TABLE 1.10
Induction times (in years) for AIDS in adults and children

Infection Time	*Adult Induction Time*	*Child Induction Time*
0.00	5	
0.25	6.75	
0.75	5, 5, 7.25	
1.00	4.25, 5.75, 6.25, 6.5	5.5
1.25	4, 4.25, 4.75, 5.75	
1.50	2.75, 3.75, 5, 5.5, 6.5	2.25
1.75	2.75, 3, 5.25, 5.25	
2.00	2.25, 3, 4, 4.5, 4.75, 5, 5.25, 5.25, 5.5, 5.5, 6	
2.25	3, 5.5	3
2.50	2.25, 2.25, 2.25, 2.25, 2.5, 2.75, 3, 3.25, 3.25, 4, 4, 4	
2.75	1.25, 1.5, 2.5, 3, 3, 3.25, 3.75, 4.5, 4.5, 5, 5, 5.25, 5.25, 5.25, 5.25, 5.25	1
3.00	2, 3.25, 3.5, 3.75, 4, 4, 4.25, 4.25, 4.25, 4.75, 4.75, 4.75, 5	1.75
3.25	1.25, 1.75, 2, 2, 2.75, 3, 3, 3.5, 3.5, 4.25, 4.5	
3.50	1.25, 2.25, 2.25, 2.5, 2.75, 2.75, 3, 3.25, 3.5, 3.5, 4, 4, 4.25, 4.5, 4.5	0.75
3.75	1.25, 1.75, 1.75, 2, 2.75, 3, 3, 3, 4, 4.25, 4.25	0.75, 1, 2.75, 3, 3.5, 4.25
4.00	1, 1.5, 1.5, 2, 2.25, 2.75, 3.5, 3.75, 3.75, 4	1
4.25	1.25, 1.5, 1.5, 2, 2, 2, 2.25, 2.5, 2.5, 2.5, 3, 3.5, 3.5	1.75
4.50	1, 1.5, 1.5, 1.5, 1.75, 2.25, 2.25, 2.5, 2.5, 2.5, 2.5, 2.75, 2.75, 2.75, 2.75, 3, 3, 3, 3.25, 3.25	3.25
4.75	1, 1.5, 1.5, 1.5, 1.75, 1.75, 2, 2.25, 2.75, 3, 3, 3.25, 3.25, 3.25, 3.25, 3.25, 3.25	1, 2.25
5.00	0.5, 1.5, 1.5, 1.75, 2, 2.25, 2.25, 2.25, 2.5, 2.5, 3, 3, 3	0.5, 0.75, 1.5, 2.5
5.25	0.25, 0.25, 0.75, 0.75, 0.75, 1, 1, 1.25, 1.25, 1.5, 1.5, 1.5, 1.5, 2.25, 2.25, 2.5, 2.5, 2.75	0.25, 1, 1.5
5.50	1, 1, 1, , 1.25, 1.25, 1.75, 2, 2.25, 2.25, 2.5	.5, 1.5, 2.5
5.75	0.25, 0.75, 1, 1.5, 1.5, 1.5, 2, 2, 2.25	1.75
600	0.5, 0.75, 0.75, 0.75, 1, 1, 1, 1.25, 1.25, 1.5, 1.5, 1.75, 1.75, 1.75, 2	0.5, 1.25
6.25	0.75, 1, 1.25, 1.75, 1.75	0.5, 1.25
6.50	0.25, 0.25, 0.75, 1, 1.25, 1.5	0.75
6.75	0.75, 0.75, 0.75, 1, 1.25, 1.25, 1.25	0.5, 0.75
7.00	0.75	0.75
7.25	0.25	0.25

2 Basic Quantities and Models

2.1 Introduction

In this chapter we consider the basic parameters used in modeling survival data. We shall define these quantities and show how they are interrelated in sections 2.2–2.4. In section 2.5 some common parametric models are discussed. The important application of regression to survival analysis is covered in section 2.6, where both parametric and semiparametric models are presented. Models for competing risks are discussed in section 2.7.

