CHAPTER 15

Rates and Proportions

15.1 INTRODUCTION

In this chapter and the next we want to study in more detail some of the topics dealing with counting data introduced in Chapter 6. In this chapter we want to take an epidemiological approach, studying populations by means of describing incidence and prevalence of disease. In a sense this is where statistics began: with a numerical description of the characteristics of a state, frequently involving mortality, fecundity, and morbidity. We call the occurrence of one of those outcomes an *event*. In the next chapter we deal with more recent developments, which have focused on a more detailed modeling of survival (hence also death, morbidity, and fecundity) and dealt with such data obtained in experiments rather than observational studies. An implication of the latter point is that sample sizes have been much smaller than used traditionally in the epidemiological context. For example, the evaluation of the success of heart transplants has, by necessity, been based on a relatively small set of data.

We begin the chapter with definitions of incidence and prevalence rates and discuss some problems with these "crude" rates. Two methods of standardization, direct and indirect, are then discussed and compared. In Section 15.4, a third standardization procedure is presented to adjust for varying exposure times among individuals. In Section 15.5, a brief tie-in is made to the multiple logistic procedures of Chapter 13. We close the chapter with notes, problems, and references.

15.2 RATES, INCIDENCE, AND PREVALENCE

The term *rate* refers to the amount of change occurring in a quantity with respect to time. In practice, *rate* refers to the amount of change in a variable over a specified time interval divided by the length of the time interval.

The data used in this chapter to illustrate the concepts come from the Third National Cancer Survey [National Cancer Institute, 1975]. For this reason we discuss the concepts in terms of incidence rates. The *incidence* of a disease in a fixed time interval is the number of new cases diagnosed during the time interval. The *prevalence* of a disease is the number of people with the disease at a fixed time point. For a chronic disease, incidence and prevalence may present markedly different ideas of the importance of a disease.

Consider the Third National Cancer Survey [National Cancer Institute, 1975]. This survey examined the incidence of cancer (by site) in nine areas during the time period 1969–1971.

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

The areas were the Detroit SMSA (Standard Metropolitan Statistical Area); Pittsburgh SMSA, Atlanta SMSA, Birmingham SMSA, Dallas–Fort Worth SMSA, state of Iowa, Minneapolis–St. Paul SMSA, state of Colorado, and the San Francisco–Oakland SMSA. The information used in this chapter refers to the combined data from the Atlanta SMSA and San Francisco–Oakland SMSA. The data are abstracted from tables in the survey. Suppose that we wanted the rate for all sites (of cancer) combined. The rate per year in the 1969–1971 time interval would be simply the number of cases divided by 3, as the data were collected over a three-year interval. The rates are as follows:

$$\text{Combined area :} \quad \frac{181{,}027}{3} = 60{,}342.3$$

$$\text{Atlanta :} \quad \frac{9{,}341}{3} = 3{,}113.7$$

$$\text{San Francisco–Oakland :} \quad \frac{30{,}931}{3} = 10{,}310.3$$

Can we conclude that cancer incidence is worse in the San Francisco–Oakland area than in the Atlanta area? The answer is "yes and no." Yes, in that there are more cases to take care of in the San Francisco–Oakland area. If we are concerned about the chance of a person getting cancer, the numbers would not be meaningful. As the San Francisco–Oakland area may have a larger population, the number of cases per number of the population might be less. To make comparisons taking the population size into account, we use

$$\text{incidence per time interval} = \frac{\text{number of new cases}}{\text{total population} \times \text{time interval}} \tag{1}$$

The result of equation (1) would be quite small, so that the number of cases per 100,000 population is used to give a more convenient number. The rate per 100,000 population per year is then

$$\text{incidence per 100,000 per time interval} = \frac{\text{number of new cases}}{\text{total population} \times \text{time interval}} \times 100{,}000$$

For these data sets, the values are:

$$\text{Combined area :} \quad \frac{181{,}027 \times 100{,}000}{21{,}003{,}451 \times 3} = 287.3 \text{ new cases per } 100{,}000 \text{ per year}$$

$$\text{Atlanta :} \quad \frac{9{,}341 \times 100{,}000}{1{,}390{,}164 \times 3} = 224.0 \text{ new cases per } 100{,}000 \text{ per year}$$

$$\text{San Francisco-Oakland :} \quad \frac{30{,}931 \times 100{,}000}{3{,}109{,}519 \times 3} = 331.6 \text{ new cases per } 100{,}000 \text{ per year}$$

Even after adjusting for population size, the San Francisco–Oakland area has a higher overall rate.

Note several facts about the estimated rates. The estimates are binomial proportions times a constant (here 100,000/3). Thus, the rate has a standard error easily estimated. Let N be the total population and n the number of new cases; the rate is $n/N \times C$ ($C = 100{,}000/3$ in this example) and the standard error is estimated by

$$\sqrt{C^2 \frac{1}{N} \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

or

$$\text{standard error of rate per time interval} = C\sqrt{\frac{1}{N} \frac{n}{N} \left(1 - \frac{n}{N}\right)}$$

For example, the combined area estimate has a standard error of

$$\frac{100{,}000}{3}\sqrt{\frac{1}{21{,}003{,}451}\frac{181{,}027}{21{,}003{,}451}\left(1-\frac{181{,}027}{21{,}003{,}451}\right)}=0.67$$

As the rates are assumed to be binomial proportions, the methods of Chapter 6 may be used to get adjusted estimates or standardized estimates of proportions.

Rates computed by the foregoing methods,

$$\frac{\text{number of new cases in the interval}}{\text{population size} \times \text{time interval}}$$

are called *crude* or *total rates*. This term is used in distinction to *standardized* or *adjusted rates*, as discussed below.

Similarly, a *prevalence rate* can be defined as

$$\text{prevalence} = \frac{\text{number of cases at a point in time}}{\text{population size}}$$

Sometimes a distinction is made between *point prevalence* and *prevalence* to facilitate discussion of chronic disease such as epilepsy and a disease of shorter duration, for example, a common cold or even accidents. It is debatable whether the word *prevalence* should be used for accidents or illnesses of short duration.

15.3 DIRECT AND INDIRECT STANDARDIZATION

15.3.1 Problems with the Use of Crude Rates

Crude rates are useful for certain purposes. For example, the crude rates indicate the load of new cases per capita in a given area of the country. Suppose that we wished to use the cancer rates as epidemiologic indicators. The inference would be that it was likely that environmental or genetic differences were responsible for a difference, if any. There may be simpler explanations, however. Breast cancer rates would probably differ in areas that had differing gender proportions. A retirement community with an older population will tend to have a higher rate. To make fair comparisons, we often want to adjust for the differences between populations in one or more factors (covariates). One approach is to find an index that is adjusted in some fashion. We discuss two methods of adjustment in the next two sections.

15.3.2 Direct Standardization

In direct standardization we are interested in adjusting by one or more variables that are divided (or naturally fall) into discrete categories. For example, in Table 15.1 we adjust for gender and for age divided into a total of 18 categories. The idea is to find an answer to the following question: Suppose that the distribution with regard to the adjusting factors was not as observed, but rather, had been the same as this other (reference) population; what would the rate have been? In other words, we apply the risks observed in our study population to a reference population.

In symbols, the adjusting variable is broken down into I cells. In each cell we know the number of events (the numerator) n_i and the total number of individuals (the denominator) N_i:

Level of adjusting factor, i:	1	2	$\cdots$	i	$\cdots$	I
Proportion observed in study population:	$\frac{n_1}{N_1}$	$\frac{n_2}{N_2}$	$\cdots$	$\frac{n_i}{N_i}$	$\cdots$	$\frac{n_1}{N_1}$

Table 15.1 Rate for Cancer of All Sites for Blacks in the San Francisco–Oakland SMSA and Reference Population

	Study Population n_i/N_i		Reference Population M_i	
Age	Females	Males	Females	Males
<5	8/16,046	6/16,493	872,451	908,739
5–9	6/18,852	7/19,265	1,012,554	1,053,350
10–14	6/19,034	3/19,070	1,061,579	1,098,507
15–19	7/16,507	6/16,506	971,894	964,845
20–24	16/15,885	9/14,015	919,434	796,774
25–29	27/12,886	19/12,091	755,140	731,598
30–34	28/10,705	18/10,445	620,499	603,548
35–39	46/9,580	25/8,764	595,108	570,117
40–44	83/9,862	47/8,858	650,232	618,891
45–49	109/10,341	108/9,297	661,500	623,879
50–54	125/8,691	131/8,052	595,876	558,124
55–59	120/6,850	189/6,428	520,069	481,137
60–64	102/5,017	158/4,690	442,191	391,746
65–69	119/3,806	159/3,345	367,046	292,621
70–74	75/2,264	154/1,847	300,747	216,929
75–79	44/1,403	72/931	224,513	149,867
80–84	28/765	51/471	139,552	84,360
>85	25/629	26/416	96,419	51,615
Subtotal	974/169,123	1,188/160,984	10,806,804	10,196,647
Total	2,162/330,107		21,003,451	

Source: National Cancer Institute [1975].

Both numerator and denominator are presented in the table. The crude rate is estimated by

$$C\frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i}$$

Consider now a *standard or reference population*, which instead of having N_i persons in the ith cell has M_i.

	Reference Population					
Level of adjusting factor	1	2	⋯	i	⋯	I
Number in reference population	M_1	M_2	⋯	M_i	⋯	M_I

The question now is: If the study population has M_i instead of N_i persons in the ith cell, what would the crude rate have been? We cannot determine what the crude rate was, but we can estimate what it might have been. In the ith cell the proportion of observed deaths was n_i/N_i. If the same proportion of deaths occurred with M_i persons, we would expect

$$n_i^* = \frac{n_i}{N_i} M_i \text{ deaths}$$

Thus, if the adjusting variables had been distributed with M_i persons in the ith cell, we estimate that the data would have been:

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Expected proportion of cases:	$\frac{n_1 M_1/N_1}{M_1}$	$\frac{n_2 M_2/N_2}{M_2}$	$\cdots$	$\frac{n_i^*}{M_i}$	$\cdots$	$\frac{n_I M_I/N_I}{M_I}$

The *adjusted rate*, r, is the crude rate for this estimated standard population:

$$r = \frac{C \sum_{i=1}^{I} n_i M_i/N_i}{\sum_{i=1}^{I} M_i} = \frac{C \sum_{i=1}^{I} n_i^*}{\sum_{i=1}^{I} M_i}$$

As an example, consider the rate for cancer for all sites for blacks in the San Francisco–Oakland SMSA, adjusted for gender and age to the total combined sample of the Third Cancer Survey, as given by the 1970 census. There are two gender categories and 18 age categories, for a total of 36 cells. The cells are laid out in two columns rather than in one row of 36 cells. The data are given in Table 15.1.

The crude rate for the San Francisco–Oakland black population is

$$\frac{100{,}000}{3} \frac{974 + 1188}{169{,}123 + 160{,}984} = 218.3$$

Table 15.2 gives the values of $n_i M_i/N_i$.

The gender- and age-adjusted rate is thus

$$\frac{100{,}000}{3} \frac{193{,}499.42}{21{,}003{,}451} = 307.09$$

Note the dramatic change in the estimated rate. This occurs because the San Francisco–Oakland SMSA black population differs in its age distribution from the overall sample.

The variance is estimated by considering the denominators in the cell as fixed and using the binomial variance of the n_i's. Since the cells constitute independent samples,

$$\begin{aligned} \operatorname{var}(r) &= \operatorname{var}\left(C \sum_{i=1}^{I} \frac{n_i M_i}{N_i} \Big/ \sum_{i=1}^{I} M_i \right) \\ &= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \left(\frac{M_i}{N_i} \right)^2 \operatorname{var}(n_i) \end{aligned}$$

Table 15.2 Estimated Number of Cases per Cell ($n_i M_i/N_i$) if the San Francisco–Oakland Area Had the Reference Population Age and Gender Distribution

Age	Females	Males	Age	Females	Males
<5	434.97	330.59	55–59	9,110.70	14,146.69
5–9	322.26	382.74	60–64	8,990.13	13,197.41
10–14	334.64	172.81	65–69	11,476.21	13,909.34
15–19	412.14	350.73	70–74	9,962.91	18,087.20
20–24	926.09	511.66	75–79	7,041.03	11,590.14
25–29	1,582.24	1,149.65	80–84	5,107.79	9,134.52
30–34	1,622.98	1,040.10	>85	3,832.23	3,225.94
35–39	2,857.51	1,629.30			
40–44	5,472.45	3,283.80	Subtotal	85,029.16	108,470.26
45–49	6,972.58	7,247.38	Total	193,499.42	
50–54	8,570.30	9,080.26			

$$= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \left(\frac{M_i}{N_i}\right)^2 N_i \frac{n_i}{N_i} \left(1 - \frac{n_i}{N_i}\right)$$

$$= \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \frac{M_i}{N_i} \frac{n_i M_i}{N_i} \left(1 - \frac{n_i}{N_i}\right)$$

where $M_\cdot = \sum_{i=1}^{I} M_i$.

If n_i/N_i is small, then $1 - n_i/N_i \doteq 1$ and

$$\text{var}(r) \doteq \frac{C^2}{M_\cdot^2} \sum_{i=1}^{I} \frac{M_i}{N_i} \left(\frac{n_i M_i}{N_i}\right) \tag{2}$$

We use this to compute a 95% confidence interval for the adjusted rate computed above. Using equation (2), the standard error is

$$\text{SE}(r) = \frac{C}{M_\cdot} \sqrt{\sum_{i=1}^{I} \frac{M_i}{N_i} \left(\frac{n_i M_i}{N_i}\right)}$$

$$= \frac{100{,}000}{3} \frac{1}{21{,}003{,}451} \left(\frac{872{,}451}{16{,}046} 434.97 + \cdots\right)^{1/2}$$

$$= 7.02$$

The quantity r is approximately normally distributed, so that the interval is

$$307.09 \pm 1.96 \times 7.02 \quad \text{or} \quad (293.3, 320.8)$$

If adjusted rates are estimated for two different populations, say r_1 and r_2, with standard errors $\text{SE}(r_1)$ and $\text{SE}(r_2)$, respectively, equality of the adjusted rates may be tested by using

$$z = \frac{r_1 - r_2}{\sqrt{\text{SE}(r_1)^2 + \text{SE}(r_2)^2}}$$

The $N(0,1)$ critical values are used, as z is approximately $N(0,1)$ under the null hypothesis of equal rates.

15.3.3 Indirect Standardization

In indirect standardization, the procedure of direct standardization is used in the opposite direction. That is, we ask the question: What would the mortality rate have been for the study population if it had the same rates as the population reference? That is, we apply the observed risks in the reference population to the study population.

Let m_i be the number of deaths in the reference population in the ith cell. The data are:

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Observed proportion in reference population:	$\frac{m_1}{M_1}$	$\frac{m_2}{M_2}$	$\cdots$	$\frac{m_i}{M_i}$	$\cdots$	$\frac{m_I}{M_I}$

where both numerator and denominators are presented in the table. Also,

Level of adjusting factor:	1	2	$\cdots$	i	$\cdots$	I
Denominators in study population:	N_1	N_2	$\cdots$	N_i	$\cdots$	N_I

The estimate of the rate the study population would have experienced is (analogous to the argument in Section 15.3.2)

$$r_{\text{REF}} = \frac{C \sum_{i=1}^{I} N_i (m_i / M_i)}{\sum_{i=1}^{I} N_i}$$

The crude rate for the study population is

$$r_{\text{STUDY}} = \frac{C \sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i}$$

where n_i is the observed number of cases in the study population at level i. Usually, there is not much interest in comparing the values r_{REF} and r_{STUDY} as such, because the distribution of the study population with regard to the adjusting factors is not a distribution of much interest. For this reason, attention is usually focused on the *standardized mortality ratio* (SMR), when death rates are considered, or the *standardized incidence ratio* (SIR), defined to be

$$\text{standardized ratio} = s = \frac{r_{\text{STUDY}}}{r_{\text{REF}}} = \frac{\sum_{i=1}^{I} n_i}{\sum_{i=1}^{I} N_i m_i / M_i} \tag{3}$$

The main advantage of the indirect standardization is that the SMR involves only the total number of events, so you do not need to know in which cells the deaths occur for the study population. An alternative way of thinking of the SMR is that it is the observed number of deaths in the study population divided by the expected number if the cell-specific rates of the reference population held.

As an example, let us compute the SIR of cancer in black males in the Third Cancer Survey, using white males of the same study as the reference population and adjusting for age. The data are presented in Table 15.3. The standardized incidence ratio is

$$s = \frac{8793}{7474.16} = 1.17645 = 1.18$$

One reasonable question to ask is whether this ratio is significantly different from 1. An approximate variance can be derived as follows:

$$s = \frac{O}{E} \qquad \text{where} \quad O = \sum_{i=1}^{I} n_i = n. \quad \text{and} \quad E = \sum_{i=1}^{I} N_i \left(\frac{m_i}{M_i} \right)$$

The variance of s is estimated by

$$\text{var}(s) = \frac{\text{var}(O) + s^2 \text{ var}(E)}{E^2} \tag{4}$$

The basic "trick" is to (1) assume that the number of cases in a particular cell follows a Poisson distribution and (2) to note that the sum of independent Poisson random variables is Poisson. Using these two facts yields

$$\text{var}(O) \doteq \sum_{i=1}^{I} n_i = n \tag{5}$$

Table 15.3 Cancer of All Areas Combined, Number of Cases, Black and White Males by Age and Number Eligible by Age

Age	Black Males		White Males			
	n_1	N_1	m_1	M_1	$\frac{N_i m_i}{M_i}$	$\left(\frac{N_i}{M_i}\right)^2 m_i$
<5	45	120,122	450	773,459	69.89	10.85
5–9	34	130,379	329	907,543	47.26	6.79
10–14	39	134,313	300	949,669	42.43	6.00
15–19	45	112,969	434	837,614	58.53	7.89
20–24	49	86,689	657	694,670	81.99	10.23
25–29	63	71,348	688	647,304	75.83	8.36
30–34	84	57,844	724	533,856	78.45	8.50
35–39	129	54,752	1,097	505,434	118.83	12.87
40–44	318	57,070	2,027	552,780	209.27	21.61
45–49	582	56,153	3,947	559,241	396.31	39.79
50–54	818	48,753	6,040	503,163	585.23	56.71
55–59	1,170	42,580	8,711	432,982	856.65	84.24
60–64	1,291	33,892	10,966	352,315	1,054.91	101.48
65–69	1,367	27,239	11,913	261,067	1,242.97	129.69
70–74	1,266	17,891	11,735	196,291	1,069.59	97.49
75–79	788	9,827	10,546	138,532	748.10	53.07
80–84	461	4,995	6,643	78,044	425.17	27.21
>85	244	3,850	3,799	46,766	312.75	25.75
Total	8,793	1,070,700	81,006	8,970,730	7,474.16	708.53

and

$$\operatorname{var}(E) \doteq \operatorname{var}\left(\sum_{i=1}^{I} \frac{N_i}{M_i} m_i\right)$$

$$= \sum_{i=1}^{I} \left(\frac{N_i}{M_i}\right)^2 m_i \tag{6}$$

The variance of s is estimated by using equations (4), (5), and (6):

$$\operatorname{var}(s) = \frac{n_\cdot + s^2 \sum (N_i/M_i)^2 m_i}{E^2}$$

A test of the hypothesis that the population value of s is 1 is obtained from

$$z = \frac{s - 1}{\sqrt{\operatorname{var}(s)}}$$

and $N(0, 1)$ critical values.

For the example,

$$\sum_{i=1}^{I} n_i = n_\cdot = 8793$$

$$E = \sum_{i=1}^{I} \frac{N_i}{M_i} m_i = 7474.16$$

$$\text{var}(E) \doteq \sum_{i=1}^{I} \left(\frac{N_i}{M_i}\right)^2 m_i = 708.53$$

$$\text{var}(s) \doteq \frac{8793 + (1.17645)^2 \times 708.53}{(7474.16)^2} = 0.000174957$$

From this and a standard error of $s \doteq 0.013$, the ratio is significantly different from one using

$$z = \frac{s - 1}{\text{SE}(s)} = \frac{0.17645}{0.013227} = 13.2$$

and $N(0, 1)$ critical values.

If the reference population is much larger than the study population, $\text{var}(E)$ will be much less than $\text{var}(O)$ and you may approximate $\text{var}(s)$ by $\text{var}(O)/E^2$.

15.3.4 Drawbacks to Using Standardized Rates

Any time a complex situation is summarized in one or a few numbers, considerable information is lost. There is always a danger that the lost information is crucial for understanding the situation under study. For example, two populations may have almost the same standardized rates but may differ greatly within the different cells; one population has much larger values in one subset of the cells and the reverse situation in another subset of cells. Even when the standardized rates differ, it is not clear if the difference is somewhat uniform across cells or results mostly from one or a few cells with much larger differences.

The moral of the story is that whenever possible, the rates in the cells used in standardization should be examined individually in addition to working with the standardized rates.

15.4 HAZARD RATES: WHEN SUBJECTS DIFFER IN EXPOSURE TIME

In the rates computed above, each person was exposed (eligible for cancer incidence) over the same length of time (three years, 1969–1971). (This is not quite true, as there is some population mobility, births, and deaths. The assumption that each person was exposed for three years is valid to a high degree of approximation.) There are other circumstances where people are observed for varying lengths of time. This happens, for example, when patients are recruited sequentially as they appear at a medical care facility. One approach would be to restrict the analysis to those who had been observed for at least some fixed amount of time (e.g., for one year). If large numbers of persons are not observed, this approach is wasteful by throwing away valuable and needed information. This section presents an approach that allows the rates to use all the available information if certain assumptions are satisfied.

Suppose that we observe subjects over time and look for an event that occurs only once. For definiteness, we speak about observing people where the event is death. Assume that over the time interval observed, if a subject has survived to some time t_0, the probability of death in a short interval from t_0 to t_1 is almost $\lambda(t_1 - t_0)$. The quantity λ is called the *hazard rate, force of mortality*, or *instantaneous death rate*. The units of λ are deaths per time unit.

How would we estimate λ from data in a real-life situation? Suppose that we have n individuals and begin observing the ith person at time B_i. If the person dies, let the time of death be D_i. Let the time of last contact be C_i for those people who are still alive. Thus, the time we are observing each person at risk of death is

$$O_i = \begin{cases} C_i - B_i & \text{if the subject is alive} \\ D_i - B_i & \text{if the subject is dead} \end{cases}$$

An unbiased estimate of λ is

$$\text{estimated hazard rate} = \widehat{\lambda} = \frac{\text{number of observed deaths}}{\sum_{i=1}^{n} O_i} = \frac{L}{\sum_{i=1}^{n} O_i} \tag{7}$$

As in the earlier sections of this chapter, $\widehat{\lambda}$ is often normalized to have different units. For example, suppose that $\widehat{\lambda}$ is in deaths per day of observation. That is, suppose that O_i is measured in days. To convert to deaths per 100 observation years, we use

$$\widehat{\lambda} \times 365 \frac{\text{days}}{\text{year}} \times 100$$

As an example, consider the paper by Clark et al. [1971]. This paper discusses the prognosis of patients who have undergone cardiac (heart) transplantation. They present data on 20 transplanted patients. These data are presented in Table 15.4. To estimate the deaths per year of exposure, we have

$$\frac{12 \text{ deaths}}{3599 \text{ exposure days}} \frac{365 \text{ days}}{\text{year}} = 1.22 \frac{\text{deaths}}{\text{exposure year}}$$

To compute the variance and standard error of the observed hazard rate, we again assume that L in equation (7) has a Poisson distribution. So conditional on the total observation period, the variability of the estimated hazard rate is proportional to the variance of L, which is estimated by L itself. Let

$$\widehat{\lambda} = \frac{CL}{\sum_{i=1}^{n} O_i}$$

where C is a constant that standardizes the hazard rate appropriately.

