(b) Using a beta prior for $H(t)$ with $q = 4$ and $H_0(t) = \ln(1 + 0.15t^{0.5})$, find the Bayes estimate of the survival function under squared-error loss.

(c) Compare the estimates found in parts a and b to the usual Kaplan–Meier estimate of the survival function.

7 Hypothesis Testing

7.1 Introduction

As we have seen in Chapters 4–6, the Nelson–Aalen estimator of the cumulative hazard rate is a basic quantity in describing the survival experience of a population. In Chapter 4, we used this estimator along with the closely related Product-Limit estimator to make crude comparisons between the disease-free survival curves of bone marrow transplant patients with different types of leukemia, and in section 6.3, we used this statistic as the basis for estimating excess mortality of Iowa psychiatric patients.

In this chapter, we shall focus on hypothesis tests that are based on comparing the Nelson–Aalen estimator, obtained directly from the data, to an expected estimator of the cumulative hazard rate, based on the assumed model under the null hypothesis. Rather than a direct comparison of these two rates, we shall examine tests that look at weighted differences between the observed and expected hazard rates. The weights will allow us to put more emphasis on certain parts of the curves. Different weights will allow us to present tests that are most sensitive to early or late departures from the hypothesized relationship between samples as specified by the null hypothesis.

In section 7.2, we shall consider the single sample problem. Here, we wish to test if the sample comes from a population with a prespecified hazard rate $h_0(t)$. In section 7.3, we will look at tests of the null hypothesis of no difference in survival between K treatments against a global alternative that at least one treatment has a different survival rate. Here, for example, we will discuss censored data versions of the Wilcoxon or Kruskal–Wallis test and log-rank or Savage test. In section 7.4, we look at K sample tests that have power to detect ordered alternatives. A censored data version of the Jonckheere–Terpstra test will be presented. In section 7.5, we will see how these tests can be modified to handle stratification on covariates which may confound the analysis. We shall see how this approach can be used to handle matched data, and we will have a censored-data version of the sign test. In section 7.6, we will look at tests based on the maximum of the sequential evaluation of these tests at each death time. These tests have the ability to detect alternatives where the hazard rates cross and are extensions of the usual Kolmogorov–Smirnov test. Finally, in section 7.7, we present three other tests which have been proposed to detect crossing hazard rates, a censored-data version of the Cramer–von Mises test, a test based on weighted differences in the Kaplan–Meier estimators, and a censored-data version of the median test.

The methods of this chapter can be applied equally well to right-censored data or to samples that are right-censored and left-truncated. As we shall see, the key statistics needed to compute the tests are the number of deaths at each time point and the number of individuals at risk at these death times. Both quantities are readily observed with left-truncated and right-censored data.

7.2 One-Sample Tests

Suppose that we have a censored sample of size n from some population. We wish to test the hypothesis that the population hazard rate is $h_0(t)$ for all $t \leq \tau$ against the alternative that the hazard rate is not $h_0(t)$ for some $t \leq \tau$. Here $h_0(t)$ is a completely specified function over the range 0 to τ. Typically, we shall take τ to be the largest of the observed study times.

An estimate of the cumulative hazard function $H(t)$ is the Nelson–Aalen estimator, (4.2.3), given by $\sum_{t_i \leq t} \frac{d_i}{Y(t_i)}$, where d_i is the number of events at the observed event times, $t_1, \ldots, t_D$ and $Y(t_i)$ is the number of individuals under study just prior to the observed event time t_i. The quantity $\frac{d_i}{Y(t_i)}$ gives a crude estimate of the hazard rate at an event time t_i. When the null hypothesis is true, the expected hazard rate at t_i is $h_0(t_i)$. We shall compare the sum of weighted differences between the observed and expected hazard rates to test the null hypothesis.

Let $W(t)$ be a weight function with the property that $W(t)$ is zero whenever $Y(t)$ is zero. The test statistic is

$$Z(\tau) = O(\tau) - E(\tau) = \sum_{i=1}^{D} W(t_i)\frac{d_i}{Y(t_i)} - \int_0^\tau W(s)h_0(s)ds. \qquad (7.2.1)$$

When the null hypothesis is true, the sample variance of this statistic is given by

$$V[Z(\tau)] = \int_0^\tau W^2(s)\frac{h_0(s)}{Y(s)}ds. \qquad (7.2.2)$$

For large samples, the statistic $Z(\tau)^2/V[Z(\tau)]$ has a central chi-squared distribution when the null hypothesis is true.

The statistic $Z(\tau)/V[Z(\tau)]^{1/2}$ is used to test the one sided alternative hypothesis that $h(t) > h_0(t)$. When the null hypothesis is true and the sample size is large, this statistic has a standard normal distribution. The null hypothesis is rejected for large values of the statistic.

The most popular choice of a weight function is the weight $W(t) = Y(t)$ which yields the one-sample log-rank test. To allow for possible left truncation, let T_j be the time on study and L_j be the delayed entry time for the jth patient. When τ is equal to the largest time on study,

$$O(\tau) = \text{observed number of events at or prior to time } \tau, \qquad (7.2.3)$$

and

$$E(\tau) = V[Z(\tau)] = \sum_{j=1}^{n}[H_0(T_j) - H_0(L_j)] \qquad (7.2.4)$$

where $H_0(t)$ is the cumulative hazard under the null hypothesis.

Other weight functions proposed in the literature include the Harrington and Fleming (1982) family $W_{HF}(t) = Y(t)S_0(t)^p[1 - S_0(t)]^q$, $p \geq 0$, $q \geq 0$, where $S_0(t) = \exp[-H_0(t)]$ is the hypothesized survival function. By choice of p and q, one can put more weight on early departures from the null hypothesis (p much larger than q), late departures from the null hypothesis (p much smaller than q), or on departures in the mid-range ($p = q > 0$). The log-rank weight is a special case of this model with $p = q = 0$.

EXAMPLE 7.1

In section 6.3, we examined models for excess and relative mortality in a sample of 26 Iowa psychiatric patients described in section 1.15. We shall now use the one-sample log-rank statistic to test the hypothesis that the hazard rate of this group of patients is the same as the hazard rate in the general Iowa population, given by the standard 1960 Iowa

mortality table. To perform this test, we will use the sex specific survival rates. Time T_j is taken as the jth individual's age at death or the end of the study, and the left-truncation time L_j, is this individual's age at entry into the study. We obtain $H(t)$ as $-\ln[S(t)]$ from the appropriate column of Table 6.2. Table 7.1 shows the calculations to compute $O(71)$ and $E(71)$.

The test statistic is $\chi^2 = (15-4.4740)^2/4.4740 = 24.7645$ which has a chi-squared distribution with one degree of freedom. Here the p-value of this test is close to zero, and we can conclude that the mortality rates of the psychiatric patients differ from those of the general public.

TABLE 7.1
Computation of One-Sample, Log-Rank Test

Subject j	*Sex*	*Status* d_i	*Age at Entry* L_i	*Age at Exit* T_j	$H_0(L_j)$	$H_0(T_j)$	$H_0(T_j) - H_0(L_j)$
1	f	1	51	52	0.0752	0.0797	0.0045
2	f	1	58	59	0.1131	0.1204	0.0073
3	f	1	55	57	0.0949	0.1066	0.0117
4	f	1	28	50	0.0325	0.0711	0.0386
5	m	0	21	51	0.0417	0.1324	0.0907
6	m	1	19	47	0.0383	0.1035	0.0652
7	f	1	25	57	0.0305	0.1066	0.0761
8	f	1	48	59	0.0637	0.1204	0.0567
9	f	1	47	61	0.0606	0.1376	0.0770
10	f	1	25	61	0.0305	0.1376	0.1071
11	f	0	31	62	0.0347	0.1478	0.1131
12	m	0	24	57	0.0473	0.1996	0.1523
13	m	0	25	58	0.0490	0.2150	0.1660
14	f	0	30	67	0.0339	0.2172	0.1833
15	f	0	33	68	0.0365	0.2357	0.1992
16	m	1	36	61	0.0656	0.2704	0.2048
17	m	0	30	61	0.0561	0.2704	0.2143
18	m	1	41	63	0.0776	0.3162	0.2386
19	f	1	43	69	0.0503	0.2561	0.2058
20	f	1	45	69	0.0548	0.2561	0.2013
21	f	0	35	65	0.0384	0.1854	0.1470
22	m	0	29	63	0.0548	0.3162	0.2614
23	m	0	35	65	0.0638	0.3700	0.3062
24	m	1	32	67	0.0590	0.4329	0.3739
25	f	1	36	76	0.0395	0.4790	0.4395
26	m	0	32	71	0.0590	0.5913	0.5323
Total		15					4.4740

Practical Notes

1. An alternate estimator of the variance of $Z(\tau)$ is given by $V[Z(\tau)] = \sum_{i=1}^{D} W(t_i)^2 \frac{d_i}{Y(t_i)^2}$ which uses the empirical estimator of $h_0(t)$ rather than the hypothesized value. When the alternative hypothesis $h(t) > h_0(t)$ is true, for some $t \leq \tau$, this variance estimator is expected to be larger than (7.2.2), and the test is less powerful using this value. On the other hand, if $h(t) < h_0(t)$, then, this variance estimator will tend to be smaller, and the test will be more powerful.
2. The statistic $O(\tau)/E(\tau)$ based on the log-rank weights is called the standardized mortality ratio (SMR).
3. A weight function suggested by Gatsonis et al. (1985) is $W(t) = (1 + \{\log[1 - S_0(t)]\}/S_0(t))Y(t)$.

Theoretical Notes

1. In this class of tests, the one-sample, log-rank test is the locally most powerful test against a shift alternative of the extreme value distribution. The weight function $W_{HF}(t) = Y(t)S_0(t)$ is the locally most powerful test for the logistic distribution. Because the one-sample Wilcoxon test also has this property, this choice of weights leads to a censored-data, one-sample, Wilcoxon test. See Andersen et al. (1993) for details.
2. These one-sample tests arise quite naturally from the theory of counting processes. Under the null hypothesis, using the notation in section 3.6, $\int_0^\tau [J(u)/Y(u)]dN(u) - \int_0^\tau J(u)h_0(u)\,du$ is a martingale. The statistic $Z(\tau)$ is a stochastic integral of the weight function $W(t)$ with respect to this martingale, and $\mathrm{Var}[Z(\tau)]$ is the predictable variation process of this stochastic integral. The asymptotic chi-squared distribution follows by the martingale central limit theorem.
3. The one-sample, log-rank test was first proposed by Breslow (1975) and generalized to left truncation by Hyde (1977) and Woolson (1981).

7.3 Tests for Two or More Samples

In section 7.2, we looked at one-sample tests that made a weighted comparison between the estimated hazard rate and the hypothesized hazard rates. We now shall extend these methods to the problem of comparing hazard rates of K ($K \geq 2$) populations, that is, we shall test

the following set of hypotheses:

$$H_0 : h_1(t) = h_2(t) = \cdots = h_K(t), \text{ for all } t \le \tau, \text{ versus} \quad (7.3.1)$$
$$H_A : \text{at least one of the } h_j(t)\text{'s is different for some } t \le \tau.$$

Here τ is the largest time at which all of the groups have at least one subject at risk.

As in section 7.2, our inference is to the hazard rates for all time points less than τ, which is, typically, the smallest of the largest time on study in each of the k groups. The alternative hypothesis is a global one in that we wish to reject the null hypothesis if, at least, one of the populations differs from the others at some time. In the next section, we will present tests that are more powerful in the case of ordered alternatives.

The data available to test the hypothesis (7.3.1) consists of independent right-censored and, possibly, left-truncated samples for each of the K populations. Let $t_1 < t_2 < \cdots < t_D$ be the distinct death times in the pooled sample. At time t_i we observe d_{ij} events in the jth sample out of Y_{ij} individuals at risk, $j = 1, \ldots, K$, $i = 1, \ldots, D$. Let $d_i = \sum_{j=1}^{K} d_{ij}$ and $Y_i = \sum_{j=1}^{K} Y_{ij}$ be the number of deaths and the number at risk in the combined sample at time t_i, $i = 1, \ldots, D$.

The test of H_0 is based on weighted comparisons of the estimated hazard rate of the jth population under the null and alternative hypotheses, based on the Nelson–Aalen estimator (4.2.3). If the null hypothesis is true, then, an estimator of the expected hazard rate in the jth population under H_0 is the pooled sample estimator of the hazard rate d_i/Y_i. Using only data from the jth sample, the estimator of the hazard rate is d_{ij}/Y_{ij}. To make comparisons, let $W_j(t)$ be a positive weight function with the property that $W_j(t_i)$ is zero whenever Y_{ij} is zero. The test of H_0 is based on the statistics

$$Z_j(\tau) = \sum_{i=1}^{D} W_j(t_i)\left\{\frac{d_{ij}}{Y_{ij}} - \frac{d_i}{Y_i}\right\}, \quad j = 1, \ldots, K. \quad (7.3.2)$$

If all the $Z_j(\tau)$'s are close to zero, then, there is little evidence to believe that the null hypothesis is false, whereas, if one of the $Z_j(\tau)$'s is far from zero, then, there is evidence that this population has a hazard rate differing from that expected under the null hypothesis.