Let X be the time until some specified event. This event may be death, the appearance of a tumor, the development of some disease, recurrence of a disease, equipment breakdown, cessation of breast feeding, and so forth. Furthermore, the event may be a good event, such as remission after some treatment, conception, cessation of smoking, and so forth. More precisely, in this chapter, X is a nonnegative random variable from a homogeneous population. Four functions characterize the distribution of X, namely, the *survival function*, which is the probability of an individual surviving to time x; the *hazard rate (function)*, sometimes termed *risk function*, which is the chance an individual of age x experiences the event in the next instant in time; the *probability density (or probability mass) function*, which is the unconditional probability of the event's occurring at time x; and the *mean residual life* at time x, which is the mean time to the event of interest, given the event has not occurred at x. If we know any one of these four

functions, then the other three can be uniquely determined. In practice, these four functions, along with another useful quantity, the *cumulative hazard function*, are used to illustrate different aspects of the distribution of X. In the competing risk context, the *cause-specific hazard rate*, which is the rate at which subjects who have yet to experience any of the competing risks are experiencing the ith competing cause of failure, is often used. This quantity and other competing risk quantities are discussed in detail in section 2.7. In Chapters 4–6, we shall see how these functions are estimated and how inferences are drawn about the survival (or failure) distribution.

2.2 The Survival Function

The basic quantity employed to describe time-to-event phenomena is the survival function, the probability of an individual surviving beyond time x (experiencing the event after time x). It is defined as

$$S(x) = Pr(X > x). \tag{2.2.1}$$

In the context of equipment or manufactured item failures, $S(x)$ is referred to as the reliability function. If X is a continuous random variable, then, $S(x)$ is a continuous, strictly decreasing function.

When X is a continuous random variable, the survival function is the complement of the cumulative distribution function, that is, $S(x) = 1 - F(x)$, where $F(x) = Pr(X \leq x)$. Also, the survival function is the integral of the probability density function, $f(x)$, that is,

$$S(x) = Pr(X > x) = \int_x^\infty f(t)\,dt. \tag{2.2.2}$$

Thus,

$$f(x) = -\frac{dS(x)}{dx}.$$

Note that $f(x)\,dx$ may be thought of as the "approximate" probability that the event will occur at time x and that $f(x)$ is a nonnegative function with the area under $f(x)$ being equal to one.

EXAMPLE 2.1 The survival function for the Weibull distribution, discussed in more detail in section 2.5, is $S(x) = \exp(-\lambda x^\alpha)$, $\lambda > 0$, $\alpha > 0$. The exponential distribution is a special case of the Weibull distribution when $\alpha = 1$. Survival curves with a common median of 6.93 are exhibited in Figure 2.1 for $\lambda = 0.26328$, $\alpha = 0.5$; $\lambda = 0.1$, $\alpha = 1$; and $\lambda = 0.00208$, $\alpha = 3$.

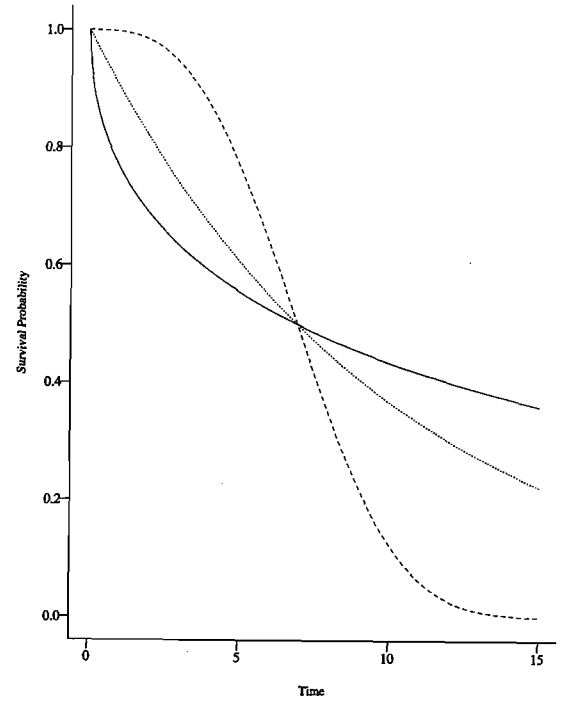

Figure 2.1 *Weibull Survival functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (———); $\alpha = 1.0$, $\lambda = 0.1$ (·····); $\alpha = 3.0$, $\lambda = 0.00208$ (——).