Table 15.4 Stanford Heart Transplant Data

i	Date of Transplantation	Date of Death	Time at Risk in Days (*if alive)[a]
1	1/6/68	1/21/68	15
2	5/2/68	5/5/68	3
3	8/22/68	10/7/68	46
4	8/31/68	—	608*
5	9/9/68	1/14/68	127
6	10/5/68	12/5/68	61
7	10/26/68	—	552*
8	11/20/68	12/14/68	24
9	11/22/68	8/30/69	281
10	2/8/69	—	447*
11	2/15/69	2/25/69	10
12	3/29/69	5/7/69	39
13	4/13/69	—	383*
14	5/22/69	—	344*
15	7/16/69	11/29/69	136
16	8/16/69	8/17/69	1
17	9/3/69	—	240*
18	9/14/69	11/13/69	60
19	1/3/70	—	118*
20	1/16/70	—	104*

[a] Total exposure days = 3599, $L = 12$.

Then the standard error of $\widehat{\lambda}$, $\mathrm{SE}(\widehat{\lambda})$, is approximately

$$\mathrm{SE}(\widehat{\lambda}) \doteq \frac{C}{\sum_{i=1}^{n} O_i}\sqrt{L}$$

A confidence interval for λ can be constructed by using confidence limits (L_1, L_2) for $E(L)$ as described in Note 6.8:

$$\text{confidence interval for } \lambda = \left(\frac{CL_1}{\sum_{i=1}^{n} O_i}, \frac{CL_2}{\sum_{i=1}^{n} O_i}\right)$$

For the example, a 95% confidence interval for the number of deaths is (6.2–21.0). A 95% confidence interval for the hazard rate is then

$$\left(\frac{6.2}{3599} \times 365, \frac{21.0}{3599} \times 365\right) = (0.63, 2.13)$$

Note that this assumes a constant hazard rate from day of transplant; this assumption is suspect. In Chapter 16 some other approaches to analyzing such data are given.

As a second more complicated illustration, consider the work of Bruce et al. [1976]. This study analyzed the experience of the Cardiopulmonary Research Institute (CAPRI) in Seattle, Washington. The program provided medically supervised exercise programs for diseased subjects. Over 50% of the participants dropped out of the program. As the subjects who continued participation and those who dropped out had similar characteristics, it was decided to compare the mortality rates for men to see if the training prevented mortality. It was recognized that subjects might drop out because of factors relating to disease, and the inference would be weak in the event of an observed difference.

The interest of this example is in the appropriate method of calculating the rates. All subjects, *including the dropouts*, enter into the computation of the mortality for active participants! The reason for this is that had they died during training, they would have been counted as active participant deaths. Thus, training must be credited with the exposure time or observed time when the dropouts were in training. For those who did not die and dropped out, the date of last contact *as an active participant* was the date at which the subjects left the training program. (Topics related to this are dealt with in Chapter 16).

In summary, to compute the mortality rates for active participants, all subjects have an observation time. The times are:

1. O_i = (time of death − time of enrollment) for those who died as active participants
2. O_i = (time of last contact − time of enrollment) for those in the program at last contact
3. O_i = (time of dropping the program−time of enrollment) for those who dropped whether or not a subsequent death was observed

The rate $\widehat{\lambda}_A$ for active participants is then computed as

$$\widehat{\lambda}_A = \frac{\text{number of deaths observed during training}}{\sum_{\text{all individuals}} O_i} = \frac{L_A}{\sum O_i}$$

To estimate the rate for dropouts, only those who drop out have time at risk of dying as a dropout. For those who have died, the time observed is

$$O_i' = (\text{time of death} - \text{time the subject dropped out})$$

For those alive at the last contact,

$$O_i' = (\text{time of last contact} - \text{time the subject dropped out})$$

The hazard rate for the dropouts, $\widehat{\lambda}_D$, is

$$\widehat{\lambda}_D = \frac{\text{number of deaths observed during dropout period}}{\sum_{\text{dropouts}} O_i'} = \frac{L_D}{\sum O_i'}$$

The paper reports rates of 2.7 deaths per 100 person-years for the active participants based on 16 deaths. The mortality rate for dropouts was 4.7 based on 34 deaths.

Are the rates statistically different at a 5% significance level? For a Poisson variable, L, the variance equals the expected number of observations and is thus estimated by the value of the variable itself. The rates $\widehat{\lambda}$ are of the form

$$\widehat{\lambda} = CL \qquad (L \text{ the number of events})$$

Thus, $\text{var}(\widehat{\lambda}) = C^2 \text{ var}(L) \doteq C^2 L = \widehat{\lambda}^2/L$.

To compare the two rates,

$$\text{var}(\widehat{\lambda}_A - \widehat{\lambda}_D) = \text{var}(\widehat{\lambda}_A) + \text{var}(\widehat{\lambda}_D) = \frac{\widehat{\lambda}_A^2}{L_A} + \frac{\widehat{\lambda}_D^2}{L_D}$$

The approximation is good for large L.

An approximate normal test for the equality of the rates is

$$z = \frac{\widehat{\lambda}_A - \widehat{\lambda}_D}{\sqrt{\widehat{\lambda}_A^2/L_A + \widehat{\lambda}_D^2/L_D}}$$

For the example, $L_A = 16, \widehat{\lambda}_A = 2.7$, and $L_D = 34, \widehat{\lambda}_D = 4.7$, so that

$$\begin{aligned} z &= \frac{2.7 - 4.7}{\sqrt{(2.7)^2/16 + (4.7)^2/34}} \\ &= -1.90 \end{aligned}$$

Thus, the difference between the two groups was not statistically significant at the 5% level.

15.5 MULTIPLE LOGISTIC MODEL FOR ESTIMATED RISK AND ADJUSTED RATES

In Chapter 13 the linear discriminant model or multiple logistic model was used to estimate the probability of an event as a function of covariates, $X_1, \ldots, X_n$. Suppose that we want a direct adjusted rate, where $X_1(i), \ldots, X_n(i)$ was the covariate value at the midpoints of the ith cell. For the study population, let p_i be the adjusted probability of an event at $X_1(i), \ldots, X_n(i)$. An adjusted estimate of the probability of an event is

$$\widehat{p} = \frac{\sum_{i=1}^{I} M_i p_i}{\sum_{i=1}^{I} M_i}$$

where M_i is the number of reference population subjects in the ith cell. This equation can be written as

$$\widehat{p} = \sum_{i=1}^{I} \left(\frac{M_i}{M.} p_i \right)$$

where $M. = \sum_{i=1}^{I} M_i$.

If the study population is small, it is better to estimate the p_i using the approach of Chapter 13 rather than the direct standardization approach of Section 15.3. This will usually be the case when there are several covariates with many possible values.

NOTES

15.1 More Than One Event per Subject

In some studies, each person may experience more than one event: for example, seizures in epileptic patients. In this case, each person could contribute more than once to the numerator in the calculation of a rate. In addition, exposure time or observed time would continue beyond an event, as the person is still at risk for another event. You need to check in this case that there are not people with "too many" events; that is, events "cluster" in a small subset of the population. A preliminary test for clustering may then be called for. This is a complicated topic. See Kalbfleisch and Prentice [2002] for references. One possible way of circumventing the problem is to record the time to the second or kth event. This builds a certain robustness into the data, but of course, makes it not possible to investigate the clustering, which may be of primary interest.

15.2 Standardization with Varying Observation Time

It is possible to compute standardized rates when the study population has the rate in each cell determined by the method of Section 15.4; that is, people are observed for varying lengths of time. In this note we discuss only the method for direct standardization.

Suppose that in each of the i cells, the rates in the study population is computed as CL_i/O_i, where C is a constant, L_i the number of events, and O_i the sum of the times observed for subjects in that cell. The adjusted rate is

$$\frac{\sum_{i=1}^{I} (M_i/L_i)\, O_i}{\sum_{i=1}^{I} M_i} = \frac{C \sum_{i=1}^{I} M_i \widehat{\lambda}_i}{M.} \qquad \text{where} \quad \widehat{\lambda}_i = \frac{L_i}{O_i}$$

The standard error is estimated to be

$$\frac{C}{M.} \sqrt{\sum_{i=1}^{I} \left(\frac{M_i}{O_i} \right) L_i}$$

15.3 Incidence, Prevalence, and Time

The *incidence* of a disease is the rate at which new cases appear; the *prevalence* is the proportion of the population that has the disease. When a disease is in a steady state, these are related via the average duration of disease:

$$\text{prevalence} = \text{incidence} \times \text{duration}$$

That is, if you catch a cold twice per year and each cold lasts a week, you will spend two weeks per year with a cold, so $2/52$ of the population should have a cold at any given time.

This equation breaks down if the disease lasts for all or most of your life and does not describe transient epidemics.

15.4 Sources of Demographic and Natural Data

There are many government sources of data in all of the Western countries. Governments of European countries, Canada, and the United States regularly publish vital statistics data as well as results of population surveys such as the Third National Cancer Survey [National Cancer Institute, 1975]. In the United States, the National Center for Health Statistics (*http://www.cdc.gov/nhcs*) publishes more than 20 series of monographs dealing with a variety of topics. For example, Series 20 provides natural data on mortality; Series 21, on natality, marriage, and divorce. These reports are obtainable from the U.S. government.

15.5 Binomial Assumptions

There is some question whether the binomial assumptions (see Chapter 6) always hold. There may be "extrabinomial" variation. In this case, standard errors will tend to be underestimated and sample size estimates will be too low, particularly in the case of dependent Bernoulli trials. Such data are not easy to analyze; sometimes a logarithmic transformation is used to stabilize the variance.

PROBLEMS

15.1 This problem will give practice by asking you to carry out analyses similar to the ones in each of the sections. The numbers from the National Cancer Institute [1975] for lung cancer cases for white males in the Pittsburgh and Detroit SMSAs are given in Table 15.5.

Table 15.5 Lung Cancer Cases by Age for White Males in the Detroit and Pittsburgh SMSAs

	Detroit		Pittsburgh	
Age	Cases	Population Size	Cases	Population Size
<5	0	149,814	0	82,242
5–9	0	175,924	0	99,975
10–14	2	189,589	1	113,146
15–19	0	156,910	0	100,139
20–24	5	113,003	0	68,062
25–29	1	113,919	0	61,254
30–34	10	92,212	7	53,289
35–39	24	90,395	21	55,604
40–44	101	108,709	56	70,832
45–49	198	110,436	148	74,781
50–54	343	98,756	249	72,247
55–59	461	82,758	368	64,114
60–64	532	63,642	470	50,592
65–69	572	47,713	414	36,087
70–74	473	35,248	330	26,840
75–79	365	25,094	259	19,492
80–84	133	12,577	105	10,987
>85	51	6,425	52	6,353
Total	3271	1,673,124	2480	1,066,036

(a) Carry out the analyses of Section 15.2 for these SMSAs.

(b) Calculate the direct and indirect standardized rates for lung cancer for white males adjusted for age. Let the Detroit SMSA be the study population and the Pittsburgh SMSA be the reference population.

(c) Compare the rates obtained in part (b) with those obtained in part (a).

15.2 **(a)** Calculate crude rates and standardized cancer rates for the white males of Table 15.5 using black males of Table 15.3 as the reference population.

(b) Calculate the standard error of the indirect standardized mortality rate and test whether it is different from 1.

(c) Compare the standardized mortality rates for blacks and whites.

15.3 The data in Table 15.6 represent the mortality experience for farmers in England and Wales 1949–1953 as compared with national mortality statistics.

Table 15.6 Mortality Experience Data for Problem 15.3

Age	National Mortality (1949–1953) Rate per 100,000/Year	Population of Farmers (1951 Census)	Deaths in 1949–1953
20–24	129.8	8,481	87
25–34	152.5	39,729	289
35–44	280.4	65,700	733
45–54	816.2	73,376	1,998
55–64	2,312.4	58,226	4,571

(a) Calculate the crude mortality rates.

(b) Calculate the standardized mortality rates.

(c) Test the significance of the standardized mortality rates.

(d) Construct a 95% confidence interval for the standardized mortality rates.

(e) What are the units for the ratios calculated in parts (a) and (b)?

15.4 Problems for discussion and thought:

(a) Direct and indirect standardization permit comparison of rates in two populations. Describe in what way this can also be accomplished by multiway contingency tables.

(b) For calculating standard errors of rates, we assumed that events were binomially (or Poisson) distributed. State the assumption of the binomial distribution in terms of, say, the event "death from cancer" for a specified population. Which of the assumptions is likely to be valid? Which is not likely to be invalid?

(c) Continuing from part (b), we calculate standard errors of rates that are population based; hence the rates are not samples. Why calculate standard errors anyway, and do significance testing?

15.5 This problem deals with a study reported in Bunker et al. [1969]. Halothane, an anesthetic agent, was introduced in 1956. Its early safety record was good, but reports of massive hepatic damage and death began to appear. In 1963, a Subcommittee on the National Halothane Study was appointed. Two prominent statisticians, Frederick Mosteller and Lincoln Moses, were members of the committee. The committee designed a large cooperative retrospective study, ultimately involving 34 institutions

Table 15.7 Mortality Data for Problem 15.5

	Number of Operations			Number of Deaths		
Physical Status	Total	Halothane	Cyclopropane	Total	Halothane	Cyclopropane
Unknown	69,239	23,684	10,147	1,378	419	297
1	185,919	65,936	27,444	445	125	91
2	104,286	36,842	14,097	1,856	560	361
3	29,491	8,918	3,814	2,135	617	403
4	3,419	1,170	681	590	182	127
5	21,797	6,579	7,423	314	74	101
6	11,112	2,632	3,814	1,392	287	476
7	2,137	439	749	673	111	253
Total	427,400	146,200	68,169	8,783	2,375	2,109

that completed the study. "The primary objective of the study was to compare halothane with other general anesthetics as to incidence of fatal massive hepatic necrosis within six weeks of anesthesia." A four-year period, 1959–1962, was chosen for the study. One categorization of the patients was by physical status at the time of the operation. Physical status varies from good (category 1) to moribund (category 7). Another categorization was by mortality level of the surgical procedure, having values of low, middle, high. The data in Table 15.7 deal with middle-level mortality surgery and two of the five anesthetic agents studied, the total number of administrations, and the number of patients dying within six weeks of the operation.

(a) Calculate the crude death rates per 100,000 per year for total, halothane, and cyclopropane. Are the crude rates for halothane and cyclopropane significantly different?

(b) By direct standardization (relative to the total), calculate standardized death rates for halothane and cyclopropane. Are the standardized rates significantly different?

(c) Calculate the standardized mortality rates for halothane and cyclopropane and test the significance of the difference.

(d) The calculations of the standard errors of the standardized rates depend on certain assumptions. Which assumptions are likely not to be valid in this example?

15.6 In 1980, 45 SIDS (sudden infant death syndrome) deaths were observed in King County. There were 15,000 births.

(a) Calculate the SIDS rate per 100,000 births.

(b) Construct a 95% confidence interval on the SIDS rate per 100,000 using the Poisson approximation to the binomial.

(c) Using the normal approximation to the Poisson, set up the 95% limits.

(d) Use the square root transformation for a Poisson random variable to generate a third set of 95% confidence intervals. Are the intervals comparable?

(e) The SIDS rate in 1970 in King County is stated to be 250 per 100,000. Someone wants to compare this 1970 rate with the 1980 rate and carries out a test of two proportions, $p_1 = 300$ per 100,000 and $p_2 = 250$ per 100,000, using the binomial distributions with $N_1 = N_2 = 100{,}000$. The large-sample normal approximation is used. What part of the Z-statistic: $(p_1 - p_2)$/standard error$(p_1 - p_2)$ will be right? What part will be wrong? Why?

Table 15.8 Heart Disease Data for Problem 15.7

Gender	Age	Epileptics: Person-Years at Risk	New and Nonfatal IHD Cases	Incidence in General Population per 100,000/year
Male	30–39	354	2	76
	40–49	303	2	430
	50–59	209	3	1291
	60–69	143	4	2166
	70+	136	4	1857
Female	30–39	534	0	9
	40–49	363	1	77
	50–59	218	3	319
	60–69	192	4	930
	70+	210	2	1087

15.7 Annegers et al. [1976] investigated ischemic heart disease (IHD) in patients with epilepsy. The hypothesis of interest was whether patients with epilepsy, particularly those on long-term anticonvulsant medication, were at less than expected risk of ischemic heart disease. The study dealt with 516 cases of epilepsy; exposure time was measured from time of diagnosis of epilepsy to time of death or time last seen alive.

(a) For males aged 60 to 69, the number of years at risk was 161 person-years. In this time interval, four IHD deaths were observed. Calculate the hazard rate for this age group in units of 100,000 persons/year.

(b) Construct a 95% confidence interval.

(c) The expected hazard rate in the general population is 1464 per 100,000 persons/year. How many deaths would you have expected in the age group 60 to 69 on the basis of the 161 person-years experience?

(d) Do the number of observed and expected deaths differ significantly?

(e) The raw data for the incidence of ischemic heart disease are given in Table 15.8. Calculate the expected number of deaths for males and the expected number of deaths for females by summing the expected numbers in the age categories (for each gender separately). Treat the total observed as a Poisson random variable and set up 95% confidence intervals. Do these include the expected number of deaths? State your conclusion.

(f) Derive a formula for an indirect standardization of these data (see Note 15.2) and apply it to these data.

15.8 A random sample of 100 subjects from a population is divided into two age groups, and for each age group the number of cases of a certain disease is determined. A reference population of 2000 persons has the following age distribution:

	Sample		Reference Population
Age	Total Number	Number of Cases	Total Number
1	80	8	1000
2	20	8	1000

(a) What is the crude case rate per 1000 population for the sample?

(b) What is the standard error of the crude case rate?

(c) What is the age-adjusted case rate per 1000 population using direct standardization and the reference population above?

(d) How would you test the hypothesis that the case rate at age 1 is not significantly different from the case rate at age 2?

15.9 The data in Table 15.9 come from a paper by Friis et al. [1981]. The mortality among male Hispanics and non-Hispanics was as shown.

Table 15.9 Mortality Data for Problem 15.9

	Hispanic Males		Non-Hispanic Males	
Age	Number	Number of Deaths	Number	Number of Deaths
0–4	11,089	0	51,250	0
5–14	18,634	0	120,301	0
15–24	10,409	0	144,363	2
25–34	16,269	2	136,808	9
35–44	11,050	0	106,492	46
45–54	6,368	7	91,513	214
55–64	3,228	8	70,950	357
65–74	1,302	12	34,834	478
75+	1,104	27	16,223	814
Total	79,453	56	772,734	1,920

(a) Calculate the crude death rate among Hispanic males.

(b) Calculate the crude death rate among non-Hispanic males.

(c) Compare parts (a) and (b) using an appropriate test.

(d) Calculate the SMR using non-Hispanic males as the reference population.

(e) Test the significance of the SMR as compared with a ratio of 1. Interpret your results.

15.10 The data in Table 15.10, abstracted from National Center for Health Statistics [1976], deal with the mortality experience in poverty and nonpoverty areas of New York and Seattle.

(a) Using New York City as the "standard population," calculate the standardized mortality rates for Seattle taking into account race and poverty area.

(b) Estimate the variance of this quantity and calculate 99% confidence limits.

(c) Calculate the standardized death rate per 100,000 population.

Table 15.10 Mortality Data for Problem 15.10

		New York City		Seattle	
Area	Race	Population	Death Rate per 1000	Population	Death Rate per 1000
Poverty	White	974,462	9.9	29,016	22.9
	All others	1,057,125	8.5	14,972	12.5
Nonpoverty	White	5,074,379	11.6	434,854	11.7
	All other	788,897	6.4	51,989	6.5

(d) Interpret your results.

(e) Why would you caution a reviewer of your analysis about the interpretation?

15.11 In a paper by Foy et al. [1983] the risk of getting *Mycoplasma* pneumonia in a two-year interval was determined on the basis of an extended survey of schoolchildren. Of interest was whether children previously exposed to *Mycoplasma pneumoniae* had a smaller risk of recurrence. In the five- to nine-year age group, the following data were obtained:

	Exposed Previously	**Not Exposed Previously**
Person-years at risk	680	134
Number with *Mycoplasma* pneumonia	7	8

(a) Calculate 95% confidence intervals for the infection rate per 100 person-years for each of the two groups.

(b) Test the significance of the difference between the infection rates.

***(c)** A statistician is asked to calculate the study size needed for a new prospective study between the two groups. He assumes that $\alpha = 0.05$, $\beta = 0.20$, and a two-tailed, two-sample test. He derives the formula

$$\lambda_2 = \sqrt{\lambda_1} - \frac{2.8}{\sqrt{n}}$$

where λ_i is the two-year infection rate for group i and n is the number of persons per group. He used the fact that the square root transformation of a Poisson random variable stabilizes the variance (see Section 10.6). Derive the formula and calculate the infection rate in group 2, λ_2 for $\lambda_1 = 10$ or 6, and sample sizes of 20, 40, 60, 80, and 100.

15.12 In a classic paper dealing with mortality among women first employed before 1930 in the U.S. radium dial–painting industry, Polednak et al. [1978] investigated 21 malignant neoplasms among a cohort of 634 women employed between 1915 and 1929. The five highest mortality rates (observed divided by expected deaths) are listed in Table 15.11.

(a) Test which ratios are significantly different from 1.

(b) Assuming that the causes of death were selected without a particular reason, adjust the observed p-values using an appropriate multiple-comparison procedure.

(c) The painters had contact with the radium through the licking of the radium-coated paintbrush to make a fine point with which to paint the dial. On the basis of this

Table 15.11 Mortality Data for Problem 15.12

Ranked Cause of Death	Observed Number	Expected Number	Ratio
Bone cancer	22	0.27	81.79
Larynx	1	0.09	11.13
Other sites	18	2.51	7.16
Brain and CNS	3	0.97	3.09
Buccal cavity, pharynx	1	0.47	2.15

information, would you have "preselected" certain malignant neoplasms? If so, how would you "adjust" the observed p-value?

15.13 Consider the data in Table 15.12 (from Janerich et al. [1974]) listing the frequency of infants with Simian creases by gender and maternal smoking status.

Table 15.12 Influence of Smoking on Development of Simian Creases

Gender of Infant	Maternal Smoking	Birthweight Interval (lb)			
		<6	6–6.99	7–7.99	≥8
Female	No	2/45	5/156	9/242	11/216
	Yes	4/48	8/107	6/110	3/44
Male	No	5/40	5/109	23/265	18/278
	Yes	10/55	6/84	10/106	6/74

(a) These data can be analyzed by the multidimensional contingency table approach of Chapter 7. However, we can also treat it as a problem in standardization. Describe how indirect standardization can be carried out using the total sample as the reference population, to compare "risk" of Simian creases in smokers and nonsmokers adjusted for birthweight and gender of the infants.

(b) Carry out the indirect standardization procedure and compare the standardized rates for smokers and nonsmokers. State your conclusions.

(c) Carry out the logistic model analysis of Chapter 7.

***15.14** Show that the variance of the standardized mortality ratio, equation (3), is approximately equal to equation (4).

REFERENCES

Annegers, J. F., Elveback, L. R., Labarthe, D. R., and Hauser, W. A. [1976]. Ischemic heart disease in patients with epilepsy. *Epilepsia*, **17**: 11–14.

Bruce, E., Frederick, R., Bruce, R., and Fisher, L. D. [1976]. Comparison of active participants and dropouts in CAPRI cardiopulmonary rehabilitation programs. *American Journal of Cardiology*, **37**: 53–60.

Bunker, J. P., Forest, W. H., Jr., Mosteller, F., and Vandam, L. D. [1969]. *The National Halothane Study: A Study of the Possible Association between Halothane Anesthesia and Postoperative Hepatic Necrosis*. National Institute of Health/National Institute of Several Medical Sciences, Bethesda, MD.