Although the general theory allows for different weight functions for each of the comparisons in (7.3.2), in practice, all of the commonly used tests have a weight function $W_j(t_i) = Y_{ij}W(t_i)$. Here, $W(t_i)$ is a common weight shared by each group, and Y_{ij} is the number at risk in the jth group at time t_i. With this choice of weight functions

$$Z_j(\tau) = \sum_{i=1}^{D} W(t_i)\left[d_{ij} - Y_{ij}\left(\frac{d_i}{Y_i}\right)\right], \quad j = 1, \ldots, K. \quad (7.3.3)$$

Note that with this class of weights the test statistic is the sum of the weighted difference between the observed number of deaths and the expected number of deaths under H_0 in the jth sample. The expected number of deaths in sample j at t_i is the proportion of individuals at risk Y_{ij}/Y_i that are in sample j at time t_i, multiplied by the number of deaths at time t_i.

The variance of $Z_j(\tau)$ in (7.3.3) is given by

$$\hat{\sigma}_{jj} = \sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\left(1 - \frac{Y_{ij}}{Y_i}\right)\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad j = 1, \ldots, K \quad (7.3.4)$$

and the covariance of $Z_j(\tau)$, $Z_g(\tau)$ is expressed by

$$\hat{\sigma}_{jg} = -\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}\left(\frac{Y_i - d_i}{Y_i - 1}\right) d_i, \quad g \ne j. \quad (7.3.5)$$

The term $(Y_i - d_i)/(Y_i - 1)$, which equals one if no two individuals have a common event time, is a correction for ties. The terms $\frac{Y_{ij}}{Y_i}(1 - \frac{Y_{ij}}{Y_i})d_i$ and $-\frac{Y_{ij}}{Y_i}\frac{Y_{ig}}{Y_i}d_i$ arise from the variance and covariance of a multinomial random variable with parameters d_i, $p_j = Y_{ij}/Y_i$, $j = 1, \ldots, K$.

The components vector $(Z_1(\tau), \ldots, Z_K(\tau))$ are linearly dependent because $\sum_{j=1}^{K} Z_j(\tau)$ is zero. The test statistic is constructed by selecting any $K - 1$ of the Z_j's. The estimated variance-covariance matrix of these statistics is given by the $(K - 1) \times (K - 1)$ matrix $\boldsymbol{\Sigma}$, formed by the appropriate $\hat{\sigma}_{jg}$'s. The test statistic is given by the quadratic form

$$\chi^2 = (Z_1(\tau), \ldots, Z_{K-1}(\tau))\boldsymbol{\Sigma}^{-1}(Z_1(\tau), \ldots, Z_{K-1}(\tau))'. \quad (7.3.6)$$

When the null hypothesis is true, this statistic has a chi-squared distribution, for large samples with $K - 1$ degrees of freedom. An α level test of H_0 rejects when χ^2 is larger than the αth upper percentage point of a chi-squared, random variable with $K - 1$ degrees of freedom.

When $K = 2$ the test statistic can be written as

$$Z = \frac{\sum_{i=1}^{D} W(t_i)[d_{i1} - Y_{i1}(\frac{d_i}{Y_i})]}{\sqrt{\sum_{i=1}^{D} W(t_i)^2 \frac{Y_{i1}}{Y_i}(1 - \frac{Y_{i1}}{Y_i})(\frac{Y_i - d_i}{Y_i - 1})d_i}}, \quad (7.3.7)$$

which has a standard normal distribution for large samples when H_0 is true. Using this statistic, an α level test of the alternative hypothesis $H_A : h_1(t) > h_2(t)$, for some $t \le \tau$, is rejected when $Z \ge Z_\alpha$, the αth upper percentage point of a standard normal distribution. The test of $H_A : h_1(t) \ne h_2(t)$, for some t, rejects when $|Z| > Z_{\alpha/2}$.

A variety of weight functions have been proposed in the literature. A common weight function, leading to a test available in most statistical packages, is $W(t) = 1$ for all t. This choice of weight function leads to

the so-called log-rank test and has optimum power to detect alternatives where the hazard rates in the K populations are proportional to each other. A second choice of weights is $W(t_i) = Y_i$. This weight function yields Gehan's (1965) generalization of the two-sample Mann–Whitney–Wilcoxon test and Breslow's (1970) generalization of the Kruskal–Wallis test. Tarone and Ware (1977) suggest a class of tests where the weight function is $W(t_i) = f(Y_i)$, and f is a fixed function. They suggest a choice of $f(y) = y^{1/2}$. This class of weights gives more weight to differences between the observed and expected number of deaths in sample j at time points where there is the most data.

An alternate censored-data version of the Mann–Whitney–Wilcoxon test was proposed by Peto and Peto (1972) and Kalbfleisch and Prentice (1980). Here, we define an estimate of the common survival function by

$$\tilde{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{Y_i + 1}\right), \tag{7.3.8}$$

which is close to the pooled Product-Limit estimator. They suggest using $W(t_i) = \tilde{S}(t_i)$. Andersen et al. (1982) suggest that this weight should be modified slightly as $W(t_i) = \tilde{S}(t_i)Y_i/(Y_i + 1)$ (see Theoretical Note 2). Either of the weights depends on the combined survival experience in the pooled sample whereas the weight $W(t_i) = Y_i$ depends heavily on the event times and censoring distributions. Due to this fact, the Gehan-Breslow weights can have misleading results when the censoring patterns are different in the individual samples (see Prentice and Marek (1979) for a case study).

Fleming and Harrington (1981) propose a very general class of tests that includes, as special cases, the log-rank test and a version of the Mann–Whitney–Wilcoxon test, very close to that suggested by Peto and Peto (1972). Here, we let $\hat{S}(t)$ be the Product-Limit estimator (3.2.1) based on the combined sample. Their weight function is given by

$$W_{p,q}(t_i) = \hat{S}(t_{i-1})^p[1 - \hat{S}(t_{i-1})]^q, \; p \ge 0, \; q \ge 0. \tag{7.3.9}$$

Here, the survival function at the previous death time is used as a weight to ensure that these weights are known just prior to the time at which the comparison is to be made. Note that $S(t_0) = 1$ and we define $0^0 = 1$ for these weights. When $p = q = 0$ for this class, we have the log-rank test. When $p = 1$, $q = 0$, we have a version of the Mann–Whitney–Wilcoxon test. When $q = 0$ and $p > 0$, these weights give the most weight to early departures between the hazard rates in the K populations, whereas, when $p = 0$ and $q > 0$, these tests give most weight to departures which occur late in time. By an appropriate choice of p and q, one can construct tests which have the most power against alternatives which have the K hazard rates differing over any desired region. This is illustrated in the following example.

EXAMPLE 7.2 In section 1.4, data on a clinical trial of the effectiveness of two methods for placing catheters in kidney dialysis patients was presented. We are interested in testing if there is a difference in the time to cutaneous exit-site infection between patients whose catheter was placed surgically (group 1) as compared to patients who had their catheters placed percutaneously (group 2).

Figure 7.1 shows the survival curves for the two samples. Table 7.2 shows the calculations needed to construct the log-rank test. Here, $Z_{obs} = 3.964/\sqrt{6.211} = 1.59$ which has a p-value of $2Pr[Z > 1.59] = 0.1117$, so the log-rank test suggests no difference between the two procedures in the distribution of the time to exit-site infection.

To further investigate these two treatments, we shall apply some of the other weight functions discussed earlier. Table 7.3 summarizes

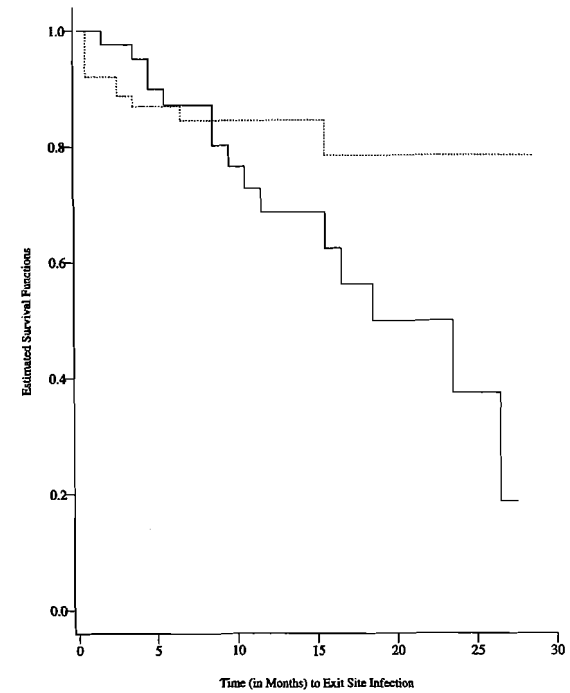

Figure 7.1 *Estimated (Infection-free) survival function for kidney dialysis patients with percutaneous (——) and surgical (——) placements of catheters.*

TABLE 7.2
Construction of Two-Sample, Log-Rank Test

t_i	Y_{i1}	d_{i1}	Y_{i2}	d_{i2}	Y_i	d_i	$Y_{i1}\left(\frac{d_i}{Y_i}\right)$	$d_{i1} - Y_{i1}\left(\frac{d_i}{Y_i}\right)$	$\frac{Y_{i1}}{Y_i}\left(1-\frac{Y_{i1}}{Y_i}\right)\left(\frac{Y_i-d_i}{Y_i-1}\right)d_i$
0.5	43	0	76	6	119	6	2.168	−2.168	1.326
1.5	43	1	60	0	103	1	0.417	0.583	0.243
2.5	42	0	56	2	98	2	0.857	−0.857	0.485
3.5	40	1	49	1	89	2	0.899	0.101	0.489
4.5	36	2	43	0	79	2	0.911	1.089	0.490
5.5	33	1	40	0	73	1	0.452	0.548	0.248
6.5	31	0	35	1	66	1	0.470	−0.470	0.249
8.5	25	2	30	0	55	2	0.909	1.091	0.487
9.5	22	1	27	0	49	1	0.449	0.551	0.247
10.5	20	1	25	0	45	1	0.444	0.556	0.247
11.5	18	1	22	0	40	1	0.450	0.550	0.248
15.5	11	1	14	1	25	2	0.880	0.120	0.472
16.5	10	1	13	0	23	1	0.435	0.565	0.246
18.5	9	1	11	0	20	1	0.450	0.550	0.248
23.5	4	1	5	0	9	1	0.444	0.556	0.247
26.5	2	1	3	0	5	1	0.400	0.600	0.240
SUM		15		11		26	11.036	3.964	6.211

TABLE 7.3
Comparison of Two-Sample Tests

Test	$W(t_i)$	$Z_1(\tau)$	σ_{11}^2	χ^2	*p-value*
Log-Rank	1.0	3.96	6.21	2.53	0.112
Gehan	Y_i	−9	38862	0.002	0.964
Tarone–Ware	$Y_i^{1/2}$	13.20	432.83	0.40	0.526
Peto–Peto	$\tilde{S}(t_i)$	2.47	4.36	1.40	0.237
Modified Peto–Peto	$\tilde{S}(t_i)Y_i/(Y_i+1)$	2.31	4.20	1.28	0.259
Fleming–Harrington $p = 0$, $q = 1$	$[1 - \hat{S}(t_{i-1})]$	1.41	0.21	9.67	0.002
Fleming–Harrington $p = 1$, $q = 0$	$\hat{S}(t_{i-1})$	2.55	4.69	1.39	0.239
Fleming–Harrington $p = 1$, $q = 1$	$\hat{S}(t_{i-1})[1 - \hat{S}(t_{i-1})]$	1.02	0.11	9.83	0.002
Fleming–Harrington $p = 0.5$, $q = 0.5$	$\hat{S}(t_{i-1})^{0.5}[1 - \hat{S}(t_{i-1})]^{0.5}$	2.47	0.66	9.28	0.002
Fleming–Harrington $p = 0.5$, $q = 2$	$\hat{S}(t_{i-1})^{0.5}[1 - \hat{S}(t_{i-1})]^2$	0.32	0.01	8.18	0.004

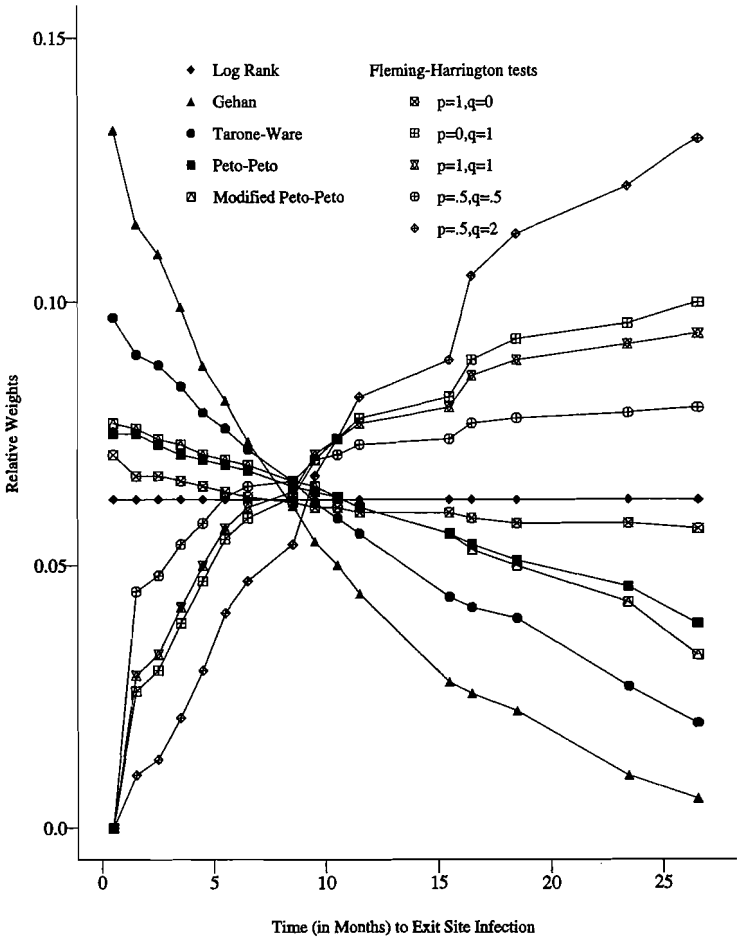

Figure 7.2 *Relative weights for comparison of observed and expected numbers of deaths for kidney dialysis patients.*

the results of these tests. Figure 7.2 shows the relative weights these tests give to the comparisons at each time point. $W(t_i)/\sum_{i=1}^{D} W(t_i)$ is plotted here. Note that Gehan's weight function gives very heavy weight to early comparisons at $t_i = 0.5$ and leads to a negative test statistic. The Fleming and Harrington tests, with $q > 0$, put more weight on

late comparisons and lead to significant tests because the two survival curves diverge for larger values of t.