Many types of survival curves can be shown but the important point to note is that they all have the same basic properties. They are monotone, nonincreasing functions equal to one at zero and zero as the time approaches infinity. Their rate of decline, of course, varies according to the risk of experiencing the event at time x but it is difficult to determine the essence of a failure pattern by simply looking at the survival curve. Nevertheless, this quantity continues to be a popular description of survival in the applied literature and can be very useful in comparing two or more mortality patterns. Next, we present one more survival curve, which will be discussed at greater length in the next section.

EXAMPLE 2.2 The U.S. Department of Health and Human Services publishes yearly survival curves for all causes of mortality in the United States and each of the fifty states by race and sex in their Vital Statistics of the United

TABLE 2.1
Survival Functions of U.S. Population By Race and Sex in 1989

Age	White Male	White Female	Black Male	Black Female	Age	White Male	White Female	Black Male	Black Female
0	1.00000	1.00000	1.00000	1.00000	43	0.93771	0.97016	0.85917	0.93361
1	0.99092	0.99285	0.97996	0.98283	44	0.93477	0.96862	0.85163	0.92998
2	0.99024	0.99232	0.97881	0.98193	45	0.93161	0.96694	0.84377	0.92612
3	0.98975	0.99192	0.97792	0.98119	46	0.92820	0.96511	0.83559	0.92202
4	0.98937	0.99160	0.97722	0.98059	47	0.92450	0.96311	0.82707	0.91765
5	0.98905	0.99134	0.97664	0.98011	48	0.92050	0.96091	0.81814	0.91300
6	0.98877	0.99111	0.97615	0.97972	49	0.91617	0.95847	0.80871	0.90804
7	0.98850	0.99091	0.97571	0.97941	50	0.91148	0.95575	0.79870	0.90275
8	0.98825	0.99073	0.97532	0.97915	51	0.90639	0.95273	0.78808	0.89709
9	0.98802	0.99056	0.97499	0.97892	52	0.90086	0.94938	0.77685	0.89103
10	0.98782	0.99041	0.97472	0.97870	53	0.89480	0.94568	0.76503	0.88453
11	0.98765	0.99028	0.97449	0.97847	54	0.88810	0.94161	0.75268	0.87754
12	0.98748	0.99015	0.97425	0.97823	55	0.88068	0.93713	0.73983	0.87000
13	0.98724	0.98999	0.97392	0.97796	56	0.87250	0.93222	0.72649	0.86190
14	0.98686	0.98977	0.97339	0.97767	57	0.86352	0.92684	0.71262	0.85321
15	0.98628	0.98948	0.97258	0.97735	58	0.85370	0.92096	0.69817	0.84381
16	0.98547	0.98909	0.97145	0.97699	59	0.84299	0.91455	0.68308	0.83358
17	0.98445	0.98862	0.97002	0.97658	60	0.83135	0.90756	0.66730	0.82243
18	0.98326	0.98809	0.96829	0.97612	61	0.81873	0.89995	0.65083	0.81029
19	0.98197	0.98755	0.96628	0.97559	62	0.80511	0.89169	0.63368	0.79719
20	0.98063	0.98703	0.96403	0.97498	63	0.79052	0.88275	0.61584	0.78323
21	0.97924	0.98654	0.96151	0.97429	64	0.77501	0.87312	0.59732	0.76858
22	0.97780	0.98607	0.95873	0.97352	65	0.75860	0.86278	0.57813	0.75330
23	0.97633	0.98561	0.95575	0.97267	66	0.74131	0.85169	0.55829	0.73748
24	0.97483	0.98514	0.95267	0.97174	67	0.72309	0.83980	0.53783	0.72104
25	0.97332	0.98466	0.94954	0.97074	68	0.70383	0.82702	0.51679	0.70393
26	0.97181	0.98416	0.94639	0.96967	69	0.68339	0.81324	0.49520	0.68604
27	0.97029	0.98365	0.94319	0.96852	70	0.66166	0.79839	0.47312	0.66730
28	0.96876	0.98312	0.93989	0.96728	71	0.63865	0.78420	0.45058	0.64769
29	0.96719	0.98257	0.93642	0.96594	72	0.61441	0.76522	0.42765	0.62723
30	0.96557	0.98199	0.93273	0.96448	73	0.58897	0.74682	0.40442	0.60591
31	0.96390	0.98138	0.92881	0.96289	74	0.56238	0.72716	0.38100	0.58375
32	0.96217	0.98073	0.92466	0.96118	75	0.53470	0.70619	0.35749	0.56074
33	0.96038	0.98005	0.92024	0.95934	76	0.50601	0.68387	0.33397	0.53689
34	0.95852	0.97933	0.91551	0.95740	77	0.47641	0.66014	0.31050	0.51219
35	0.95659	0.97858	0.91044	0.95336	78	0.44604	0.63494	0.28713	0.48663
36	0.95457	0.97779	0.90501	0.95321	79	0.41503	0.60822	0.26391	0.46020
37	0.95245	0.97696	0.89922	0.95095	80	0.38355	0.57991	0.24091	0.43291
38	0.95024	0.97607	0.89312	0.94855	81	0.35178	0.54997	0.21819	0.40475
39	0.94794	0.97510	0.88677	0.94598	82	0.31991	0.51835	0.19583	0.37573
40	0.94555	0.97404	0.88021	0.94321	83	0.28816	0.48502	0.17392	0.34588
41	0.94307	0.97287	0.87344	0.94023	84	0.25677	0.44993	0.15257	0.31522
42	0.94047	0.97158	0.86643	0.93703	85	0.22599	0.41306	0.13191	0.28378