Clark, D. A., Stinson, E. B., Griepp, R. B., Schroeder, J. S., Shumway, N. E., and Harrison, D. C. [1971]. Cardiac transplantation: VI. Prognosis of patients selected for cardiac transplantation. *Annals of Internal Medicine*, **75**: 15–21. Used with permission.

Foy, H. M., Kenny, G. E., Cooney, M. K., Allan, I. D., and van Belle, G. [1983]. Naturally acquired immunity to mycoplasma pneumonia infections. *Journal of Infectious Diseases*, **147**: 967–973. Used with permission from University of Chicago Press.

Friis, R., Nanjundappa, G., Prendergast, J. J., Jr., and Welsh, M. [1981]. Coronary heart disease mortality and risk among hispanics and non-hispanics in Orange County, CA. *Public Health Reports*, **96**: 418–422.

Janerich, D. T., Skalko, R. G., and Porter, I. H. (eds.) [1974]. *Congenital Defects: New Directions in Research*. Academic Press, New York.

Kalbfleisch, J. D., and Prentice, R. L. [2002]. *The Statistical Analysis of Failure Time Data*, 2nd ed. Wiley, New York.

National Cancer Institute [1975]. *Third National Cancer Survey: Incidence Data*. Monograph 41. DHEW Publication (NIH) 75–787. U.S. Government Printing Office, Washington, DC.

National Center for Health Statistics [1976]. *Selected Vital and Health Statistics in Poverty and Non-poverty Areas of 19 Large Cities: United States, 1969–1971*. Series 21, No. 26. U.S. Government Printing Office, Washington, DC.

Polednak, A. P., Stehney, A. F., and Rowland, R. E. [1978]. Mortality among women first employed before 1930 in the U.S. radium dial-painting industry. *American Journal of Epidemiology*, **107**: 179–195.

CHAPTER 16

Analysis of the Time to an Event: Survival Analysis

16.1 INTRODUCTION

Many biomedical analyses study the time to an event. A cancer study of combination therapy using surgery, radiation, and chemotherapy may examine the time from the onset of therapy until death. A study of coronary artery bypass surgery may analyze the time from surgery until death. In each of these two cases, the event being used is death. Other events are also analyzed. In some cancer studies, the time from successful therapy (i.e., a patient goes into remission) until remission ends is studied. In cardiovascular studies, one may analyze the time to a heart attack or death, whichever event occurs first. A health services project may consider the time from enrollment in a health plan until the first use of the facilities. An analysis of children and their need for dental care may use the time from birth until the first cavity is filled. An assessment of an ointment for contact skin allergies may consider the time from treatment until the rash has cleared up.

In each of the foregoing situations, the data consisted of the time from a fixed or designated initial point until an event occurs. In this chapter we show how to analyze such *event data*. When the event of interest is death, the subject is called *survival analysis*. In medicine and public health this name is often used generically, even when the endpoint or event being studied is not death but something else. In industrial settings the study of the lifetime of a component (until failure) is called *reliability theory*, and social scientists use the term *event history analysis*. For concreteness, we often speak of the event as death and the time as survival time. However, it should always be kept in mind that there are other uses.

In this chapter we consider the presentation of time to event data, estimation of the time to an event, and its statistical variability. We also consider potential predictor or explanatory variables. A third topic is to compare the time to event in several different groups. For example, a study of two alternative modes of cancer therapy may examine which group has the best survival experience.

When the event is not death, there may be multiple occurrences for a given person or multiple types of event. It is usually possible to restrict the analysis to the first event as we did in the situations described above. This restriction trades a considerable gain statistical simplicity for an often modest loss in power. We discuss the analysis of multiple events only briefly.

16.2 SURVIVORSHIP FUNCTION OR SURVIVAL CURVE

In previous chapters we examined means of characterizing the distribution of a variable using, for example, the cumulative distribution function and histograms. One might take survival data

Biostatistics: A Methodology for the Health Sciences, Second Edition, by Gerald van Belle, Lloyd D. Fisher, Patrick J. Heagerty, and Thomas S. Lumley
ISBN 0-471-03185-2

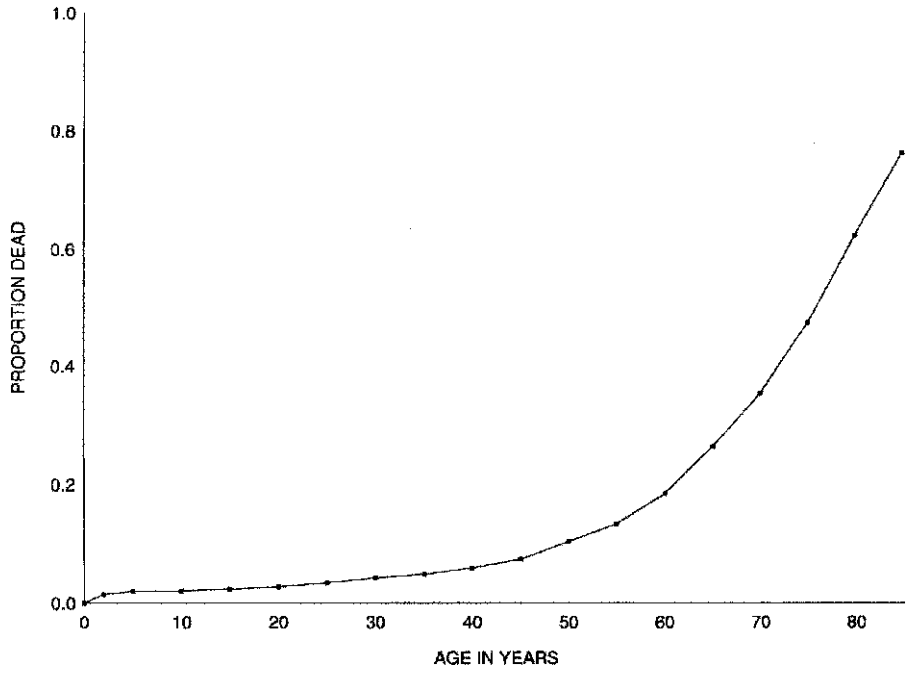

Figure 16.1 Cumulative probability of death, United States, 1974. (From U.S. Department of Health, Education, and Welfare [1976].)

and present the cumulative distribution function. Figure 16.1 shows an estimate for the U.S. population in 1974 of the probability of dying before a fixed age. This is an estimate of the cumulative distribution of survival in the United States in 1974. Note that there is an increase in deaths during the first year; after this the rate levels off but then climbs progressively in the later years. This cumulative probability of death is then an estimate of the probability that a person dies at or before the given time. That is,

$$F(t) = P[\text{person dies at a time} \leq t]$$

If we had observed the entire survival experience of the 1974 population, we would estimate this quantity as we estimated the cumulative distribution function previously. We would estimate it as

$$F(t) = \frac{\text{number of people who die at or before time } t}{\text{total number observed}} \tag{1}$$

Note, however, that we cannot estimate the survival experience of the 1974 population this way because we have not observed all of its members until death. This is a most fortunate circumstance since the population includes all of the authors of this book as well as many of its readers. In the next section, we discuss some methods of estimating survival when one does not observe the true survival of the entire population.

It is depressing to speak of death; it is more pleasant to speak of life. In analyzing survival data, the custom has grown not of using the cumulative probability of death but of using an equivalent function called the *survivorship function* or *survival curve*. This function is merely the percent of people who live to a fixed time or beyond.

Definition 16.1. The *survival curve*, or *survivorship function*, is the proportion or percent of people living to a fixed time t or beyond. The curve is then a function of t:

$$S(t) = \begin{cases} \text{percent of people surviving to time } t \text{ or beyond if} \\ \quad \text{expressed as a percent} \\ \text{proportion of people surviving to time } t \text{ or beyond} \\ \quad \text{if expressed as a proportion} \end{cases} \tag{2}$$

If we have a sample from a population, there is a distinction between the population survival curve and the sample or estimated population survival curve. In practice, there is no distinct notation unless it is necessary to emphasize the difference. The context will usually show which of the two is meant.

The cumulative distribution function of the survival and the survival curve are closely related. If the two curves are continuous, they are related by

$$S(t) = 100[1 - F(t)] \quad \text{or} \quad S(t) = 1 - F(t)$$

(When we look at the sample curves, the curves are equal at all points except for the points where the curves jump. At these points there is a slight technical problem because we have used $\leq$ in one instance and $\geq$ in the other instance. But for all practical purposes, the two curves are related by the equation above.)

Figure 16.2 shows the survival curve for the U.S. population as given in Figure 16.1. As you can see, the survival curve results by "flipping over" the cumulative probability of death and using percentages. As mentioned above, the estimate of the curve in Figure 16.2 is complicated by the fact that many people in the 1974 U.S. population are happily alive. Thus, their true

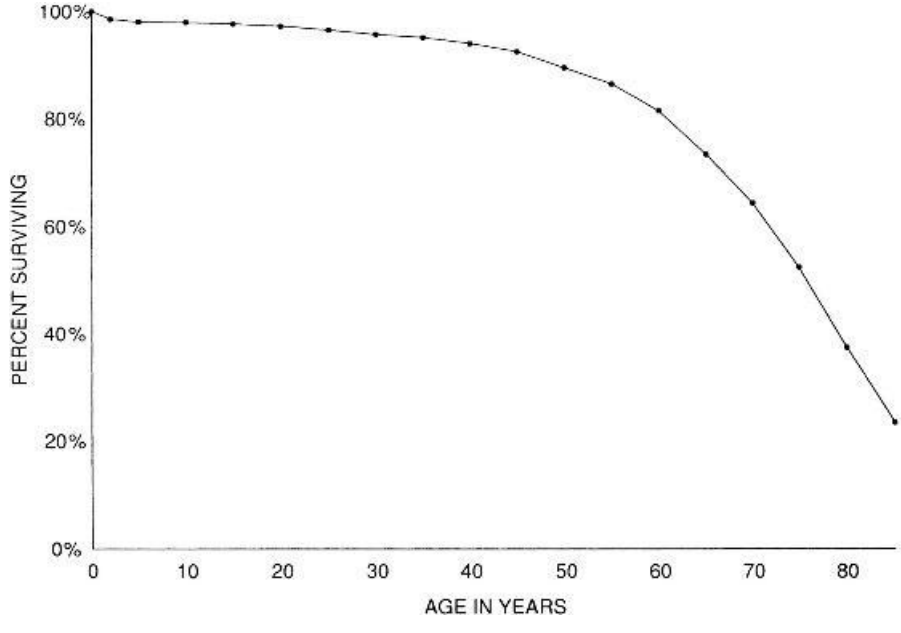

Figure 16.2 Survival curve of the U.S. population, 1974. Same data as used in Figure 16.1.

survival is not yet observed. The survival in the overall population is not yet observed. The survival in the overall population is estimated by the method discussed in the next section.

Sometimes the *proportion* surviving to time t or beyond is used. We will use them interchangeably. The two are simply related; to find the percent, merely multiply the proportion by 100.

If we observe the survival of all persons, it is easy to estimate the survival curve. In analogy with the estimate of the cumulative distribution function, the estimate of the survival curve at a fixed t is merely the percent of people whose survival was equal to the value t or greater. That is,

$$S(t) = 100\left(\frac{\text{number of people who survive to or beyond } t}{\text{total number observed}}\right) \tag{3}$$

In many instances, we are not able to observe everyone until they reach the event of interest. This makes the estimation problem more challenging. We discuss the estimates in the next section.

16.3 ESTIMATION OF THE SURVIVAL CURVE: ACTUARIAL OR LIFE TABLE METHOD

Consider a clinical study of a procedure with a high initial mortality rate: for example, very delicate high-risk surgery during its development period. Suppose that we design a study to follow a group of such people for two years. Because most of the mortality is expected during the first year, it is decided to concentrate the effort on the first year. Two thousand people are to be entered in the study; half of them will be followed for two years, while one-half will be followed only for the critical first year. The people are randomized into two groups, group 1 to be followed for one year and group 2 to be followed for both years. Suppose that the data are as follows:

	Group 1		Group 2	
Year	Number Observed	Number Who Died	Number Observed	Number Who Died
1	1000	240	1000	200
2	—	—	800	16

We wish to estimate one- and two-year survival. We consider three methods of estimation. The first two methods will not be appropriate but are used to motivate the correct life table method to follow.

One way of estimating survival might be to estimate separately the one- and two-year survival. Since it is wasteful to "throw away" data and the reason that 2000 people were observed for one year was because that year was considered crucial, it is natural to estimate the percent surviving for one year by the total population. This percentage is as follows:

$$\text{percent of one-year survival} = 100\left(\frac{2000 - 240 - 200}{2000}\right) = 78.0\%$$

To estimate two-year survival, we did not observe what happened to the subjects in group 1 during the second year. Thus, we might estimate the survival using only those in group 2. This

estimate is

$$\text{percent of two-year survival} = 100\left(\frac{1000-200-16}{1000}\right) = 78.4\%$$

There are two problems with this estimation method. The first is that we need to know the potential follow-up time (one year or two years) for everyone. In a clinical trial this is reasonable, but in a cohort study we may not know whether someone who in fact died after six months would have been followed up for one year or two years if he or she had not died. Nor is it reasonable that our estimate of the survival should depend on this unobservable potential follow-up.

More importantly, we have a problem in that the estimated percent surviving one year is less than the percent surviving two years! Clearly, as time increases, the percent surviving must decrease, but the sampling variability in the estimate has led to the second-year estimate being larger than the first-year estimate. Although this method is approximately unbiased and uses all the available data, it is not a desirable way to estimate our survival curve.

One way to get around this problem is to use only the subjects from group 2 who are observed for two years. Then we have a straightforward estimate of survival at each time period. The percent surviving one year or more is 80%, while the percent surviving two or more years is, as before, 78.4%. This gives a consistent pattern of survival but seems quite wasteful; we deliberately designed the study to allow us to observe more subjects in the first year, when the mortality was expected to be high. It does not seem appropriate to throw away the 1000 subjects who were only observed for one year. If we need to do this, we had an extremely poor experimental design.

The solution to our problem is to note that we can efficiently estimate the probability of one-year survival using both groups of people. Further, using the second group, we can estimate the probability of surviving the second year *conditionally upon having survived the first year*. The two estimates as percentages are

$$\text{percent of one-year survival} = 78.0\%$$

$$\text{percent surviving year 2} = 100\left(\frac{800-16}{800}\right) = 98.0\%$$

We can then combine these to get an estimate of the probability of surviving in the first year and the second year by using the concept of conditional probability. We see that the probability of two-year survival is the probability of one-year survival times the probability of two-year survival given one-year survival, and so cannot be larger than the probability of one-year survival. The probability of two-year survival is as follows:

$$P[A \text{ and } B] = P[A]P[B|A]$$

Let A be the survival of one year and B the survival of two years. Then

$$\begin{aligned} P[\text{one-year survival}] &= P[\text{one-year survival}] \\ &\quad \times P[\text{two-year survival}|\text{one-year survival}] \\ &= 0.78 \times 0.98 = 0.7644 \end{aligned}$$

For these probability calculations, note that it is more convenient to have probabilities than percents because the probabilities multiply. If we had percents, the formula would have an extra factor of 100. For this reason the calculations on the survival curves are usually done as probabilities and then switched to percentages for graphical presentation. We will adhere to this.

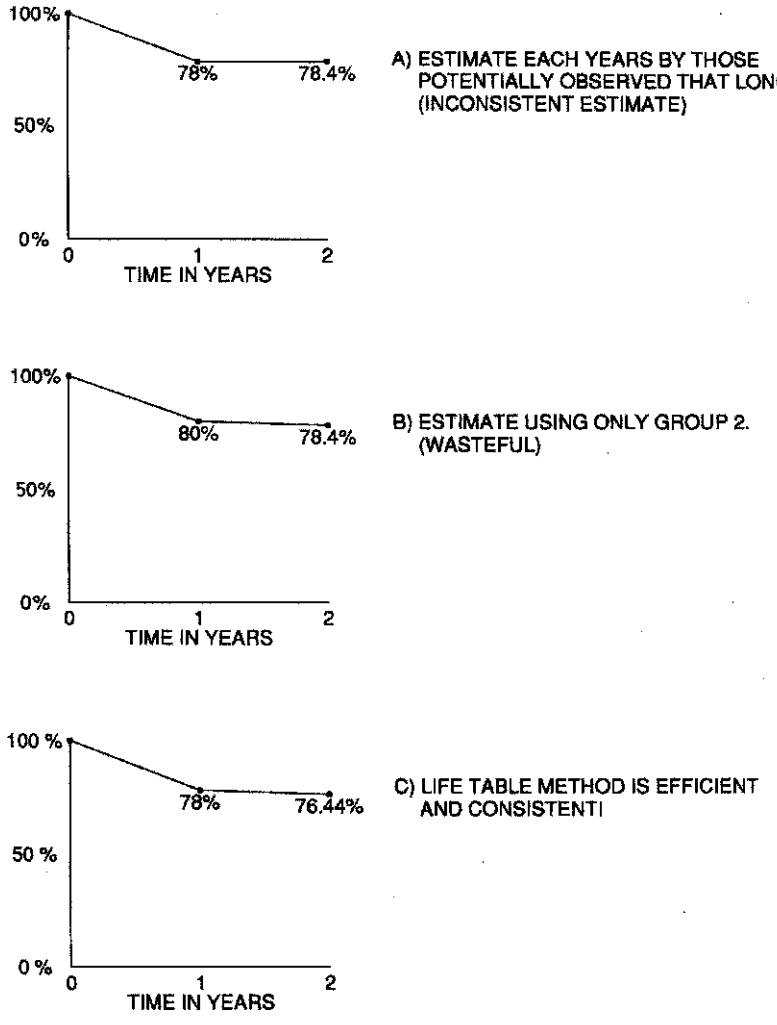

Figure 16.3 Three methods of estimating survival.

Figure 16.3 presents the three estimates; for these data they are all close. The third estimate gives a self-consistent estimate of the curve (i.e., the curve will never increase) and the estimate is efficient (because it uses all the data); it is the correct method for estimating survival. This idea can easily be generalized to more than two intervals.

When the data are grouped into time intervals, we can estimate the survival in each interval. Let x denote the lower endpoint of each interval. [x rather than t is used here to conform to standard notation in the actuarial field. When it is necessary to index the intervals, we will use $i(x)$ to denote the inverse relationship.] Let $\prod_i$ denote the probability of surviving to $x(i)$, where $x(i)$ is the lower endpoint of the ith interval; that is,

$$\prod_i = S(x(i))$$

where S is the survival curve (expressed here as the proportion surviving). Further, let π_i be the probability of living through the interval, with lower endpoint $x(i)$, conditionally upon the event

of being alive at the beginning of the interval. Using the definition of a conditional probability,

$$\pi_i = \frac{\prod_{i+1}}{\prod_i} = \frac{P[\text{survive to the end of the } i\text{th interval}]}{P[\text{survive to the end of the } (i-1)\text{st interval}]} \tag{4}$$

From this,

$$\prod_{i+1} = \pi_i \prod_i$$

and

$$\prod_{i+1} = \pi_1 \pi_2 \cdots \pi_i \qquad \text{where} \quad \prod_1 = 1 \tag{5}$$

In presenting group data graphically, one plots points corresponding to the time of the lower endpoint of the interval and the corresponding $\prod_i$ value. The plotted points are then joined by straight-line segments, as in Figure 16.4.

There is one further complication before we present the life table estimates. If we are following people periodically (e.g., every six months or every year), it will occasionally happen that people cannot be located. Such subjects are called *lost to follow-up* in the study. Further, subjects may be withdrawn from the study for a variety of reasons. In clinical studies in the United States, all subjects have the right to withdraw from participation at any time. Or we might be trying to examine a medical survival in patients who could potentially be treated with surgery. Some of them may subsequently receive surgery; we could withdraw such patients from the analysis at the time they received surgery. The rationale for this would be that after they received surgery, their survival experience is potentially altered. Whatever the reason for a person being lost to follow-up or withdrawn, this fact must be considered in the life table analysis.

To estimate the survival curve from data, the method is to estimate the π_i and $\prod_i$ by the product of the estimates of the π_i according to equation (5). The data are usually presented in the form of Table 16.1. How might one estimate the probability of dying in the interval whose

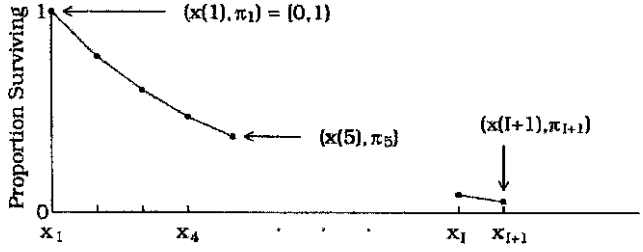

Figure 16.4 Form of the presentation of the survival curve for grouped survival data.

Table 16.1 Presentation of Life Table Data

Interval	Number of Subjects Observed Alive at Beginning of Interval	Died during Interval	Lost to Follow-up during Interval	Withdrawn Alive during Interval
x to $x + \Delta x$	l_x	d_x	u_x	w_x
$x(1) - x(2)$	$l_{x(1)}$	$d_{x(1)}$	$u_{x(1)}$	$w_{x(1)}$
$x(2) - x(3)$	$l_{x(2)}$	$d_{x(2)}$	$u_{x(2)}$	$w_{x(2)}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x(I) - x(I+1)$	$l_{x(I)}$	$d_{x(I)}$	$u_{x(I)}$	$w_{x(I)}$

lower endpoint is x conditionally upon being alive at the beginning of the interval? At first glance one might reason that there were l_x subjects, of whom (a binomial) d_x died, so that the estimate should be d_x/l_x. The problem is that those who were lost to follow-up or withdrew during the interval might have died during the interval *after* withdrawing, and this would not be counted. If such persons were equally likely to withdraw at any time during the interval, on the average they would be observed only one-half of the time. Thus, they really represent only one-half a person at risk. Thus the effective number of persons at risk, l'_x, is

$$\begin{aligned} l'_x &= \underbrace{l_x - (u_x + w_x)}_{\substack{\text{number observed} \\ \text{over entire interval}}} + \underbrace{\tfrac{1}{2}(u_x + w_x)}_{\substack{\text{number observed} \\ \text{over } \frac{1}{2} \text{ interval}}} \\ &= l_x - \tfrac{1}{2}(u_x + w_x) \end{aligned} \tag{6}$$

where the u_x is the number lost to follow-up and w_x is the number withdrawing. The estimate of the proportion dying, q_x, is thus

$$q_x = \frac{d_x}{l'_x}$$

The estimate of π_i, the probability of surviving the interval $x(i)$ to $x(i+1)$, is

$$p_{x(i)} = 1 - q_{x(i)}$$

Finally, the estimate of $\prod_i = \pi_1 \pi_2 \cdots \pi_{i-1}$, $\prod_1 = 1$ is

$$P_{x(i)} = p_{x(1)} p_{x(2)} \cdots p_{x(i-1)}, \quad P_{x(0)} = 1 \tag{7}$$

Note that those who are lost to follow-up and those who are withdrawn alive are treated together; that is, in the estimates, only the sum of the two is used. In many presentations such people are lumped together as *withdrawn* or *censored*.

Before presenting the estimates, it is also clear that an estimate of the survival curve will be more useful if some idea of its variability is given.

An estimate of the standard error of the P_x is given by Greenwood's formula [Greenwood, 1926]:

$$\begin{aligned} \mathrm{SE}(P_{x(i)}) &= P_{x(i)} \sqrt{\sum_{j=1}^{i-1} \frac{q_{x(j)}}{l'_{x(j)} - d_{x(j)}}} \\ &= P_{x(i)} \sqrt{\sum_{j=1}^{i-1} \frac{q_{x(j)}}{l'_{x(j)} p_{x(j)}}} \end{aligned} \tag{8}$$

Confidence intervals constructed using ± 1.96 times this standard error are valid only in relatively large samples. For example, it is easy to see that these confidence intervals could extend outside the interval [0, 1], where the probability must lie. Better confidence intervals in small samples can be obtained by transforming $P(t)$; they are discussed in the Notes to this chapter.