EXAMPLE 7.3 In section 1.16, data on 462 individuals who lived at the Channing House retirement center was reported. These data are left-truncated by the individual's entry time into the retirement center. In Example 4.3, survival curves were constructed for both males and females. We shall now apply the methods of this section to test the hypothesis that females tend to live longer than males. We test the hypothesis H_0 : $h_F(t) = h_M(t)$, 777 months $\leq t \leq$ 1152 months against the one sided hypothesis H_A : $h_F(t) \leq h_M(t)$ for all $t \in [777, 1152]$ and $h_F(t) < h_M(t)$ for some $t \in [777, 1152]$.

To perform this test, we need to compute Y_{iF} and Y_{iM} as the number of females and males, respectively, who were in the center at age t_i. The values of these quantities are depicted in Figure 4.10. The test will be based on the weighted difference between the observed and expected number of male deaths. Using the log-rank weights, we find $Z_M(1152) = 9.682$, $\hat{V}(Z_M(1152)) = 28.19$, so $Z_{\text{obs}} = 1.82$ and the one-sided p-value is 0.0341, which provides evidence that males are dying at a faster rate than females.

EXAMPLE 7.4 In Chapter 4, we investigated the relationship between the disease-free survival functions of 137 patients given a bone marrow transplant (see section 1.3 for details). Three groups were considered: Group 1 consisting of 38 ALL patients; Group 2 consisting of 54 AML low-risk patients and Group 3 consisting of 45 AML high-risk patients. The survival curves for these three groups are shown in Figure 4.2 in section 4.2.

We shall test the hypothesis that the disease-free survival functions of these three populations are the same over the range of observation, $t \leq$ 2204 days, versus the alternative that at least one of the populations has a different survival rate. Using the log-rank weights, we find $Z_1(2204) = 2.148$; $Z_2(2204) = -14.966$ and $Z_3(2204) = 12.818$, and the covariance matrix is

$$(\hat{\sigma}_{jg}, j, g = 1, \ldots, 3) = \begin{pmatrix} 15.9552 & -10.3451 & -5.6101 \\ -10.3451 & 20.3398 & -9.9947 \\ -5.6101 & -9.9947 & 15.6048 \end{pmatrix}.$$

Notice that the $Z_j(2204)$'s sum to zero and that the matrix $(\hat{\sigma}_{jg})$ is singular. The test is constructed by selecting any two of the $Z_j(2204)$'s, and constructing a quadratic form, using the appropriate rows and columns of the covariance matrix. The resulting statistic will be the same regardless of which $Z_j(2204)$'s are selected. The test statistic in this case is

$$\chi^2 = (2.148, -14.966) \begin{pmatrix} 15.9552 & -10.3451 \\ -10.3451 & 20.3398 \end{pmatrix}^{-1} \begin{pmatrix} 2.148 \\ -14.966 \end{pmatrix} = 13.8037.$$

When the null hypothesis is true, this statistic has an approximate chi-square distribution with two degrees of freedom which yields a p-value of 0.0010.

We can apply any of the weight functions discussed above to this problem. For example, Gehan's weight $W(t_i) = Y_i$ yields $\chi^2 = 16.2407$ ($p = 0.0003$); Tarone-Ware's weight $W(t_i) = Y_i^{1/2}$ yields $\chi^2 = 15.6529$ ($p = 0.0040$), Fleming and Harrington's weight, with $p = 1$, $q = 0$, yields $\chi^2 = 15.6725$ ($p = 0.0040$), with $p = 0$, $q = 1$, yields $\chi^2 = 6.1097$ ($p = 0.0471$), and with $p = q = 1$, yields $\chi^2 = 9.9331$ ($p = 0.0070$). All of these tests agree with the conclusion that the disease-free survival curves are not the same in these three disease categories.

An important consideration in applying the tests discussed in this section is the choice of the weight function to be used. In most applications of these methods, our strategy is to compute the statistics using the log-rank weights $W(t_i) = 1$ and the Gehan weight with $W(t_i) = Y_i$. Tests using these weights are available in most statistical packages which makes their application routine in most problems.

In some applications, one of the other weight functions may be more appropriate, based on the investigator's desire to emphasize either late or early departures between the hazard rates. For example, in comparing relapse-free survival between different regimes in bone marrow transplants for leukemia, a weight function giving higher weights to late differences between the hazard rates is often used. Such a function downweights early differences in the survival rates, which are often due to the toxicity of the preparative regimes, and emphasizes differences occurring late, which are due to differences in curing the leukemia. This is illustrated in the following example.

EXAMPLE 7.5 In section 1.9, data from a study of the efficacy of autologous (auto) versus allogeneic (allo) transplants for acute myelogenous leukemia was described. Of interest is a comparison of disease-free survival for these two types of transplants. Here, the event of interest is death or relapse, which ever comes first. In comparing these two types of transplants, it is well known that patients given an allogeneic transplant tend to have more complications early in their recovery process. The most critical of these complications is acute graft-versus-host disease which occurs within the first 100 days after the transplant and is often lethal. Because patients given an autologous transplant are not at risk of developing acute graft-versus-host disease, they tend to have a higher survival rate during this period. Of primary interest to most investigators in this area is comparing the treatment failure rate (death or relapse) among long-term survivors. To test this hypothesis, we shall use a test with the Fleming and Harrington weight function $W(t_i) = 1 - S(t_{i-1})$ (Eq. 7.3.9

with $p = 0$, $q = 1$). This function downweights events (primarily due to acute graft-versus-host disease) which occur early.

For these weights, we find that $Z_1(t) = -2.093$ and $\hat{\sigma}_{11}(\tau) = 1.02$, so that the chi-square statistic has a value of 4.20 which gives a p-value of 0.0404 . This suggest that there is a difference in the treatment failure rates for the two types of transplants.

By comparison, the log-rank test and Gehan's test have p-values of 0.5368 and 0.7556, respectively. These statistics have large p-values because the hazard rates of the two types of transplants cross at about 12 months, so that the late advantage of allogeneic transplants is negated by the high, early mortality of this type of transplant.

Practical Notes

1. The SAS procedure LIFETEST can be used to perform the log-rank test and Gehan's test for right-censored data. This procedure has two ways to perform the test. The first is to use the STRATA statement. This statement produces $Z_j(\tau)$, the matrix $(\hat{\sigma}_{jg})$ and the chi-square statistics. The second possibility for producing a test is to use the TEST statement. This statistic is equivalent to those obtained using the STRATA command when there is only one death at each time point. When there is more than one death at some time, it computes a statistic obtained as the average of the statistics one can obtain by breaking these ties in all possible ways. This leads to different statistics than those we present here. We recommend the use of the STRATA command for tests using SAS.
2. The S-Plus function surv.diff produces the Fleming and Harrington class of tests with $q = 0$. By choosing $p = 0$, the log-rank test can be obtained.
3. All of the tests described in this section are based on large-sample approximations to the distribution of the chi-square statistics. They also assume that the censoring distributions are independent of the failure distributions. Care should be used in interpreting these results when the sample sizes are small or when there are few events. (Cf. Kellerer and Chmelevsky 1983, Latta, 1981, or Peace and Flora, 1978, for the results of Monte Carlo studies of the small-sample power of these tests.)
4. Based on a Monte Carlo study, Kellerer and Chmelevsky (1983) conclude that, when the sample sizes are small for two-sample tests, the one-sided test must be used with caution. When the sample sizes are very different in the two groups and when the alternative hypothesis is that the hazard rate of the smaller sample is larger than the rate in the larger sample, these tests tend to falsely reject the null hypothesis too often. The tests are extremely conservative when the larger sample is associated with the larger hazard rate under the alternative hypothesis. They and Prentice and Marek (1979) strongly recommend that only two-sided tests be used in making comparisons.
5. For the two-sample tests, the log-rank weights have optimal local power to detect differences in the hazard rates, when the hazard rates are proportional. This corresponds to the survival functions satisfying a Lehmann alternative $S_j(t) = S(t)^{\theta_j}$. These are also the optimal weights for the K sample test with proportional hazards when, for large samples, the numbers at risk in each sample at each time point are proportional in size. For the two-sample case, Fleming and Harrington's class of tests with $q = 0$ has optimal local power to detect the alternatives $h_2(t) = h_1(t)e^{\theta}[S_1(t)^p + [1 - S_1(t)]^q e^{\theta}]^{-1}$. See Fleming and Harrington (1981) or Andersen et al. (1993) for a more detailed discussion.

Theoretical Notes

1. The tests discussed in this section arise naturally using counting process theory. In section 3.6, we saw that the Nelson–Aalen estimator of the cumulative hazard rate in the jth sample was a stochastic integral of a predictable process with respect to a basic martingale and, as such, is itself a martingale. By a similar argument, when the null hypothesis is true, the Nelson–Aalen estimator of the common cumulative hazard rate is also a martingale. Furthermore, the difference of two martingales can be shown to also be a martingale. If $W_j(t)$ is a predictable weight function, then, $Z_j(\tau)$ is the integral of a predictable process with respect to a martingale and is, again, a martingale when the null hypothesis is true. The estimated variance of $Z_j(\tau)$ in (7.3.4) comes from the predictable variation process of $Z_j(\tau)$ using a version of the predictable variation process for the basic martingale which allows for discrete event times. More detailed descriptions of this derivation are found in Aalen (1975) and Gill (1980) for the two-sample case and in Andersen et al. (1982) for the K sample case.
2. The modification of Andersen et al. (1982) to the Peto and Peto weight, $W(t_i) = \tilde{S}(t_i)Y_i/(Y_i + 1)$ makes the weight function predictable in the sense discussed in section 3.6.
3. The statistics presented in this section are generalizations to censored data of linear rank statistics. For uncensored data, a linear rank statistic is constructed as $Z_j = \sum_{i=1}^{n_j} a_n(R_{ij})$, $j = 1, \ldots, K$.

   Here R_{ij} is the rank of the ith observation in the jth sample among the pooled observations. The scores $a_n(i)$ are obtained from a score function Ψ defined on the unit interval by $a_n(i) = E[\Psi(T_{(i)})]$, where $T_{(i)}$ is the ith order statistic from a uniform $[0, 1]$ sample of size n or by $a_n(i) = \Psi[i/(n + 1)]$. For a censored sample, these scores are generalized as follows: An uncensored observation is given a score of

$\Psi[1-\tilde{S}(t)]$, with $\tilde{S}(t)$ given by (7.3.8); a censored observation is given a score of $\frac{1}{1-\Psi[1-\tilde{S}(t)]}\int_{\Psi[1-\tilde{S}(t)]}^{1}\Psi(u)du$. The score function $\Psi(u) = 2u-1$, for example will yield the Peto and Peto version of Gehan's test. (See Kalbfleisch and Prentice, 1980 for additional development of this concept.)

4. Gill (1980) discusses the Pitman efficiency of these tests.
5. Gehan's two-sample test can be constructed as a generalization of the Mann–Whitney test as follows. Let (T_{ij}, δ_{ij}) denote the time on study and the event indicator for the ith observation in the jth sample. Define the scoring function $\phi[(T_{i1}, \delta_{i1}), (T_{b2}, \delta_{b2})]$ as follows:

$$\phi[(T_{i1},\delta_{i1}),(T_{b2},\delta_{b2})] = \begin{pmatrix} +1, \text{ if } T_{i1} \le T_{b2},\ \delta_{i1}=1,\ \delta_{b2}=0, \\ \quad \text{or } T_{i1} < T_{b2},\ \delta_{i1}=1,\ \delta_{b2}=1, \\ -1, \text{ if } T_{i1} \ge T_{b2},\ \delta_{b2}=1,\ \delta_{i1}=0, \\ \quad \text{or } T_{i1} > T_{b2},\ \delta_{i1}=1,\ \delta_{b2}=1, \\ 0, \text{ otherwise} \end{pmatrix}$$

Then, $Z_1(\tau) = \sum_{i=1}^{n_1}\sum_{b=1}^{n_2}\phi[(T_{i1},\delta_{i1}),(T_{b2},\delta_{b2})]$ is the number of observations from the first sample that are definitely smaller than an observation in the second sample. Gehan (1965) provides a variance estimator of this statistic under the null hypothesis, based on assuming a fixed censoring pattern and that all possible permutations of the two samples over this pattern are equally likely. Essentially, this estimator assumes that the censoring patterns are the same for both samples. When this is not the case, this variance estimator may lead to incorrect decisions.