States Series. In Table 2.1, we present the overall survival probabilities for males and females, by race, taken from the 1990 report (U.S. Department of Health and Human Services, 1990). Figure 2.2 shows the survival curves and allows a visual comparison of the curves. We can see that white females have the best survival probability, white males and black females are comparable in their survival probabilities, and black males have the worst survival.

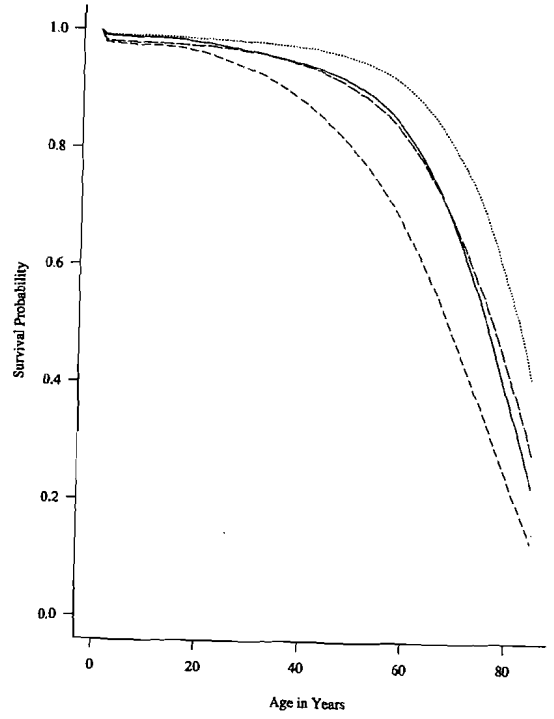

Figure 2.2 *Survival Functions for all cause mortality for the US population in 1989. White males* (———); *white females* (·····); *black males* (——); *black females* (— — —).

