. We shall again use rect-
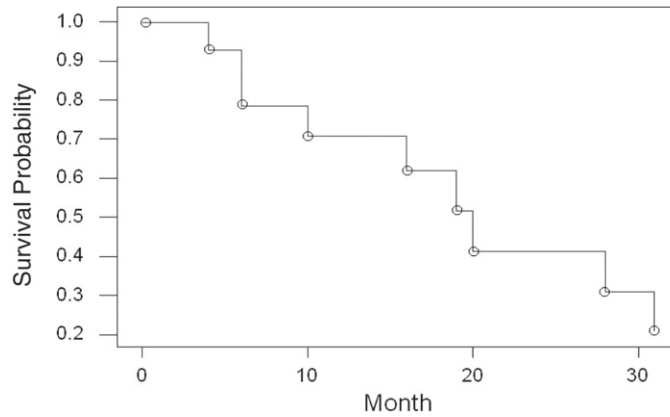
**Figure 11.4** Survival distribution estimated by the product-limit method.

angles to approximate this area. Because of the step nature of the survival curve here, the rectangles are already formed for us. Unlike the life-table method, the widths of the intervals here are usually different. The following formula shows the area of each rectangle being calculated as the product of the height of the rectangle, the estimated cumulative survival probability associated with $x_i$, by the width, $x_{i+1}$ minus $x_i$. In symbols, this is

$$\bar{x}_r = \sum_{i=0}^{k-1} P_{x_i}(x_{i+1} - x_i)$$

where $k$ is the number of distinct time points when someone relapsed, $x_0$ is defined to be zero, and $P_0$ is defined to be one.

For these data, the estimate of the restricted mean alcohol-free time, restricted to a 31-month window, is given by

$$\bar{x}_r = 1(4 - 0) + 0.929(6 - 4) + \cdots + 0.309(31 - 28) = 18.4.$$

This is an underestimate of the true mean alcohol-free time because we are restricted to the study timeframe and there were still people free of alcohol at the end of the study.

From Table 11.5, we can see that the sample median survival time, the point at which the cumulative survival probability is 0.5, occurs between the 19th and 20th months and is closer to month 19. We shall interpolate to find the sample median in the same way as in the life-table method. From our data, the sample median survival time is found as follows:

$$Sample\ median = 19 + (20 - 19)\left(\frac{0.516 - 0.5}{0.516 - 0.412}\right) = 19.2\ months.$$

We should not use interpolation to find the median if there is a large gap in time between the two survival times in which we will be using the interpolation.

The computer can be used to calculate the entries in Table 11.5 as well as the sample mean, the median, and the graph in Figure 11.4 (see **Program Note 11.2** on the website).

Because the product-limit method is based on the ranking of individual survival times, it is cumbersome to apply with a large data set. We would not consider using it

with the ovarian cancer data from the California Tumor Registry that had over 2000 observations. For a large data set, the life-table method simplifies the calculation and gives results similar to the product-limit method.

So far we have focused on describing the survival experience of a single population. However, we are often interested in comparing the survival experiences of two or more groups of subjects who differ on some account — for example, patients who have received different therapies for cancer or patients who belong to different age or sex groups. The comparison of two survival distributions is the topic of the following section.

# 11.4    Comparison of Two Survival Distributions

When comparing the survival experience of two or more groups, the description of the differences in the estimated survival distributions and the plot of the survival curves are only the beginning of the analysis. In addition to these descriptive techniques, researchers require a statistical test to determine whether the observed differences are statistically significant or due to chance variation.

In the analysis of survival data, we generally do not assume that the data follow any particular probability distribution. In the analysis, we also use the median survival time, rather than the mean, to summarize the survival experience. Because of these features, it seems as if a nonparametric test should be used when comparing survival distributions. If we know that the survival data follow a particular distribution, we should take advantage of that knowledge. There are parametric tests available that can be used when we know the probability distribution of the survival data (Lee 1992).

For small data sets with no censored observations, the Wilcoxon rank sum test (the Mann-Whitney test) can be used to test the null hypothesis of no difference in survival distributions for two independent samples. However, since survival time data usually contain censored observations, the Wilcoxon test cannot be directly applied. In this section, we show how the Cochran-Mantel-Haenszel (CMH) test statistic, described in Chapter 10, can be used in testing the hypothesis of no difference between two survival distributions (Mantel 1966). There are a number of other tests, extensions of the Wilcoxon and other rank tests, that could be used as well, but the CMH test seems to perform as well, if not better, than these other tests.