7.4 Tests for Trend

In the previous section, we examined tests, based on a weighted comparison of the observed and expected number of deaths in each of K samples, to test the null hypothesis that the K population hazard rates are the same versus the global alternative hypothesis that, at least, one of the hazard rates is different. In this section, we shall use the statistics developed in section 7.3 to develop a test statistic with power to detect ordered alternatives, that is, we shall test

$$H_0 : h_1(t) = h_2(t) = \cdots = h_K(t), \text{ for } t \le \tau, \tag{7.4.1}$$

against

$$H_A : h_1(t) \le h_2(t) \le \cdots \le h_K(t) \text{ for } t \le \tau, \text{ with at least one strict inequality.}$$

The alternative hypothesis is equivalent to the hypothesis that $S_1(t) \ge S_2(t) \ge \cdots \ge S_K(t)$.

The test will be based on the statistic $Z_j(\tau)$ given by (7.3.3). Any of the weight functions discussed in section 7.3 can be used in constructing the test. As discussed earlier these various weight functions give more or less weight to the time points at which the comparison between the observed and expected number of deaths in each sample is made. We let $\hat{\Sigma}$ be the full $K \times K$ covariance matrix, $\hat{\Sigma} = (\hat{\sigma}_{jg}, j, g = 1, \ldots, K)$. Here, $\hat{\sigma}_{jg}$ is given by Eqs. (7.3.4) and (7.3.5).

To construct the test, a sequence of scores $a_1 < a_2 < \cdots < a_K$ is selected. Any increasing set of scores can be used in constructing the test, and the test is invariant under linear transformations of the scores. In most cases, the scores $a_j = j$ are used, but one may take the scores to be some numerical characteristic of the jth population. The test statistic is

$$Z = \frac{\sum_{j=1}^{K} a_j Z_j(\tau)}{\sqrt{\sum_{j=1}^{K}\sum_{g=1}^{K} a_j a_g \hat{\sigma}_{jg}}}. \tag{7.4.2}$$

When the null hypothesis is true and the sample sizes are sufficiently large, then, this statistic has a standard normal distribution. If the alternative hypothesis is true, then, the $Z_j(\tau)$ associated with larger values of a_j should tend to be large, and, thus, the null hypothesis is rejected in favor of H_A at an α Type I error rate when the test statistic is larger than the αth upper percentile of a standard normal distribution.

EXAMPLE 7.6

In section 1.8, a study of 90 patients diagnosed with cancer of the larynx in the 70s at a Dutch hospital was reported. The data consists of the times between first treatment and either death or the end of the study. Patients were classified by the stage of their disease using the American Joint Committee for Cancer Staging. We shall test the hypothesis that there is no difference in the death rates among the four stages of the disease versus the hypothesis that, the higher the stage, the higher the death rate. The data is found on our web site. The four survival curves are shown in Figure 7. 3. We shall use the scores $a_j = j$, $j = 1, \ldots, 4$ in constructing our tests.

Using the log-rank weights,

$$\mathbf{Z}(10.7) = (-7.5660, -3.0117, 2.9155, 7.6623) \text{ and}$$

$$\hat{\Sigma} = \begin{pmatrix} 12.0740 & -4.4516 & -6.2465 & -1.3759 \\ -4.4516 & 7.8730 & -2.7599 & -0.6614 \\ -6.2465 & -2.7599 & 9.9302 & -0.9238 \\ -1.3759 & -0.6614 & -0.9238 & 2.9612 \end{pmatrix}.$$

The value of the test statistic (7.4.2) is 3.72 and the p-value of the test is less than 0.0001.

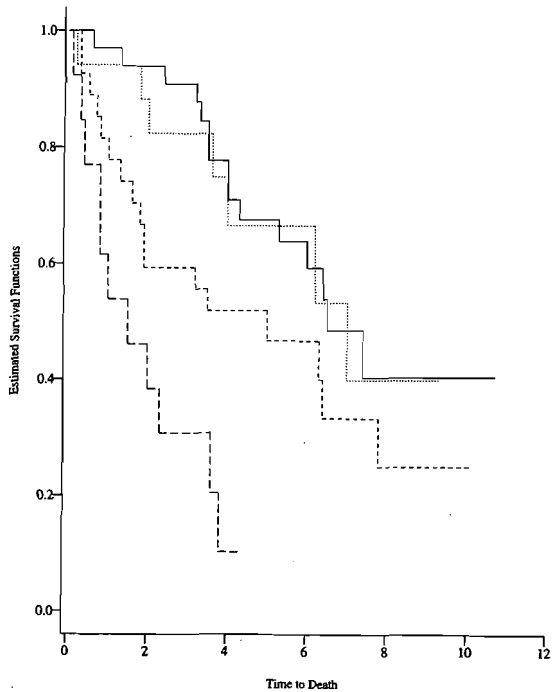

Figure 7.3 *Estimated survival function for larynx cancer patients. Stage I (——) Stage II (——) Stage III (— — —) Stage IV (— — —)*

Using the Tarone and Ware weights, we find that the value of the test statistic is 4.06. Using Gehan's weights, the value of the test statistic is 4.22, and using the Peto and Peto weights, the value of the test statistic is 4.13. All three tests have a p-value of less than 0.0001, providing strong evidence that the hazard rates are in the expected order.

Practical Notes

1. The SAS procedure LIFETEST provides the statistics $Z_j(t)$ and $\hat{\Sigma}$ based on the log-rank weights and Gehan's weights.
2. This test should be applied only when there is some a priori information that the alternatives are ordered.

Theoretical Note

1. When there is no censoring, the test using Gehan or Peto–Peto weights reduces to the Jonckheere–Terpstra test.

7.5 Stratified Tests

Often in applying the tests discussed in section 7.3, an investigator is faced with the problem of adjusting for some other covariates that affect the event rates in the K populations. One approach is to imbed the problem of comparing the K populations into a regression formulation, as described in section 2.6, and perform the test to compare the populations, after an adjustment for covariates, by using one of the techniques described in Chapters 8–10. An alternative approach is to base the test on a stratified version of one of the tests discussed in section 7.3. This approach is feasible when the number of levels of the covariate is not too large or when a continuous covariate is discretized into a workable number of levels. In this section, we shall discuss how such stratified tests are constructed, and we shall see how these tests can be used to analyze the survival of matched pairs.

We assume that our test is to be stratified on M levels of a set of covariates. We shall test the hypothesis

$$H_0 : h_{1s}(t) = h_{2s}(t) = \cdots = h_{Ks}(t), \text{ for } s = 1, \ldots, M, \ t < \tau. \quad (7.5.1)$$

Based only on the data from the sth strata, let $Z_{js}(\tau)$ be the statistic (7.3.3) and $\hat{\Sigma}_s$ be the variance-covariance matrix of the $Z_{js}(\tau)$'s obtained from (7.3.4) and (7.3.5). As in the previous two sections, any choice of weight functions can be used for Z_{js}. These statistics can be used to test the hypothesis of difference in the hazard rates within stratum s by constructing the test statistic (7.3.6). A global test of (7.5.1) is constructed as follows:

$$\text{Let } Z_{j\cdot}(\tau) = \sum_{s=1}^{M} Z_{js}(\tau) \text{ and } \hat{\sigma}_{jg\cdot} = \sum_{s=1}^{M} \hat{\sigma}_{jgs}. \quad (7.5.2)$$

The test statistic, as in (7.3.6), is

$$(Z_{1\cdot}(\tau), \ldots, Z_{K-1\cdot}(\tau))\hat{\Sigma}_{\cdot}^{-1}(Z_{1\cdot}(\tau), \ldots, Z_{K-1\cdot}(\tau))^t \quad (7.5.3)$$

where $\hat{\Sigma}_{\cdot}$ is the $(K-1)\times(K-1)$ matrix obtained from the $\hat{\sigma}_{jg\cdot}$'s. When the total sample size is large and the null hypothesis is true, this statistic has a chi-squared distribution with $K-1$ degrees of freedom. For the

two-sample problem, the stratified test statistic is

$$\frac{\sum_{s=1}^{M} Z_{1s}(\tau)}{\sqrt{\sum_{s=1}^{M} \hat{\sigma}_{11s}}} \tag{7.5.4}$$

which is asymptotically standard normal when the null hypothesis is true.

EXAMPLE 7.7 In section 1.10, the results of a small study comparing the effectiveness of allogeneic (allo) transplants versus autogeneic (auto) transplants for Hodgkin's disease (HOD) or non-Hodgkin's lymphoma (NHL) was presented. Of interest is a test of the null hypothesis of no difference in the leukemia-free-survival rate between patients given an allo ($j = 1$) or auto ($j = 2$) transplant, adjusting for the patient's disease state.

From only the data on Hodgkin's patients, we find $Z_{1\text{HOD}}(2144) = 3.1062$ and $\hat{\sigma}_{11\text{HOD}} = 1.5177$ using log-rank weights. For the non-Hodgkin's lymphoma patients, we find $Z_{1\text{NHL}}(2144) = -2.3056$ and $\hat{\sigma}_{11\text{NHL}} = 3.3556$. The stratified log-rank test is

$$Z = \frac{3.1062 + (-2.3056)}{\sqrt{1.5177 + 3.3556}} = 0.568,$$

which has a p-value of 0.5699.

In this example, if we perform the test only on the Hodgkin's disease patients, we find that the test statistic has a value of 2.89 ($p = 0.004$), whereas using only non-Hodgkin's patients, we find that the test statistic is -1.26 ($p = 0.2082$). The small value of the combined statistic is due, in part, to the reversal of the relationship between transplant procedure and disease-free survival in the Hodgkin's group from that in the non-Hodgkin's group.

EXAMPLE 7.4 *(continued)* In Example 7.4, we found that there was evidence of a difference in disease-free survival rates between bone marrow patients with ALL, low-risk AML and high-risk AML. A proper comparison of these three disease groups should account for other factors which may influence disease-free survival. One such factor is the use of a graft-versus-host prophylactic combining methotretexate (MTX) with some other agent. We shall perform a stratified Gehan weighted test for differences in the hazard rates of the three disease states. Using (7.3.2)–(7.3.5), for the no MTX strata,

$$Z_{1\text{NOMTX}} = -103,\ Z_{2\text{NOMTX}} = -892,\ Z_{3\text{NOMTX}} = 995,$$

$$\hat{\Sigma}_{\text{NOMTX}} = \begin{pmatrix} 49366.6 & -32120.6 & -17246.0 \\ -32120.6 & 69388.9 & -37268.2 \\ -17246.0 & -37268.2 & 54514.2 \end{pmatrix},$$

and, for the MTX strata,

$$Z_{1\text{MTX}} = 20,\ Z_{2\text{NOMTX}} = -45,\ Z_{3\text{NOMTX}} = 25,$$

and

$$\hat{\Sigma}_{\text{NOMTX}} = \begin{pmatrix} 5137.1 & -2685.6 & -2451.6 \\ -2685.6 & 4397.5 & -1711.9 \\ -2451.6 & -1711.9 & 4163.5 \end{pmatrix}.$$

Pooling the results in the two strata,

$$Z_{1\cdot} = -83,\ Z_{2\cdot} = -937,\ Z_{3\cdot} = 1020, \text{ and}$$

$$\hat{\Sigma}_{\cdot} = \begin{pmatrix} 54503.7 & -34806.2 & -19697.6 \\ -34806.2 & 73786.1 & -38980.1 \\ -19697.6 & -38980.1 & 58677.7 \end{pmatrix}.$$

The stratified Gehan test statistic is

$$(-83,\ -937)\begin{pmatrix} 54503.7 & -34806.2 \\ -34806.2 & 73786.1 \end{pmatrix}^{-1}\begin{pmatrix} -83 \\ -937 \end{pmatrix} = 19.14$$

which has a p-value of less than 0.0001 when compared to a chi-square with two degrees of freedom. The tests on the individual strata give test statistics of $\chi^2 = 19.1822$ ($p = 0.0001$) for the no MTX group and $\chi^2 = 0.4765$ ($p = 0.7880$) in the MTX arm. The global test, ignoring MTX, found a test statistic of 16.2407 with a p-value of 0.0003.

