Applied Survival Analysis

Regression Modeling of Time to Event Data

DAVID W. HOSMER, Jr.

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

STANLEY LEMESHOW

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

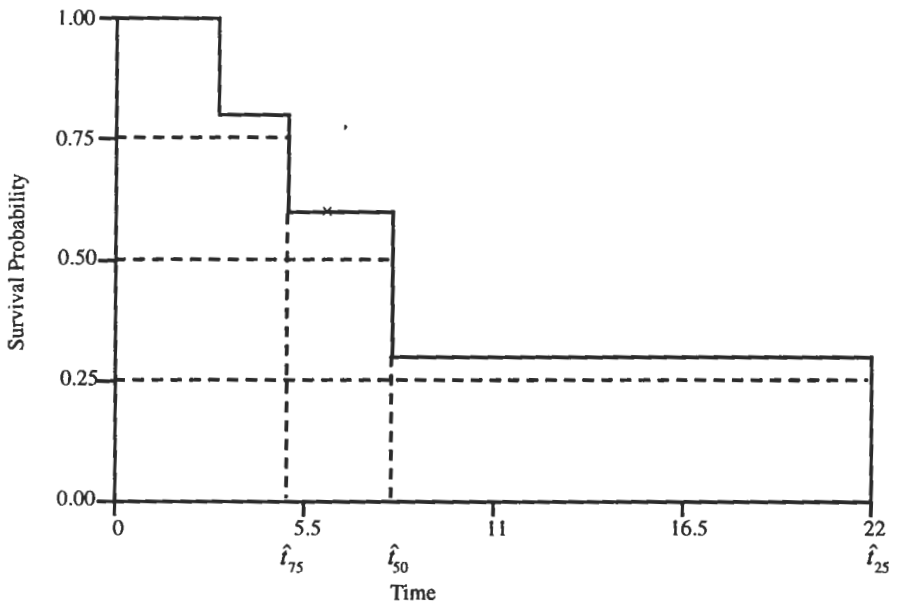

Figure 2.6 Kaplan–Meier estimate of the survivorship function for the data in Table 2.1 and graphically determined estimates of the quartiles.

ion at the riser connecting steps ending, respectively (looking right to left), at 8 and 5 months. The vertical line hits at exactly 8 months. Thus the estimated median survival time in this example is $\hat{t}_{50} = 8$. A formula to describe this estimator is

$$\hat{t}_{50} = \min\{t : \hat{S}(t) \leq 0.50\}.$$

The formula says to proceed as if you are walking up a set of stairs from the right to the riser where the horizontal line hits. The estimate is the time value defining the left-most point of the step you're standing on. If we assume that the riser is attached at the top and bottom, then the description also works when the horizontal line hits one of the steps. The estimate is, again, the value of time defining the left-most point of the step. In general, the estimate of the pth percentile is

$$\hat{t}_p = \min\{t : \hat{S}(t) \leq (p/100)\}.$$

The estimates of the other quartiles from Table 2.2 are $\hat{t}_{75} = 5$ and $\hat{t}_{25} = 22$.

For the full data set for the HMO-HIV+ study, the estimates of the three quartiles are $\hat{t}_{75}=3$, $\hat{t}_{50}=7$ and $\hat{t}_{25}=15$. The interpretation of these values is that we estimate that 75 percent will live at least three months, half are estimated to live at least 7 months, and only 25 percent are estimated to live at least 15 months.

A confidence interval estimate for the quantiles can add further understanding about possible values for the parameter being estimated. Approximate confidence intervals may be obtained by appealing to the theory that, for large samples, the quantile estimator is normally distributed with mean equal to the quantile being estimated. An estimator of the variance of this distribution may be obtained from an application of the delta method, as outlined in Collet (1994) and discussed in greater detail from the counting process approach in Andersen, Borgan, Gill and Keiding (1993). The suggested estimator for the variance of the estimator of the pth percentile is

$$\hat{\mathrm{Var}}\left(\hat{t}_p\right)=\frac{\hat{\mathrm{Var}}\left(\hat{S}(\hat{t}_p)\right)}{\left[\hat{f}(\hat{t}_p)\right]^2}. \tag{2.11}$$

The numerator of (2.11) is Greenwood's estimator and the denominator is an estimator of the density function of the distribution of survival time. The estimator of the density function used by many software packages is

$$\hat{f}(\hat{t}_p)=\frac{\hat{S}(\hat{u}_p)-\hat{S}(\hat{l}_p)}{\hat{l}_p-\hat{u}_p}. \tag{2.12}$$

The values $\hat{u}_p$ and $\hat{l}_p$ are chosen such that $\hat{u}_p<\hat{t}_p<\hat{l}_p$ and most often are obtained from the equations shown below:

$$\hat{u}_p=\max\left\{t:\hat{S}(t)\geq(p/100)+0.05\right\} \text{ and } \hat{l}_p=\min\left\{t:\hat{S}(t)\leq(p/100)-0.05\right\}. \tag{2.13}$$

While values other than 0.05 could have been used in (2.13), 0.05 seems to work well in practice and is used by a number of statistical packages. The endpoints of a $100(1-\alpha)$ percent confidence interval are

$$\hat{t}_p \pm z_{1-\alpha/2}\hat{\mathrm{SE}}(\hat{t}_p), \tag{2.14}$$

where $\hat{\mathrm{SE}}(\hat{t}_p) = \sqrt{\hat{\mathrm{Var}}(\hat{t}_p)}$.