## 11.4.1    The CMH Test

The key to the use of the CMH method with survival data is to realize that the data in each time interval can be formulated as a 2 by 2 table. The number of deaths and the number of survivors (the number exposed minus the number of deaths) for the two groups can be put in a 2 by 2 table for each time interval as shown next.

|  | Number of Deaths | Number of Survivors | Total |
|---|---|---|---|
| Group 1 | $d_{1i}$ | $(n'_{1i} - d_{1i})$ | $n'_{1i}$ |
| Group 2 | $d_{2i}$ | $(n'_{2i} - d_{2i})$ | $n'_{2i}$ |
| Total | $d_{.i}$ | $(n'_{.i} - d_{.i})$ | $n'_{.i}$ |

It can be shown that the time intervals are uncorrelated with one another, which allows us to use the CMH statistic here.

Let us consider an example.

---

**Example 11.3**

The Hypertension Detection and Follow-up Program examined the effect of serum creatinine on eight-year mortality among hypertensive persons under care (Shulman et al. 1989). We are interested in testing whether or not the survival experience of persons with a serum creatinine concentration less than 1.7 mg/dL at the time of screening is more favorable than those with a serum creatinine concentration greater than or equal to 1.7 mg/dL. The data for testing this hypothesis are shown in Table 11.6.

Table 11.6   Sample sizes and numbers of deaths by year and level of serum creatinine concentration in the HDFP study.

| Year under Care | Serum Creatinine (mg/dL) | | | | | |
| | <1.7 | | ≥1.7 | | Total | |
| | $d_{1i}$ | $n'_{1i}$ | $d_{2i}$ | $n'_{2i}$ | $d_i$ | $n'_i$ |
|---|---|---|---|---|---|---|
| 0–1 | 93 | 10,469.5 | 21 | 297.0 | 114 | 10,766.5 |
| 1–2 | 115 | 10,374.5 | 16 | 276.0 | 131 | 10,650.5 |
| 2–3 | 125 | 10,254.0 | 13 | 260.0 | 138 | 10,514.0 |
| 3–4 | 181 | 10,121.5 | 14 | 246.5 | 195 | 10,368.0 |
| 4–5 | 160 | 9,930.5 | 17 | 232.0 | 177 | 10,162.5 |
| 5–6 | 212 | 9,763.0 | 10 | 215.0 | 222 | 9,978.0 |
| 6–7 | 191 | 9,551.0 | 14 | 205.0 | 205 | 9,756.0 |
| 7–8 | 203 | 9,147.5 | 8 | 186.5 | 211 | 9,334.0 |
| Total | 1,270 | | 113 | | 1,393 | |

---

First, we use the data in Example 11.3 to estimate the cumulative survival probabilities for the two groups, applying the methods discussed earlier. The estimated cumulative survival probabilities and their standard errors are shown in Table 11.7.

The estimated cumulative survival probabilities are also shown graphically in Figure 11.5. The survival distribution appears to be more favorable for the hypertensive persons

Table 11.7   Cumulative survival probabilities and standard errors by year and level of serum creatinine concentration in the HDFP study.

| Year under Care | Creatinine Level <1.7 | | Creatinine Level ≥1.7 | |
| | Survival Probability | Standard Error | Survival Probability | Standard Error |
|---|---|---|---|---|
| 0–1 | 1.0000 | 0 | 1.0000 | 0 |
| 1–2 | 0.9911 | 0.0009 | 0.9295 | 0.0148 |
| 2–3 | 0.9801 | 0.0014 | 0.8758 | 0.0191 |
| 3–4 | 0.9682 | 0.0017 | 0.8322 | 0.0216 |
| 4–5 | 0.9509 | 0.0021 | 0.7851 | 0.0238 |
| 5–6 | 0.9355 | 0.0024 | 0.7279 | 0.0258 |
| 6–7 | 0.9151 | 0.0027 | 0.6942 | 0.0267 |
| 7–8 | 0.8969 | 0.0028 | 0.6470 | 0.0277 |
| 8 | 0.8770 | 0.0032 | 0.6194 | 0.0282 |
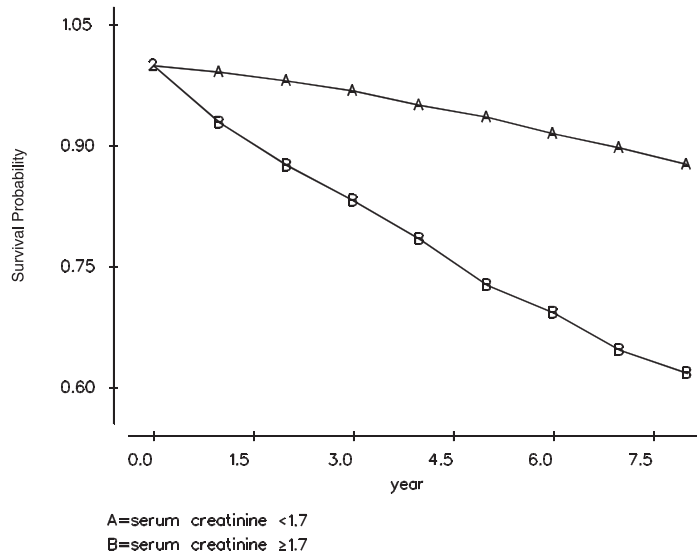
**Figure 11.5** Estimated survival distributions by level of serum creatinine concentration.