Example 16.1. The method is illustrated by data of Parker et al. [1946], as discussed in Gehan [1969]. Those data are from 2418 males with a diagnosis of angina pectoris (chest pain thought to be of cardiac origin) at the Mayo Clinic between January 1, 1927 and December 31, 1936. The life table of survival time from diagnosis (in yearly intervals) is shown in Table 16.2.

Table 16.2 Life Table Analysis of 2418 Males with Angina Pectoris

x to $x + \Delta x$ (yr)	l_x	d_x	u_x	w_x	l'_x	q_x	p_x	P_x	SE(P_x)
0–1	2418	456	0	0	2418	0.1886	0.8114	1.0000	—
1–2	1962	226	39	0	1942.5	0.1163	0.8837	0.8114	0.0080
2–3	1697	152	22	0	1686.0	0.0902	0.9098	0.7170	0.0092
3–4	1523	171	23	0	1511.5	0.1131	0.8869	0.6524	0.0097
4–5	1329	135	24	0	1317.0	0.1025	0.8975	0.5786	0.0101
5–6	1170	125	107	0	1116.5	0.1120	0.8880	0.5139	0.0103
6–7	938	83	133	0	871.5	0.0952	0.9048	0.4611	0.0104
7–8	722	74	102	0	671.0	0.1103	0.8897	0.4172	0.0105
8–9	546	51	68	0	512.0	0.0996	0.9004	0.3712	0.0106
9–10	427	42	64	0	395.0	0.1063	0.8937	0.3342	0.0107
10–11	321	43	45	0	298.5	0.1441	0.8559	0.2987	0.0109
11–12	233	34	53	0	206.5	0.1646	0.8354	0.2557	0.0111
12–13	146	18	33	0	129.5	0.1390	0.8610	0.2136	0.0114
13–14	95	9	27	0	81.5	0.1104	0.8896	0.1839	0.0118
14–15	59	6	23	0	47.5	0.1263	0.8737	0.1636	0.0123

Source: Data from Gehan [1969].

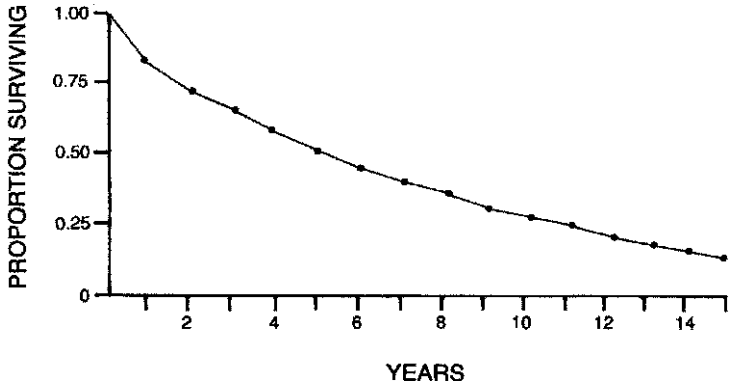

Figure 16.5 Survivorship function. (Data from Gehan [1969]; see Table 16.2.)

The survival data are given graphically in Figure 16.5. Note that in this case the proportion rather than the percent is presented.

As a second example, we consider patients with the same diagnosis, angina pectoris; these data are more recent.

Example 16.2. Passamani et al. [1982] studied patients with chest pain who were studied for possible coronary artery disease. Chest pain upon exertion is often associated with coronary artery disease. The chest pain was evaluated by a physician as definitely angina, probably angina, probably not angina, and definitely not angina. The definitions of these four classes were:

- *Definitely angina:* a substantial discomfort that is precipitated by exertion, relieved by rest and/or nitroglycerin in less than 10 minutes, and has a typical radiation to either shoulder, jaw, or the inner aspect of the arm. At times, definite angina may be isolated to the shoulder, jaw, arm, or upper abdomen.

- *Probably angina:* has most of the features of definite angina but may not be entirely typical in some aspects.
- *Probably not angina:* an atypical overall pattern of chest pain symptoms which does not fit the description of definite angina.
- *Definitely not angina:* a pattern of chest pain symptoms that are unrelated to activity, unrelieved by nitroglycerin and/or rest, and appear clearly noncardiac in origin.

The data are plotted in Figure 16.6. Note how much improved the survival of the angina patients (definite and probable) is compared with the Mayo data of Figure 16.5. Those data had a 52% five-year survival. These data have 91% and 85% five-year survival! This indicates the great difficulty of using historical control data. A statistic and p-value for testing differences among the four groups is discussed in Section 16.6.

Table 16.3 gives the calculation using 91-day intervals and four intervals to approximate a year for one of the four groups, the definite angina patients. As a sample calculation, consider the interval from 637 to 728 days. We see that

$$
\begin{aligned}
l_x &= 2704, \qquad u_x + d_x = 281 \\
l'_x &= 2704 - \frac{281}{2} = 2563.5 \\
q_x &= \frac{12}{2563.5} = 0.0047 \\
p_x &= 1 - 0.0047 = 0.9953 \\
P_x &= 0.9350 \times 0.9953 = 0.9306
\end{aligned}
$$

Note that the definite angina cases have the worst survival, followed by the probable angina cases (91%). The other two categories are almost indistinguishable.

As we have seen, in the life table method we have some data for which the event in question is not observed, often because at the time of the end of data collection and analysis, patients are still alive. One term used for such data is *censoring*, a term that brings to mind a powerful, possibly sinister figure throwing away data to mislead one in the data analysis. In this context

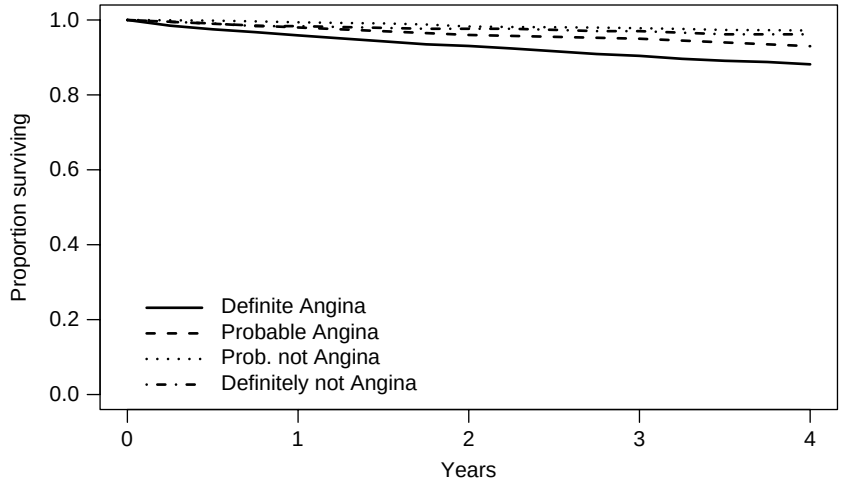

Figure 16.6 Survival by classification of chest pain. (Data from Passamani et al. [1982].)

Table 16.3 Life Table for Definite Angina Patients. Time in Days

$t(i)$	Enter	At Risk	Dead	Withdrawn Alive	Proportion Dead	Cumulative Survival of the End of Interval	SE	Effective Sample Size
0.0–90.9	2894	2894.0	44	0	0.0152	0.9848	0.002	2893.99
91.0–181.9	2850	2850.0	28	0	0.0098	0.9751	0.003	2893.99
182.0–272.9	2822	2822.0	22	0	0.0078	0.9675	0.003	2894.00
273.0–363.9	2800	2799.0	25	2	0.0089	0.9589	0.004	2893.77
364.0–454.9	2773	2773.0	23	0	0.0083	0.9509	0.004	2893.46
455.0–545.9	2750	2750.0	23	0	0.0084	0.9430	0.004	2893.23
546.0–636.9	2727	2727.0	23	0	0.0084	0.9350	0.005	2893.06
637.0–727.9	2704	2563.5	12	281	0.0047	0.9306	0.005	2882.32
728.0–818.9	2411	2394.0	17	34	0.0071	0.9240	0.005	2850.22
819.0–909.9	2360	2359.0	19	2	0.0081	0.9166	0.005	2818.52
910.0–1000.9	2339	2336.5	19	5	0.0081	0.9091	0.005	2792.12
1001.0–1091.9	2315	2035.5	11	559	0.0054	0.9042	0.006	2753.73
1092.0–1182.9	1745	1722.5	15	45	0.0087	0.8963	0.006	2654.36
1183.0–1273.9	1685	1685.0	19	0	0.0059	0.8910	0.006	2596.11
1274.0–1364.9	1675	1670.5	6	9	0.0036	0.8878	0.006	2564.52
1365.0–1455.9	1660	1274.5	9	771	0.0071	0.8816	0.007	2449.65

it refers to the fact that although one is interested in survival times, the actual survival times are not observed for all the subjects. We have seen several sources of censored data. Subjects may be alive at the time of analysis; (subjects) may be lost to follow-up; (subjects) may refuse to participate further in research; or (subjects) may undergo a different therapy which removes them from estimates of the survival in a particular therapeutic group.

The *life table* or *actuarial method* that we have used above has the strength of allowing censored data and also uses the data with maximum efficiency. There is an important underlying assumption if we are to get unbiased estimates of the survival in a population from which such subjects may be considered to come. *It is necessary that the withdrawal or censoring not be associated with the endpoint*. Obviously, if everyone is withdrawn because their situation deteriorates, one would expect a bias in the estimation of death. Let us emphasize this again. The life table estimate gives *biased* estimates if subjects who are censored at a given time have higher or lower chance of failure than those not censored at that time. The assumption we need is technically called *noninformative censoring*; the term *independent censoring* is also used.

We return later in the chapter to the related but distinct problem of competing causes of death, for example, examining the differences in death from cardiovascular causes in an elderly population where many people die of cancer or infectious disease during the study.

16.4 HAZARD FUNCTION OR FORCE OF MORTALITY

In the analysis of survival data, one is often interested in examining which periods have the highest or lowest risk of death. By risk of death, one has in mind the risk or probability among those alive at that time. For example, in very old age there is a high risk of dying each year *among* those reaching that age. The probability of any person dying, say, in the 100th year is small because so few people live to be 100 years old.

This concept is made rigorous by the idea of the hazard function or *hazard rate*. (A very precise definition of the hazard function requires ideas beyond the scope of this book and is discussed briefly in the Notes at the end of this chapter.) The hazard function is also called the *force of mortality*, *age-specific death rate, conditional failure rate*, and *instantaneous death rate*.

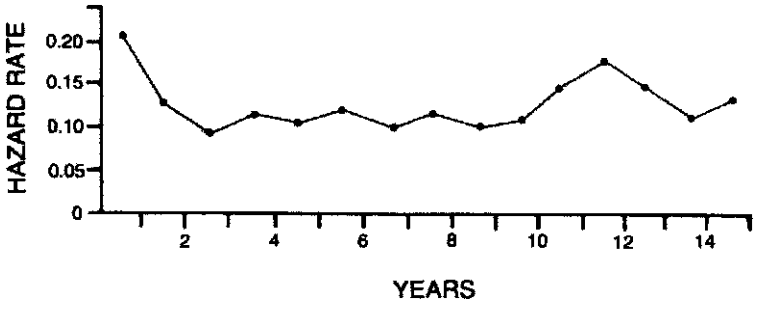

Figure 16.7 Hazard function for Example 16.1. (Data from Parker et al. [1946].)

Definition 16.2. In a life table situation, the (*interval or actuarial*) *hazard rate* is the expected number dying in the interval, divided by the product of the average number exposed in the interval and the interval width.

In other words, the hazard rate, λ, is the probability of dying per unit time given survival to the time point in question. The estimate h of the hazard function is given by

$$h_x = \frac{d_x}{l'_x - d_x/2} \frac{1}{\Delta x} \tag{9}$$

where $\Delta x(1) = x(i+1) - x(i)$, the interval width. This is an estimate of the form

$$\frac{\text{number dying}}{\text{total exposure time}}$$

l'_x is an estimate of the number at risk of death. Note that this estimate is analogous to the definition in Section 15.4. Those who die will on average have been exposed for approximately one-half of the time interval, so the number of intervals of observed time is approximately $(l'_x - d_x/2)\Delta x$. Thus, the hazard rate is a death rate; its units are proportion per unit time (e.g., percent per year). If the hazard rate has a constant value λ over time, the survival is exponential, that is, $S(t) = 100e^{-\lambda t}$, a point returned to later. The estimated hazard rate for Parker's data of Example 16.1 is given in Figure 16.7.

A large-sample approximation from Gehan [1969] for the SE of h is

$$\text{SE}(h_x) = \left\{ \frac{h_x^3}{l_x q_x} \left[1 - \left(\frac{h_x \Delta x}{2} \right)^2 \right] \right\}^{1/2} \tag{10}$$

For the data of Example 16.1, we compute the hazard function for the second interval. We find that

$$h_1 = \left(\frac{226}{1942.5 - 226/2} \right) \left(\frac{1}{1} \right) = 0.124$$

16.5 PRODUCT LIMIT OR KAPLAN–MEIER ESTIMATE OF THE SURVIVAL CURVE

If survival data are recorded in great detail, accuracy is preserved by placing the data into smaller rather than larger intervals. Obviously, if data are grouped, for example, into five-year

intervals while the time of death is recorded to the nearest day, considerable detail is lost. The *product limit* or *Kaplan–Meier estimate* is based on the idea of taking more and more intervals. In the limit, the intervals become arbitrarily small.

Suppose in the following that the time at which data are censored (lost to follow-up or withdrawn from the study) and the time of death (when observed) are measured to a high degree of accuracy. The product limit *or* Kaplan–Meier (see Kaplan and Meier [1958]) estimate (KM estimate) results from the actuarial or life table method of Section 16.4 as the number of intervals increases in such a way that the maximum interval width approaches zero. In this case it can be seen that the estimated survival curve is constant except for jumps at the observed times of death. The values of the survival probability before a time of death(s) is multiplied by the estimated probability of surviving past the time of death to find the new value of the survival curve.

To be more precise, suppose that n persons are observed. Further, suppose that the time of death is observed in l of the subjects at k distinct times $t_1 < t_2 < \cdots < t_k$. Let m_i be the number of deaths at time t_i. The other $n - l$ subjects are censored observations. If a censoring time and a death occur at the same time, it is assumed that the true time of death for the censored subject is greater than the censoring time observed. Let n_i be the number of subjects at risk of dying at time t_i. That is, $n_i = n$ minus the number of deaths prior to t_i and minus the number of subjects whose observations were censored prior to time t_i. The product limit estimate of the survival curve expressed as a proportion is

$$S(t) = \begin{cases} 1 & \text{for } t < t_1 \\ \prod_{j=1}^{i} \dfrac{n_i - m_i}{n_i}, & t_i \le t < t_{i+1} (i < k) \\ 0 & \text{for } t_k \le t \quad \text{if } m_k = n_k \text{ (i.e., no one survives past time } t_k) \\ \prod_{j=1}^{k} \dfrac{n_i - m_i}{n_i} & \text{for } t_k \le t \le \text{ largest observed censored observation} \end{cases} \tag{11}$$

If $m_k < n_k$, then $S(t)$ is undefined for $t >$ largest observed censored observation. Some software will report either $S(t) = 0$ or $S(t) = S(t_k)$ for times after the last censored observation, but this should not be encouraged.

We illustrate the method with an example.

Example 16.3. We again use the Stanford heart transplant data discussed in Section 15.4. Suppose that we wished to estimate the survival of these patients given medical treatment only. A complication is that when a donor heart becomes available, the patient has a heart transplant; we can no longer observe what the survival without a transplant would have been. One *incorrect* way to analyze such data would be the following. Since we are interested in medical survival, we should not worry about patients who have had surgery. We should go through the records and look at the survival curves only for patients who did not have surgery. Since by definition such people died awaiting the donor heart, their early survival experience would be quite poor.

At the time of the Stanford study, waiting lists were short and a donor heart was transplanted to the best-matching recipient on the waiting list [Crowley and Hu, 1977]. Thus, we may use surgery for heart transplantation as a source of censoring for medical survival: The availability of a heart should not be related to the severity of illness of the recipient. Current practice is quite different; more seriously ill patients are more likely to receive a transplant (*http://www.optn.org/*), so the censoring by surgery would be *informative* (biased) in a modern study.

Table 16.4 presents the medical survival data using surgery as the source of censoring for the Stanford heart transplant patients. The computations as described above are given. The product limit estimate of the correct survival curve is shown by solid lines in Figure 16.8. Lines with x's is the incorrect curve if one ignores the effect of surgery as censoring and totally eliminates

Table 16.4 Survival Data for Heart Transplant Patients

t (days)	Death (*)	n_i	$(n_i - m_i)/n_i$	$S(t),\ t_i \le t < t_{i+1}$
1	*	34	33/34	0.971
1		33		
2		32		
5	*	31	30/31	0.939
7	*	30	29/30	0.908
7		29		
11		28		
11		27		
12	*	26	25/26	0.873
15	*	25	24/25	0.838
15		24		
16		23		
17	*	22	21/22	0.800
17		21		
17		20		
19		19		
22		18		
24		17		
24		16		
26		15		
34	*	14	13/14	0.743
34		13		
35	*	12	11/12	0.681
36	*	11	10/11	0.619
36		10		
40	*	9	8/9	0.550
49	*	8	7/8	0.481
49		7		
50		6		
69		5		
81		4		
84	*	3	2/3	0.321
111	*	2	1/2	0.160
480		1		

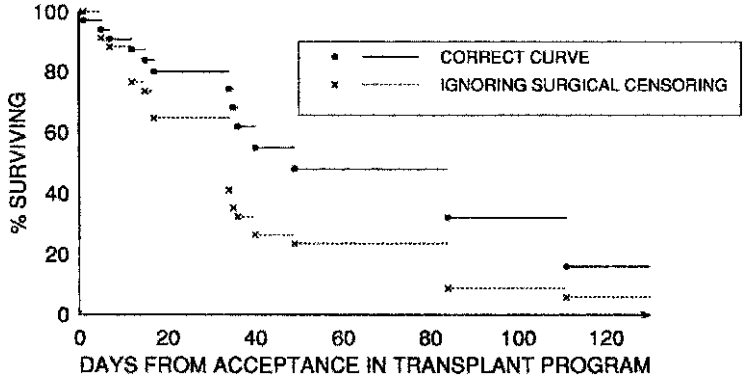

Figure 16.8 Days from acceptance in transplant program. Kaplan–Meier survival curve.

such subjects from the analysis. Finally, note that there was one patient who spontaneously improved under medical treatment and was reported alive at 16 months. The data of that subject are reported in the medical survival data as a 480-day survivor. As before, an asymptotic formula for the standard error of the estimate may be given. Greenwood's formula for the approximate standard error of the estimate also holds in this case. The form it takes is

$$\mathrm{SE}(S(t)) \doteq S(t)\sqrt{\sum_{j=1}^{i}\frac{m_j}{n_j(n_j-m_j)}} \qquad \text{for } t_i \le t < t_{i+1} \tag{12}$$

16.6 COMPARISON OF DIFFERENT SURVIVAL CURVES: LOG-RANK TEST

In this section we consider a test statistic for comparing two or more survival curves for different groups of subjects. This statistic is based on the following idea. Take a particular interval in which deaths occur, or in the case of the product limit curve, a time when one or more deaths occur. Suppose that the first group considered has one-third of the subjects being observed. How many deaths would we expect in the first group if, in fact, the survival experience is the same for all the groups? We expect the number of deaths to be proportional to the fraction of the people at risk of dying in the group. That is, for the first group the expected number of deaths would be the observed number of deaths at that time divided by 3. The log-rank test uses this simple fact. At each interval or time of death we take the observed number of deaths and calculate the expected number of deaths that would occur in each of the groups if all had the same risk of dying. For each group, the expected number of deaths is summed over all intervals and then compared to the observed number of deaths. Using this comparison, we get a statistic, the *log-rank statistic*, which has approximately a chi-square distribution with $k-1$ degrees of freedom when k groups are observed. We formalize this.

Suppose that one is interested in comparing the survival experience of k populations. Suppose that there are M different times at which deaths appear. For the life table method, this will usually be each interval. In the product limit approach, each death observed will be associated with a unique time. At the mth time, let d_{im} be the number of deaths observed in the ith population and l_{im} be the number at risk of dying. (For the life table approach with withdrawals, l_{ij} is the appropriate l'_x.) The data may be presented in M $2 \times k$ contingency tables with totals:

1	2	$\cdots$	k		
d_{1m}	d_{2m}	$\cdots$	d_{km}	D_m	dying
$l_{1m}-d_{1m}$	$l_{2m}-d_{2m}$	$\ldots$	$l_{km}-d_{km}$	A_m	alive
l_{1m}	l_{2m}	$\cdots$	l_{km}	T_m	total

$$m = 1, 2, \ldots, M$$

If all of the k populations are at equal risk of death, the probability of death will be the same in each population, and conditionally upon the row and column totals,

$$E(d_{im}) = \frac{l_{im}D_m}{T_m} \tag{13}$$

as in the chi-square test for contingency tables.

In the ith population, the total number of deaths observed is

$$O_i = \sum_{m=1}^{M} d_{im} \tag{14}$$

Examining all of the times of death, the expected number of deaths in the ith population is

$$E_i = \sum_{m=1}^{M} E(d_{im}) = \sum_{m=1}^{M} \frac{l_{im} D_m}{T_m} \tag{15}$$

The test statistic is then computed from the observed minus expected values. A simple approximate statistic suitable for hand calculation is

$$X^2 = \sum_{i=1}^{k} (O_i - E_i)^2 \Big/ E_i \tag{16}$$

The statistic is written in the familiar form of the chi-square test for comparing observed and expected values. [If any $E_i = 0$, define $(O_i - E_i)^2/E_i = 0$.] Under the null hypothesis of equal survival curves in the k groups this statistic will have approximately a chi-square distribution with $k-1$ degrees of freedom. The approximation is good when the subjects at risk are distributed over the k groups in roughly the same proportions at all times. The complete formulas for the log-rank test, which is implemented in most major statistics packages, are given in Note 16.3.

The log-rank test is illustrated by using the data of the Stanford transplant patients (Table 16.4) and comparing them with the data of Houston heart transplant patients, as reported in Messmer et al. [1969]. The time of survival for 15 Houston patients is read from Figure 16.9 and therefore has some inaccuracy.

Ordering both the Stanford and Houston transplant patients by their survival time after transplantation and status (dead or alive) gives Table 16.5. The dashes for the d_{im} values indicate where withdrawals occur, and those lines could have been omitted in the calculation. One stops when there are no future deaths at a time when members of both populations are present.

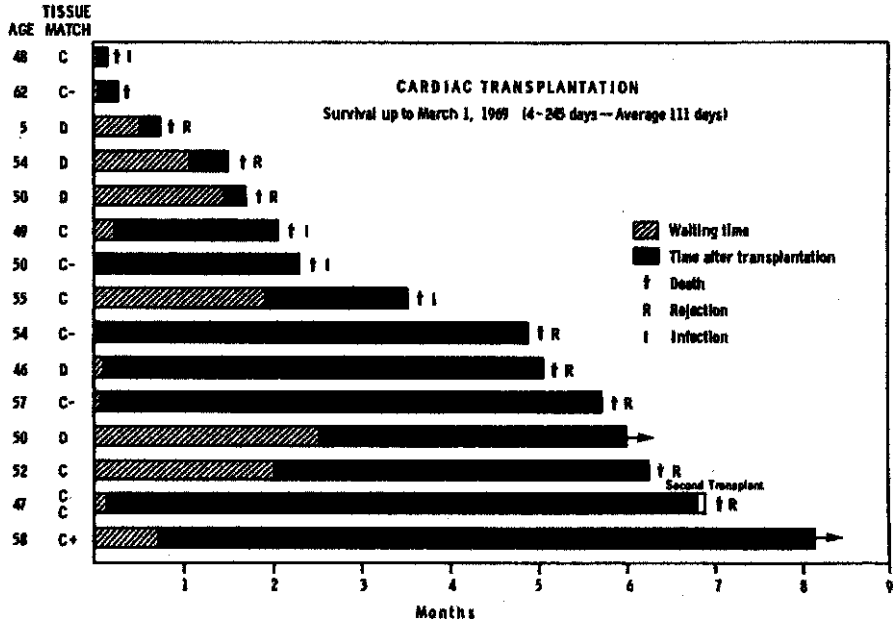

Figure 16.9 Survival of 15 patients given a cardiac allograft. Arrows indicate patients still alive on March 1, 1969. (Data from Messmer et al. [1969].)