Evaluation of (2.11) through (2.14) is most easily illustrated with an example. In the HMO-HIV+ study, the estimated median survival time is $\hat{t}_{50} = 7$ months. The value of $\hat{u}_{50}$ is the largest value of time, t, such that $\hat{S}(t) \geq 0.55$. After sorting on survival time and listing the values of the Kaplan–Meier estimator, we find that $\hat{S}(5) = 0.56$ and $\hat{S}(6) = 0.54$, hence $\hat{u}_{50} = 5$. The value of $\hat{l}_{50}$ is the smallest value of t, time, such that $\hat{S}(t) \leq 0.45$. From the same listing we find that $\hat{S}(7) = 0.47$ and $\hat{S}(8) = 0.42$, hence $\hat{l}_{50} = 8$. Thus the estimate of the density function in (2.12) is

$$\hat{f}(\hat{t}_{50}) = \frac{\hat{S}(5) - \hat{S}(8)}{8-5} = \frac{0.56 - 0.42}{8-5} = 0.0467.$$

The value of Greenwood's estimator at $t = 7$ months is

$$\hat{\mathrm{Var}}\left(\hat{S}(7)\right) = 0.002672$$

and evaluation of (2.11) yields

$$\hat{\mathrm{Var}}\left(\hat{t}_{50}\right) = \frac{0.00267}{[0.0467]^2} = 1.224.$$

The end points of the 95 percent confidence interval for median survival time are

$$7 \pm 1.96 \times \sqrt{1.224} = (4.8,\ 9.2).$$

Table 2.5 Estimated Quartiles, Estimated Standard Errors and 95% Confidence Intervals for Survival Time in the HMO-HIV+ Study

Quantile	Estimate	Std. Err.	95% CIE
75	3	0.59	1.8, 4.2
50	7	1.11	4.8, 9.2
25	15	7.45	1.4, 29.6

Table 2.5 presents the estimated survival times for the quartiles, their estimated standard errors, and 95 percent confidence intervals. The results in Table 2.5 further quantify our previous observation of many early deaths with a few at nearly the maximum of follow-up. We note that the confidence interval is quite wide for the 25th percentile. After 15 months only 17 subjects remained at risk. The lack of precision in the confidence interval estimate for this percentile is due to the smaller number of subjects at risk. In general, the right tail of the survivorship function is estimated with considerably less precision than the left tail.

The confidence interval estimator in (2.14) requires that we compute an estimator of the density function at the estimator of the quantile, and the endpoints depend on the assumption that the distribution of the estimated quantile is normal. The sensitivity of the confidence interval to the choice of estimator of the density and the assumption of normality has not been studied. Brookmeyer and Crowley (1982) proposed an alternative method which does not require estimation of the density function [this is discussed in general terms in Andersen, Borgan, Gill and Keiding (1993)]. In this method, the confidence interval for a quantile consists of the values t such that

$$\frac{\left|\hat{S}(t) - p/100\right|}{\hat{\mathrm{SE}}\left(\hat{S}(t)\right)} \leq z_{1-\alpha/2}.$$

The expression on the left side is a test statistic for the hypothesis $H_o: S(t) = p/100$. The confidence interval is the set of values of t for which we would fail to reject the hypothesis. In other words, it is the set of observed survival times for which the confidence interval estimates for the survivorship function contain the quantile. This interval may be determined graphically in a manner similar to Figure 2.6 by drawing a horizontal line from $p/100$ to where it intersects the step functions defining the upper and lower pointwise confidence intervals. The endpoints of the confidence interval are found by drawing vertical lines down to the time axis. If the software package provides the capability to list the endpoints of the confidence intervals for the estimated survivorship function, then the upper and lower endpoints can be precisely determined. Alternative test statistics based on transformations of the estimated survivorship function, such as the log-log transformation, could be used equally well. Brookmeyer and Crowley recommend that this interval be used when there are no tied survival times. The data from the HMO-HIV+ study contain many tied survival times and thus it

would be inappropriate to use the Brookmeyer–Crowley limits in a definitive analysis. However, these data may be used to illustrate the calculations for the median survival time.

Table 2.6 lists the values of the estimated survivorship function and the endpoints of 95 percent confidence intervals determined by the log-log transformation for survival times around the median value of 7 months.