A=serum creatinine <1.7
B=serum creatinine ≥1.7

with a serum creatinine concentration less than 1.7 mg/dL than those with a serum creatinine concentration greater than or equal to 1.7 mg/dL. The two survival curves are consistently diverging, suggesting that the odds ratios in each time interval are similar to one another. Therefore, we do not have any problem using the CMH test to compare the two survival distributions.

To apply this test to the data in Table 11.6, we need to find the expected number of deaths and the variance for the (1, 1) cell in each of the eight 2 by 2 tables. For example, the 2 by 2 table for the year 0–1 is shown next.

| Creatinine Level | Number of Deaths | Number of Survivors | Total |
|---|---|---|---|
| <1.7 mg/dL | 93 | 10,376.5 | 10,469.5 |
| ≥1.7 mg/dL | 21 | 276.0 | 297.0 |
| Total | 114 | 10,652.5 | 10,766.5 |

The expected number of deaths in the (1, 1) cell is the product of the total of the first row and the first column divided by the table total. Thus, the expected value is

$$10469.5 \ (114) \ / \ 10766.5 = 110.86.$$

The estimated sample variance of the (1, 1) cell is the product of the four marginal totals divided by the square of the table total times the table total minus one. Thus, the sample variance is

$$\frac{10469.5(297)(114)(10652.5)}{10766.5^2(10766.5-1)} = 3.03.$$

Table 11.8 shows the expected number of deaths and the estimated variances for the eight (1, 1) cells based on the data in Table 11.6. The observed number of deaths in Group 1 (creatinine less than 1.7 mg/dL) is 1280 and the expected number of deaths is 1361, suggesting that Group 1 has a favorable survival experience. We shall test the hypothesis of no difference in the survival distributions of the two groups at the 0.01

Table 11.8   Expected values and variances of the (1, 1) cells.

| Year under Care | Expected Value | Variance |
|---|---|---|
| 0–1 | 110.86 | 3.03 |
| 1–2 | 127.61 | 3.27 |
| 2–3 | 134.59 | 3.28 |
| 3–4 | 190.36 | 4.44 |
| 4–5 | 172.96 | 3.88 |
| 5–6 | 217.22 | 4.58 |
| 6–7 | 200.69 | 4.13 |
| 7–8 | 206.78 | 4.04 |
| Total | 1,361.07 | 30.65 |

significance level. The test statistic, $X^2_{CMH}$, is calculated based on the data in Tables 11.6 and 11.8 as follows:

$$X^2_{CMH} = \frac{(|O - E| - 0.5)}{V} = \frac{(|1289 - 1361.07| - 0.5)^2}{30.65} = 211.80.$$

Since the test statistic is greater than 6.63 ($= \chi^2_{1,0.99}$), we reject the null hypothesis and conclude that persons with a serum creatinine concentration less than 1.7 mg/dL had a more favorable survival distribution than those with a higher creatinine value at the time of screening.

## 11.4.2   The Normal Distribution Approach

The individual survival probabilities of the two groups can be compared using the method discussed in Chapter 8. But this approach has the disadvantage that it focuses on a particular point in time and does not use all the information in the data set. For example, the two-year survival probability of the group with serum creatinine level less than 1.7 is 98 percent compared with 88 percent for the group with serum creatinine level greater than or equal to 1.7. Let us test whether these probabilities are significantly different at the 0.01 level. The test statistic for this comparison can be calculated from the data in Table 11.7 as follows:

$$z = \frac{p_1 - p_2}{\sqrt{[s.e.(p_1)]^2 + [s.e.(p_2)]^2}} = \frac{0.9801 - 0.8758}{\sqrt{0.0014^2 + 0.0191^2}} = 5.45.$$

The $p$-value for the calculated $z$ statistic is 0.0001, which is statistically significant. If after two years the survival experience changed, this test would not provide any information about that change. One could use multiple tests but doing that has the disadvantage of not yielding a single overall test.