When X is a discrete, random variable, different techniques are required. Discrete, random variables in survival analyses arise due to rounding off measurements, grouping of failure times into intervals, or

when lifetimes refer to an integral number of units. Suppose that X can take on values x_j, $j = 1, 2, \ldots$ with probability mass function (p.m.f.) $p(x_j) = Pr(X = x_j)$, $j = 1, 2, \ldots$, where $x_1 < x_2 < \cdots$.

The survival function for a discrete random variable X is given by

$$S(x) = Pr(X > x) = \sum_{x_j > x} p(x_j). \tag{2.2.3}$$

EXAMPLE 2.3 Consider, for pedagogical purposes, the lifetime X, which has the p.m.f. $p(x_j) = Pr(X = j) = 1/3$, $j = 1, 2, 3$, a simple discrete uniform distribution. The corresponding survival function, plotted in Figure 2.3, is expressed by

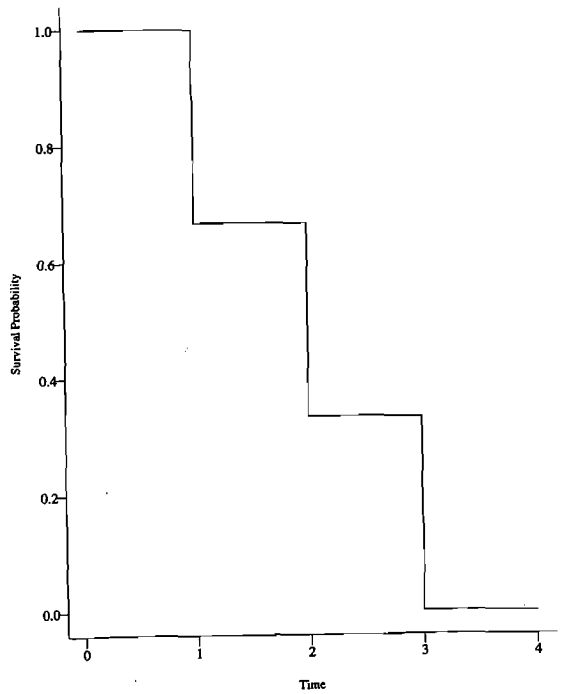

Figure 2.3 *Survival function for a discrete random lifetime*

$$S(x) = Pr(X > x) = \sum_{x_j > x} p(x_j) = \begin{cases} 1 & \text{if } 0 \le x < 1, \\ 2/3 & \text{if } 1 \le x < 2, \\ 1/3 & \text{if } 2 \le x < 3, \\ 0 & \text{if } x \ge 3. \end{cases}$$

Note that, when X is discrete, the survival function is a nonincreasing step function.

2.3 The Hazard Function

A basic quantity, fundamental in survival analysis, is the hazard function. This function is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, the inverse of the Mill's ratio in economics, or simply as the hazard rate. The hazard rate is defined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \le X < x + \Delta x \mid X \ge x]}{\Delta x}. \tag{2.3.1}$$

If X is a continuous random variable, then,

$$h(x) = f(x)/S(x) = -d \ln[S(x)]/dx. \tag{2.3.2}$$

A related quantity is the cumulative hazard function $H(x)$, defined by

$$H(x) = \int_0^x h(u)\, du = -\ln[S(x)]. \tag{2.3.3}$$

Thus, for continuous lifetimes,

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)\, du\right]. \tag{2.3.4}$$

From (2.3.1), one can see that $h(x)\Delta x$ may be viewed as the "approximate" probability of an individual of age x experiencing the event in the next instant. This function is particularly useful in determining the appropriate failure distributions utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time. There are many general shapes for the hazard rate. The only restriction on $h(x)$ is that it be nonnegative, i.e., $h(x) \ge 0$.