In Table 2.6, we see that the confidence interval estimate at 4 months does not contain 0.5, while at 5 months it does contain 0.5. Thus the lower endpoint of the Brookmeyer–Crowley interval is 5 months. We see that the confidence interval at 9 months does not contain 0.5, while the interval at 8 months does contain it. Hence, the upper limit is 8 months. Brookmeyer–Crowley limits could be determined in a similar manner for other quantiles, though those for the median are most often calculated and reported by software packages. The Brookmeyer–Crowley confidence interval for the median of (5, 8) is comparable to the interval (4.8, 9.2) from Table 2.5, which was based on the large sample distribution of the estimator of the median.

In the analysis of survival time, the sample mean is not as important a measure of central tendency as it is in other settings. (The exception is in fully parametric modeling of survival times when the estimator of the mean, or a function of it, provides an estimator of a parameter vital to the analysis and interpretation of the data. We discuss parametric modeling in Chapter 8.) This is due to the fact that censored survival time data are most often skewed to the right and, in these situations, the median usually provides a more intuitive measure of central tendency. For the sake of completeness, we describe how the estimator of the mean

Table 2.6 Listing of Observed Survival Times, the Estimated Survivorship Function and Individual 95% Confidence Limits for Values of Time near the Estimated Median Survival Time of 7 months for the HMO-HIV+ Data

Time	Estimate	95% CIE
4	0.64	0.54, 0.66
5	0.56	0.46, 0.66
6	0.54	0.43, 0.64
7	0.47	0.36, 0.57
8	0.42	0.32, 0.52
9	0.39	0.28, 0.49

and the estimator of its variance are calculated and illustrate their use with examples from the HMO-HIV+ study.

Computational questions arise if the largest observation is censored, in which case one has two choices: (1) Use only the observed survival times (in which case the estimator is biased downwards) or (2) use all observations (in which case one "pretends" that the largest observation was actually a survival time, but the estimator is interpreted conditionally on the observed range). There is no uniform agreement on which is the best approach. For example, SAS (PROC LIFETEST) uses the former approach while BMDP (1L) uses the latter approach. In the absence of censoring, both approaches yield the usual arithmetic mean.

The estimator used for the mean is obtained from a mathematical result which states that, for a positive continuous random variable, the mean is equal to the area under the survivorship function. From mathematical methods of calculus this may be represented as the integral of the survivorship function over the range, that is,

$$\mu = \int_0^\infty S(u)\, du.$$

If we restrict the variable to the interval $[0,t^*]$, then the mean of the variable in this interval is

$$\mu(t^*) = \int_0^{t^*} S(u)du.$$

The estimator is obtained by using the Kaplan–Meier estimator of the survivorship function. The reason for restricting the range over which the mean is calculated is that the Kaplan–Meier estimator is undefined beyond the largest value of time. The value of t^* used depends on which of the two previously described approaches is chosen. Recall that the observed ordered survival times are denoted $t_{(i)}$, $i = 1,\ldots,m$. We denote the largest observed value of time in the sample as $t_{(n)}$. The two approaches to calculating the estimator of the mean correspond to defining $t^* = t_{(m)}$, that is, using the interval $[0,t_{(m)}]$, or defining $t^* = t_{(n)}$, i.e., using the interval $[0,t_{(n)}]$. In situations where the largest observed value of time is an observed failure time, the two approaches yield identical estimators.

The value of the estimator is the area under the step function defined by the Kaplan–Meier estimator and the particular interval chosen.

To illustrate the calculation, consider the data in Table 2.1 for which the estimated survivorship function is presented in Table 2.2 and is graphed in Figure 2.1. In this example, the largest observed value of time is 22 months and it represents a survival time. Thus, the value of the estimated mean is the area under the step function shown in Figure 2.1. This area is the sum of the areas of four rectangles defined by the heights of the four steps and the four observed survival times. The actual calculation is performed as follows (refer to Table 2.2):

$$\begin{aligned}\hat{\mu}(22) &= 1.0\times[3-0]+0.8\times[5-3]+0.6\times[8-5]+0.3\times[22-8]\\ &= 10.6.\end{aligned}$$

This is the value which would be reported by both BMDP and SAS.

For sake of illustration, suppose that the value recorded at 22 months was a censored observation. If we use the interval [0,22] (BMDP's method), we would report the estimated mean as $\hat{\mu}(22)=10.6$. If we use the interval [0,8] (SAS's method), then we would report the estimated mean as $\hat{\mu}(8)=6.4$. This is the area of the first three rectangles in Figure 2.1. In this example, the two estimates of the mean are quite different since the largest observation, 22 months, is much larger than the largest observed survival time, 8 months.