## 11.4.3   The Log-Rank Test

The CMH test for the comparison of survival curves is often called the log-rank test because of the similarity of these two test statistics. Peto and Peto's log-rank statistic is based on a set of scores derived from the logarithm of the survival function (Lee 1992; Peto and Peto 1972). Because of its complexity in calculation, researchers often use an approximate log-rank chi-square statistic that is easier to compute (Matthews and Farewell 1985, Chapter 7; Peto et al. 1977). Just as in the CMH approach, the approxi-

mate log rank test is based on the individual 2 by 2 tables, but it looks at the number of deaths and expected number of deaths for each group. The approximate log-rank statistic is calculated by

$$X^2_{LR} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

where $O_1$ is the sum of the observed numbers of deaths across the time points in the (1, 1) cell and $O_2$ is the corresponding sum for the (2, 1) cell. $E_1$ and $E_2$ are the corresponding sums of the expected number of deaths. The approximate log-rank test statistic looks like the goodness-of-fit chi-square statistic. Applying the approximate log-rank test chi-square procedure to the preceding data, we get

$$X^2_{LR} = \frac{(1280 - 1361.07)^2}{1361.07} + \frac{(113 - 31.93)^2}{31.93} = 210.66.$$

It gives practically an identical result to the CMH chi-square value just shown. One advantage of the approximate log-rank test is that it can be extended to more than two group comparisons. The exact calculation of the statistic is more involved than we wish to present in this text, but different software packages often report the exact value.

### 11.4.4  Use of the CMH Approach with Small Data Sets

The CMH test statistic can also be used with a smaller data set along with the product-limit method. Let us consider an example.

**Example 11.4**

We reexamine the data used in Table 11.4 in comparing the survival distributions of male and female patients at the 0.05 significance level. The male and female survival distributions are shown in Figure 11.6. The median survival time for males is about 20 months, and it is 16 months for females.
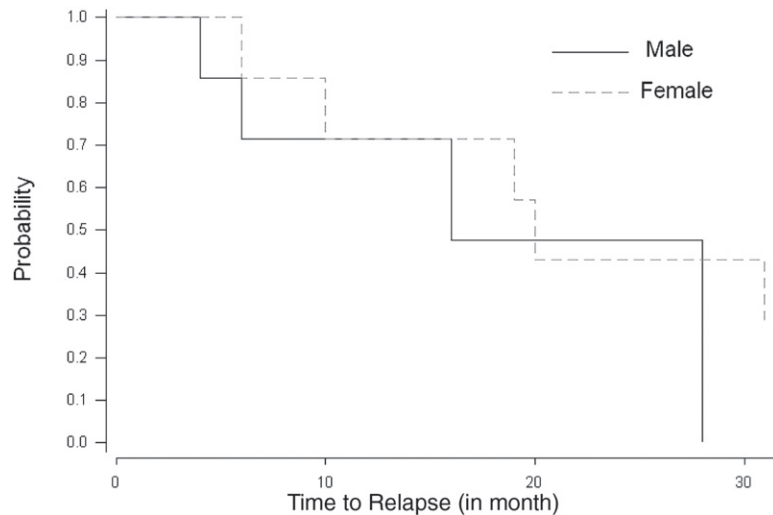


**Figure 11.6**   Estimated survival distributions by gender for the data in Table 11.4.

Table 11.9   Comparison of alcohol-free time distributions for females and males.

| Survival Time (1) $x_i$ | | Number of Subjects | | | Observed No. of Relapses | | | Expected Relapses (8) | Variance (9) |
|---|---|---|---|---|---|---|---|---|---|
| | | Total (2) $n'_i$ | Female (3) $n'_{1i}$ | Male (4) $n'_{2i}$ | Total (5) $d_i$ | Female (6) $d_{1i}$ | Male (7) $d_{2i}$ | | |
| 4 | M | 14 | 7 | 7 | 1 | 0 | 1 | 0.50 | 0.25 |
| 6 | MF | 13 | 7 | 6 | 2 | 1 | 1 | 1.08 | 0.46 |
| 9* | M | 11 | 6 | 5 | 0 | 0 | 0 | 0 | 0 |
| 10 | F | 10 | 6 | 4 | 1 | 1 | 0 | 0.60 | 0.24 |
| 14* | M | 9 | 5 | 4 | 0 | 0 | 0 | 0 | 0 |
| 16 | M | 8 | 5 | 3 | 1 | 0 | 1 | 0.63 | 0.23 |
| 17* | M | 7 | 5 | 2 | 0 | 0 | 0 | 0 | 0 |
| 19 | F | 6 | 5 | 1 | 1 | 1 | 0 | 0.83 | 0.14 |
| 20 | F | 5 | 4 | 1 | 1 | 1 | 0 | 0.80 | 0.16 |
| 28 | M | 4 | 3 | 1 | 1 | 0 | 1 | 0.75 | 0.19 |
| 31 | F | 3 | 3 | 0 | 1 | 1 | 0 | 1.00 | 0.00 |
| 34* | F | 2 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47* | F | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| Total | | | | | 9 | 5 | 4 | 6.19 | 1.67 |