Another use for the stratified test is for matched pairs, censored, survival data. Here we have paired event times (T_{1i}, T_{2i}) and their corresponding event indicators $(\delta_{1i}, \delta_{2i})$, for $i = 1, \ldots, M$. We wish to test the hypothesis $H_0 : h_{1i}(t) = h_{2i}(t)$, $i = 1, \ldots, M$. Computing the statistics (7.3.3) and (7.3.4),

$$Z_{1i}(\tau) = \begin{cases} W(T_{1i})(1 - 1/2) = W(T_{1i})/2 & \text{if } T_{1i} < T_{2i},\ \delta_{1i} = 1 \\ & \text{or } T_{1i} = T_{2i},\ \delta_{1i} = 1,\ \delta_{2i} = 0 \\ W(T_{2i})(0 - 1/2) = -W(T_{2i})/2 & \text{if } T_{2i} < T_{1i},\ \delta_{2i} = 1 \\ & \text{or } T_{2i} = T_{1i},\ \delta_{2i} = 1,\ \delta_{1i} = 0 \\ 0 & \text{otherwise} \end{cases}, \tag{7.5.5}$$

and

$$\hat{\sigma}_{11i} = \begin{cases} W(T_{1i})^2(1/2)(1 - 1/2) = W(T_{1i})^2/4 & \text{if } T_{1i} < T_{2i},\ \delta_{1i} = 1 \\ & \text{or } T_{1i} = T_{2i},\ \delta_{1i} = 1,\ \delta_{2i} = 0 \\ W(T_{2i})^2(1/2)(1 - 1/2) = W(T_{2i})^2/4 & \text{if } T_{2i} < T_{1i},\ \delta_{2i} = 1 \\ & \text{or } T_{2i} = T_{1i},\ \delta_{2i} = 1,\ \delta_{1i} = 0 \\ 0 & \text{otherwise} \end{cases}.$$

For any of the weight functions we have discussed,

$$Z_{1\cdot}(\tau) = w\frac{D_1 - D_2}{2} \tag{7.5.6}$$

and

$$\hat{\sigma}_{11.} = w^2 \frac{D_1 + D_2}{4},$$

where D_1 is the number of matched pairs in which the individual from sample 1 experiences the event first and D_2 is the number in which the individual from sample 2 experiences the event first. Here w is the value of the weight function at the time when the smaller of the pair fails. Because these weights do not depend on which group this failure came from, the test statistic is

$$\frac{D_1 - D_2}{\sqrt{D_1 + D_2}}, \tag{7.5.7}$$

which has a standard normal distribution when the number of pairs is large and the null hypothesis is true. Note that matched pairs, where the smaller of the two times corresponds to a censored observation, make no contribution to $Z_{1.}$ or $\hat{\sigma}_{11.}$.

EXAMPLE 7.8

In section 1.2, the results of a clinical trial of the drug 6-mercaptopurine (6-MP) versus a placebo in 42 children with acute leukemia was described. The trial was conducted by matching pairs of patients at a given hospital by remission status (complete or partial) and randomizing within the pair to either a 6-MP or placebo maintenance therapy. Patients were followed until their leukemia returned (relapse) or until the end of the study.

Survival curves for the two groups were computed in Example 4.1. We shall now test the hypothesis that there is no difference in the rate of recurrence of leukemia in the two groups. From Table 1.1, we find $D_{\text{Placebo}} = 18$ and $D_{\text{6-MP}} = 3$, so that the test statistic is $(18 - 3)/(18 + 3)^{1/2} = 3.27$. The p-value of the test is $2Pr[Z \geq 3.27] = 0.001$, so that there is sufficient evidence that the relapse rates are different in the two groups.

Practical Notes

1. The test for matched pairs uses only information from those pairs where the smaller of the two times corresponds to an event time. The effective sample size of the test is the number of such pairs.
2. The test for matched pairs is the censored-data version of the sign test.
3. The stratified tests will have good power against alternatives that are in the same direction in each stratum. When this is not the case, these statistics may have very low power, and separate tests for each stratum are indicated. (See Example 7.5.)

Theoretical Notes

1. The test for matched pairs relies only on intrapair comparisons. Other tests for matched pairs have been suggested which assume a bivariate model for the paired times, but make interpair comparisons.
2. The asymptotic chi-squared distribution of the stratified tests discussed in this section is valid when either the number of strata is fixed and the number within each stratum is large or when the number in each stratum is fixed and the number of strata is large. See Andersen et al. (1993) for details on the asymptotics of these tests.

7.6 Renyi Type Tests

In section 7.3, a series of tests to compare the survival experience of two or more samples were presented. All of these tests were based on the weighted integral of estimated differences between the cumulative hazard rates in the samples. When these tests are applied to samples from populations where the hazard rates cross, these tests have little power because early differences in favor of one group are canceled out by late differences in favor of the other treatment. In this section, we present a class of tests with greater power to detect crossing hazards. We will focus on the two sample versions of the test.

The test statistics to be used are called Renyi statistics and are censored-data analogs of the Kolmogorov–Smirnov statistic for comparing two uncensored samples. To construct the tests, we will find the value of the test statistic (7.3.3) for some weight function at each death time. When the hazard rates cross, the absolute value of these sequential evaluations of the test statistic will have a maximum value at some time point prior to the largest death time. When this value is too large, then, the null hypothesis of interest $H_0 : h_1(t) = h_2(t)$, $t < \tau$, is rejected in favor of $H_A : h_1(t) \neq h_2(t)$, for some t. To adjust for the fact that we are constructing multiple test statistics on the same set of data, a correction is made to the critical value of the test.

To construct the test, suppose that we have two independent samples of size n_1 and n_2, respectively. Let $n = n_1 + n_2$. Let $t_1 < t_2 < \cdots < t_D$ be the distinct death times in the pooled sample. In sample j let d_{ij} be the number of deaths and Y_{ij} the number at risk at time t_i, $i = 1, \ldots, D$, $j = 1, 2$. Let $Y_i = Y_{i1} + Y_{i2}$ be the total number at risk in both samples and $d_i = d_{i1} + d_{i2}$ be the total number of deaths in the combined sample at time t_i. Let $W(t)$ be a weight function. For example, for the "log-rank" version of this test $W(t) = 1$ and, for the "Gehan–Wilcoxon" version, $W(t_i) = Y_{i1} + Y_{i2}$. For each value of t_i we compute, $Z(t_i)$, which is the value of the numerator of the statistic (7.3.7) using only

the death times observed up to time t_i, that is,

$$Z(t_i) = \sum_{t_k \le t_i} W(t_k)\left[d_{k1} - Y_{k1}\left(\frac{d_k}{Y_k}\right)\right], \quad i = 1, \ldots, D. \tag{7.6.1}$$

Let $\sigma(\tau)$ be the standard error of $Z(\tau)$ which, from (7.3.7), is given by

$$\sigma^2(\tau) = \sum_{t_k \le \tau} W(t_k)^2 \left(\frac{Y_{k1}}{Y_k}\right)\left(\frac{Y_{k2}}{Y_k}\right)\left(\frac{Y_k - d_k}{Y_k - 1}\right)(d_k); \tag{7.6.2}$$

where τ is the largest t_k with $Y_{k1}, Y_{k2} > 0$.

The test statistic for a two-sided alternative is given by

$$Q = \sup\{|Z(t)|,\ t \le \tau\}/\sigma(\tau). \tag{7.6.3}$$

When the null hypothesis is true, then, the distribution of Q can be approximated by the distribution of the $\sup(|B(x)|,\ 0 \le x \le 1)$ where B is a standard Brownian motion process. Critical values of Q are found in Table C.5 in Appendix C.

The usual weighted log rank test statistic is $Z(\tau)/\sigma(\tau)$. For this test, when the two hazard rates cross, early positive differences between the two hazard rates are canceled out by later differences in the rates, with opposite signs. The supremum version of the statistic should have greater power to detect such differences between the hazard rates.

EXAMPLE 7.9 A clinical trial of chemotherapy against chemotherapy combined with radiotherapy in the treatment of locally unresectable gastric cancer was conducted by the Gastrointestinal Tumor Study Group (1982). In this trial, forty-five patients were randomized to each of the two arms and followed for about eight years. The data, found in Stablein and Koutrouvelis (1985), is as follows:

Chemotherapy Only

1	63	105	129	182	216	250	262	301	301	342	354
356	358	380	383	383	388	394	408	460	489	499	523
524	535	562	569	675	676	748	778	786	797	955	968
1000	1245	1271	1420	1551	1694	2363	2754*	2950*			

Chemotherapy Plus Radiotherapy

17	42	44	48	60	72	74	95	103	108	122	144
167	170	183	185	193	195	197	208	234	235	254	307
315	401	445	464	484	528	542	547	577	580	795	855
1366	1577	2060	2412*	2486*	2796*	2802*	2934*	2988*			

*Denotes censored observation.

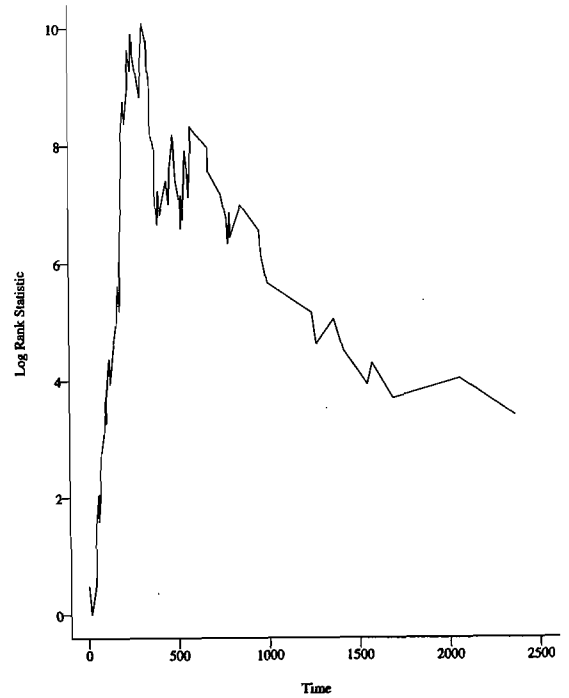

Figure 7.4 *Values of $|Z(t_i)|$ for the gastrointestinal tumor study*

We wish to test the hypothesis that the survival rate of the two groups is the same by using the log rank version ($W(t_i) = 1$) of the Renyi statistics. Figure 7.4 shows the value of $|Z(t_i)|$. Here, the maximum occurs at $t_i = 315$ with a value of 9.80. The value of $\sigma(2363) = 4.46$, so $Q = 2.20$. From Table C.5 in Appendix C we find that the p-value of this test is 0.053 so the null hypothesis of no difference in survival rates between the two treatment groups is not rejected at the 5% level. If we had used the nonsequential log-rank test, we have $Z(2363) = -2.15$, yielding a p-value of 0.6295 which is not significant. From Figure 7.5, which plots the Kaplan–Meier curves for the two samples, we see that the usual log rank statistic has a small value because early differences in favor of the chemotherapy only group are negated by a late survival advantage for the chemotherapy plus radiotherapy group.

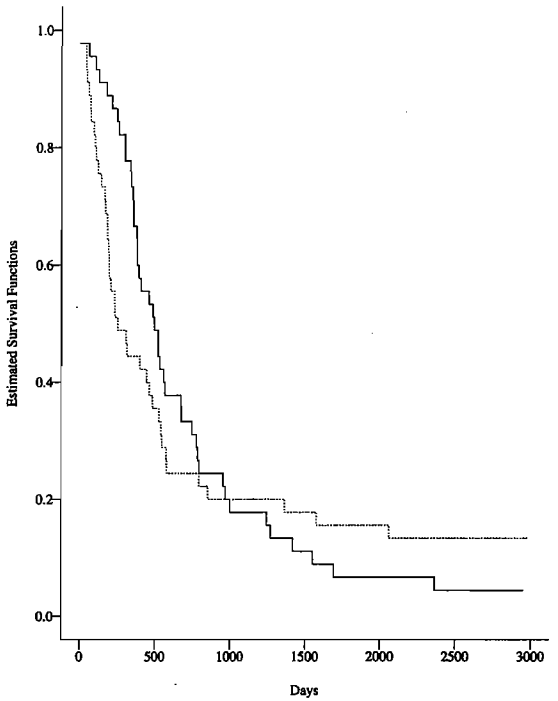

Figure 7.5 *Estimated survival functions for the gastrointestinal tumor study. Chemotherapy only (———) Chemotherapy plus radiation (——)*

Practical Notes

1. A one-sided test of the hypothesis $H_0 : S_1(t) = S_2(t)$ against $H_A : S_1(t) < S_2(t)$ can be based on $Q^* = \sup[Z(t), t \leq \tau]/\sigma(\tau)$. When H_0 is true, Q^* converges in distribution to the supremum of a Brownian motion process $B(t)$ (see Theoretical Note 3 above). The p-value of a one-sided test is given by $Pr[\sup B(t) > Q^*] = 2[1 - \Phi(Q^*)]$, where $\Phi()$ is the standard normal cumulative distribution function.

Theoretical Notes

1. The supremum versions of the weighted, log-rank tests were proposed by Gill (1980). He calls the statistic (7.6.3) a "Renyi-type"

statistic. Further development of the statistical properties of the test can be found in Fleming and Harrington (1991) and Fleming et al. (1980). Schumacher (1984) presents a comparison of this class of statistics to tests based on the complete test statistic and other versions of the Kolmogorov–Smirnov statistic.