The equation defining the estimator based on the observed range of survival times only is

$$\hat{\mu}(t_{(m)}) = \sum_{i=1}^{m} \hat{S}(t_{(i-1)})\left(t_{(i)} - t_{(i-1)}\right), \tag{2.15}$$

where $\hat{S}(t_{(0)})=1.0$ and $t_{(0)}=0.0$. The equation defining the estimator for the entire observed range of data is

$$\hat{\mu}(t_{(n)}) = \hat{\mu}(t_{(m)}) + (1-c_{(n)})\hat{S}(t_{(m)})(t_{(n)} - t_{(m)}), \tag{2.16}$$

where $c_{(n)}$ denotes the censoring status, (0, 1), of this observation. Each term in the summation in (2.15) denotes the calculation of the area of one of the rectangles defined by the Kaplan–Meier estimator and two observed times. Note that the estimators in (2.15) and (2.16) are identical when the largest observation and the largest observed survival time are equal.

We recommend that the estimator based on the entire observed range of the data (2.16) be used since the one based on the observed

range of survival times (2.15) does not use the information on survival available in times larger than the largest survival time. We note that if those observations that are long and censored had actually been observed survival times, then the estimated mean survival time would have been increased substantially. However, there may be situations (e.g., when there is considerable uncertainty in measuring the longest censored time [$t_{(n)}$ in (2.16)], when the estimator based on survival times only is preferred.

The estimator of the variance of the sample mean is neither particularly intuitive nor easy to motivate, so we just provide it and demonstrate the calculation. In the case of no censored data, it reduces to the usual "sample variance divided by the sample size" estimator. Andersen, Borgan, Gill and Keiding (1993) present a mathematical derivation of the estimator of the mean and its variance, as well as results which show that the standard normal distribution may be used to form a confidence interval estimator. The equation defining the estimator of the variance of the sample mean computed using (2.15) is as follows:

$$\hat{\mathrm{Var}}(\hat{\mu}(t_{(m)})) = \frac{n_d}{n_d - 1} \sum_{i=1}^{m-1} \frac{A_i^2 d_i}{n_i(n_i - d_i)}, \tag{2.17}$$

where $n_d = \sum_{i=1}^{m} d_i$ denotes the total number of subjects with an observed survival time and

$$A_i = \sum_{j=i}^{m-1} \hat{S}(t_{(j)})(t_{(j+1)} - t_{(j)}).$$

The estimator of the variance using (2.16) is obtained by "pretending" that the largest observed time is an observed survival time for purposes of the summation in (2.17), but n_d is not changed. An example will help distinguish between the two cases. The data in Table 2.1 yielded an estimated mean $\hat{\mu}(22) = 10.6$. Evaluation of the estimator in (2.17) yields

$$\begin{aligned} \hat{\mathrm{Var}}[\hat{\mu}(22)] &= \frac{4}{4-1}\left[\frac{7.6^2}{5(5-1)} + \frac{6.0^2}{4(4-1)} + \frac{4.2^2}{2(2-1)}\right] \\ &= 19.61 \end{aligned}$$

where

$$7.6 = A_1 = 0.8(5-3) + 0.6(8-5) + 0.3(22-8)\,,$$
$$6.0 = A_2 = 0.6(8-5) + 0.3(22-8)$$

and

$$4.2 = A_3 = 0.3(22-8)\,.$$

Assume for the moment that the largest value, 22 months, is a censored observation and that we use (2.16) to estimate the mean. Then the estimate of the variance is

$$\hat{\mathrm{Var}}[\hat{\mu}(22)] = \frac{3}{3-1}\left[\frac{7.6^2}{5(5-1)} + \frac{6.0^2}{4(4-1)} + \frac{4.2^2}{2(2-1)}\right]$$
$$= 22.06\,.$$

If we restrict estimation of the mean to observed survival times and estimate the mean using (2.15), then the estimate of the variance obtained by evaluating (2.17) is

$$\hat{\mathrm{Var}}[\hat{\mu}(8)] = \frac{3}{3-1}\left[\frac{3.4^2}{5(5-1)} + \frac{1.8^2}{4(4-1)}\right]$$
$$= 1.27\,,$$

where

$$3.4 = A_1 = 0.8(5-3) + 0.6(8-5)$$

and

$$1.8 = A_2 = 0.6(8-5)\,.$$

Approximate confidence intervals are obtained using percentiles from the standard normal distribution. Using the data in Table 2.1, the endpoints of a 95 percent confidence interval are $10.6 \pm 1.96\sqrt{19.61}$. This is shown only for purposes of illustration since the sample size is only five with four survival times and any asymptotic theory will not hold. In practice, the estimated mean and its estimated standard error would typically be included in the table containing the estimates of the key quantiles and their estimated standard errors.

For the whole HMO-HIV+ study the estimate of the mean using all of the observed times is $\hat{\mu}(60) = 14.67$ and the estimated variance from (2.17) is 3.93, yielding a 95 percent confidence interval of (10.78,

18.56). We note that, in this example, the largest survival time was 58 months and $\hat{\mu}(58) = 14.59$. Thus, the means from the two approaches are not too different. The right skewness evident in the plot of the survivorship function shown in Figure 2.2 is further quantified by the difference between the estimate of the median (7 months) and the estimate of the mean (approximately 15 months). In these data, as is the case with most analyses of survival time, the median is the better measure of central tendency.