*Censored observations

We wish to determine whether or not there is a significant difference between these two distributions. The data and the calculation of the test statistic for making this comparison are shown in Table 11.9.

The first column of the table shows the observed alcohol-free times ($x_i$) with the censoring status and gender indicated. The second column is the total number of subjects under observation at time $x$. The third and fourth columns show, respectively, the number of females (Group 1) and the number of males (Group 2) under observation at time $x$. The fifth column shows the observed number of relapses at time $x$. The numbers of relapses at time $x$ in Group 1 and in Group 2 are shown in columns 6 and 7, respectively.

The eighth column shows the expected number of relapses at time $x_i$ for females. It is calculated in the same manner as before. For example, at 6 months, two relapses are recorded. The proportion of females under observation at 6 months is 7/13. Therefore, the expected number of relapses for females is $2*(7/13)$, or 1.08. The variances of the observed numbers of relapses for females at time $x_i$ are shown in column 9. These calculations are performed only for the uncensored survival times. The values are next summed and the CMH chi-square statistic is calculated as follows:

$$X^2_{CMH} = \frac{(|O-E|-0.5)^2}{V} = \frac{(|5-6.19|-0.5)^2}{1.67} = 0.29.$$

Since the test statistic is smaller than 3.84 ($= \chi^2_{1,0.95}$), we fail to reject the null hypothesis.

The approximate log-rank chi-square statistic gives

$$X_{LR}^2 = \frac{(5-6.19)^2}{6.19} + \frac{(4-2.81)^2}{2.81} = 0.73.$$

Computer programs provide the exact log-rank chi-square value of 0.84. Although the CMH chi-square value is smaller than the log-rank chi-square values (due mainly to the correction for continuity), we draw the same conclusion. Note that the CMH chi-square without the correction for continuity is 0.85.

In Chapter 10, we indicated that the CMH test statistic should be used only when the odds ratios are similar across the subtables. The same idea applies here, and the plot of the two survival functions gives a rough way of assessing the validity of this assumption. If the assumption is true, the plot of the two survival functions should be roughly parallel. If the lines representing the two survival functions cross one another, this definitely means that the assumption does not hold and the CMH test statistic should not be used. The reason for this is that one group has a better survival experience during part of the study period, and the other group has a better experience during another part of the period. Thus, it is difficult to say that one group has a better overall experience. The log-rank test also has the same requirement.

Comparison of two consistently different survival curves can be done by the computer (see **Program Note 11.3** on the website). Most statistics packages provide the log-rank chi-square and options for creating graphs of the survival functions.

## Conclusion

In this chapter, we presented two methods for analyzing survival data: the life-table and product-limit methods. The life-table method is generally used for large data sets and the product-limit method for smaller data sets. In addition, we demonstrated the calculation of the sample median and restricted mean survival times. We also discussed why the median is preferred to the mean as a single summary statistic for use with survival data. We highly recommended the plotting of the survival distribution for a more complete description of survival data. Finally, we showed the use of the Cochran-Mantel-Haenszel test for comparing the equality of two survival distributions.

### EXERCISES

**11.1** In an effort to understand employment experience of nurses, personnel records of two large hospitals were reviewed (Benedict, Glasser, and Lee 1989). A total of 3221 nurses were hired during a 10-year period from 1970 to 1979 and employment records were reviewed 18 months beyond the end of 1979. In this cohort, only 780 nurses worked more than 33 months. The length of employment was presented by 3-month intervals as follows:

| Month after Employment | Number Terminated | Number Censored | Number at Beginning of Interval |
|---|---|---|---|
| 0–3 | 582 | 0 | 3,221 |
| 3–6 | 369 | 0 | |
| 6–9 | 247 | 0 | |
| 9–12 | 212 | 0 | |
| 12–15 | 182 | 0 | |
| 15–18 | 144 | 0 | |
| 18–21 | 129 | 75 | |
| 21–24 | 99 | 74 | |
| 24–27 | 85 | 59 | |
| 27–30 | 51 | 53 | |
| 30–33 | 45 | 35 | |
| 33+ | 0 | 780 | |

a. Prepare a worksheet for a life-table analysis and estimate the cumulative survival probabilities, the restricted mean, and the median length of employment. Also estimate the probability of termination for each of the intervals.

b. Estimate the standard errors of (1) the estimated cumulative survival probabilities and (2) the probability of termination for each interval.

c. Calculate 95% confidence intervals for (1) the 24-month cumulative survival probability and (2) the probability of termination during the first three months of employment.

d. What additional data, if any, do you need and what further analyses would you perform to assess the nursing employment situation?