Table 16.5 Stanford and Houston Survival Data

	Stanford		Houston			
Day	l_{1m}	d_{1m}	l_{2m}	d_{2m}	$E(d_{1m})$	$E(d_{2m})$
1	20	1	15	0	0.571	0.429
3	19	1	15	0	0.559	0.441
4	18	0	15	1	0.545	0.455
6	18	0	14	2	1.125	0.875
7	18	0	12	1	0.600	0.400
10	18	1	11	0	0.621	0.379
12	17	0	11	1	0.607	0.393
15	17	1	10	0	0.630	0.370
24	16	1	10	0	0.615	0.385
39	15	1	10	0	0.600	0.400
46	14	1	10	0	0.583	0.417
48	13	0	10	1	0.565	0.435
54	13	0	9	1	0.591	0.409
60	13	1	8	0	0.619	0.381
61	12	1	8	1	1.200	0.800
102	11	0	7	0	—	—
104	10	0	6	0	—	—
110	10	0	6	1	0.625	0.375
118	10	0	5	0	—	—
127	9	1	5	0	0.643	0.357
136	8	1	5	0	0.615	0.385
146	7	0	5	1	0.583	0.417
148	7	0	4	1	0.636	0.364
169	7	0	3	1	0.700	0.300
200	7	0	2	1	0.778	0.222

Summing the appropriate columns, one finds that

$$O_1 = \sum_m d_{1m} = 11$$

$$E_1 = \sum_m E(d_{1m}) = 14.611$$

$$O_2 = \sum_m d_{2m} = 13$$

$$E_2 = \sum_m E(d_{2m}) = 9.389$$

The log-rank statistic is 2.32. The simple, less powerful approximation is $X^2 = (11 - 14.611)^2/14.611 + (13 - 9.389)^2/9.389 = 2.28$. Looking at the critical values of the chi-square distribution with one degree of freedom, there is not a statistically significant difference in the survival experience of the two populations.

Another approach is to look at the difference between survival curves at a fixed time point. Using either the life table or Kaplan–Meier product limit estimate at a fixed time T_o, one can estimate the probability of survival to T_o, say, $S(T_o)$ and the standard error of $S(T_o)$, $\mathrm{SE}(S(T_o))$, as described in the sections above. Suppose that a subscript is used on S to denote estimates for different populations. To compare the survival experience of two populations with regard to

surviving to T_o, the following statistic is $N(0, 1)$, as the sample sizes become large [when the null hypothesis of $S_1(T_o) = S_2(T_o)$ is valid]:

$$Z = \frac{S_1(T_o) - S_2(T_o)}{\sqrt{\text{SE}(S_1(T_o))^2 + \text{SE}(S_2(T_o))^2}} \tag{17}$$

A one- or two-sided test may be performed, depending on the alternative hypothesis of interest. For k groups, to compare the probability of survival to time T_o, the estimated values may be compared by constructing multiple comparison confidence intervals.

16.7 ADJUSTMENT FOR CONFOUNDING FACTORS BY STRATIFICATION

In Example 16.2, in the Coronary Artery Surgery Study (Passamani et al., 1982), the degree of impairment due to chest pain pattern was related to survival. Patients with pain definitely not angina had a better survival pattern than patients with definite angina. The chest pain status is predictive of survival. These patients were studied by coronary angiography; the amount of disease in their coronary arteries as well as their left ventricular performance (the performance of the pumping part of the heart) were also evaluated. One might argue that the amount of disease is a more fundamental predictor than type of chest pain. If the pain results from coronary artery disease that affects the arteries and ventricle, the latter affects survival more fundamentally. We might ask the question: Is there additional prognostic information in the type of chest pain if one takes into account, or adjusts for, the angiographic findings?

We have used various methods of adjusting for variables. As discussed in Chapter 2, twin studies adjust for genetic variation by matching people with the same genetic pattern. Analogously, matched-pairs studies match people to be (effectively) twins in the pertinent variables; this adjusts for covariates. One step up from this is *stratified analysis*. In this case, the strata are to be quite homogeneous. People in the same strata are (to a good approximation) the same with respect to the variable or variables used to define the strata. One example of stratified analysis occurred with the Mantel–Haenszel procedure for summing 2×2 tables. The point of the stratification was to adjust for the variable or variables defining the strata. In this section we consider the same approach to the analysis of the life table or actuarial method of comparing survival curves from different groups.

16.7.1 Stratification of Life Table Analyses: Log-Rank Test

To extend the life table approach to stratification is straightforward. The first step is to perform the life table survival analysis *within each stratum*. If we do this for the four chest pain classes as discussed in Example 16.2 to adjust for angiographic data, we would use strata that depend on the angiographic findings. This is done below. Within each of the strata, we will be comparing persons with the same angiographic findings but different chest pain status. The log-rank statistic may be computed *separately* for each of the strata, giving us an observed and expected number of deaths for each group being studied. Somehow we want to combine the information across all the strata. This was done, for example, in the Mantel–Haenszel approach to 2×2 tables. We do this by summing the values for each group of the observed and expected numbers of deaths for the different strata. These observed and expected numbers are then combined into a final log-rank statistic. Note 16.3 gives the details of the computation of the statistic. Because it is based on many more subjects, the final statistic will be much more powerful then the log-rank statistic for any one stratum, *provided* that there is a consistent trend in the same direction within strata. We illustrate this by example.

Example 16.2. (*continued*) We continue with our study of chest pain groups. We would like to adjust for angiographic variables. A study of the angiographic variables showed that most of the prognostic information is contained within these variables:

1. The number of vessels diseased of the three major coronary vessels
2. The number of proximal vessels diseased (i.e., the number of diseased vessels where the disease is near the point where the blood pumps into the heart)
3. The left ventricular function, measured by a variable called LVSCORE

Various combinations of these three variables were used to define 30 different strata. Table 16.6 gives the values of the variables and the strata. Separate survival curves result in the differing strata. Figures 16.10 and 16.11 present the survival curves for two of the different strata used.

Note that the overall p-value is 0.69, a result that is not statistically significant. Thus although the survival patterns differ among chest pain categories, the differences may be explained by different amounts of underlying coronary artery disease. In other words, adjustment for the arteriographic and ventriculographic findings removed the group differences.

Note that of 30 strata, one p-value, that of stratum 25, is less than 0.05. Because of the multiple comparison problem, this is not a worry. Further, in this stratum, the definite angina cases have one observed and 0.03 expected deaths. As the log-rank statistic has an *asymptotic* chi-square distribution, the small expected number of deaths make the asymptotic distribution inappropriate in this stratum.

16.8 COX PROPORTIONAL HAZARD REGRESSION MODEL

In earlier work on the life table method, we observed various ways of dealing with factors that were related to survival. One method is to plot data for different groups, where the groups were defined by different values on the factor(s) being analyzed. When we wanted to adjust for covariates, we examined stratified life table analyses. These approaches are limited, however, by the numbers involved. If we want to divide the data into strata on 10 variables simultaneously, there will be so many strata that most strata will contain no one or at most one person. This makes comparisons impossible. One way of getting around the number problem is to have an appropriate mathematical model with covariates. In this section we consider the *Cox proportional hazards regression model*. This model is a mathematical model of survival that allows covariate values to be taken into account. Use of the model in survival analysis is quite similar to the multiple regression analysis of Chapter 11. We first turn to examination of the model itself.

16.8.1 Cox Proportional Hazard Model

Suppose that we want to examine the survival pattern of two people, one of whom initially is at higher risk than the other. A natural way to quantify the idea of risk is the hazard function discussed previously. We may think of the hazard function as the instantaneous probability of dying given that a person has survived to a particular time. The person with the higher risk will have a higher value for the hazard function than a person who has lower risk at the particular time. The Cox proportional hazard model works with covariates; the model expresses the hazard as a function of the covariate values. The major assumption of the model is that if the first person has a risk of death at the initial time point that is, say, twice as high as that of a second person, the risk of death at later times is also twice as large. We now express this mathematically.

Suppose that at the average value of all of our covariates in the population, the hazard at time t, is denoted by $h_0(t)$. Any other person whose values on the variables being considered are not equal to the mean values will have a hazard function proportional to $h_0(t)$. This proportionality constant varies from person to person depending on the values of the variables. We develop this

Table 16.6 Stratified Analysis of Survival by Chest Pain Classification

Stratification Variables				Deaths								
				Definite Angina		Probable Angina		Probably Not Angina		Definitely Not Angina		Log-Rank Statistic
Stratum Number	Number of Vessels	Number of Prox. Vessels	Left Ventricular Score	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	Obs.	Exp.	*p*-Value
1	0	0	5–11	9	10.07	42	38.33	39	43.35	9	7.25	0.74
2	0	0	12–16	0	0.79	2	1.25	1	0.87	0	0.09	0.73
3	0	0	17–30	0	0.00	0	0.00	0	0.00	0	0.00	1.00
4	1	0	5–11	19	18.88	26	23.84	5	6.71	0	0.56	0.85
5	1	0	12–16	3	3.46	5	3.25	0	1.06	0	0.23	0.52
6	1	0	17–30	1	0.31	0	0.62	0	0.08	—	—	0.43
7	1	1	5–11	14	13.36	13	13.19	2	2.00	0	0.45	0.96
8	1	1	12–16	1	2.53	3	2.05	0	0.27	1	0.15	0.15
9	1	1	17–30	4	3.49	2	2.22	0	0.30	—	—	0.93
10	2	0	5–11	17	18.54	16	14.62	2	2.29	1	0.55	0.93
11	2	0	12–16	7	6.81	2	3.90	3	1.11	0	0.18	0.20
12	2	0	17–30	5	3.49	3	3.99	1	0.98	0	0.53	0.72
13	2	1	5–11	18	15.50	10	14.91	1	1.07	2	0.24	0.11
14	2	1	12–16	9	9.06	6	4.99	0	0.80	0	0.14	0.77
15	2	1	17–30	3	3.40	3	2.38	0	0.22	—	—	0.93
16	2	2	5–11	18	17.36	13	13.56	1	0.92	0	0.16	0.99
17	2	2	12–16	19	6.70	4	5.98	0	0.32	—	—	0.59
18	2	2	17–30	3	4.67	4	2.33	—	—	—	—	0.62
19	3	0	5–11	11	11.75	9	7.44	0	0.72	0	0.10	0.76
20	3	0	12–16	8	7.49	7	6.69	—	—	0	0.83	0.83
21	3	0	17–30	4	4.31	1	0.69	—	—	—	—	0.98
22	3	1	5–11	28	23.67	15	17.78	0	1.54	—	—	0.37
23	3	1	12–16	17	16.66	6	6.34	—	—	—	—	1.00
24	3	1	17–30	9	7.32	5	6.15	0	0.53	—	—	0.72
25	3	2	5–11	36	32.08	11	17.55	2	0.34	1	0.03	0.01
26	3	2	12–16	20	16.48	6	8.45	0	1.07	—	—	0.42
27	3	2	17–30	8	9.34	7	5.17	—	—	0	0.49	0.72
28	3	3	5–11	17	22.42	19	14.36	1	0.22	—	—	0.11
29	3	3	12–16	16	14.62	6	8.24	1	0.14	—	—	0.09
30	3	3	17–30	11	12.93	4	2.07	—	—	—	—	0.56
Total				325	317.49	250	251.63	59	66.91	14	11.97	0.69[a]

[a] A dash indicates no individuals in the group in the given stratum. Obs., observed; Exp., expected; log-rank statistic = 1.47 with 3 degrees of freedom.

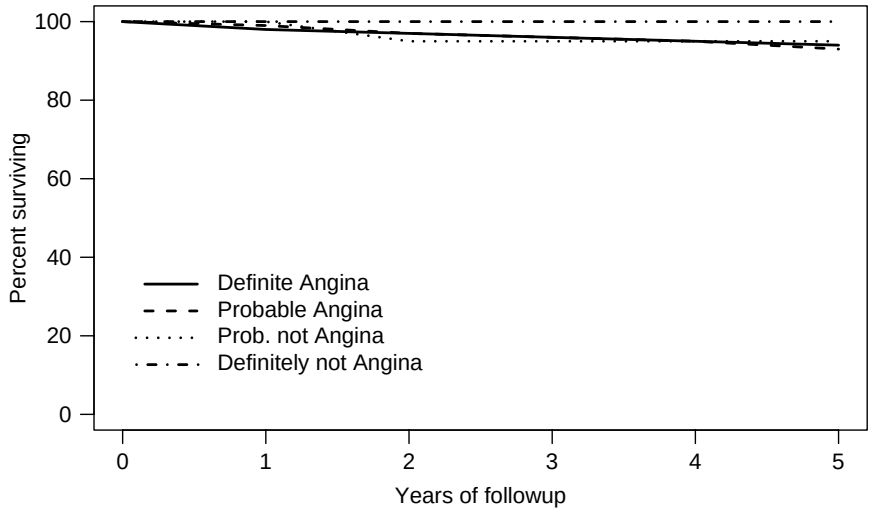

Figure 16.10 Example 16.4: survival curves for stratum 7. Cases have one proximal vessels diseased with good ventricular function (LVSCORE of 5–11).

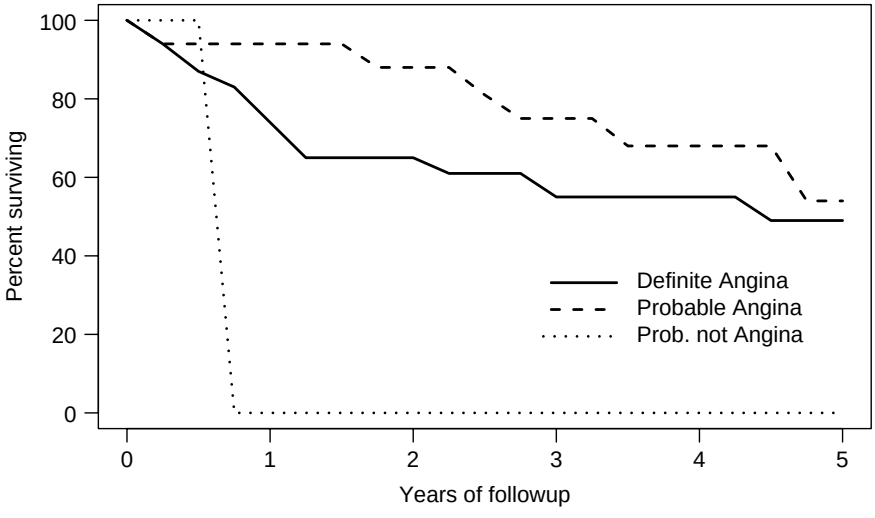

Figure 16.11 Example 16.4: survival curves for stratum 29. Cases have three proximal vessels diseased with impaired ventricular function (LVSCORE of 12–17).

algebraically. There are variables $X_1, \ldots, X_p$ to be considered. Let $\mathbf{X}$ denote the values of all the X_i, that is, $\mathbf{X} = (X_1, \ldots, X_p)$.

1. If a person has $\mathbf{X} = \overline{\mathbf{X}} = (\overline{X}_1, \ldots, \overline{X}_p)$, the hazard function is $h_0(t)$.
2. If a person has different values for $\mathbf{X}$, the hazard function is $h_0(t)C$, where C is a constant that depends on the values of $\mathbf{X}$. If we think of the hazard as depending on $\mathbf{X}$, as well as t, the hazard is

$$h_0(t)C(\mathbf{X})$$

3. For any two people with values of $\mathbf{X} = \mathbf{X}(1)$ and $\mathbf{X} = \mathbf{X}(2)$, respectively, the ratio of their two hazard functions is

$$\frac{h_0(t)C(\mathbf{X}(1))}{h_0(t)C(\mathbf{X}(2))} = \frac{C(\mathbf{X}(1))}{C(\mathbf{X}(2))} \tag{18}$$

The hazard functions are *proportional*; the ratio does not depend on t.

Let us reiterate this last point. Given two people, if one has one-half as much risk initially as a second person, then at all time points, risk is one-half that of the second person. Thus, the two hazard functions are proportional, and such models are called *proportional hazard models*.

Note that proportionality of the hazard function is an assumption that does not necessarily hold. For example, if two people were such that one is to be treated medically and the second surgically by open heart surgery, the person being treated surgically may be at higher risk initially because of the possibility of operative mortality; later, however, the risk may be the same or even less than that of the equivalent person being treated medically. In this case, if one of the covariate values indicates whether a person is treated medically or surgically, the proportional hazards model will not hold. In a given situation you need to examine the plausibility of the assumption. The model has been shown empirically to hold reasonably well for many populations over moderately long periods, say five to 10 years. Still, proportional hazards is an assumption.

As currently used, one particular parametric form has been chosen for the proportionality constant $C(\mathbf{X})$. Since it multiplies a hazard function, this constant must always be positive because the resulting hazard function is an instantaneous probability of an endpoint and consequently must be nonnegative. A convenient functional form that reasonably fits many data sets is

$$C(\mathbf{X}) = e^{\alpha+\beta_1 X_1+\cdots+\beta_p X_p}, \qquad \text{where} \quad \alpha = -\beta_1 \overline{X}_1 - \cdots - \beta_p \overline{X}_p \tag{19}$$

In this parameterization, the unknown population parameters β_i are to be estimated from a data set at hand.

With hazard $h_0(t)$, let $S_{0,\text{pop}}(t)$ be the corresponding survival curve. For a person with covariate values $\mathbf{X} = (X_1, \ldots, X_p)$, let the survival be $S(t|\mathbf{X})$. Using the previous equations, the survival curve is

$$S(t|\mathbf{X}) = (S_{0,\text{pop}}(t))^{\exp(\alpha+\beta_1 X_1+\cdots+\beta_p X_p)} \tag{20}$$

That is, the survival curve for any person is obtained by raising a standard survival curve $[S_{0,\text{pop}}(t)]$ to an appropriate power. To estimate this quantity, the following steps are performed:

1. Estimate $S_{0,\text{pop}}$ and $\alpha, \beta_1, \ldots, \beta_p$ by $S_0(t), a, b_1, \ldots, b_p$. This is done by a computer program. The estimation is too complex to do by hand.
2. Compute $Y = a + b_1 X_1 + \cdots + b_p X_p$ [where $\mathbf{X} = (X_1, \ldots, X_p)$].
3. Compute $k = e^Y$.
4. Finally, compute $S_0(t)^k$.

The estimated survival curve is the population curve (the curve for the mean covariate values) raised to a power. If the power k is equal to 1, corresponding to e^0, the underlying curve for S_0 results. If k is greater than 1, the curve lies below S_0, and if k is less than 1, the curve lies above S_0. This is presented graphically in Figure 16.12.

Note several factors about these curves:

1. The curves do not cross each other. This means that a procedure having a high initial mortality, such as a high dose of radiation in cancer therapy, but better long-term survival,

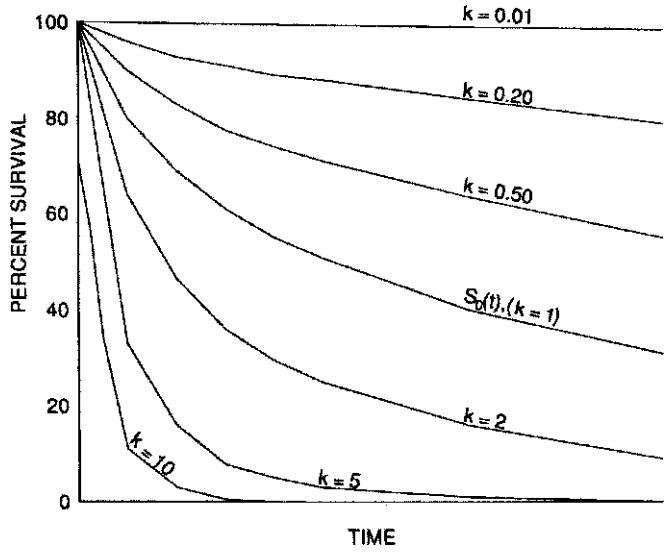

Figure 16.12 Proportional hazard survival curves as a function of $k = e^{a+b_1x_1+\cdots+b_px_p}$.

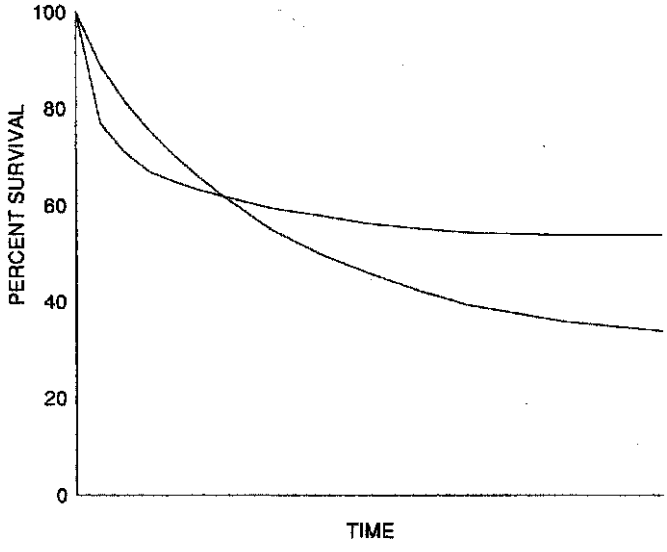

Figure 16.13 Two survival curves without proportional hazards.

as in Figure 16.13, could not be modeled by the proportional hazard model with one of the variables, say X_1, equal to 1 if the therapy were radiation and 0 if an alternative therapy were used.

2. The proportionality constant in the proportional hazard model,

$$e^{\alpha+\beta_1x_1+\cdots+\beta_px_p}$$

is parametric. We have not specified the form of the underlying survival S_0. This curve is not estimated by a parametric model but by other means.

3. Where there is a plateau in one curve, the other curve has a plateau at the same time points. The proportional hazards assumption implies that covariates do not affect the timing of plateaus or other distinctive features of the curves, only their height.

16.8.2 Example of the Cox Proportional Hazard Regression Model

The *Cox proportional hazard model* is also called the *Cox proportional regression model* or the *Cox regression model*. The reason for calling this model a regression model is that the dependent variable of interest, survival, is modeled upon or "regressed upon" the values of the covariates or independent variables. The analogies between multiple regression and the Cox regression are quite good, although there is not a one-to-one correspondence between the techniques. Computer software for Cox regression typically produces at least the quantities shown in Table 16.7.

The following example illustrates the use of the Cox proportional hazards model.

Example 16.4. The left main coronary artery is a short segment of the arteries delivering blood to the heart. Two of the three major arterial systems branch off the left main coronary artery. If this artery should close, death is almost certain. Two randomized clinical trials (Veterans' Administration Study Group, Takaro et al. [1976] and the European Coronary Surgery Study Group [1980]) reported superior survival in patients undergoing coronary artery bypass surgery. Chaitman et al. [1981] examined the observational data of the Coronary Artery Surgery Study (CASS), registry. Patients were analyzed as being in the medical group until censored at the time of surgery. They were then entered into the surgical survival experience at the day of surgery.