2. For a standard Brownian motion process $B(t)$, Billingsly (1968) shows that

$$Pr[\sup |B(t)| > y] = 1 - \frac{4}{\pi}\sum_{k=0}^{\infty}\frac{(-1)^k}{2k+1}\exp[-\pi^2(2k+1)^2/8y^2].$$

3. Using the counting process methods introduced in section 3.6, one can show, under H_0, that, if $\sigma^2(t)$ converges in probability to a limiting variance $\sigma_0^2(t)$ on $[0, \infty)$ then, $Z(t)$ converges weakly to the process $B[\sigma_0^2(t)]$ on the interval $[0, \infty]$. This implies that $\sup[Z(t)/\sigma(\infty) : 0 < t < \infty]$ converges in distribution to $[\sup B(t) : t \in A]$ where $A = \{\sigma_0^2(t)/\sigma_0^2(\infty), 0 \leq t \leq \infty\}$. When the underlying common survival function is continuous, then, A is the full unit interval, so the asymptotic p-values are exact. When the underlying common survival function is discrete, then, A is a subset of the interval (0, 1), and the test is a bit conservative. See Fleming et al. (1987) for details of the asymptotics.
4. Other extensions of the Kolmogorov–Smirnov test have been suggested in the literature. Schumacher (1984) provides details of tests based on the maximum value of either $\log[\hat{H}_1(t)] - \log[\hat{H}_2(t)]$ or $\hat{H}_1(t) - \hat{H}_2(t)$. In a Monte Carlo study, he shows that the performance of these statistics is quite poor, and they are not recommended for routine use.
5. Both Schumacher (1984) and Fleming et al. (1987) have conducted simulation studies comparing the Renyi statistic of this section to the complete test statistic of section 7.3. For the log-rank test, they conclude there is relatively little loss of power for the Renyi statistics when the hazard rates are proportional and there is little censoring. For nonproportional or crossing hazards the Renyi test seems to perform much better than the usual log-rank test for light censoring. The apparent advantage of the Renyi statistic for light censoring diminishes as the censoring fraction increases.

7.7 Other Two-Sample Tests

In this section, we present three other two-sample tests which have been suggested in the literature. These tests are constructed to have greater power than the tests in section 7.3 to detect crossing hazard

rates. All three tests are analogs of common nonparametric tests for uncensored data.

The first test is a censored-data version of the Cramer–von Mises test. For uncensored data, the Cramer–Von Mises test is based on the integrated, squared difference between the two empirical survival functions. For right-censored data, it is more appropriate to base a test on the integrated, squared difference between the two estimated hazard rates. This is done to obtain a limiting distribution which does not depend on the relationship between the death and censoring times and because such tests arise naturally from counting process theory. We shall present two versions of the test.

To construct the test, recall, from Chapter 4, that the Nelson–Aalen estimator of the cumulative hazard function in the jth sample is given by

$$\tilde{H}_j(t) = \sum_{t_i \le t} \frac{d_{ij}}{Y_{ij}}, \quad j = 1, 2. \tag{7.7.1}$$

An estimator of the variance of $\tilde{H}_j(t)$ is given by

$$\sigma_j^2(t) = \sum_{t_j \le t} \frac{d_{ij}}{Y_{ij}(Y_{ij} - 1)}, \quad j = 1, 2. \tag{7.7.2}$$

Our test is based on the difference between $\tilde{H}_1(t)$ and $\tilde{H}_2(t)$, so that we need to compute $\sigma^2(t) = \sigma_1^2(t) + \sigma_2^2(t)$, which is the estimated variance of $\tilde{H}_1(t) - \tilde{H}_2(t)$. Also let $A(t) = n\sigma^2(t)/[1 + n\sigma^2(t)]$.

The first version of the Cramer-von Mises statistic is given by

$$Q_1 = \left(\frac{1}{\sigma^2(\tau)}\right)^2 \int_0^\tau [\tilde{H}_1(t) - \tilde{H}_2(t)]^2 d\sigma^2(t)$$

which can be computed as

$$Q_1 = \left(\frac{1}{\sigma^2(\tau)}\right)^2 \sum_{t_i \le \tau} [\tilde{H}_1(t_i) - \tilde{H}_2(t_i)]^2 [\sigma^2(t_i) - \sigma^2(t_{i-1})], \tag{7.7.3}$$

where $t_0 = 0$, and the sum is over the distinct death times less than τ. When the null hypothesis is true, one can show that the large sample distribution of Q_1 is the same as that of $R_1 = \int_0^1 [B(x)]^2 dx$, where $B(x)$ is a standard Brownian motion process. The survival function of R_1 is found in Table C.6 of Appendix C.

An alternate version of the Cramer–von Mises statistic is given by

$$Q_2 = n \int_0^\tau \left[\frac{\tilde{H}_1(t) - \tilde{H}_2(t)}{1 + n\sigma^2(t)}\right]^2 dA(t)$$

which is computed as

$$Q_2 = n \sum_{t_i \le \tau} \left[\frac{\tilde{H}_1(t_i) - \tilde{H}_2(t_i)}{1 + n\sigma^2(t_i)}\right]^2 [A(t_i) - A(t_{i-1})]. \tag{7.7.4}$$

When the null hypothesis is true, the large sample distribution of Q_2 is the same as that of $R_2 = \int_0^{A(\tau)} [B^0(x)]^2 dx$, where $B^0()$ is a Brownian bridge process. Table C.7 of Appendix C provides critical values for the test based on Q_2.

EXAMPLE 7.2 *(continued)* We shall apply the two Cramer–von Mises tests to the comparison of the rate of cutaneous exit-site infections for kidney dialysis patients whose catheters were placed surgically (group 1) as compared to patients who had percutaneous placement of their catheters (group 2). Routine calculations yield $Q_1 = 1.8061$ which, from Table C.6 of Appendix C, has a p-value of 0.0399. For the second version of the Cramer–von Mises test, $Q_2 = 0.2667$ and $A(\tau) = 0.99$. From Table C.7 of Appendix C, we find that this test has a p-value of 0.195.

A common test for uncensored data is the two-sample t-test, based on the difference in sample means between the two populations. The second test we present is a censored-data version of this test based on the Kaplan–Meier estimators in the two samples, $\hat{S}_1(t)$ and $\hat{S}_2(t)$. In section 4.5, we saw that the population mean can be estimated by the area under the Kaplan–Meier curve $\hat{S}(t)$. This suggests that a test based on the area under the curve $\hat{S}_1(t) - \hat{S}_2(t)$, over the range where both of the two samples still have individuals at risk, will provide a censored data analog of the two-sample t-test. For censored data, we have seen that the estimate of the survival function may be unstable in later time intervals when there is heavy censoring, so that relatively small differences in the Kaplan–Meier estimates in the tail may have too much weight when comparing the two survival curves. To handle this problem, the area under a weighted difference in the two survival functions is used. The weight function, which downweights differences late in time when there is heavy censoring, is based on the distribution of the censoring times.

To construct the test, we pool the observed times in the two samples. Let $t_1 < t_2 < \cdots < t_n$ denote the ordered times. Notice that, here, as opposed to the other procedures where only event times are considered, the t_i's consist of both event and censoring times. Let d_{ij}, c_{ij}, Y_{ij} be, respectively, the number of events, the number of censored observations, and the number at risk at time t_i in the jth sample, $j = 1, 2$. Let $\hat{S}_j(t)$ be the Kaplan–Meier estimator of the event distribution using data in the jth sample and let $\hat{G}_j(t)$ be the Kaplan–Meier estimator of the time to censoring in the jth sample, that is, $\hat{G}_j(t) = \prod_{t_i \le t}[1 - c_{ij}/Y_{ij}]$.

Finally, let $\hat{S}_p(t)$ be the Kaplan–Meier estimator based on the combined sample.

To construct the test statistic, we define a weight function by

$$w(t) = \frac{n\hat{G}_1(t)\hat{G}_2(t)}{n_1\hat{G}_1(t) + n_2\hat{G}_2(t)}, \quad 0 \le t \le t_D \tag{7.7.5}$$

where n_1 and n_2 are the two sample sizes and $n = n_1 + n_2$. Notice that $w(t)$ is constant between successive censored observations and, when there is heavy censoring in either sample, $w(t)$ is close to zero. When there is no censoring, $w(t)$ is equal to 1. The test statistic is given by

$$W_{KM} = \sqrt{\frac{n_1 n_2}{n}} \int_0^{t_D} w(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt$$

which can be computed by

$$W_{KM} = \sqrt{\frac{n_1 n_2}{n}} \sum_{i=1}^{D-1} [t_{i+1} - t_i]w(t_i)[\hat{S}_1(t_i) - \hat{S}_2(t_i)]. \tag{7.7.6}$$

To find the variance of W_{KM}, first, compute

$$A_i = \int_{t_i}^{t_D} w(u)\hat{S}_p(u)du = \sum_{k=i}^{D-1}(t_{k+1} - t_k)w(t_k)\hat{S}_p(t_k). \tag{7.7.7}$$

The estimated variance of $W_{KM} = \hat{\sigma}_p^2$ is given by

$$\hat{\sigma}_p^2 = \sum_{i=1}^{D-1} \frac{A_i^2}{\hat{S}_p(t_i)\hat{S}_p(t_{i-1})} \frac{n_1\hat{G}_1(t_{i-1}) + n_2\hat{G}_2(t_{i-1})}{n\hat{G}_1(t_{i-1})\hat{G}_2(t_{i-1})}[\hat{S}_p(t_{i-1}) - \hat{S}_p(t_i)]. \tag{7.7.8}$$

Note that the sum in (7.7.8) has only nonzero contributions when t_i is a death time, because, at censored observations, $\hat{S}_p(t_{i-1}) - \hat{S}_p(t_i) = 0$. When there is no censoring, $\hat{\sigma}_p^2$ reduces to the usual sample variance on the data from the combined sample.

To test the null hypothesis that $S_1(t) = S_2(t)$, the test statistic used is $Z = W_{KM}/\hat{\sigma}_p$ which has an approximate standard normal distribution when the null hypothesis is true. If the alternative hypothesis is that $S_1(t) > S_2(t)$, then, the null hypothesis is rejected when Z is larger than the αth upper percentage point of a standard normal, whereas, for a two-sided alternative, the null hypothesis is rejected when the absolute value of Z is larger than $\alpha/2$ upper percentage point of a standard normal.

EXAMPLE 7.5 *(continued)* We shall calculate the weighted difference of Kaplan–Meier estimators statistic for the comparison of auto versus allo transplants. The calculations yield a value of 5.1789 for W_{KM} and 141.5430 for $\hat{\sigma}_p^2$, so $Z = 0.4353$. The p-value of the two-sided test of equality of the two survival curves is $2Pr[Z \ge 0.4553] = 0.6634$.

The final test we shall present is a censored-data version of the two-sample median test proposed by Brookmeyer and Crowley (1982b). This test is useful when one is interested in comparing the median survival times of the two samples rather than in comparing the difference in the hazard rates or the survival functions over time. The test has reasonable power to detect differences between samples when the hazard rates of the two populations cross. It is sensitive to differences in median survival for any shaped survival function.

To construct the test, we have two, independent, censored samples of sizes n_1 and n_2, respectively. Let $n = n_1 + n_2$ be the total sample size. For each sample, we construct the Kaplan–Meier estimator (4.2.1), $\hat{S}_j(t)$, $j = 1, 2$. When the null hypothesis of no difference in survival between the two groups is true, an estimator of the common survival function is the weighted Kaplan–Meier estimator,

$$\hat{S}_W(t) = \frac{n_1\hat{S}_1(t) + n_2\hat{S}_2(t)}{n}. \tag{7.7.9}$$

This weighted Kaplan–Meier estimator represents the survival function of an average individual on study and is a function only of the survival experiences in the two samples. It does not depend on the censoring patterns in the two samples, as would the Kaplan–Meier estimate based on the combined sample.

Using the weighted Kaplan–Meier estimator, an estimate of the pooled sample median $\hat{M}$ is found as follows. Let $t_1 < t_2 < \cdots < t_D$ be the event times in the pooled sample. If $\hat{S}_W(t_i) = 0.5$ for some death time, set $\hat{M} = t_i$. If no event time gives a value of $\hat{S}_W$ equal to 0.5, set M_L as the largest event time with $\hat{S}_W(M_L) > 0.5$ and M_U as the smallest event time with $\hat{S}_W(M_U) < 0.5$. The pooled median must lie in the interval (M_L, M_U) and is found by linear interpolation, that is,

$$\hat{M} = M_L + \frac{(0.5 - \hat{S}_W(M_L))(M_U - M_L)}{\hat{S}_W(M_U) - \hat{S}_W(M_L)}. \tag{7.7.10}$$

To compute this median, we are using a version of the weighted Kaplan–Meier estimator, which connects the values of the estimator at death times by a straight line, rather than the usual estimator which is a step function.