**11.2** The Hypertension Detection and Follow-up Program collected mortality data for eight years (Shulman et al. 1989). The following data show the survival experience of two subgroups formed by the level of serum creatinine concentration:

| | Serum Creatinine Concentration (mg/dL) | | | | | |
|---|---|---|---|---|---|---|
| Year | 2.00–2.49 | | | ≥2.5 | | |
| Care | Alive | Died | Censored | Alive | Died | Censored |
| 0–1 | 78 | 3 | 0 | 72 | 8 | 0 |
| 1–2 | 75 | 4 | 0 | 64 | 8 | 0 |
| 2–3 | 71 | 6 | 0 | 56 | 3 | 0 |
| 3–4 | 65 | 3 | 0 | 53 | 3 | 0 |
| 4–5 | 62 | 5 | 0 | 50 | 8 | 0 |
| 5–6 | 57 | 4 | 0 | 42 | 3 | 0 |
| 6–7 | 53 | 2 | 0 | 39 | 5 | 0 |
| 7–8 | 51 | 3 | 3 | 34 | 1 | 1 |
| 8+ | 45 | 0 | 45 | 32 | 0 | 32 |

a. Analyze the survival pattern of each group using the life-table method: Estimate the cumulative survival probabilities and their standard errors, and compare the survival curves of these two groups graphically.

b. If it is appropriate, determine whether or not the two survival distributions are equal at the 0.01 significance level.

c. Comment on what factors may have confounded the preceding comparison and what further analyses you think are necessary before you can draw more defensible conclusions.

**11.3** The SHEP (Systolic Hypertension in the Elderly Program) Cooperative Research Group (1991) assessed the ability of antihypertensive drug treatment to reduce the risk of stroke (nonfatal and fatal) in a randomized, double-blind, placebo-controlled experiment. A total of 4736 persons with systolic hypertension (systolic blood pressure 160 mmHg and above and diastolic blood pressure less than 90 mmHg) were screened from 447,921 elderly persons aged 60 years and above. During the study period, 213 deaths occurred in the treatment group and 242 deaths in the placebo group. The average follow-up period was 4.5 years. Total stroke was the primary end point and the following data were reported:

| | Treatment Group | | | Placebo Group | | |
|---|---|---|---|---|---|---|
| Year | Number Started | Strokes | Lost | Number Started | Strokes | Lost |
| 0–1 | 2,365 | 28 | 0 | 2,371 | 34 | 0 |
| 1–2 | 2,316 | 22 | 0 | 2,308 | 42 | 0 |
| 2–3 | 2,264 | 21 | 0 | 2,229 | 22 | 2 |
| 3–4 | 2,153 | 18 | 0 | 2,193 | 34 | 2 |
| 4–5 | 1,438 | 13 | 5 | 1,393 | 24 | 1 |
| 5–6* | 613 | 1 | 0 | 584 | 3 | 0 |

*The last stroke occurred during the 67th month of follow-up.

a. To analyze the above data by the life-table method, how would you set up the worksheet? It is obvious that there were censored observations other than the lost-to-follow-up, such as deaths and withdrawn alive. This can be seen since the difference in the number of persons starting one interval and the number starting the following interval decreased by more than the number of strokes in the interval. Would you include or exclude the data in the last reported interval?

b. If it is appropriate, test the hypothesis of the equality of the two survival distributions at the 0.05 significance level.