A Cox model using a therapy indicator variable was used to examine the effect of therapy. Eight variables were used in this model:

- *CHFSCR:* a score for congestive heart failure (CHF). The score ranged from 0 to 4; 0 indicated no CHF symptoms. A score of 4 was indicative of severe, treated CHF.
- *LMCA:* the percent of diameter narrowing of the left main coronary artery due to atherosclerotic heart disease. By selection, all cases had at least 50% narrowing of the left main coronary artery (LMCA).
- *LVSCR:* a measure of ventricular function, the pumping action of the heart. The score ranged from 5 (normal) to a potential maximum of 30 (not attained). The higher the score, the worse the ventricular function.
- *DOM:* the dominance of the heart shows whether the right coronary artery carries the usual amount of blood; there is great biological variability. Patients are classed as right or balanced dominance (DOM = 0). A left-dominant subject has a higher proportion of blood flow through the LMCA, making left main disease even more important (DOM = 1).
- *AGE:* the patient's age in years.
- *HYPTEN:* Is there a history of hypertension? HYPTEN = 1 for yes and HYPTEN = 0 for no.
- *THRPY:* This is 1 for medical therapy and 2 for surgical therapy.
- *RCA:* This variable is 1 if the right coronary artery has $\geq$ 70% stenosis and is zero otherwise.

The Cox model produces the results shown in Table 16.8. The chi-square value for CHFSCR is found by the square of β divided by the standard error. For example, $(0.2985/0.0667)^2 = 20.03$, which is the chi-square value to within the numerical accuracy. The underlying survival curve (at the mean covariate values) has probabilities 0.944 and 0.910 of one- and two-year survival, respectively. The first case in the file has values CHFSCR = 3, LMCA = 90, LVSCR = 18, DOM = 0, AGE = 49, HYPTEN = 1, THRPY = 1, and RCA = 1. What is the estimated

Table 16.7 Computer Output for Cox Regression

Output	Description	Use of Output
b_i	Estimate of the regression coefficient β_i	**1.** The b_i give an estimate of the increase in risk (the hazard function) for different values of $X_1, \ldots, X_p$. **2.** The regression coefficients allow estimation of $e^{\alpha+\beta_1 X_1+\cdots+\beta_p X_p}$ by $e^{a+b_1 x_1+\cdots+b_p x_p}$. By using this and the estimate of $S_0(t)$, we can estimate survival for any person in terms of the values of $X_1, \ldots, X_p$ for each time t.
$\text{SE}(b_i)$	Estimated standard error of b_i	**1.** The distribution of b_i is approximately $N(\beta_i, \text{SE}(b_i)^2)$ for large sample sizes. We can obtain $100(1-\alpha)\%$ confidence intervals for β_i as $(b_i - z_{1-\alpha/2}\text{SE}(b_i), b_i + z_{1-\alpha/2}\text{SE}(b_i))$. **2.** We test for statistical significance of β_i (in a model with the other X_j's) by rejecting $\beta_i = 0$ if $b_i^2/[\text{SE}(b_i)]^2 \geq \chi^2_{1,1-\alpha}$. $\chi^2_{1,1-\alpha}$ is the $1-\alpha$ percentile of the χ^2 distribution with one degree of freedom. This χ^2 test or the equivalent z test is also given by most software.
Model chi-square	Chi-square value for the entire model with p degrees of freedom	**1.** For nested models the chi-square values may be subtracted (as are the degrees of freedom) to give a chi-square test. **2.** For a single model this chi-square statistic tests for *any* relationships among the $X_1, \ldots, X_p$ and the survival experience. The null hypothesis tested is $\beta_1 = \cdots = \beta_p = 0$, which is only occasionally an interesting null hypothesis. This is analogous to testing for zero multiple correlation between survival and $(X_1, \ldots, X_p)$ in a multiple regression setting.
$S_0(t)$ and α, or $S_0(t)^\alpha$	Estimate of the survival function for a person with covariate values equal to the mean of each variable, or for a person with zero values of the covariate	**1.** With $S_0(t)$ and a, or $S_0(t)^\alpha$ and the b_i, we may plot the estimated survival experience of the population for any fixed value of the covariates. **2.** For a fixed time, say t_0, by varying the values of the covariates $\mathbf{X}$, we may present the effect of combinations of the covariate values (see Example 16.5).

probability of one- and two-year survival for this person?

$$\begin{aligned} a + b_1 X_1 + \cdots + b_n X_n &= -2.8968 + (0.2985 \times 3) + (0.0178 \times 90) \\ &\quad + (0.1126 \times 18) + (1.2331 \times 0) + (0.0423 \times 49) \\ &\quad + (-0.5428 \times 1) + (-1.0777 \times 1) \\ &\quad + (0.5285 \times 1) \\ &= 2.6622 \end{aligned}$$

Table 16.8 Results of Cox Model Fitting

Variable	Beta	Standard Error	Chi-Square	Probability
CHFSCR	0.2985	0.0667	20.01	0.0000
LMCA	0.0178	0.0049	13.53	0.0002
LVSCR	0.1126	0.0182	38.41	0.0000
DOM	1.2331	0.3564	11.97	0.0006
AGE	0.0423	0.0098	18.75	0.0000
HYPTEN	−0.5428	0.1547	12.31	0.0005
THRPY	−1.0777	0.1668	41.77	0.0000
RCA	0.5285	0.2923	3.27	0.0706
Constant	−2.8968			

$$\begin{aligned}\text{estimated probability of one-year survival} &= 0.944^{e^{2.6622}}\\ &= 0.944^{14.328}\\ &= 0.438\\ \text{estimated probability of two-year survival} &= 0.910^{14.328}\\ &= 0.259\end{aligned}$$

The estimated probability of survival under medical therapy is 44% for one year and 26% for two years. This bad prognosis is due largely to heart failure (CHFSCR) and very poor ventricular function (LVSCR).

16.8.3 Interpretation of the Regression Coefficients β_i

In the multiple regression setting, the regression coefficients may be interpreted as the average difference in the response variables between cases where the predictor variable differs by one unit, with everything else the same. In this section we look at the interpretation of the β_i for the Cox proportional hazard model. Recall that the hazard function is proportional to the probability of failure in a short time interval. Suppose that we have two patients whose covariate values are the same on all the p regression variables for the Cox model with the exception of the ith variable. If we take the ratio of the hazard functions for the two people at some time t, we have the ratio of the probability of an event in a short interval after time t. The ratio of these two probabilities is the relative risk of an event during this time period. This is also called the *instantaneous relative risk*. For the Cox proportional hazards model, we find that

$$\begin{aligned}\text{instantaneous relative risk (RR)} &= \frac{h_0(t)e^{\alpha+\beta_1 X_1+\cdots+\beta_i X_i^{(1)}+\cdots+\beta_p X_p}}{h_0(t)e^{\alpha+\beta_1 X_1+\cdots+\beta_i X_i^{(2)}+\cdots+\beta_p X_p}}\\ &= e^{\beta_i(X_i^{(1)}-X_i^{(2)})}\end{aligned} \tag{21}$$

An equivalent formulation is to take the logarithm of the instantaneous relative risk (RR). The logarithm is given by

$$\ln(\text{RR}) = \beta_i(X_i^{(1)} - X_i^{(2)}) \tag{22}$$

In words, the regression coefficients β of the Cox proportional hazard model are equal to the logarithm of the relative risk if the variable X is increased by one unit.

16.8.4 Evaluating the Proportional Hazards Assumption

One graphical assessment of the proportional hazards assumption for binary (or categorical) variables plots the cumulative hazard in each group on a logarithmic scale. Under the proportional hazards assumption, the resulting curves should be parallel, that is, separated by a constant vertical difference (the reason is given in Section 16.8.3). Although popular, these *log-log plots* are not particularly useful. Judging whether two curves (as opposed to straight lines) are parallel is difficult, and the problem is compounded by the fact that the uncertainty in the estimated log hazard varies substantially along the curves.

An better approach to judging proportional hazards involves smoothed plots of the *scaled Schoenfeld residuals*, proposed by Therneau and Grambsch [2000]. These plots, available in Stata and S, estimate how a coefficient β_i varies over time. In addition to an easier visual interpretation, the Schoenfeld residual methods provide a formal test of the proportional hazards assumption and are valid for continuous as well as categorical variables.

The technical details of the Schoenfeld residual methods are complex, but there is a simple underlying heuristic. Suppose that the hazard ratio for, say, hypertension is greater than unity. Hypertensive persons will be overrepresented among the deaths in any given period. If, in addition, the hazard ratio increases with time, overrepresentation of hypertensives among the deaths will increase with time. By calculating the proportion of hypertensives among the deaths and the population at risk in each short interval of time, we should be able to detect the increasing hazard ratio.

If there is substantial nonproportionality of hazards, it may be desirable to stratify the model (see Section 16.10.2) on the variable in question, or to define a time-dependent variable as in Example 16.7 in Section 16.10.2.

Example 16.5. Primary biliary cirrhosis is a rare, autoimmune disease of the liver. Until the advent of liver transplantation, it was untreatable and eventually fatal. The Mayo Clinic performed a randomized trial of one proposed treatment, D–penicillamine, in 312 patients. The treatment was not effective, but the data from the trial have been used to develop a widely used prognostic model for survival of this disease. The data for this model have been made available on the Web by Terry Therneau of the Mayo Clinic and are linked in the Web appendix.

The Mayo model includes five covariates:

- *BILI:* logarithm of serum bilirubin concentration. Bilirubin is excreted in the bile and accumulates in liver disease.
- *PROTIME:* logarithm of the prothrombin time, a measure of blood clotting. Prothrombin time is increased when the liver fails to produce certain clotting factors.
- *ALBUMIN:* logarithm of serum albumin concentration. The liver produces albumin to prevent blood plasma from leaking out of capillaries.
- *EDTRT:* edema (fluid retention), coded as 0 for no edema, $\frac{1}{2}$ for untreated edema or edema resolved by treatment, 1 for edema present despite treatment.
- *AGE:* in tens of years. Age affects the risk for almost any cause of death.

Figure 16.14 shows scatter plots for two of these covariates against survival time. The censored observations are indicated by open triangles, the deaths by filled circles. There is clearly a relationship with both variables. It is also interesting to note that according to Fleming and Harrington [1991, Chap. 5], the outlying value of 18 for prothrombin time was a data-entry error; it should be 11.

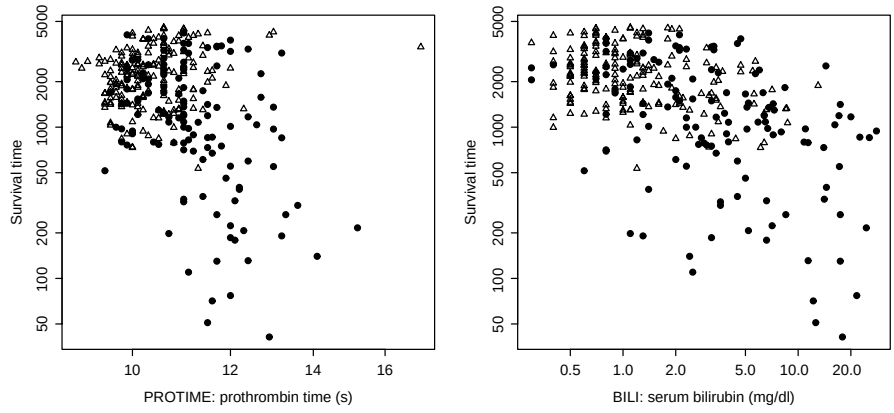

Figure 16.14 Scatter plots of survival time vs. PROTIME and BILI in the Mayo PBC data. Triangles indicate censored times.

The Mayo model has the following coefficients:

Variable	b	SE(b)
BILI	0.88	0.10
EDTRT	0.79	0.30
ALBUMIN	−3.06	0.72
PROTIME	3.01	1.02
AGE	0.33	0.08

The survival function for someone with no edema, albumin of 3.5 mg/dL, prothrombin time of 10 seconds, bilirubin of 1.75 mg/dL, and age 50 is:

t (yr)	$S(t)$ (%)	t (yr)	$S(t)$ (%)
1	98	6	80
2	97	7	74
3	92	8	68
4	88	9	61
5	84	10	51

Figure 16.15 shows a scaled Schoenfeld residual plot for PROTIME. The smooth curve that estimates $\beta(t)$ shows that the logarithm of the hazard ratio for elevated prothrombin time is very high initially and then decreases to near zero over the first three to four years. That is, a patient with high prothrombin time is at greatly increased risk of death, but a patient who had a high prothrombin time four years ago and is still alive is not at particularly high risk. The p-value for nonproportionality for PROTIME is 0.055, so there is moderately strong evidence that the pattern we see in Figure 16.15 is real.

16.8.5 Use of the Cox Model as a Method of Adjustment

In Section 16.7 we considered stratified life table analyses to adjust for confounding factors or covariates. The Cox model may be used for the same purpose. As in the multiple linear regression model, there are two ways in which we may adjust. One is to consider a variable whose effect we want to study in relationship to survival. Suppose that we want adjust for

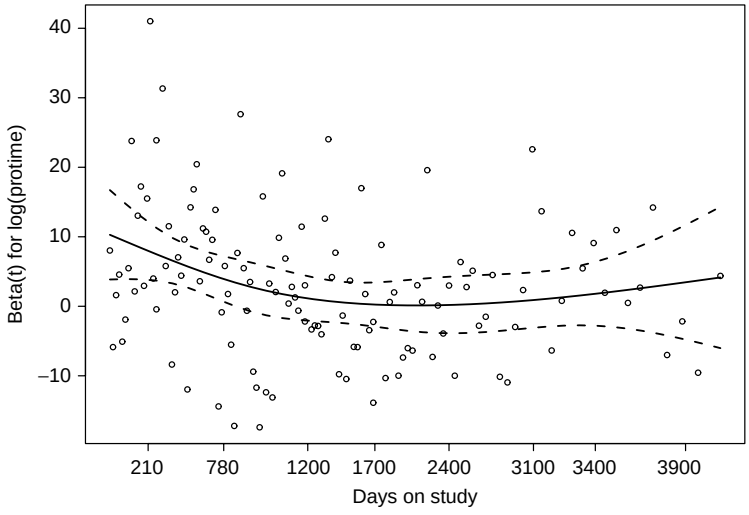

Figure 16.15 Assessing proportional hazards for PROTIME with scaled Schoenfeld residuals.

variables $X_1, \ldots, X_k$. We run the Cox proportional hazards regression model with the variable of interest and the adjustment covariates in the model. The statistical significance of the variable of interest may be tested by taking its estimated regression coefficient, dividing by its standard error and using a normal probability critical value. An equivalent approach, similar to nested hypotheses in the multiple linear regression model, is to run the Cox proportional hazards model with only the adjusting covariates. This will result in a chi-square statistic for the entire model. A second Cox proportional hazards model may be run with the variable of interest in the model in addition to the adjustment covariates. This will result in a second chi-square statistic for the model. The chi-square statistic for the second model minus the chi-square statistic for the first model will have approximately a chi-square distribution with one degree of freedom if the variable of interest has no effect on the survival after adjustment for the covariates $X_1, \ldots, X_p$.

Example 16.5. (*continued*) Of the 418 patients in the Mayo Clinic PBC data set, 312 agreed to participate in the randomized trial and 106 refused. As the data from the randomized trial were used to develop a predictive model for survival, it is important to know whether the randomized and nonrandomized patients differ in important ways.

A simple comparison of survival times in these two groups does not answer quite the right question. Suppose that patients agreeing to be randomized had longer survival times but also had lower levels of bilirubin that were sufficient to explain their improved survival. This discrepancy in survival times does not invalidate the model. Conversely, if the two groups had very similar survival times despite a difference in average bilirubin levels, this would be evidence against the model.

We can estimate the adjusted difference between the randomized and nonrandomized patients by fitting a Cox model that has the five Mayo model predictors and an additional variable indicating which group the patient is in. The estimated hazard ratio for nonrandomized patients is 0.97, with a 95% confidence interval from 0.66 to 1.41. We would not typically report coefficients and confidence intervals for the other adjusting covariates; their associations with survival are not of direct interest in this analysis.

There is no evidence of any difference in survival between randomized and nonrandomized patients in this study, but the confidence intervals are quite wide, so these differences have not been ruled out.

Other examples of estimating adjusted contrasts using the Cox model appear in Section 16.10.

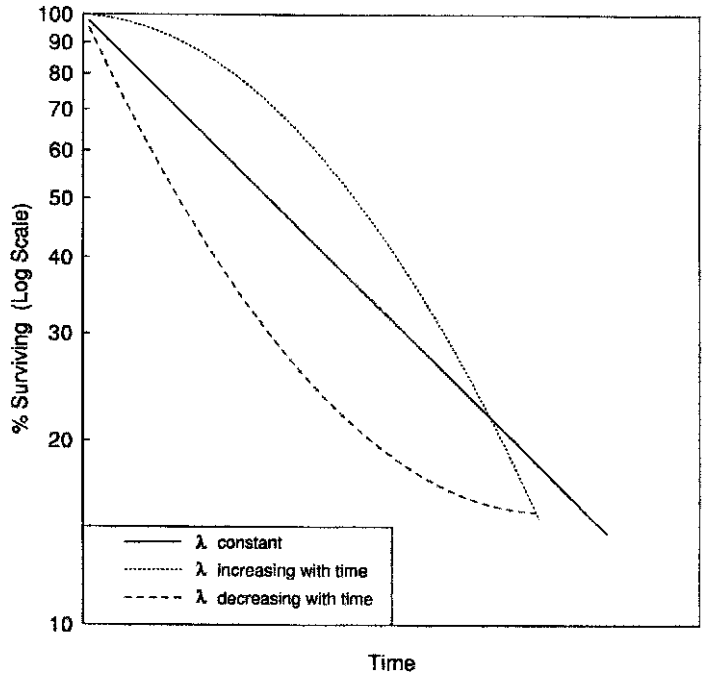

Figure 16.16 Log plot for exponential survival.

16.9 PARAMETRIC MODELS

16.9.1 Exponential Model; Rates

Suppose that at each instant of time, the instantaneous probability of death is the same. That is, suppose that the hazard rate or force of mortality is constant. Although in human populations this is not a useful assumption over a wide time interval, it may be a valid assumption over a five- or 10-year interval, say.

If the constant hazard rate is λ, the survival curve is $S(t) = e^{-\lambda t}$. From this expression the term *exponential survival* arises. The expected length of survival is $1/\lambda$. If the exponential situation holds, the parameter λ is estimated by the number of events divided by total exposure time. The methods and interpretation of rates are then appropriate. If $S(t)$ is exponential, $\log S(t) = -\lambda t$ is a straight line with slope $-\lambda$. Plotting an estimate of $S(t)$ on a logarithmic scale is one way of visually examining the appropriateness of assuming an exponential model. Figure 16.16 shows some of the patterns that one might observe.

To illustrate this we return to the Mayo primary biliary cirrhosis data set but now consider an analysis of time until loss to follow-up, that is, a survival analysis where the event is loss to follow-up. To avoid confusing patients lost to follow-up with those alive and under observation at the end of the study, we look at just the first eight years of the study. From the plot one sees that the data do *not* look exponential (Figure 16.17). Rather, it appears that the hazard of dropping out is initially very low and increases progressively.

16.9.2 Two Other Parametric Models for Survival Analysis

There are a variety of parametric models for survival distributions. In this section, two are mentioned. For details of the distributions and parameter estimates, the reader is referred to

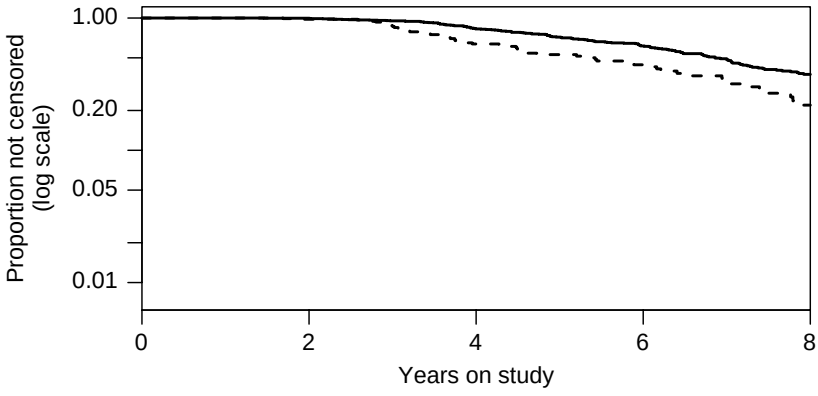

Figure 16.17 Loss to follow-up of 312 randomized and 106 nonrandomized patients with primary biliary cirrhosis.

texts by Mann et al. [1974] and Gross and Clark [1975]. These books also present a variety of models not touched on here.

The two-parameter *Weibull distribution* has a survival curve of the form

$$S(t) = e^{-\alpha t^{\beta}} \qquad \text{for } t > 0\,(\alpha > 0, \beta > 0) \tag{23}$$

If $\beta = 1$, the Weibull distribution is the exponential model with constant hazard rate. The hazard rate decreases with time if $\beta < 1$ and increases with time if $\beta > 1$. Often, if the time of survival is measured from diagnosis of a disease, a Weibull with $\beta > 1$ will reasonably model the situation. Estimates are made by computer.

Another distribution, the *lognormal distribution*, assumes that the logarithm of the survival time is normally distributed. If there is no censoring of data, one may work with the logarithm of the survival times and use methods appropriate for the normal distribution.

Regression versions of the exponential, lognormal, Weibull, and other parametric survival models are also available in many statistical packages. The exponential and Weibull models are special cases of the Cox proportional hazards model and have little advantage over the Cox model. The lognormal model is not related to the Cox model.

16.10 EXTENSIONS

16.10.1 Cox Model with Time-Dependent Covariates

If two groups are defined by some baseline measurement, such as smokers and nonsmokers, their hazard ratio would be expected to change over time simply because some of the smokers will stop smoking and lower their risk of death. For this reason it may be desirable to base the hazard ratio at time t on the most recent available values of covariates rather than on the values at the start of follow-up. The Cox model is then most naturally written in terms of the hazard rather than the survival:

$$\text{hazard at time } t = h_0(t) \exp\left[\alpha + \beta_1 X_1(t) + \beta_2 X_2(t) + \cdots + \beta_p X_p(t)\right]$$

and we write $X_1(t)$ for the value of X_1 at time t.

The hazard ratio between two subjects with covariates $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ is then

$$\begin{aligned}\frac{h(t;\mathbf{X}^{(1)})}{h(t;\mathbf{X}^{(2)})} &= \frac{h_0(t)\exp[\alpha+\beta_1 X_1(t)^{(1)}+\beta_2 X_2(t)^{(1)}+\cdots+\beta_p X_p(t)^{(1)}]}{h_0(t)\exp[\alpha+\beta_1 X_1(t)^{(2)}+\beta_2 X_2(t)^{(2)}+\cdots+\beta_p X_p(t)^{(2)}]} \\ &= \frac{\exp[\beta_1 X_1(t)^{(1)}+\beta_2 X_2(t)^{(1)}+\cdots+\beta_p X_p(t)^{(1)}]}{\exp[\beta_1 X_1(t)^{(1)}+\beta_2 X_2(t)^{(1)}+\cdots+\beta_p X_p(t)^{(1)}]} \\ &= \exp\left\{\beta_1\left[X_1(t)^{(1)}-(X_1(t)^{(2)}\right]+\beta_2\left[X_2(t)^{(1)}-(X_2(t)^{(2)}\right]\right. \\ &\quad \left.+\cdots+\beta_p\left[X_p(t)^{(1)}-(X_p(t)^{(2)}\right]\right\}\end{aligned}$$

In the constant-covariate situation, the proportional hazards assumption means that the hazard ratio does not change over time; in the time-dependent situation, it means that the hazard ratio changes only due to changes in the covariates over time.

Example 16.6. An example of time-dependent covariates comes from a study by Holt and colleagues [2002] that examined the effects of court protective orders on abuse of women by their domestic partners. In this study the time-dependent covariates were the presence (1) or absence (0) of temporary restraining orders and permanent restraining orders. At the start of the study, after the first police report of abuse, both variables would be zero. Most of the women in the study (2366) never obtained a protection order, so the variable remained at zero. Of those who obtained a two-week temporary order (325), about half (185) later obtained a permanent order. The time-dependent Cox model compares the risk of abuse in women who do and do not have each type of protective order *at the same time after their initial incident*. Cox models thus reduce the potential for confounding by time since the initial incident: Since permanent protective orders tend to happen later in time, when risks are already lower, they might appear protective even if they actually had no effect.