Once the pooled sample median is found, the estimated probability that a randomly selected individual in the jth sample exceeds this value is computed from each sample's Kaplan–Meier estimator. Again, a smoothed version of the Kaplan–Meier estimator, which connects the values of the estimator at each death time, is used in each sample. We find the two death times in the jth sample that bracket $\hat{M}$, $T_{Lj} \le \hat{M} < T_{Uj}$. The estimated probability of survival beyond $\hat{M}$ in the jth sample,

found by linear interpolation, is given by

$$\hat{S}_j(\hat{M}) = \hat{S}_j(T_{Lj}) + \frac{(\hat{S}_j(T_{Uj}) - \hat{S}_j(T_{Lj}))(\hat{M} - T_{Lj})}{(T_{Uj} - T_{Lj})}, \; j = 1, 2. \quad (7.7.11)$$

The test is based on comparing this value to 0.5, the expected survival if the null hypothesis of no difference in median survival between the two groups is true, that is, the test is based on the statistic $n^{1/2}[S_1(\hat{M}) - 0.5]$. For sufficiently large samples, this statistic has an approximate normal distribution with a variance found as follows. As usual, let t_{ij} denote the distinct death times in the jth sample, d_{ij} the number of deaths at time t_{ij} and Y_{ij} the number at risk at time t_{ij}. For $j = 1, 2$, define

$$V_j = \left[\hat{S}_j(T_{Uj})\left(\frac{\hat{M} - T_{Lj}}{T_{Uj} - T_{Lj}}\right)\right]^2 \sum_{t_{ij} \le T_{Uj}} \frac{d_{ij}}{Y_{ij}(Y_{ij} - d_{ij})} \quad (7.7.12)$$

$$+ \left\{\left[\hat{S}_j(T_{Lj})\left(\frac{T_{Uj} - \hat{M}}{T_{Uj} - T_{Lj}}\right)\right]^2 + \frac{2(\hat{M} - T_{Lj})(T_{Uj} - \hat{M})}{(T_{Uj} - T_{Lj})^2}\hat{S}(T_{Uj})\hat{S}(T_{Lj})\right\}$$

$$\times \sum_{t_{ij} \le T_{Lj}} \frac{d_{ij}}{Y_{ij}(Y_{ij} - d_{ij})}.$$

Then, the variance of $n^{1/2}[S_1(\hat{M}) - 0.5]$ is estimated consistently by

$$\sigma^2 = \frac{n_2^2}{n}\{V_1 + V_2\}, \quad (7.7.13)$$

and the (two-sided) test statistic is

$$\chi^2 = n\frac{[S_1(\hat{M}) - 0.5]^2}{\sigma^2}, \quad (7.7.14)$$

which has a chi-squared distribution with one degree of freedom when the null hypothesis is true.

EXAMPLE 7.5 *(continued)* We shall illustrate the median test on the data comparing allo and auto transplants. Here the estimated median from the pooled sample $\hat{M} = 17.9225$. Using the data from the first sample, we find $\hat{S}_1(\hat{M}) = 0.5409$ and $\hat{S}_2(\hat{M}) = 0.4395$. Routine calculations find that $V_1 = 0.0122$ and $V_2 = 0.0140$, so $\sigma^2 = 0.6754$. Thus, $\chi^2 = 101(0.5409 - 0.5)^2/0.6754 = 0.2496$. The p-value of the test is the probability that a chi-squared, random variable with one degree of freedom will exceed this value, which is 0.6173.

Practical Notes

1. Schumacher (1984) has studied the small-sample properties of the two Cramer–von Mises type tests. He concludes that there is some loss of power using these tests, compared to the log-rank test of section 7.3, when the hazard rates in the two samples are proportional. However the test based on Q_1 seems to perform quite well in this case. Tests based on Q_2 perform extremely well compared to other tests, when there is a large early difference in the hazard rates or when the hazard rates cross.
2. An alternative Cramer–von Mises test was proposed by Koziol (1978). This test is based on the statistic

$$-\frac{n_1 n_2}{n_1 + n_2}\int_0^\tau [\hat{S}_1(t) - \hat{S}_2(t)]^2 d[\hat{S}_p(t)],$$

where $\hat{S}_1(t)$, $\hat{S}_2(t)$, and $\hat{S}_p(t)$ are the Kaplan–Meier estimates of the survival function from the first, second, and pooled samples, respectively. This statistic reduces to the usual Cramer–von Mises test when there is no censoring. The asymptotic distribution of the statistic, however, can only be derived under the additional assumption that the two censoring distributions are equal and that the hazard rate of the censoring distribution is proportional to the hazard rate of the death distribution. In his Monte Carlo study, Schumacher (1984) shows that the performance of this test is similar to that of Q_2.
3. Any weight function with the property that $|w(t)| \le \gamma \hat{G}_j^{(1/2)+\delta}$, for $\gamma, \delta > 0$ can be used in the calculation of W_{KM}. Another choice of $w(t)$, suggested by Pepe and Fleming (1989), is the square root of equation (7.7.5). If one wishes to compare the two survival curves over some range, say $t \ge t_0$, the weight function $w(t)I[t \ge t_0]$ may be appropriate. Other choices of weights could be motivated by some measure of quality of life or the cost of treatment.
4. When there is no censoring, the estimated variance based on (7.7.8) is equal to the sample variance based on the pooled sample. This is a different estimator of the common variance in the two samples than the usual pooled sample variance constructed as a weighted average of the individual sample variances found in most elementary text books.
5. An alternate estimator of the variance of W_{KM} can be constructed by using an unpooled variance estimator. Here, let

$$A_{ij} = \int_{t_i}^{t_D} w(u)\hat{S}_j(u)du, \; j = 1, 2, \; i = 1, \ldots, D - 1.$$

The unpooled estimator of the variance is

$$\hat{\sigma}_{up}^2 = \frac{n_1 n_2}{n}\left\{\sum_{j=1}^{2}\frac{1}{n_j - 1}\sum_{i=1}^{D-1}\frac{A_{ij}^2}{\hat{S}_j(t_i)\hat{S}_j(t_{i-1})\hat{G}_j(t_{i-1})}[\hat{S}_j(t_{i-1}) - \hat{S}_j(t_i)]\right\},$$

which, in the uncensored-data case, reduces to $(n_2/n)S_1^2+(n_1/n)S_2^2$, with S_j^2 the usual sample variance. Monte Carlo studies by Pepe and Fleming (1989) show that this variance estimator performs poorer than the pooled estimator (7.7.8) and that its performance is poor when the censoring patterns are different in the two samples.

6. For n_j moderately large (> 50), one can approximate V_j in (7.7.12) by the simple expression

$$\hat{S}_j(\hat{M})^2 \sum_{t_{ij} \le \hat{M}} \frac{d_{ij}}{Y_{ij}(Y_{ij} - d_{ij})}.$$

7. Brookmeyer and Crowley (1982b) present an extension of this two-sample median test to the K-sample problem.

Theoretical Notes

1. The weighted Kaplan–Meier test was proposed by Pepe and Fleming (1989) who developed its properties. This statistic can not be derived using counting process techniques. Pepe and Fleming (1991) give details of the asymptotic properties of the test.
2. A small-sample Monte Carlo study reported in Pepe and Fleming (1989) shows that, when the hazard rates in the two populations are proportional, the power of W_{KM} is slightly less than that of the log-rank test. The test performs substantially better when the two hazard rates cross. This observation is confirmed in Pepe and Fleming (1991) who base these observations on the asymptotic relative efficiency of the two tests.
3. Brookmeyer and Crowley (1982b) discuss the asymptotic relative efficiency of the median test as compared to the log-rank and Wilcoxon tests. They show that the asymptotic relative efficiency is about half of that of the log-rank test for proportional hazards alternatives, but about twice that of the log-rank test for a translation alternative. The performance of these tests is also presented for small-sample sizes based on a Monte Carlo study.

7.8 Test Based on Differences in Outcome at a Fixed Point in Time

Up to this point we have considered tests that compare hazard rates or survival functions over a range of time points. Occasionally we are interested in comparing K survival curves or K cumulative incidence curves

at a predetermined fixed point in time, t_0. It is important to emphasize that the fixed point, t_0, must be selected before the data is examined. It would make the p-values invalid if the curves were compared for a variety of points. For example, one may wish to compare the survival curves at 1 year or the cumulative incidence curves at 3 years. We have available to us Kaplan–Meier estimators of the survival function or estimated cumulative incidence functions as well as estimates of the variances of these statistics (See Chapter 4 for the calculation of these quantities).

The tests statistics we use are special cases of tests for contrasts between a set of parameters. If we let $\boldsymbol{\Theta}^t = (\theta_1, \ldots, \theta_p)$ be a p-parameter λ_j vector, then a contrast is a linear combination of the covariates. A contrast is a set of coefficients $\mathbf{c} = (c_1 \ldots c_p)$ which define a new parameter $\theta^c = \mathbf{c}\boldsymbol{\Theta} = c_1\theta_1 + \cdots + c_p\theta_p$. For example, if $p = 3$, then the vector $\mathbf{c} = (1, -1, 0)$ yields $\theta^c = \theta_1 - \theta_2$ and a test of the hypothesis that $\theta^C = 0$ is a test that $\theta_1 = \theta_2$.

Suppose that we have q contrasts $\mathbf{c}_k = (c_{k1}, \ldots, c_{kp}), k = 1, \ldots, q$, and we wish to test the hypothesis that $\mathbf{c}_k\boldsymbol{\Theta} = 0$ for all k, then the test statistic will be a *quadratic form* constructed using the estimates of $\theta_1, \ldots, \theta_p$. To construct the quadratic form we define a $q \times p$ contrast matrix

$$\mathbf{C} = \begin{pmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_q \end{pmatrix}. \tag{7.8.1}$$

We compute an estimate of $\theta_j, \hat{\theta}_j$ and the variance matrix, $\mathbf{V}$, with elements, $\hat{\text{V}}\text{ar}[\hat{\theta}_j, \hat{\theta}_k]$. To test the hypothesis $H_0 : \mathbf{C}\boldsymbol{\Theta}^t = \mathbf{0}$, the test statistic is

$$\chi^2 = [\mathbf{C}\hat{\boldsymbol{\Theta}}]^t[\mathbf{C}\mathbf{V}\mathbf{C}^t]^{-1}\mathbf{C}\hat{\boldsymbol{\Theta}}. \tag{7.8.2}$$

When the estimators are approximately normally distributed, this form has an asymptotically chi-squared with q degrees of freedom.

In a survival analysis context we wish to test

$$H_0 : S_1(t_0) = S_2(t_0) = \cdots = S_K(t_0) \quad \text{versus} \tag{7.8.3}$$

H_A : at least one of the $S_j(t_0)$'s is different, for predetermined t_0,

or

$$H_0 : \text{CI}_1(t_0) = \text{CI}_2(t_0) = \cdots = \text{CI}_K(t_0) \quad \text{versus} \tag{7.8.4}$$

H_A : at least one of the $\text{CI}_j(t_0)$'s is different, for predetermined t_0.

Notation similar to that used in sections 4.2, 4.7, and 7.3 will be used and the groups will be assumed to be independent. Let $\hat{\theta}_j$ be the Kaplan–Meier estimate of the jth survival curve or the estimate of the

jth cumulative incidence curve at the predetermined time point t_0. $\mathbf{C}$ will be taken to be the $p-1 \times p$ matrix

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 & -1 \\ 0 & 1 & 0 & \dots & 0 & -1 \\ & & & \cdot & & \\ & & & \cdot & & \\ & & & \cdot & & \\ 0 & 0 & 0 & \dots & 1 & -1 \end{pmatrix}$$

Here $\mathbf{V}$ is a diagonal matrix with elements $V_k = \hat{V}(\hat{\theta}_k(t_0)), k = 1, \dots, p$.

The quadratic form (7.8.2) is

$$\chi^2 = \begin{pmatrix} \hat{\theta}_1 - \hat{\theta}_p \\ \hat{\theta}_2 - \hat{\theta}_p \\ \cdot \\ \cdot \\ \cdot \\ \hat{\theta}_{p-1} - \hat{\theta}_p \end{pmatrix}' \begin{pmatrix} V_1 + V_p & V_p & \dots & V_p \\ V_p & V_2 + V_p & \dots & V_p \\ \cdot & \cdot & \dots & \\ \cdot & \cdot & \dots & \\ \cdot & \cdot & \dots & \\ V_p & V_p & \dots & V_{p-1} + V_p \end{pmatrix}^{-1} \begin{pmatrix} \hat{\theta}_1 - \hat{\theta}_p \\ \hat{\theta}_2 - \hat{\theta}_p \\ \cdot \\ \cdot \\ \cdot \\ \hat{\theta}_{p-1} - \hat{\theta}_p \end{pmatrix} \tag{7.8.5}$$

EXAMPLE 7.2 *(continued)* In example 7.2, data from a clinical trial of the effectiveness of two methods for placing catheters in kidney dialysis patients was used to illustrate various two-sample tests over the entire range of time. The estimated survival functions for the two groups are given in Figure 7.1. Suppose an investigator is interested in comparing the survival functions at 3 months (short duration of time to infection). Thus, using the Kaplan–Meier estimates of the survival functions from (4.2.1) and the estimated variances of these survival functions from (4.2.2) for the jth goup, we obtain the Z test statistic as the square root of the chi-squared quadratic form with one degree of freedom from (7.8.5) to be

$$Z = \frac{\hat{S}_1(t_0) - \hat{S}_2(t_0)}{\sqrt{\hat{V}[\hat{S}_1(t_0)] + \hat{V}[\hat{S}_2(t_0)]}} \tag{7.8.6}$$

which, when H_0 is true, has a standard normal ditribution for large samples. Using this statistic, an α level test of the alternative hypothesis $H_A : S_1(t_0) > S_2(t_0)$ is rejected when $Z \geq Z_\alpha$, the αth upper percentage point of a standard normal distribution. The test of $H_A : S_1(t_0) \neq S_2(t_0)$ rejects when $|Z| > Z_{\alpha/2}$.