Some generic types of hazard rates are plotted in Figure 2.4. For example, one may believe that the hazard rate for the occurrence of a particular event is increasing, decreasing, constant, bathtub-shaped,

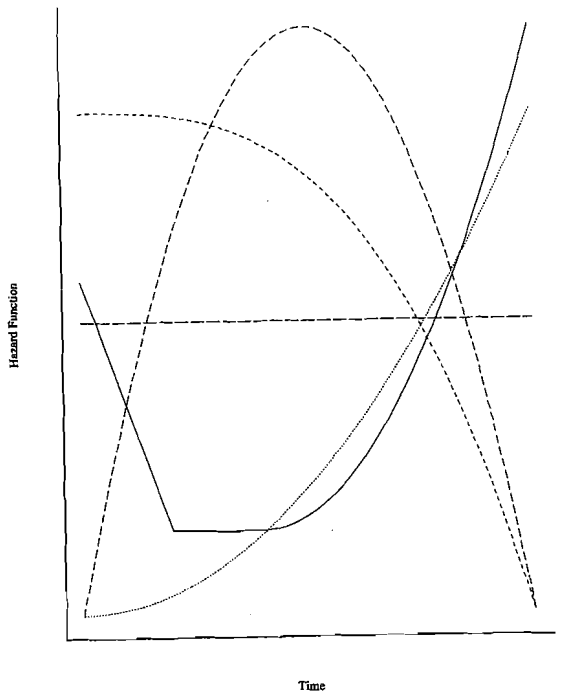

Figure 2.4 *Shapes of hazard functions. Constant hazard* (——); *increasing hazard* (——); *decreasing hazard* (- - - - -); *bathtub shaped* (——); *humpshaped* (——).

hump-shaped, or possessing some other characteristic which describes the failure mechanism.

Models with increasing hazard rates may arise when there is natural aging or wear. Decreasing hazard functions are much less common but find occasional use when there is a very early likelihood of failure, such as in certain types of electronic devices or in patients experiencing certain types of transplants. Most often, a bathtub-shaped hazard is appropriate in populations followed from birth. Similarly, some manufactured equipment may experience early failure due to faulty parts, followed by a constant hazard rate which, in the later stages of equipment life, increases. Most population mortality data follow this type of hazard function where, during an early period, deaths result, primarily, from infant diseases, after which the death rate stabilizes, followed by

an increasing hazard rate due to the natural aging process. Finally, if the hazard rate is increasing early and eventually begins declining, then, the hazard is termed hump-shaped. This type of hazard rate is often used in modeling survival after successful surgery where there is an initial increase in risk due to infection, hemorrhaging, or other complications just after the procedure, followed by a steady decline in risk as the patient recovers. Specific distributions which give rise to these different types of failure rates are presented in section 2.5.

EXAMPLE 2.1 *(continued)* One particular distribution, which is flexible enough to accommodate increasing ($\alpha > 1$), decreasing ($\alpha < 1$), or constant hazard rates ($\alpha = 1$), is the Weibull distribution introduced in Example 2.1. Hazard rates, $h(x) = \alpha\lambda x^{\alpha-1}$, are plotted for the same values of the parameters used in Figure 2.1, namely, $\lambda = 0.26328$, $\alpha = 0.5$; $\lambda = 0.1$, $\alpha = 1$; and $\lambda = 0.00208$, $\alpha = 3$ in Figure 2.5. One can see