Temporary restraining orders were associated with an increase in the hazard of psychological abuse (hazard ratio 4.9, 95% confidence interval 2.6 to 8.6) and no change in the hazard of physical abuse (hazard ratio 1.6, 95% CI 0.6 to 4.4). Permanent restraining orders appeared to reduce physical abuse (hazard ratio 0.2, 95% CI 0.1 to 0.8) and have no effect on psychological abuse (hazard ratio 0.9, 95% CI 0.5 to 1.7).

In some settings it may be more appropriate to use values of covariates for some short or long period in the past rather than the instantaneously updated values. These time-dependent variables reflect the history of exposure rather than just the current status.

Example 16.7. Heckbert et al. [2001] studied how the risk of a recurrent heart attack changed over time in women who had already had one heart attack and were taking hormone replacement therapy (HRT). Estrogen, the active ingredient of HRT, is known to improve cholesterol levels but also to increase blood clotting, and so might have positive or negative effects on heart disease. A recent randomized trial, HERS [Hulley et al., 1998], suggested that the balance of risk and benefit might change over time.

The researchers hypothesized that having recently started hormone replacement therapy would increase the risk of heart attack, but that long-term therapy might not increase the risk. They defined three time-dependent exposure variables:

- *STARTING:* 1 for women taking HRT who started less than 60 days ago, 0 otherwise
- *RECENT:* 1 for women taking HRT who started between 60 and 365 days previously, 0 otherwise
- *LONGTERM:* 1 for women taking HRT who started more than a year ago, 0 otherwise

The hypothesis was that the coefficients for STARTING would be positive (increased risk), but that coefficients for RECENT and LONGTERM would be lower, and possibly negative. They found that the hazard ratio e^b for STARTING was 2.16, with a 95% confidence interval, 0.94 to 4.95, not quite excluding 1. The hazard ratio for LONGTERM was 0.76, a with 95% confidence interval 0.42 to 1.36.

Time-dependent covariates are not always appropriate. In particular, they do not result in useful predictive models: In order to estimate the chance of surviving for the next five years, it is necessary to have covariate values for the next five years to plug into the model.

Even when time-dependent models are appropriate, they involve significantly more complex computation, however, good facilities for time-dependent Cox models are now available in many major statistics packages. Computational details vary between packages, and between versions of the same package, but the basic approach is to break each person's data into many short time intervals on which their covariates are constant. These time intervals are treated as if they came from separate people, which is valid as long as each person can have only one event.

Time-dependent covariates are discussed in many of the recent textbooks on survival analysis, including Therneau and Grambsch [2000], Klein and Moeschberger [1997], and Kleinbaum [1996] and in older references such as Kalbfleisch and Prentice [1980] and Breslow and Day [1987].

16.10.2 Stratification in the Cox Model

The Cox model, which assumes that hazards are proportional over time, can be extended to a stratified model in which hazards need only be proportional within the same stratum and can differ arbitrarily between strata. Stratification can be useful when a small number of important variables do not satisfy the proportional hazards assumption. In addition to the usual difficulties that occur with stratifying on too many variables, the stratified model also suffers from the fact that it is not possible to test the effects of the stratifying variables.

For example, Lumley et al. [2002] constructed a predictive model for the risk of stroke in elderly people. The rates of stroke were not proportional between men and women, so a model stratified by gender was used. Instead of a single underlying survival curve $S_o(t)$, the model has curves $S_m(t)$ for men and $S_w(t)$ for women. The hazard ratio for other covariates, such as diabetes or smoking, is assumed to be constant over time within each stratum. The hazard ratio may be constrained to be the same for women and men or allowed to differ. As Table 16.9 shows, the stroke prediction model used a common hazard ratio for diabetes in men and women, but the hazard ratio for history of heart disease was allowed to differ between men and women. A Java applet showing this model is linked from the Web appendix.

Table 16.9 Stratified Cox Model for Risk of Stroke

	Mean		Coefficient	
	2495 Men	3393 Women	Men	Women
Left ventricular hypertrophy by ECG (%)	5.1	4.9	0.501	
Diabetes (%)	14.9	12.5	0.521	
Elevated fasting glucose (%)	19.0	14.4	0.347	
Creatinine >1.25 mg/dL (%)	39.6	8.1	0.141	
Time to walk 15 ft (s)	5.5	6.0	0.099	
Systolic blood pressure (mmHg)	143	144	172/10	
History of heart disease (%)	26.5	16.1	0.445	0.073
Atrial fibrillation by ECG (%)	3.5	2.1	0.4097	1.346
Age (yr)	73	73	0.382/10	0.613/10

16.10.3 Left Truncation

In the examples discussed so far, the survival time has been measured from the beginning of the study, so that all subjects are under observation from time 0 under they die or are censored. There are situations where this is not feasible. Consider a study of occupational exposure to a potential carcinogen, where workers at a factory are interviewed about their past exposure and other risk factors such as cancer, and then followed up.

It would be desirable to set time zero to be when each worker was first employed at the factory rather than the date when the study was performed. This would more accurately approximate the ideal study that recruited everyone as they entered employment and followed them for the rest of their lives. There is a serious complication, however. Workers who died before the study started will not be included, making the sample biased. This phenomenon is called *left truncation*. Truncation is not quite the same as censoring, although both involve incomplete information. With censoring, we have information on only part of a person's life. With truncation, we have no information on some people and complete information on others.

The solution to left truncation is similar to the solution to right censoring. If we break time up into short intervals, each person contributes information about the probability of surviving through an interval given that one is alive at the start of the interval. These probabilities can be multiplied to give an overall survival probability. Most statistical software will allow you to specify an *entry* time as well as a survival or censoring time, and will fit Cox regression models to data specified in this way.

In the occupational exposure example, consider a worker who started at the factory in 1955, who entered the study in 1985, and who died in 1995. We want to take time to be 0 in 1955, so the *entry* time is $1985 - 1955$, or 30 years, and the survival time is $1995 - 1955$, or 40 years. Another worker might have started at the factory in 1975, been recruited in 1985, and still be alive at the end of the study in 2000. This would give an *entry* time of $1985 - 1975$, or 10 years, and a censoring time of $2000 - 1975$, or 25 years.

Breslow and Day [1987] discuss an example of this sort in some detail, comparing the effects of placing time zero at different events in analyzing the cancer risks of workers at a nickel refinery.

16.10.4 Other References Dealing with Survival Analysis and Heart Transplant Data

The first heart transplant data has been used extensively as an illustration in the development of survival techniques. Further references are Mantel and Byar [1974], Turnbull et al. [1974], and Crowley and Hu [1977].

NOTES

16.1 Recurrent Events

Some events can occur more than once for the same person. Although it is usually possible to study just the time until the first event, it may be useful to incorporate subsequent events to increase the information available. The hazard formulation of survival analysis extends naturally to recurrent events. The hazard (now often called the *intensity*) is still defined in terms of the probability of having an event in a small interval of time, conditional on being alive and under observation. The difference is that now a person can still be alive and under observation after an event occurs. Although computation for recurrent event models is fairly straightforward, there are a number of important methodologic issues that need to be considered. In particular, there is no really satisfactory way to handle recurrent events and deaths in the same analysis. Volume 16, No. 18 of *Statistics in Medicine* (April 30, 1997) has a number of papers discussing these issues. The Web appendix to this chapter includes some examples of analyses of recurrent infections in children with chronic granulomatous disease, a genetic immune deficiency.

16.2 More on the Hazard Rate and Proportional Hazards

Many of the concepts presented in this chapter are analogs of continuous quantities that are best defined in terms of calculus. If the survival function is $S(t)$, its probability density function is

$$f(t) = -\frac{dS(t)}{dt}$$

The hazard rate is then

$$h(t) = \frac{f(t)}{S(t)}$$

From this it follows that the survival is found from the hazard rate by the equation

$$S(t) = e^{-\int_0^t h(x)\,dx}$$

The quantity

$$H(t) = \int_0^t h(x)\ dx = -\log S(t)$$

is called the *cumulative hazard*. Under the proportional hazards assumption, the cumulative hazards H_1 and H_2 for two groups of cases are related by

$$H_1(t) = \lambda \times H_2(t)$$

so

$$\log H_1(t) = \log \lambda + \log H_2(t)$$

16.3 Log-Rank Statistic and Log-Rank Statistic for Stratified Data

We present the statistic using some matrix ideas. The notation is that of Section 16.6 on the log-rank test. For the ith group at the mth time of a death (or deaths), there were d_{im} deaths and l_{im} persons at risk. Suppose that we have k groups and M times of death. For $i, j = 1, \ldots, k$, let

$$V_{ij} = \begin{cases} \sum_{m=1}^{M} \dfrac{l_{im}(T_m - l_{im})D_m(T_m - D_m)}{T_m^2(T_m - 1)}, & i = j \\ \sum_{m=1}^{M} \dfrac{-l_{im}l_{jm}D_m(T_m - D_m)}{T_m^2(T_m - 1)}, & i \neq j \end{cases}$$

Define the $(k-1) \times (k-1)$ matrix V by

$$V = \begin{pmatrix} V_{11} & V_{12} & \cdots & V_{1,k-1} \\ V_{21} & & & \vdots \\ \vdots & & & \vdots \\ V_{k-1,1} & \cdots & \cdots & V_{k-1,k-1} \end{pmatrix}$$

Define vectors of observed and expected number of deaths in groups $1, 2, \ldots, k-1$ by

$$\mathbf{O} = \begin{pmatrix} O_1 \\ \vdots \\ O_{k-1} \end{pmatrix}, \qquad \mathbf{E} = \begin{pmatrix} E_1 \\ \vdots \\ E_{k-1} \end{pmatrix}$$

The log-rank statistic is

$$(\mathbf{O} - \mathbf{E})' V^{-1} (\mathbf{O} - \mathbf{E})$$

where $'$ denotes a transpose and -1 a matrix inverse. If there are $s = 1, \ldots, S$ strata, for each stratum we have **O, E,** and V. Let these values be indexed by s to denote the strata. The log-rank statistic is

$$\left[\sum_{s=1}^{S} (\mathbf{O}_s - \mathbf{E}_s)\right]' \left(\sum_{s=1}^{S} V_s\right)^{-1} \left[\sum_{s=1}^{S} (\mathbf{O}_s - \mathbf{E}_s)\right]$$

16.4 Estimating the Probability Density Function in Life Table Methods

The density function in the interval from $x(i)$ to $x(i+1)$ for the life table is estimated by

$$f_i = \frac{P_i - P_{i+1}}{x(i+1) - x(i)}$$

The standard error of f_i is estimated by

$$\frac{p_i q_i}{\sqrt{x(i+1) - x(i)}} \left(\sum_{j=1}^{i-1} \frac{q_j}{l'_j p_j} + \frac{p_i}{l'_i q_i}\right)^{1/2}$$

16.5 Other Confidence Intervals for the Survival Function

Direct use of Greenwood's formula to construct confidence intervals in small samples can lead to confidence intervals that cross 0% or 100% survival. Even when this does not occur, the confidence intervals do not perform very well. Better confidence intervals are obtained by multiplying, rather than adding, the same quantity above and below the estimated survival function. That is, the confidence interval is given by

$$\left[\hat{S}(t) \times \exp\left(-z_{\alpha/2} \frac{\text{SE}(\hat{S}(t))}{\hat{S}(t)}\right), \hat{S}(t) \times \exp\left(z_{\alpha/2} \frac{\text{SE}(\hat{S}(t))}{\hat{S}(t)}\right)\right]$$

Bie et al. [1987] studied this interval and a more complicated one based on transforming $S(t)$ to $\arcsin\{\exp[-S(t)/2]\}$ and found that both performed well even with only 25 observations, half of which were censored.

16.6 Group Expected Survival

The baseline survival curve $S_0(t)$ estimates the survival probability at time t for a person whose covariates equal the average of the population. This is not the same as the survival curve expected for the population $S(t)$ as estimated by the Kaplan–Meier method. The population curve $S(t)$

decreases faster than $S_0(t)$ initially, as those with worse-than-average covariates die and then flattens out relative to $S_0(t)$, as the remaining sample has better-than-average covariates. The difference between $S(t)$ and $S_0(t)$ is more pronounced when covariate effects are strong and when there is little censoring.

The relationship between the curves is that the population curve is the average of all the predicted individual survival curves:

$$S(t) = \sum_i S_0(t)^{e^{\alpha+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_p X_p}}$$

This relationship can be used to predict the population curve for a new population and compare it to the expected population, an extension of the direct standardization of rates in Chapter 15. For example, the predictions of a Cox model can be validated in a new population by dividing the new population into groups and comparing the expected $S(t)$ for each group with the observed survival curve calculated by the Kaplan–Meier method.

Example 16.5. (*continued*) Figure 16.18 compares the expected and observed survival rates for the 106 nonrandomized patients from the Mayo Clinic PBC data. These patients were divided into three equal groups based on the risk predicted by the Mayo model. The Kaplan–Meier survival curve and the group expected survival curve were calculated for each of the three groups. The relatively smooth lines are the expected survival; the stepped lines are the Kaplan–Meier estimates. There is no suggestion that the expected and observed curves differ importantly.

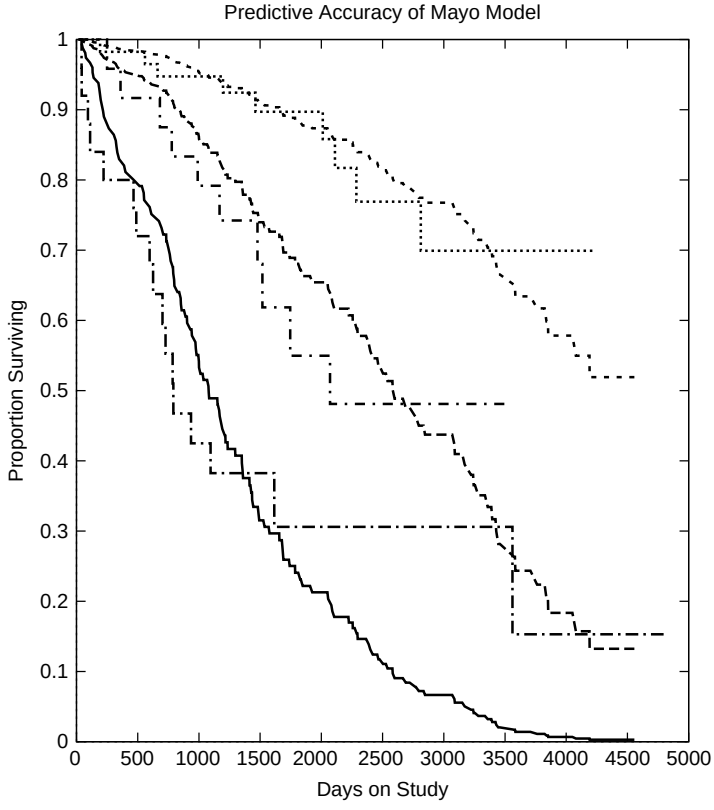

Figure 16.18 Expected and observed survival curves for three groups of nonrandomized patients.

For a stratified life table analysis, the same calculation of expected survival can be done more easily. In this context it is called the method of *direct adjustment*. Suppose that we want to compare survival in treatment groups $j = 1, 2$ and we have strata $i = 1, 2, \ldots, m$. We calculate the survival curve for each treatment group in each stratum $S_{ij}(t)$ and then add up over strata

$$S_j(t) = \sum_{i=1}^{m} S_{ij}(t) r_i$$

where r_i is the proportion of subjects in stratum i.

16.7 Competing Risks

In certain situations one is only interested in certain causes of death that may be linked to the disease in question. For example, in a study of heart disease a death in a plane crash might be considered an unreasonable endpoint to attribute to the disease. It is tempting to censor people who die of genuinely unrelated causes. This cannot be true *noninformative censoring*, as someone who dies in a plane crash certainly has a reduced (zero) risk of heart disease in the future. On the other hand, there seems to be no way that these deaths would bias the remaining sample. It turns out that conclusions from Cox regression in this case are basically valid but that estimated survival curves need to be rethought. Such endpoints are called *competing risks*.

In a more complicated version of the problem, there is often interest in the effects of a treatment on more than one type of event. Lowering blood pressure reduces the risk of death from stroke, heart attack, cardiac arrest, and congestive heart failure, but different drugs may affect these events differently. Inference for these *dependent* competing risks is much more difficult and is complicated further by the fact that it is theoretically impossible to determine whether competing risks are dependent or independent. When all the events are rare, as in primary prevention of cardiovascular disease, ignoring the competing-risks problem may be a satisfactory practical approach. With more common events, this is not possible.

In some cases it is appropriate to treat deaths from other causes as indicating indefinitely long "survival" for the cause of interest. For example, consider a study of time to stroke in elderly people (e.g., Section 16.10.2). If a subject dies from breast cancer at 3.5 years follow-up, her chance of ever having a stroke is known exactly: She never will. This can be represented by censoring her observation time not at the time of death but at a time after the end of the study. The resulting survival curve will estimate the proportion of people who have not had strokes, which will not decrease to zero as follow-up time increases. In other cases this approach is undesirable because decreases in stroke risk and increases in other risks have the same impact—in a clinical trial of stroke prevention one would not want to declare the treatment successful just because it made people die of other causes.

Kalbfleisch and Prentice [2003], Gross and Clark [1975], and Prentice et al. [1978] discuss such issues. Pepe and Mori [1993] discuss alternatives to estimating the cause-specific survival function. Misuse of the cause-specific survival function has been an important issue in radiation oncology and is discussed by Gelman et al. [1990]. The impossibility of testing for dependent competing risks was shown by Tsiatis [1978]. The proof is highly technical but the result should be intuitively plausible: No data are available after censoring, so there should be no way to tell if survival is the same as for noncensored people.

A related issue is multivariate failure time, where events of different types can be observed for the same person. These could be ordered events, such as cancer recurrence and death; multiple versions of the same event, such as time to vision impairment in left and right eyes; or separate events, such as time to marriage and time to having children. Therneau and Grambsch [2000] discuss multivariate failure times, as does Lin [1994]. Somewhat surprisingly, this is a more tractable problem than competing risks.

16.8 Counting Process Notation

Many modern books on survival analysis and most recent statistical papers on the subject use a different mathematical notation from ours, the *counting process notation*. We have described each person's data by a covariate vector $\mathbf{X}_i$, an observation time T_i, and a censoring indicator Δ_i. The counting process notation replaces the time and censoring indicator with two functions of time: $N_i(t)$, which counts the number of times the person has been observed to "die" by time t, and $Y_i(t)$, which is 1 when the person is under observation and 0 otherwise. The covariate vector is usually called $\mathbf{Z}_i(t)$ rather than $\mathbf{X}_i$.

For ordinary survival data this means $N_i(t) = 0$ and $Y_i(t) = 1$ for $t < T_i$, $N_i(t) = \Delta_i$ and $Y_i(t) = 1$ for $t = T_i$, and $N_i(t) = \Delta_i$ and $Y_i(t) = 0$ for $t > T_i$. The notation $dN_i(t)$ means the jump in N_i at time t. This is zero except at the time of a death, when it is 1.

As a final complication, integral notation is used to indicate sums over a time point. For example, the notation $\int Z_i(t)\, dNi(t)$ means the sum of $Z_i(t) \times dN_i(t)$ over all time points. As $dNi(t) = 0$ except at the time of death, this is 0 if the person is censored and is $Z_i(T_i)$ if the person dies at time T_i.

This apparently cumbersome notation was introduced initially for purely mathematical reasons. It becomes more obviously useful when handling recurrent events [when $N_i(t)$ counts the number of events that have occurred], or left-truncation, when $Y_i(t) = 0$ before entry into the study to indicate that a death at that time would not have been observed. Klein and Moeschberger [1997] provide a reasonably accessible treatment of survival analysis using counting process notation.

PROBLEMS

The first four problems deal with the life table or actuarial method of estimating the survival curve. In each case, fill in the question marks from the other numbers given in the table.

16.1 Example 16.2 deals with chest pain in groups in the Coronary Artery Surgery Study; all times are in days. The life table for the individuals with chest pain thought probably not to be angina is given in Table 16.10.

16.2 From Example 16.2 for patients with chest pain thought definitely to be angina the life table is as given in Table 16.11.

16.3 Patients from Example 16.4 on a beta-blocking drug are used here and those not on a beta-blocking drug in Problem 16.4. The life table for those using such drugs at enrollment is given in Table 16.12.

16.4 Those not using beta-blocking drugs have the survival experience shown in Table 16.13.

16.5 Take the Stanford heart transplant data of Example 16.3. Place the data in a life table analysis using 50-day intervals. Plot the data over the interval from zero to 300 days. (Do not compute the Greenwood standard errors.)