**11.4** A group of 31 patients diagnosed with lymphoma and presenting with clinical symptoms ("B" symptoms) was compared with another group of 33 lymphoma patients diagnosed without symptoms ("A" symptoms) (Mattews and Farewell 1985, page 89). The recorded survival times (in months) for the 64 patients are as follows:

| A symptoms: | 3.2* | 4.4* | 6.2 | 9.0 | 9.9 | 14.4 | 15.8 | 18.5 | 27.6* | 28.5 | 30.1* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 31.5* | 32.2* | 41.0 | 41.8* | 44.5* | 47.8* | 50.6* | 54.3* | 55.0 | 60.0* | 60.4* |
| | 63.6* | 63.7* | 63.8* | 66.1* | 68.0* | 68.7* | 68.8* | 70.9* | 71.5* | 75.3* | 75.7* |
| B symptoms: | 2.5 | 4.1 | 4.6 | 6.4 | 6.7 | 7.4 | 7.6 | 7.7 | 7.8 | 8.8 | 13.3 |
| | 13.4 | 18.3 | 19.7 | 21.9 | 24.7 | 27.5 | 29.7 | 30.1* | 32.9 | 33.5 | 35.4* |
| | 37.7* | 40.9* | 42.6* | 45.4* | 48.5 | 48.9* | 60.4* | 64.4* | 66.4* | | |

Asterisks indicate censored observations.

a. Estimate the survival probabilities, plot the survival curves, and determine whether the use of the CMH or log-rank test is appropriate in comparing the two survival curves.

b. Carry out the test at the 0.01 significance level and interpret the results. How would you interpret the prolonged horizontal survival curve at the end of survival curves in both groups?

**11.5** The following data were abstracted from the records of the neonatal intensive care unit (NICU) in a hospital during the month of February 1993 (day and 24-hour clock time are used to describe the timing of events — e.g., 0102 indicates the first day of February, 2/AM):

| No. | Sex | Born | Last Observed | Status |
|-----|------|------|---------------|--------------|
| 1 | Boy | 0102 | 2210 | Discharged |
| 2 | Girl | 0306 | 1722 | Died |
| 3 | Boy | 0309 | 1517 | Died |
| 4 | Boy | 0523 | 2609 | Discharged |
| 5 | Boy | 0918 | 1001 | Died |
| 6 | Girl | 1004 | 2411 | Died |
| 7 | Boy | 1107 | 2512 | Discharged |
| 8 | Girl | 1110 | 1815 | Discharged |
| 9 | Boy | 1206 | 1408 | Died |
| 10 | Girl | 1307 | 2320 | Died |
| 11 | Girl | 1412 | 2823 | Still in NICU |
| 12 | Boy | 1500 | 1510 | Died |
| 13 | Boy | 1607 | 2220 | Died |
| 14 | Girl | 1819 | 2823 | Still in NICU |
| 15 | Boy | 1903 | 2009 | Died |
| 16 | Boy | 2009 | 2711 | Discharged |
| 17 | Boy | 2110 | 2823 | Still in NICU |
| 18 | Girl | 2208 | 2320 | Died |
| 19 | Girl | 2321 | 2823 | Still in NICU |
| 20 | Girl | 2323 | 2810 | Discharged |
| 21 | Boy | 2402 | 2823 | Still in NICU |
| 22 | Girl | 2509 | 2823 | Still in NICU |
| 23 | Boy | 2620 | 2823 | Still in NICU |
| 24 | Girl | 2701 | 2822 | Died |

    a. Estimate the neonatal survival function for these NICU infants, estimate the median survival time, and form the 90 percent confidence interval for the 50-hour survival probability.

    b. Plot the estimated neonatal survival functions separately for boys and girls and test the equality of the two survival distributions at the 0.10 significance level.

**11.6** Quality of care for colorectal cancer was evaluated by comparing the survival experience of patients in two types of health plans (fee-for-service and health maintenance organization) offered by the same health care provider (Vemon et al. 1992). The following data were generated from the reported survival curves:

| Practice | Survival Times in Months | | | | | | | | | | | |
|----------|----|-----|----|-----|----|-----|----|-----|----|-----|----|----|
| Fee-for- | 2 | 5 | 10 | 12* | 14 | 14 | 16 | 18 | 23 | 26* | 27 | 31 |
| Service | 34 | 37* | 39 | 42* | 46 | 47* | 50 | 53* | | | | |
| HMO | 4 | 10* | 12 | 15 | 19 | 25 | 30* | 35 | 38 | 43* | 49 | 54* |

Asterisks indicate censored observations.

    a. Estimate the survival distributions by the product-limit method and graphically compare the survival curves.

    b. Compare the equality of the survival distributions of the two medical services at the 0.01 significance level.