The estimates of the survival functions are

$$\hat{S}_1(3) = 0.9767 \quad \text{and} \quad \hat{S}_2(3) = 0.882,$$

and estimates of the variances are

$$\hat{V}[\hat{S}_1(3)] = 0.00053 \quad \text{and} \quad \hat{V}[\hat{S}_2(3)] = 0.00141.$$

This leads to a test statistic of

$$Z = \frac{0.9767 - 0.8882}{\sqrt{0.00053 + 0.00141}} = 2.01,$$

which leads to a two-sided p-value of 0.044. This difference is due to the first group's (surgical placement of catheters) having a smaller probability of infection at three months than the second group (percutaneous placement of catheters).

It should be noted that another investigator comparing the survival function of the different placement of the catheters at a large time period would get the opposite conclusion. This again emphasizes the need to preselect t_0 before examining the data.

This example is also an illustration of what can occur when the hazards are not proportional (an assumption formally tested in Chapter 9, Example 9.2).

EXAMPLE 7.4 *(continued)* In the example of section 4.7 the relapse cumulative incidence curve for 39 ALL patients was calculated as shown in Table 4.8. At one year the estimated cumulative incidence was 0.2380 (variance = 0.0048). In this data set there are two additional groups, AML low-risk and AML high-risk, whose relapse cumulative incidences are 0.0556 (variance = 0.0010) and 0.3556 (variance = 0.0054), respectively. A test of the hypothesis of no difference in three-year cumulative incidence between the three disease groups at one year has a $\chi^2 = 17.32$, which has a p-value of 0.0002 when compared to a chi-square distribution with 2 degrees of freedom.

If one is interested in comparing K groups in a pairwise simultaneous manner then an adjustment for multiple tests must be made. One such method that can be used is the Bonferroni method of multiple comparisons. Here if $K(K-1)/2$ pairwise comparisons are made and one still wants to maintain an overall α-level test, then each individual test is carried out at an $\alpha^* = \alpha/K(K-1)/2$ (or $\alpha/2K(K-1)/2 = \alpha/K(K-1)$ for two-sided tests) level of significance and if all null hypotheses are actually true, then the probability is at least $1-\alpha$ that none of the null hypotheses will be wrongly rejected. This method is somewhat conservative and becomes more conserative as the number of comparisons increases.

EXAMPLE 7.4 *(continued)* For our previous example of the three groups (ALL, AML low-risk, AML high-risk) when the Bonferroni method of multiple com-

parisons (in our case, $K = 3$) is used to make pairwise comparisons of the cumulative incidence curves, each test needs to be carried out at the $0.05/3 = 0.017$ level of significance. The contrasts $(1, -1, 0)$, $(1, 0, -1)$, and $(0, 1, -1)$ may be used to test each of the individual pairwise comparisons. Using the appropriate variances in (7.8.5), we get

$$\text{for} \quad H_0 : CI_1(t_0) = CI_2(t_0) \text{ at } t_0 = 1$$

we have

$$Z = 2.41, \ p\text{-value} = 0.016,$$

$$\text{for} \quad H_0 : CI_1(t_0) = CI_3(t_0) \text{ at } t_0 = 1$$

we have

$$Z = -1.17, \ p\text{-value} = 0.242,$$

and

$$\text{for} \quad H_0 : CI_2(t_0) = CI_3(t_0) \text{ at } t_0 = 1$$

we have

$$Z = -3.76, \ p\text{-value} = 0.0002.$$

Thus we conclude that the AML high-risk group is statistically different from the other two groups and that the ALL and AML low-risk groups are not statistically different from each other.

Practical Notes

1. One may test a hypothesis for any linear combination of several groups. For example, if one wants to test whether the cumulative incidence curves for the ALL patients are different than those for the AML (both high-risk and low-risk) patients, then one may select the linear contrast $(2, -1, -1)$ and use the quadratic form (7.8.5).

7.9 Exercises

7.1 In a study of the effectiveness of a combined drug regimen for the treatment of rheumatoid arthritis, 40 white patients were followed for a period ranging from 1 to 18 years. During the course of the study, 9 patients died. The ages at entry into the study and at death for these 9 patients were as follows:

Female deaths: (66, 74), (60, 76), (70, 77), (71, 81)
Male deaths: (50, 59), (60, 66), (51, 69), (69, 71), (58, 71)

For the 31 patients still alive at the end of the study their ages at entry and last follow-up were as follows:

Female Survivors: (50, 68), (55, 72), (56, 60), (45, 55), (48, 51), (44, 55), (33, 51), (44, 50), (60, 70), (55, 60), (60, 72), (77, 80), (70, 75), (66, 70), (59, 63), (62, 63)
Male Survivors: (53, 68), (55, 62), (56, 63), (45, 51), (48, 61), (49, 55), (43, 51), (44, 54), (61, 70), (45, 60), (63, 72), (74, 80), (70, 76), (66, 72), (54, 70)

Using the all-cause U.S. mortality table for 1989 (Table 2.1) test the hypothesis that the death rate of these rheumatoid arthritis patients is not different from that in the general population using the log-rank test.

7.2 In Exercise 5 of Chapter 6, the survival experience of patients given an autologous transplant was compared to a postulated exponential survival rate with a hazard rate of 0.045. Using the data in Table 1.4 of Chapter 1, test the hypothesis that the hazard rate of these auto transplant patients is equal to 0.045 against the alternative that it is larger than 0.045 using the one-sample, log-rank test. Repeat this test using a weight function which gives heavier weight to departures early in time from this hazard rate.

7.3 Consider the data reported in section 1.6 on the times until staphylococcus infection of burn patients (see our web page).

(a) Using the log-rank test, test the hypothesis of no difference in the rate of staphylococcus infection between patients whose burns were cared for with a routine bathing care method versus those whose body cleansing was initially performed using 4% chlorhexidine gluconate. Use a two-sided test and a 0.05 significance level.

(b) Repeat the test using Gehan's test.

(c) Repeat the test using the Tarone and Ware weights.

7.4 In section 1.11, data from a study of the effect of ploidy on survival for patients with tumors of the tongue was reported.

(a) Test the hypothesis that the survival rates of patients with cancer of the tongue are the same for patients with aneuploid and diploid tumors using the log-rank test.

(b) If primary interest is in detecting differences in survival rates between the two types of cancers which occur soon after the diagnosis of the cancer, repeat part a using a more appropriate test statistic.

7.5 Using the data on laryngeal cancers in Example 7.6, test, by the log-rank statistic, the null hypothesis of no difference in death rates among the four stages of cancer against the global alternative that at least one of the death rates differs from the others. Compare your results to those found in Example 7.6.

7.6 One of the goals of recent research is to explore the efficacy of triple-drug combinations of antiretroviral therapy for treatment of HIV-infected patients. Because of limitations on potency and the continuing emergence of drug resistance seen with the use of currently available antiretroviral agents in monotherapy and two-drug regimens, triple-combination regimens should represent a more promising approach to maximize antiviral activity, maintain long-term efficacy, and reduce the incidence of drug resistance. Towards this end, investigators performed a randomized study comparing AZT + zalcitabine (ddC) versus AZT + zalcitabine (ddC) + saquinavir. The data, time from administration of treatment (in days) until the CD4 count reached a prespecified level, is given below for the two groups.

AZT + zalcitabine (ddC): 85, 32, 38+, 45, 4+, 84, 49, 180+, 87, 75, 102, 39, 12, 11, 80, 35, 6
AZT + zalcitabine (ddC) + saquinavir: 22, 2, 48, 85, 160, 238, 56+, 94+, 51+, 12, 171, 80, 180, 4, 90, 180+, 3

Use the log rank statistic to test if there is a difference in the distribution of the times at which patient's CD4 reaches the prespecified level for the two treatments.

7.7 A study was performed to determine the efficacy of boron neutron capture therapy (BNCT) in treating the therapeutically refractory F98 glioma, using boronophenylalanine (BPA) as the capture agent. F98 glioma cells were implanted into the brains of rats. Three groups of rats were studied. One group went untreated, another was treated only with radiation, and the third group received radiation plus an appropriate concentration of BPA. The data for the three groups lists the death times (in days) and is given below:

Untreated	*Radiated*	Radiated + BPA
20	26	31
21	28	32
23	29	34
24	29	35
24	30	36
26	30	38
26	31	38
27	31	39
28	32	42^+
30	35^+	42^+

$^+$Censored observation

(a) Compare the survival curves for the three groups.

(b) Perform pairwise tests to determine if there is any difference in survival between pairs of groups.

(c) There is a priori evidence that, if there is a difference in survival, there should be a natural ordering, namely, untreated animals will have the worst survival, radiated rats will have slightly improved survival, and the radiated rats + BPA should have the best survival. Perform the test for trend which would test this ordered hypothesis.

7.8 In Example 7.4, we compared the disease-free survival rates of ALL patients with those of high-risk and low risk AML patients. Because acute graft-versus-host (aGVHD) disease is considered to have an antileukemic effect, one would expect lower relapse rates for patients who have developed aGVHD than for those that do not develop aGVHD. Using the data on out web page, examine the validity of this finding by

(a) testing if the hazard rate for the occurrence of aGVHD is the same for the three groups,

(b) testing if the hazard rate for relapse is the same in all three groups, and

(c) testing if the hazard rate for relapse in the three disease groups is the same for patients who have developed aGVHD. (Hint: For this test, the data is left-truncated at the time of aGVHD).

7.9 On our web page, data is reported on the death times of 863 kidney transplant patients (see section 1.7). Here, patients can be classified by race and sex into one of four groups.

(a) Test the hypothesis that there is no difference in survival between the four groups.

(b) Provide individual tests, for each sex, of the hypothesis of no racial differences in survival rates. Also, adjusting by stratification for the sex of the patient, test the hypothesis that blacks have a higher mortality rate than whites.

7.10 In Example 7.6 we found that the four populations of cancer patients had ordered hazard rates. Of interest is knowing which pairs of the hazard rates are different. Using the log-rank test, perform the three pairwise tests of the hypothesis $H_{0j} : h_j(t) = h_{j+1}(t)$ versus $H_{Aj} : h_j(t) < h_{j+1}(t)$, for $j = 1, 2, 3$. For each test, use only those individuals with stage j or $j+1$ of the disease. Make an adjustment to your critical value for multiple testing to give an approximate 0.05 level test.

One method to making the pairwise comparisons is to base the pairwise tests on the full $\mathbf{Z}(\tau)$ vector. To perform this test, recall that this vector has an asymptotic K variate normal distribution with mean 0 and covariance matrix $\hat{\boldsymbol{\Sigma}}$ under the null hypothesis. Thus, the statistic $Z_j(\tau) - Z_{j+1}(\tau)$ has a normal distribution with mean 0 and variance $\hat{\sigma}_{jj} + \hat{\sigma}_{j+1j+1} - 2\hat{\sigma}_{jj+1}$ when the null hypothesis is true. Large negative values of this test statistic will suggest that the hazard rate in

sample j is smaller than in sample $j+1$, so the hypothesis $H_{0j}: h_j(t) = h_{j+1}(t)$ is rejected in favor of $H_{Aj}: h_j(t) < h_{j+1}(t)$ when $[Z_j(\tau) - Z_{j+1}(\tau)]/[\hat{\sigma}_{jj} + \hat{\sigma}_{j+1j+1} - 2\hat{\sigma}_{jj+1}]^{1/2}$ is smaller than the αth lower percentile of a standard normal. Use the information in Example 7.6 and this statistic to make the multiple comparisons.

7.11 The data on laryngeal cancer patients was collected over the period 1970–1978. It is possible that the therapy used to treat laryngeal cancer may have changed over this nine year period. To adjust for this possible confounding fact, test the hypothesis of no difference in survival between patients with different stages of disease against a global alternative using a test which stratifies on the cancer being diagnosed prior to 1975 or not. Also perform a separate test of the hypothesis of interest in each stratum.

7.12 (a) Repeat Exercise 3 using the log-rank version of the Renyi statistic.

(b) Repeat Exercise 4 using the Gehan version of the Renyi statistic.

7.13 In Table 1.3 of section 1.5, the data on time to death for breast cancer-patients who where classed as lymph node negative by standard light microscopy (SLM) or by immunohistochemical (IH) examination of their lymph nodes is reported. Test the hypothesis that there is no difference in survival between theses two groups using

(a) the log-rank test,

(b) the Renyi statistic based on the log-rank test,

(c) the Cramer-von Mises statistic, and

(d) the weighted difference in the Kaplan–Meier statistic W_{KM}.

7.14 Repeat Exercise 7 using

(a) the Renyi statistic based on the log-rank test,

(b) the Cramer-von Mises statistic, and

(c) the weighted difference in the Kaplan–Meier statistic W_{KM}.

7.15 Using the data of section 1.3,

(a) compare the three survival functions for ALL, AML low-risk, and AML high-risk at one year;

(b) perform pairwise multiple comparisons for the three groups employing the Bonferroni correction for multiple tests.

8 Semiparametric Proportional Hazards Regression with Fixed Covariates

8.1 Introduction

Often one is interested in comparing two or more groups of times-to-event. If the groups are similar, except for the treatment under study, then, the nonparametric methods of Chapter 7 may be used directly. More often than not, the subjects in the groups have some additional characteristics that may affect their outcome. For example, subjects may have demographic variables recorded, such as age, gender, socioeconomic status, or education; behavioral variables, such as dietary habits, smoking history, physical activity level, or alcohol consumption; or physiological variables, such as blood pressure, blood glucose levels, hemoglobin levels, or heart rate. Such variables may be used as covariates (explanatory variables, confounders, risk factors, independent variables) in explaining the response (dependent) variable. After