2.4 COMPARISON OF SURVIVORSHIP FUNCTIONS

After providing a description of the overall survival experience in the study, we usually turn our attention to a comparison of the survivorship experience in key subgroups in the data. These groups might be defined by treatment arms in a clinical trial or by other key factors thought to be related to survival. The goals in this analysis are identical to those of the two sample *t*-test, the nonparametric rank sum test and the one-way analysis of variance. Namely, we wish to quantify differences between groups through point and interval estimates of key measures. Standard statistical procedures, such as those named above, may be used without modification when there are no censored observations.

Since survival data are typically right skewed, we would likely use rank-based non-parametric tests followed by estimates and confidence intervals of medians (and possibly other quantiles) within groups. Modifications of these procedures are required when censored observations are present in the data. These tests are described and illustrated with the HMO-HIV+ study data beginning with methods for comparing two groups.

When comparing groups of subjects, it is always a good idea to begin with a graphical display of the data in each group. In studies of survival time, we should graph the Kaplan–Meier estimator of the survivorship function for each of the groups. In the HMO-HIV+ study, a variable thought to be related to the survival experience of the subjects was a history of IV drug use, coded 0 = No and 1 = Yes. Figure 2.7 presents the graphs of the estimated survivorship functions for these two groups of subjects.

Both groups show a similar pattern of survival: a rapidly descending survivorship function with a long right tail. This is the result of a number of early deaths and a few subjects with survival near the maximum follow-up time. Since the estimated survivorship functions do not go to

The derivation and algebraic representation of the tests can, at times, seem complex and confusing. Lawless (1982) presents a concise summary of the traditional approach to the development of these tests, based on the theory of nonparametric tests, using exponentially ordered scores. However, in recent years, these tests have been reexamined from the counting process point of view and have been shown to be special cases of a more general class of counting process based tests. These results are summarized in Andersen, Borgan, Gill and Keiding (1993).

The calculation of each test is based on a contingency table of group by status at each observed survival time, as shown in Table 2.7. In this table, the number at risk at observed survival time $t_{(i)}$ is denoted by n_{0i} in Group 0 and by n_{1i} in group 1; the number of observed deaths in each of the these two groups is denoted by d_{0i} and d_{1i}, respectively; the total number at risk is denoted by n_i; and the total number of deaths is denoted by d_i. The contribution to the test statistic at each time is obtained by calculating the expected number of deaths in group 1 or 0, assuming that the survivorship function is the same in each of the two groups. This yields the usual *row total times column total divided by grand total* estimator. For example, using group 1, the estimator is

$$\hat{e}_{1i} = \frac{n_{1i} d_i}{n_i}. \tag{2.18}$$

Most software packages base their estimator of the variance of d_{1i} on the hypergeometric distribution, defined as follows:

$$\hat{v}_{1i} = \frac{n_{1i} n_{0i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}. \tag{2.19}$$

Table 2.7 Table Used for Test of Equality of the Survivorship Function in Two Groups at Observed Survival Time $t_{(i)}$

Event/Group	1	0	Total
Die	d_{1i}	d_{0i}	d_i
Not Die	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At Risk	n_{1i}	n_{0i}	n_i

The contribution to the test statistic depends on which of the various tests is used, but each may be expressed in the form of a ratio of weighted sums over the observed survival times. These tests may be defined in general as follows:

$$Q = \frac{\left[\sum_{i=1}^{m} w_i(d_{1i} - \hat{e}_{1i})\right]^2}{\sum_{i=1}^{m} w_i^2 \hat{v}_{1i}}. \tag{2.20}$$

Under the null hypothesis that the two survivorship functions are the same, and assuming that the censoring experience is independent of group, and that the total number of observed events and the sum of the expected number of events is large, then the significance level for Q may be obtained using the chi-square distribution with one degree-of-freedom [i.e., $p = \Pr(\chi^2(1) \geq Q)$]. Exact methods of inference for use with small samples have been implemented in the software package StatXact 3 (1995) but will not be discussed in this text.

The most frequently used test is based on weights equal to one, $w_i = 1$. In this case, the test mimics the well-known Mantel–Haenszel test of the hypothesis that the stratum specific odds-ratio is equal to one [see Mantel (1966) for further details]. However, this test is most often called the log-rank test, due to Peto and Peto (1972). The test is related to a test proposed by Savage (1956) for noncensored data, and BMDP calls it the generalized Savage test.

Gehan (1965) and Breslow (1970) generalized the Wilcoxon rank-sum test to allow for censored data. This test uses weights equal to the number of subjects at risk at each survival time, $w_i = n_i$, and is called the Wilcoxon or generalized Wilcoxon test by most software packages.