**11.7** From April 1, 1999, family physicians are required to refer all patients who have suspected breast cancer in the United Kingdom to a hospital to be seen within 14 days of referral. Data from a cancer registry were used to examine whether the survival distributions of different length of delay groups (from referral to treatment) are different (Sainsbury, Johnston, and Haward 1999). Patients diagnosed with breast cancer during the 1986–1990 period were used for this analysis. Of the 9488 patients registered, 5708 had information on dates of referral and treatment. It was stated that "survival curves were estimated by the Kaplan-Meier method." Based on a survival analysis of the following data, the authors concluded that "delays of more than 90 days are unlikely to have an impact on survival and that, if delays can be kept to within this time, efforts to shorten delays further should not have priority."

| | Number of Survivors at the Beginning of the Interval | | | |
| --- | --- | --- | --- | --- |
| | Delay Groups | | | |
| **Years of Survival** | **<30 Days** | **30–59 Days** | **60–89 Days** | **≥90 Days** |
| 0–1 | 3,534 | 1,578 | 345 | 251 |
| 1–2 | 3,113 | 1,490 | 328 | 235 |
| 2–3 | 2,743 | 1,370 | 301 | 217 |
| 3–4 | 2,470 | 1,274 | 275 | 198 |
| 4–5 | 2,235 | 1,182 | 254 | 186 |
| 5–6 | 2,062 | 1,101 | 239 | 176 |
| 6–7 | 1,897 | 1,050 | 225 | 168 |
| 7–8 | 1,769 | 982 | 212 | 157 |
| 8+ | 1,647 | 913 | 199 | 154 |

Assume that there were no censored observations.

a. More than one-half of the data were in the less than 30 days delay group. What are merits and demerits of splitting this group to <15 days and 15–29 days groups?

b. Do you think that the Kaplan-Meier method was appropriate for this analysis?

c. Estimate the survival distributions for <30 days delay group and the ≥90 days delay group, and test whether two survival distributions are significantly different at the 0.05 level.

d. Do you think the author's conclusions are supported by your analysis? Why or why not? What are possible confounders for the difference in survival distributions?

## REFERENCES

Benedict, M. B., J. H. Glasser, and E. S. Lee. "Assessing Hospital Nursing Staff Retention and Turnover: A Life Table Approach." *Evaluation and the Health Professions* 12:73–96, 1989.

California Tumor Registry. *Cancer Registration and Survival in California.* Berkeley, CA: State of California Department of Public Health, 1963, p. 258–259.

Kaplan, E. L., and P. Meier. "Nonparametric Estimation from Incomplete Observations." *Journal of the American Statistical Association* 53:457–481, 1958.

Lee, E. T. *Statistical Methods For Survival Data Analysis*, 2nd ed. New York: Wiley, 1992.

Mantel, N. "Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Considerations." *Cancer Chemotherapy Reports* 50:163–170, 1966.

Matthews, D. E., and V. Farewell. *Using and Understanding Medical Statistics*, *Basel*. New York: Karger, 1985.

National Cancer Institute. *Cancer Statistics Review: 1973–1987*. Bethesda, MD: National Institutes of Health, U.S. Public Health Service, NIH Publication No. 90–2789, 1990.

Peto, R., and J. Peto. "Asymptotically Efficient Rank Invariant Test Procedures (with Discussion)." *Journal of Royal Statistical Society* A 135:185–206, 1972.

Peto, R., M. C. Pike, P. Armitage, N. E. Breslow, D.R. Cox, S.V. Howard, N. Mantel, K. McPherson, J. Peto, and P. G. Smith. "Design and Analysis of Randomized Clinical Trials Requiring Prolonged Observation of Each Patent, II: Analysis and Examples." *British Journal of Cancer* 35:1–39, 1977.

Sainsbury, R., C. Johnston, and B. Haward. "Effect on Survival of Delays in Referral of Patients with Breast-cancer Symptoms: A Retrospective Analysis." *Lancet* 353:1132–1135, 1999.

Shulman, N. B., C. E. Ford, W. D. Hall, M. D. Blaufox, D. Simon, H. G. Langford, and K. A. Schneider. On behalf of the Hypertension Detection and Follow-up Program Cooperative Group. "Prognostic Value of Serum Creatinine and Effect of Treatment of Hypertension on Renal Function." *Hypertension* 13 (Supplement I):80–93, 1989.

Systolic Hypertension in the Elderly Program Cooperative Research Group. "Prevention of Stroke by Antihypertenstive Drug Treatment in Older Persons with Isolated Systolic Hypertension." *Journal of American Medical Association* 265:3255–3264, 1991.

Thomas, D. R., and G. L. Grunkemeier. "Confidence Interval Estimation of Survival Probabilities for Censored Data." *Journal of the American Statistical Association* 70:865–871, 1975.

Vernon, S. W., J. I. Hughes, V. M. Heckel, and G. L. Jackson. "Quality of Care for Colorectal Cancer in a Fee-for-Service and Health Maintenance Organization Practice." *Cancer* 69:2418–2425, 1992.