16.6 For Problem 16.1, compute the hazard function (in probability of dying/day) for intervals:

- **(a)** 546–637
- **(b)** 1092–1183
- **(c)** 1456–1547

16.7 For the data of Problem 16.2, compute the hazard rate for the patients:

- **(a)** 0–91
- **(b)** 91–182
- **(c)** 819–910

Table 16.10 Life Table for Patients with Chest Pain Probably Not Angina

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	2404	2404.0	2	0	0.0008	0.9992	?
91.0–181.9	2402	?	2	0	0.0008	0.9983	?
182.0–272.9	2400	2400.0	?	0	0.0021	0.9963	0.001
273.0–363.9	2395	2395.0	6	0	?	0.9938	0.002
364.0–454.9	?	2388.0	4	2	0.0017	0.9921	0.002
455.0–545.9	2383	2383.0	3	0	0.0013	?	0.002
546.0–636.9	2380	2380.0	7	0	0.0029	0.9879	0.002
637.0–727.9	2373	?	12	300	?	?	0.003
728.0–818.9	2061	2051.5	?	19	0.0015	0.9812	0.003
819.0–909.9	?	2039.0	1	0	0.0005	0.9807	0.003
910.0–1000.9	2038	2037.0	2	?	0.0010	0.9797	0.003
1001.0–1091.9	2034	?	3	517	0.0017	0.9781	0.003
1092.0–1182.9	1514	1494.0	3	40	0.0020	0.9761	0.003
1183.0–1273.9	1471	1471.0	4	0	?	0.9734	0.004
1274.0–1364.9	1467	1466.5	1	1	0.0007	0.9728	0.004
1365.0–1455.9	?	1144.0	1	642	0.0009	0.9719	0.004
1456.0–1546.9	822	777.5	1	?	0.0013	0.9707	0.004
1547.0–1637.9	732	732.0	1	0	0.0014	?	0.004
1638.0–1728.9	731	730.0	2	2	0.0027	0.9667	0.004
1729.0–1819.9	727	449.0	1	?	0.0022	0.9645	0.005

Table 16.11 Life Table for Patients with Definite Angina

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	426	426.0	2	?	0.0047	0.9953	0.003
91.0–181.9	?	424.0	2	0	0.0047	0.9906	?
182.0–272.9	422	?	3	0	?	?	0.006
273.0–363.9	419	419.0	0	0	0.0000	0.9836	0.006
364.0–454.9	419	419.0	1	0	0.0024	0.9812	0.007
455.0–545.9	418	417.5	?	1	0.0024	0.9789	0.007
546.0–636.9	416	416.0	1	0	0.0024	0.9765	0.007
637.0–727.9	415	382.0	0	?	0.0000	0.9765	0.007
728.0–818.9	349	343.0	0	11	0.0000	0.9765	0.007
819.0–909.9	338	338.0	1	0	0.0030	0.9736	0.008
910.0–1000.9	337	336.5	0	1	0.0000	0.9736	0.008
1001.0–1091.9	336	?	1	97	?	?	0.009
1092.0–1182.9	238	232.5	0	11	0.0000	0.9702	0.009
1183.0–1273.9	227	?	1	1	0.0044	0.9660	0.010
1274.0–1364.9	?	224.5	1	1	0.0045	0.9617	0.010
1365.0–1455.9	?	170.0	0	106	0.0000	0.9617	0.010
1456.0–1446.9	117	114.0	?	6	0.0000	0.9617	0.010
1547.0–1637.9	?	?	0	1	0.0000	0.9617	0.010
1638.0–1728.9	110	109.5	0	1	0.0000	0.9617	0.010
1729.0–1819.9	109	65.5	0	87	0.0000	0.9617	0.010

Table 16.12 Life Table for Patients Taking a β-Blocker

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	4942	4942.0	?	0	0.0097	0.9903	0.001
91.0–181.9	4894	4894.0	33	0	0.0067	0.9836	0.002
182.0–272.9	4861	4861.0	?	?	0.0058	0.9779	?
273.0–363.9	4833	4832.5	28	1	0.0058	0.9723	0.002
364.0–454.9	4804	4804.0	17	0	0.0035	?	0.002
455.0–545.9	4787	4786.5	29	1	?	?	0.003
546.0–636.9	4757	4757.0	22	0	0.0046	0.9585	0.003
637.0–727.9	4735	4376.0	25	718	0.0057	0.9530	0.003
728.0–818.9	?	?	?	62	0.0043	0.9489	0.003
819.0–909.9	3913	3912.0	23	2	?	0.9434	0.003
910.0–1000.9	3888	3884.5	19	7	0.0049	0.9388	0.004
1001.0–1091.9	?	?	?	1191	0.0040	0.9350	0.004
1092.0–1182.9	2658	2624.5	14	67	0.0053	0.9300	0.004
1183.0–1273.9	2577	2576.5	11	1	0.0043	0.9261	0.004
1274.0–1364.9	2565	2561.0	15	8	?	0.9206	0.004
1365.0–1455.9	2542	1849.5	12	1385	0.0065	0.9147	0.005
1456.0–1446.9	1145	1075.0	5	?	0.0047	?	0.005
1547.0–1637.9	1000	999.0	4	2	0.0040	0.9068	0.005
1638.0–1728.9	994	989.0	4	10	0.0040	0.9031	0.006
1729.0–1819.9	980	580.0	5	800	0.0086	0.8953	0.006

Table 16.13 Life Table for Patients Not Taking a β-Blocker

$t(i)$	Enter	At Risk	Dead	Withdraw Alive	Proportion Dead	Cumulative Survival	SE
0.0–90.9	6453	?	45	0	?	?	?
91.0–181.9	6408	?	28	0	?	?	?
182.0–272.9	6380	?	42	0	?	?	?
273.0–363.9	6338	?	25	2	?	?	?
364.0–454.9	6311	6310.0	24	2	0.0038	0.9746	0.002
455.0–545.9	6285	6285.0	32	0	0.0051	0.9696	0.002
546.0–636.9	6253	6253.0	?	0	0.0048	0.9650	0.002
637.0–727.9	6223	5889.0	23	668	0.0039	0.9612	0.002
728.0–818.9	?	?	23	40	0.0042	0.9572	0.003
819.0–909.9	?	5467.0	17	4	?	0.9542	0.003
910.0–1000.9	5448	5444.5	23	7	0.0042	0.9502	0.003
1001.0–1091.9	5418	4787.4	25	1261	0.0052	0.9452	0.003
1092.0–1182.9	4132	4082.0	?	100	0.0054	0.9401	0.003
1183.0–1273.9	4010	4010.0	23	0	0.0057	0.9347	0.003
1274.0–1364.9	3987	3981.0	18	?	0.0020	0.9329	0.003
1365.0–1455.9	3967	3100.0	13	1734	0.0042	0.9289	0.003
1456.0–1446.9	2220	2104.0	13	?	0.0062	0.9232	0.004
1547.0–1637.9	1975	1974.0	?	2	0.0020	0.9213	0.004
1638.0–1728.9	1969	1961.5	11	15	0.0056	0.9162	0.004
1729.0–1819.9	1943	1212.0	17	7	0.0058	0.9109	0.005

16.8 Data used by Pike [1966] are quoted in Kalbfleisch and Prentice [2003]. Two groups of rats with different pretreatment regimes were exposed to the carcinogen DBMA. The time to mortality from vaginal cancer in the two groups was: (* indicates a censored observation):

- *Group 1:* 143, 164, 188, 188, 190, 192, 206, 209, 213, 216, 216*, 220, 227, 230, 234, 244*, 246, 265, 304
- *Group 2:* 142, 156, 163, 198, 204*, 205, 232, 232, 233, 233, 233, 233, 239, 240, 261, 280, 280, 296, 296, 323, 344*

(a) Compute and graph the two product limit curves of the groups.

(b) Compute the expected number of deaths in each group and the value of the approximation $[\sum(O - E)^2/E]$ to the log-rank test. Are the survival times different in the two groups at the 5% significance level?

(c) How close is the approximate log-rank statistic to the exact value reported by your favorite statistics software?

16.9 The data of Problems 16.3 and 16.4, where stratified into the 30 strata discussed in the text, give the results shown in Table 16.14.

(a) What are the observed and expected numbers in the two groups? (Why do you have to add only three columns?)

(b) Two strata (12 and 17) are significant with $p = 0.02$. If the true survival patterns (in the conceptual underlying populations) are the same, does this surprise you?

(c) What is $\sum(O-E)^2/E$? How does this compare to the more complicated log-rank statistic which can be shown to be 6.510?

16.10 The paper by Chaitman et al. [1981] studied patients with left main coronary artery disease, as discussed in Example 16.4. Separate Cox survival runs were performed for the medical and surgical groups. The data are presented in Table 16.15. The survival, at the mean covariate values, for one, two, and three years are given by $S_0(1)$, $S_0(2)$, and $S_0(3)$, respectively. The zero–one variables are 0 for no and 1 for yes. Consider five patients with the variable values given in Table 16.16.

(a) What is the estimate of the two-year medical survival for patients 1, 2, and 3?

(b) What is the estimate of the three-year surgical survival for patients 4 and 5?

(c) What are the estimated one-year medical and one-year surgical survival rates for patient 1? For patient 3?

(d) What is the logarithm of the instantaneous relative risk for two individuals treated medically who differ by 20 years, but otherwise have the same values for the variables? What is the instantaneous relative risk?

(e) What is the instantaneous relative risk due to diabetes (yes vs. no) for surgical cases?

***(f)** What is the standard error for the LV score coefficient for the surgical group? For the age coefficient for the medical group? Form an approximate 95% confidence interval for the age coefficient in the medical group.

16.11 Alderman et al. [1983] studied the medical and surgical survival of patients with poor left ventricular function; that is, they studied patients whose hearts pumped poorly. Their model (in one analysis) included the following variables:

Table 16.14 Drug Use Data for Problem 16.9

	Drug Use		No Drug Use		
Stratum	Obs.	Exp.	Obs.	Exp.	p-Value
1	45	43.30	71	72.70	0.74
2	2	2.23	4	3.77	0.84
3	0	0.20	1	0.80	0.54
4	27	28.54	37	35.46	0.69
5	6	4.84	5	6.16	0.48
6	2	0.76	1	2.24	0.08
7	20	16.87	20	23.13	0.31
8	4	5.25	10	8.75	0.49
9	3	3.17	5	4.83	0.90
10	18	16.55	21	22.45	0.63
11	5	6.68	9	7.32	0.35
12	8	4.58	1	4.42	0.02
13	21	16.04	13	17.96	0.08
14	6	8.95	16	13.05	0.19
15	2	2.63	5	4.37	0.61
16	16	16.82	20	19.81	0.78
17	5	9.86	15	10.14	0.02
18	4	4.40	5	4.60	0.78
19	7	11.48	16	11.52	0.06
20	10	8.98	8	9.02	0.62
21	4	2.89	2	3.11	0.34
22	21	19.67	24	25.33	0.68
23	13	14.59	20	18.41	0.56
24	5	6.86	11	9.14	0.32
25	35	29.64	21	26.36	0.14
26	18	14.82	13	16.18	0.24
27	7	8.89	8	6.11	0.29
28	22	17.08	18	22.92	0.10
29	11	11.24	15	14.76	0.92
30	8	9.11	8	6.89	0.52

- *Impairment:* impairment due to congestive heart failure (CHF); 0 = never had CHF; 1 = had CHF but have no impairment; 2 = mild CHF impairment; 3 = moderate CHF impairment; and 4 = severe CHF impairment
- *Age:* in years
- *LMCA:* percent of diameter narrowing of the left main coronary artery
- *EF:* ejection fraction, the percent of the blood in the pumping chamber (left ventricle) of the heart pumped out during heartbeat
- *Digitalis:* Does the patient use digitalis? 1 = yes, 2 = no
- *Therapy:* 1 = medical; 2 = surgical
- *Vessel:* number (0 to 3) of vessels diseased with 70% or more stenosis

The β values and their standard errors are given in Table 16.17.

(a) Fill in the chi-square value column where missing.

(b) For which variables is $p < 0.10$? 0.05? 0.01? 0.001?

Table 16.15 Significant Independent Predictors of Mortality in Patients with Greater Than 50% Stenosis of the Left Main Coronary Artery

	Medical Group		Surgical Group	
Variable	X^{2a}	β_i	X^{2a}	β_i
LV score (5–30)	19.12	0.1231	18.54	0.1176
CHF score (0–4)	9.39	0.2815	8.16	0.2964
Age	14.42	0.0526	6.98	0.0402
% LMCA stenosis (50–100)	19.81	0.0293	—	—
Hypertension (0–1)	9.41	0.7067	5.74	0.5455
Left dominance (0–1)	—	—	10.23	1.0101
Smoking (1 = never, 2 = ever, 3 = present)	7.26	0.4389	—	—
MI status (0 = none, 1 = single, 2 = multiple)	4.41	−0.2842	—	—
Diabetes (0–1)	—	—	4.67	0.5934
Total chi-square	90.97	—	67.11	—
Degrees of freedom	7	—	6	—
p	<0.0001	—	<0.0001	—
Constant c		−7.2956		−3.7807
Estimated survival				
$S_0(1)$		0.90		0.97
$S_0(2)$		0.83		0.95
$S_0(3)$		0.76		0.93

[a] Adjusted chi-square (X^2) statistics were computed with all variables considered together. Chi-square >6.63 corresponds to $p < 0.01$, and chi-square >10.83, to $p < 0.001$. β, beta coefficient; CHF, congestive heart failure; LMCA, left main coronary artery; LV, left ventricular; MI, myocardial infarction. Dashes indicate a variable not in the particular model.

Table 16.16 Variable Data for Problem 16.10

	Patient Number				
Variable	1	2	3	4	5
LV score	13	5	7	8	12
CHF score	2	0	1	0	3
Age	71	62	42	55	46
Percent LMCA stenosis	75	90	50	70	95
Hypertension	No	Yes	Yes	No	No
Left dominance	No	No	No	Yes	No
Smoking	Ever	Present	Ever	Ever	Present
MI status	Multiple	None	Single	None	Single
Diabetes	No	No	No	Yes	No

Table 16.17 Data for Problem 16.11

Variable	Beta	Standard Error	Chi-Square
Impairment	0.2677	0.0505	?
Age	0.0430	0.0084	26.02
LMCA	0.0090	0.0024	?
EF	−0.0362	0.0098	?
Digitalis	−0.3802	0.1625	?
Therapy	−0.3418	0.1458	5.49
Vessel	0.2081	0.1012	4.23
Constant	−1.2873		

Table 16.18 Variable Data for Problem 16.11

	Patient Number		
Variable	1	2	3
Impairment	Severe	Mild	Moderate
Age	64	51	59
LMCA	50%	0%	0%
EF	15	32	23
Digitalis	Yes	Yes	Yes
Therapy	Medical	Surgical	Medical
Vessel	3	2	3

(c) What is the instantaneous relative risk of 70% LMCA compared to 0% LMCA?

(d) Consider three patients with the covariate values given in Table 16.18.

At the mean values of the data, the one- and two-year survival were 88.0% and 80.16%, respectively. Find the probability of one- and two-year survival for these three patients.

(e) With this model: (**i**) Can surgery be better for one person and medical treatment for another? Why? What does this say about unthinking application of the model? (**ii**) Under surgical therapy, can the curve cross over the estimated medical survival for some patients? For heavy surgical mortality, would a proportional hazard model always seem appropriate?

16.12 The Clark et al. [1971] heart transplant data were collected as follows. People with failing hearts waited for a donor heart to become available; this usually occurred within 90 days. However, some patients died before a donor heart became available. Figure 16.19 plots the survival curves of (1) those not transplanted (indicated by circles) and (2) the transplant patients from time of surgery (indicated by the triangles).

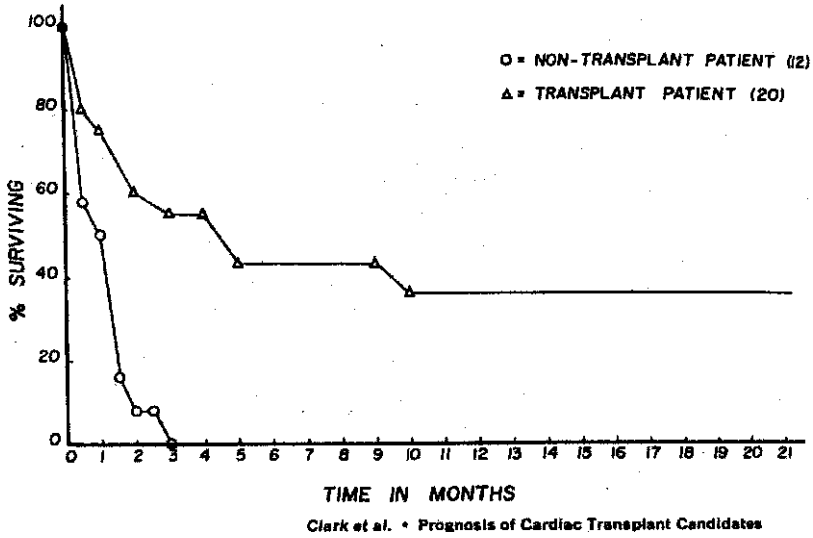

Figure 16.19 Survival calculated by the life table method. Survival for transplanted patients is calculated from the time of operation; survival of nontransplanted patients is calculated from the time of selection for transplantation.

(a) Is the survival of the nontransplanted patients a reasonable estimate of the nonoperative survival of candidates for heart transplant? Why or why not?

(b) Would you be willing to conclude from the figure (assuming a statistically significant result) that 1960s heart transplant surgery prolonged life? Why or why not?

(c) Consider a Cox model fitted with transplantation as a time-dependent covariate:

$$h_i(t) = h_0(t)e^{\exp(\alpha+\beta\times\text{TRANSPLANT}(t))}$$

The estimate of β is 0.13, with a 95% confidence interval $(-0.46, 0.72)$. (Verify this if you have access to suitable software.) What is the interpretation of this estimate? What would you conclude about whether 1960s-style heart transplant surgery prolongs life?

(d) A later, expanded version of the Stanford heart transplant data includes the age of the participant and the year of the transplant (from 1967 to 1973). Adding these variables gives the following coefficients:

Variable	β	SE(β)	p-value
Transplant	−0.030	0.318	0.92
Age	0.027	0.014	0.06
Year	−0.179	0.070	0.01

What would you conclude from these results, and why?

16.13 Simes et al. [2002] analyzed results from the LIPID trial that compared the cholesterol-lowering drug pravastatin to placebo in preventing coronary heart disease events. The outcome defined by the trial was time until fatal coronary heart disease or nonfatal myocardial infarction.

(a) The authors report that Cox model with one variable coded 1 for pravastatin and 0 for placebo gives a reduction in the risk of 24% (95% confidence interval, 15 to 32%). What is the hazard ratio? What is the coefficient for the treatment variable?

(b) A second model had three variables: treatment, HDL (good) cholesterol level after treatment, and total cholesterol level after treatment. The estimated risk reduction for the treatment variable in this model is 9% (95% confidence interval, −7 to 22%). What is the interpretation of the coefficient for treatment in this model?

16.14 In an elderly cohort, the death rate from heart disease was approximately constant at 2% per year, and from other causes was approximately constant at 3% per year.

(a) Suppose that a researcher computed a survival curve for time to heart disease death, treating deaths from other causes as censored. As described in Section 16.9.1, the survival function would be approximately $S(t) = e^{-0.02t}$. Compute this function at 1, 2, 3, . . . ,10 years.

(b) Another researcher computed a survival curve for time to non-heart-disease death, censoring deaths from heart disease. What would the survival function be? Compute it at 1, 2, 3, . . . ,10 years.

(c) What is the true survival function for deaths from all causes? Compare it to the two cause-specific functions and discuss why they appear inconsistent.

REFERENCES

Alderman, E. L., Fisher, L. D., Litwin, P., Kaiser, G. C., Myers, W. O., Maynard, C., Levine, F., and Schloss, M. [1983]. Results of coronary artery surgery in patients with poor left ventricular function (CASS). *Circulation*, **68**: 785–789. Used with permission from the American Heart Society.

Bie, O., Borgan, Ø., and Liestøl, K. [1987]. Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties. *Scandinavian Journal of Statistics*, **14**: 221–223.

Breslow, N. E., and Day, N. E. [1987]. *Statistical Methods in Cancer Research*, Vol. II. International Agency for Research on Cancer, Lyon, France.

Chaitman, B. R., Fisher, L. D., Bourassa, M. G., Davis, K., Rogers, W. J., Maynard, C., Tyras, D. H., Berger, R. L., Judkins, M. P., Ringqvist, I., Mock, M. B., and Killip, T. [1981]. Effect of coronary bypass surgery on survival patterns in subsets of patients with left main coronary disease. *American Journal of Cardiology*, **48**: 765–777.

Clark, D. A., Stinson, E. B., Grieppe, R. B., Schroeder, J. S., Shumway, N. E., and Harrison, D. B. [1971]. Cardiac transplantation in man: VI. Prognosis of patients selected for cardiac transplantation. *Annals of Internal Medicine*, **75**: 15–21.

Crowley, J., and Hu, M. [1977]. Covariance analysis of heart transplant survival data. *Journal of the American Statistical Association*, **72**: 27–36.

European Coronary Surgery Study Group [1980]. Prospective randomized study of coronary artery bypass surgery in stable angina pectoris: second interim report. *Lancet*, Sept. 6, **2**: 491–495.

Fleming, T. R., and Harrington, D. [1991]. *Counting Processes and Survival Analysis.* Wiley, New York.

Gehan, E. A. [1969]. Estimating survival functions from the life table. *Journal of Chronic Diseases*, **21**: 629–644. Copyright © 1969 by Pergamon Press, Inc. Used with permission.

Gelman, R., Gelber, R., Henderson I. C., Coleman, C. N., and Harris, J. R. [1990]. Improved methodology for analyzing local and distant recurrence. *Journal of Clinical Oncology*, **8**(3): 548–555.

Greenwood, M. [1926]. *Reports on Public Health and Medical Subjects*, No. 33, App. I, The errors of sampling of the survivorship tables. H. M. Stationary Office, London.

Gross, A. J. and Clark, V. A. [1975]. *Survival Distributions: Reliability Applications in the Biomedical Sciences*. Wiley, New York.

Heckbert, S. R., Kaplan, R. C., Weiss, N. S., Psaty, B. M., Lin, D., Furberg, C. D., Starr, J. S., Anderson, G. D., and LaCroix, A. Z. [2001]. Risk of recurrent coronary events in relation to use and recent initiation of postmenopausal hormone therapy. *Archives of Internal Medicine*, **161**(14): 1709–1713.

Holt, V. L., Kernic, M. A., Lumley, T., Wolf, M. E., and Rivara, F. P. [2002]. Civil protection orders and risk of subsequent police-reported violence. *Journal of the American Medical Association*, **288**(5): 589–594

Hulley, S., Grady, D., Bush, T., Furberg, C., Herrington, D., Riggs, B., and Vittinghoff, E. [1998]. Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. *Journal of the American Medical Association*, **280**(7): 605–613.

Kalbfleisch, J. D., and Prentice, R. L. [2003]. *The Statistical Analysis of Failure Time Data*. 2nd edition Wiley, New York.

Kaplan, E. L., and Meier, P. [1958]. Nonparametric estimation for incomplete observations. *Journal of the American Statistical Association*, **53**: 457–481.

Klein, J. P., and Moeschberger, M. L. [1997]. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer-Verlag, New York.

Kleinbaum, D. G. [1996]. *Survival Analysis: A Self-Learning Text*. Springer-Verlag, New York.

Lin, D. Y. [1994]. Cox regression analysis of multivariate failure time data: the marginal approach. *Statistics in Medicine*, **13**: 2233–2247.

Lumley, T., Kronmal, D., Cushman, M., Monolio, T. A. and Goldstein, S. [2002]. Predicting stroke in the elderly: validation and web-based application. *Journal of Clinical Epidemiology*, **55**: 129–136.

Mann, N. R., Schafer, R. C. and Singpurwalla, N. D. [1974]. *Methods for Statistical Analysis of Reliability and Life Data*. Wiley, New York.

Mantel, N., and Byar, D. [1974]. Evaluation of response time 32 data involving transient states: an illustration using heart transplant data. *Journal of the American Statistical Association*, **69**: 81–86.

Messmer, B. J., Nora, J. J., Leachman, R. E., and Cooley, D. A. [1969]. Survival times after cardiac allographs. *Lancet*, May 10, **1**: 954–956.

Miller, R. G. [1981]. *Survival Analysis*. Wiley, New York.

Parker, R. L., Dry, T. J., Willius, F. A., and Gage, R. P. [1946]. Life expectancy in angina pectoris. *Journal of the American Medical Association*, **131**: 95–100.

Passamani, E. R., Fisher, L. D., Davis, K. B., Russel, R. O., Oberman, A., Rogers, W. J., Kennedy, J. W., Alderman, E., and Cohen, L. [1982]. The relationship of symptoms to severity, location and extent of coronary artery disease and mortality. Unpublished study.

Pepe, M. S., and Mori, M. [1993]. Kaplan–Meier, marginal, or conditional probability curves in summarizing competing risks failure time data. *Statistics in Medicine*, **12:** 737–751.

Pike, M. C. [1966]. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, **26**: 579–581.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. L. [1978]. The analysis of failure times in the presence of competing risks. *Biometrics*, **34**: 541–554.

Simes, R. S., Masschner, I. C., Hunt, D., Colquhoun, D., Sullivan, D., Stewart, R. A. H., Hague, W., Kelch, A., Thompson, P., White, H., Shaw, V., and Torkin, A. [2002]. Relationship between lipid levels and clinical outcomes in the long-term intervention with Pravastatin in ischemic disease (LIPID) trial: to what extent is the reduction in coronary events with Pravastatin explained by on-study lipid levels? *Circulation*, **105**: 1162–1169.

Takaro, T., Hultgren, H. N., Lipton, M. J., Detre, K. M., and participants in the study group [1976]. The Veteran's Administration cooperative randomized study of surgery for coronary arterial occlusive disease: II. Subgroup with significant left main lesions. *Circulation Supplement 3*, **54**: III-107 to III-117.

Therneau, T. M., and Grambsch, P. [2000]. *Modelling Survival Data: Extending the Cox Model*. Springer-Verlag, New York.

Tsiatis, A. A. [1978]. An example of non-identifiability in competing risks. *Scandinavian Actuarial Journal*, 235–239.

Turnbull, B., Brown, B., and Hu, M. [1974]. Survivorship analysis of heart transplant data. *Journal of the American Statistical Association*, **69**: 74–80.

U.S. Department of Health, Education, and Welfare [1976]. *Vital Statistics of the United States, 1974*, Vol. II, Sec. 5, Life tables. U.S. Government Printing Office, Washington, DC.