SAS's lifetest procedure provides two ways of obtaining the same test, but different variance estimators are used. In SAS, if we define the grouping variable to be a stratification variable, the variance estimator $\hat{v}_{1i}$ is used. If we use SAS's test option, then the variance estimator

$$\hat{v}_{1i}^{*} = \frac{n_{1i} n_{0i}}{n_i^2}$$

is used, which assumes that $d_i = 1$; there are no tied failure times. Thus, in any one example, we may obtain test statistics of similar magnitude

but with slightly different values. Because survival time is often recorded in discrete units that may lead to ties, we recommend that the variance estimator $\hat{v}_{1i}$ be used.

The choice of weight influences the type of differences in the survivorship function the test is most apt to detect. The generalized Wilcoxon test, since it uses weights equal to the number at risk, will put relatively more weight on differences between the survivorship functions at smaller values of time. The log-rank test, since it uses weights equal to one, will place more emphasis than does the generalized Wilcoxon test on differences between the functions at larger values of time. Other tests have been proposed that use weight functions intermediate between these, for example, Tarone and Ware (1977) suggested using $w_i = \sqrt{n_i}$.

Peto and Peto (1972) and Prentice (1978) suggested using a weight function that depends more explicitly on the observed survival experience of the combined sample. The weight function is a modification of the Kaplan-Meier estimator and is defined in such a way that its value is known just prior to the observed failure. The value of any estimated survivorship function at a particular observed failure time is known only after the observation is made. The property of having the value known in advance of the actual observed failure is referred to as *predictable* in counting process terminology. This theory is needed to prove results concerning the distribution of the test statistics. The modified estimator of the survivorship function is

$$\tilde{S}(t) = \prod_{t_{(j)} \le t} \left(\frac{n_j + 1 - d_j}{n_j + 1} \right) \tag{2.21}$$

and the weight used is

$$w_i = \tilde{S}(t_{(i-1)}) \times \frac{n_i}{n_i + 1}. \tag{2.22}$$

Note that when $d_i = 1$ the weight is equal to the modified estimator, that is, $w_i = \tilde{S}(t_{(i)})$, which is an assumption made in the implementation of this test in BMDP. In the example demonstrating the calculations, we will use both the correct version of the weight given in (2.22) as well as BMDP's implementation. In subsequent examples, only the BMDP version of the Peto–Prentice test will be discussed, as it is the only software package providing this test.

Harrington and Fleming (1982) suggested a class of tests that incorporates features of both the log-rank and the Peto and Prentice tests. They suggest using the Kaplan–Meier estimator raised to a power, as the weight, namely

$$w_i = \left[\hat{S}(t_{(i-1)})\right]^{\rho}.$$

If the power is $\rho = 0$ then $w_i = 1$ and the test is the log-rank test. However, if $\rho = 1$ then the weight is the Kaplan–Meier estimator at the previous survival time, a weight similar to that of the Peto and Prentice test. This test has been implemented in the S-PLUS software package.

The principle advantage of the Peto–Prentice and Harrington-Fleming tests over the generalized Wilcoxon test is that they weight relative to the overall survival experience. The generalized Wilcoxon test uses the size of the risk set and hence weights depend both on the censoring as well as the survival experience. If the pattern of censoring is markedly different in each of the groups, then this test may either reject or fail to reject, not on the basis of similarity or differences in the survivorship functions, but on the pattern of censoring. For this reason most software packages will provide information as to the pattern of censoring in each of the two groups. This information should be checked for comparability—especially when the results of several of these tests are provided and yield markedly different significance levels.

A problem can occur if the estimated survivorship functions cross one another. This means that in some time intervals one group will have a more favorable survival experience, while in other time intervals the other group will have the more favorable experience. This situation is analogous to having interaction present when applying Mantel–Haenszel methods to a stratified contingency table. Unfortunately, tests for the homogeneity across strata may not be used in most survival time applications, because data in tables like Table 2.7 will be too thin to satisfy the necessary large sample criteria. Fleming, Harrington and O'Sullivan (1987) proposed a test that addresses the problem by using, as a test statistic, the maximum observed difference between the two survivorship functions. This test has not been implemented in any software package. We consider methods based on regression modeling to address this issue in Chapter 7. For the time being, our only check is via a visual examination of the plot of the Kaplan–Meier estimator for the two groups being compared. If one or more of the various tests fails to reject a difference, and if we see that the curves cross, then this "interaction" may be present.

It is not possible to provide a categorical rank ordering of the values of the test statistics. The actual calculated values will depend on the observed survival and censoring times.

In order to illustrate the computation of each of the tests, we have chosen a small subset of subjects in each of the two drug use groups in the HMO-HIV+ study. These data are listed in Table 2.8. Column 1 of Table 2.9 lists the eight distinct survival times. Columns 2 through 5 present the quantities defined by the notation shown in Table 2.7, and columns 6 and 7 present quantities defined in equations (2.18) and (2.19). Columns 8 through 11 present values for the weight functions for the four tests, where "LR" stands for log-rank test weights, "WL" stands for generalized Wilcoxon test weights, "TW" stands for Tarone–Ware weights and "PP" stands for Peto–Prentice weights. The calculated values of the test statistics and their respective p-values are shown in Table 2.10. The difference between the values of the log-rank and generalized Wilcoxon tests in Table 2.10 reflects the fact that the two groups differed most at the later observed survival times. The significance levels in Table 2.10 are provided only for the purpose of illustrating the calculations since, with only 4 events in each group and an expected number of events in group 1 of 5.45, the assumption that the sample sizes are large is a bit tenuous.

Recall the Kaplan–Meier estimates of the survivorship functions for the two drug groups in the whole HMO-HIV+ study, shown in Figure 2.7. Note that the two curves do not cross at any point, indicating that the previously described problem of "interaction" may not be present. An inspection of the proportion of values that are censored and the pattern of censoring (not shown) indicates that the censoring experience of the two groups is similar. Thus it would appear that the assumptions necessary for using the tests for equality of the survivorship functions seem to hold. Table 2.11 presents the values of the test statistics.

In Table 2.11, all tests are highly significant and support the impression from Figure 2.7 that those with a prior history of drug use tended

Table 2.8 Listing of Data from the Two Drug Use Groups in the HMO-HIV+ Study Used to Illustrate the Tests for the Comparison of Two Survivorship Functions

Drug Use Group	Ordered Observed Survival Times
No	3, 4*, 5, 22, 34
Yes	2, 3, 4, 7*, 11

* Denotes a censored observation.

Table 2.9 Listing of Quantities Needed to Calculate the Tests for the Equality of Two Survivorship Functions

							Weights			
Time	d_{1i}	n_{1i}	d_i	n_i	$\hat{e}_{1i}$	$\hat{v}_{1i}$	LR	WL	TW	PP
2	0	5	1	10	0.500	0.250	1	10	3.16	0.909
3	1	5	2	9	1.110	0.432	1	9	3.00	0.818
4	0	4	1	7	0.571	0.245	1	7	2.64	0.636
5	1	3	1	5	0.600	0.240	1	5	2.23	0.530
11	0	2	1	3	0.667	0.222	1	3	1.73	0.398
22	1	2	1	2	1.000	0	1	2	1.41	0.265
34	1	1	1	1	1.000	0	1	1	1.00	0.133

to die sooner than those who did not have a history of drug use. In practice, one could provide additional support for this conclusion by presenting the estimates of the within-group median survival times along with confidence interval estimates.

Each of the tests used to compare the survivorship experience in two groups may be extended to compare more than two groups. For example, the survivorship experience of three or four racial groups could be compared. In the HMO-HIV+ study, it was hypothesized that age might be related to survival. Since age is a continuous variable, one approach to assessing a potential relationship is to use regression modeling. This is discussed in detail in Chapter 3. An approach used in practice, for preliminary analyses that can yield easily understood summary measures, is to break a continuous variable into several groups of interest and use methods for grouped data on the categorized variable. We use this approach with groups based on the following intervals for age: {[20 – 29], [30 – 34], [35 – 39], [40 – 54]}. Table 2.12 presents the number of subjects, the number of deaths, the median survival time and

Table 2.10 Listing of the Test Statistics and *p*-Values for the Equality of Two Survivorship Functions Computed from Table 2.9

Statistic	Value	*p*-Value
Log-rank	1.512	0.219
Generalized Wilcoxon	1.250	0.264
Tarone–Ware	1.363	0.243
Peto-Prentice (Correct wt.)	1.327	0.249
Peto-Prentice (BMDP)	1.423	0.233

Table 2.11 Test Statistics and *p*-Values for the Equality of the Survivorship Functions for the Two Drug Use Groups in the HMO-HIV+ Study

Statistic	Value	*p*-Value
Log-rank	11.856	<0.001
Generalized Wilcoxon	10.910	<0.001
Tarone–Ware	12.336	<0.001
Peto–Prentice (BMDP)	11.497	<0.001

associated 95 percent confidence interval for each age group.

The estimated median survival time is 43 months for the youngest age group in Table 2.12, which is considerably larger than the estimated median in each of the other three groups. This suggests that these young subjects may have a more favorable survival experience than older subjects. However, the estimated standard error of the estimated median is 32.8 and the symmetric normal theory confidence interval covers the entire observed range of time. This problem arises because there are only 12 subjects in this age group, the minimum value of the estimated survivorship function is 0.24 at 58 months and the largest observations are two censored values at 60 months. The medians and confidence intervals for the other three groups suggest that survival experience worsens with age. The goal in the four-group comparison will be to evaluate whether trends seen in the medians persist when the entire survival experience of the groups is compared. Before presenting the graphs of the Kaplan–Meier estimates of the survivorship functions for the four age groups, we present the details of the extension of the two-group tests to the multiple-group situation.

If we assume that there are K groups, then the calculations of the test statistics are based on a two by K table for each observed survival time. The general form of this table is presented in Table 2.13. In a manner

Table 2.12 Number of Subjects, Events and Estimated Median Survival Time in Four Age Groups in the HMO-HIV+ Study

Age Group	Freq	Deaths	Median	95% CIE
20–29	12	8	43	*
30–34	34	29	9	6.3, 11.7
35–39	25	20	7	4.5, 9.5
40–54	29	23	4	2.5, 5.5

* Estimated standard error too large to compute a CIE.

similar to the two-group case, we estimate the expected number of events for each group under an assumption of equal survivorship functions as

$$\hat{e}_{ki} = \frac{d_i n_{ki}}{n_i}, \; k = 1,2,\ldots,K. \tag{2.23}$$

We compare the observed and expected numbers of events for $K-1$ of the K groups. The reason for this will be explained shortly. The easiest way to denote the $K-1$ comparisons is to use vector notation to represent both observed and estimated expected number of events as follows:

$$\mathbf{d}_i' = (d_{1i}, d_{2i}, \ldots, d_{K-1i}),$$

and

$$\hat{\mathbf{e}}_i' = (\hat{e}_{1i}, \hat{e}_{2i}, \ldots, \hat{e}_{K-1i}).$$

The difference between these two vectors is

$$(\mathbf{d}_i - \hat{\mathbf{e}}_i)' = (d_{1i} - \hat{e}_{1i}, d_{2i} - \hat{e}_{2i}, \ldots, d_{K-1i} - \hat{e}_{K-1i}). \tag{2.24}$$

For convenience, we have used the first $K-1$ of the K groups, but any collection of $K-1$ groups could equally well be used.

To obtain a test statistic, we need an estimator of the covariance matrix of $\mathbf{d}_i$. The elements of this matrix are obtained assuming that the observed number of events follows a multivariate central hypergeometric distribution [see Johnson and Kotz (1997)]. The diagonal elements of the $(K-1)\times(K-1)$ matrix, denoted $\hat{\mathbf{V}}_i$, are

$$\hat{v}_{kki} = \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad k = 1,2,\ldots,K-1, \tag{2.25}$$

and the off-diagonal elements are

$$\hat{v}_{kli} = -\frac{n_{ki}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad k,l = 1,2,\ldots,K-1, k \neq l. \tag{2.26}$$

The various multiple-group versions of the two-group test statistics are obtained by computing a weighted difference between the observed and expected number of events. The weights used at each distinct survival time can be any of the weights used in the two-group test, denoted in general at time $t_{(i)}$ by w_i. To obtain a formula for the test statistic, we

Table 2.13 Table Used for the Test for the Equality of the Survivorship Function in K Groups at Observed Survival Time $t_{(i)}$

Event/Group	1	2	...	k	...	K	Total
Die	d_{1i}	d_{2i}	...	d_{ki}	...	d_{Ki}	d_i
Not Die	$n_{1i}-d_{1i}$	$n_{2i}-d_{2i}$	...	$n_{ki}-d_{ki}$	...	$n_{Ki}-d_{Ki}$	n_i-d_i
At Risk	n_{1i}	n_{2i}	...	n_{ki}	...	n_{Ki}	n_i

define a $K-1$ by $K-1$ diagonal matrix denoted $\mathbf{W}_i = \text{diag}(w_i)$. This matrix has the value of the weight, w_i, at time $t_{(i)}$ in all $K-1$ positions along the diagonal of the matrix. The test statistic to compare the survivorship experience of the K groups is

$$Q = \left[\sum_{i=1}^{m} \mathbf{W}_i(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]' \left[\sum_{i=1}^{m} \mathbf{W}_i \hat{\mathbf{V}}_i \mathbf{W}_i\right]^{-1} \left[\sum_{i=1}^{m} \mathbf{W}_i(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]. \quad (2.27)$$

The reason we use only $K-1$ of the K possible observed to expected comparisons is to prevent the matrix in the center of the right-hand side of (2.27) from being singular. The value of the test statistic in (2.27) is the same, regardless of which collection of $K-1$ groups are used.

The expression on the right-hand side of (2.27) may look intimidating to those not familiar with matrix algebra calculations, but when $K=2$ it simplifies to the more easily understood statistic defined in (2.20). Most software packages providing statistics for several definitions of the weight use (2.27). These packages typically provide only the test statistic and a p-value. One exception is SAS's lifetest procedure, which provides the individual elements in (2.24)–(2.25) for the log-rank and generalized Wilcoxon tests when the group variable is defined as a stratum variable. Under the hypothesis of equal survival functions, and if the summed estimated expected number of events is large, then Q will be approximately distributed as chi-square with $K-1$ degrees-of-freedom, and the p-value is $p = \Pr(\chi^2(K-1) \geq Q)$. The remarks made earlier about how the choice of weights in the two-group case can affect the ability of the test to detect differences apply to the multiple-group case as well.

The log-rank test, $w_i = 1$, has the following easily computed, conservative, approximation:

$$Q_c = \sum_{k=1}^{K} \frac{(d_{k+} - \hat{e}_{k+})^2}{\hat{e}_{k+}} < Q,$$