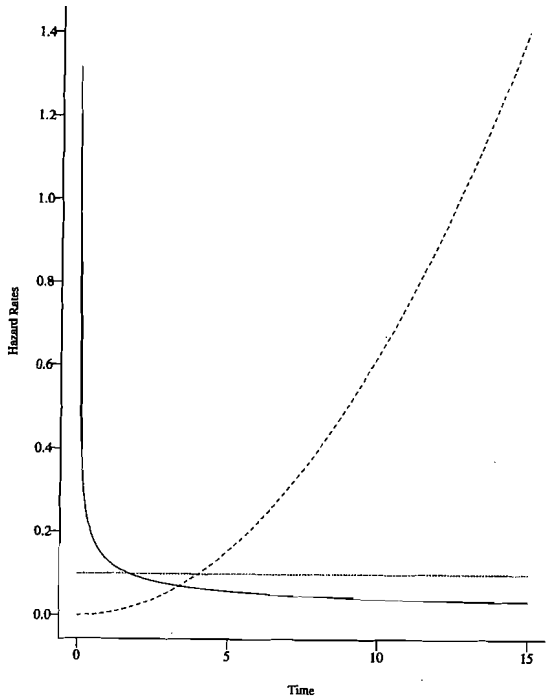

Figure 2.5 *Weibull hazard functions for* $\alpha = 0.5$, $\lambda = 0.26328$ (——); $\alpha = 1.0$, $\lambda = 0.1$ (——); $\alpha = 3.0$, $\lambda = 0.00208$ (——).

that, though the three survival functions have the same basic shape, the hazard functions are dramatically different.

An example of a bathtub-shaped hazard rate is presented in the following example.

EXAMPLE 2.2 *(continued)* The 1989 U.S. mortality hazard rates, by sex and race, are presented in Figure 2.6. One can see the decreasing hazard rates early in all four groups, followed, approximately, by, a constant hazard rate, eventually leading to an increasing hazard rate starting at different times for each group.

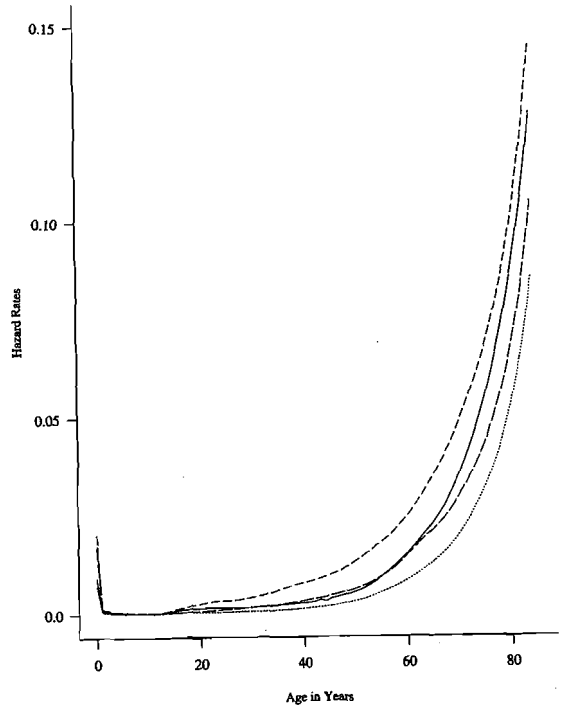

Figure 2.6 *Hazard functions for all cause mortality for the US population in 1989. White males* (———); *white females* (·····); *black males* (- - - - -); *black females* (— — —).

When X is a discrete random variable, the hazard function is given by

$$h(x_j) = Pr(X = x_j \mid X \geq x_j) = \frac{p(x_j)}{S(x_{j-1})}, \quad j = 1, 2, \ldots \tag{2.3.5}$$

where $S(x_0) = 1$. Because $p(x_j) = S(x_{j-1}) - S(x_j)$, in conjunction with (2.3.5), $h(x_j) = 1 - S(x_j)/S(x_{j-1})$, $j = 1, 2, \ldots$.

Note that the survival function may be written as the product of conditional survival probabilities

$$S(x) = \prod_{x_j \leq x} S(x_j)/S(x_{j-1}). \tag{2.3.6}$$

Thus, the survival function is related to the hazard function by

$$S(x) = \prod_{x_j \leq x} [1 - h(x_j)]. \tag{2.3.7}$$

EXAMPLE 2.3 *(continued)* Let us reconsider the discrete random variable X in Example 2.3 with $p(x_j) = Pr(X = j) = 1/3$, $j = 1, 2, 3$. The hazard function may be obtained by direct application of (2.3.5). This leads to

$$h(x_j) = \begin{cases} 1/3, \text{for } j = 1, \\ 1/2, \text{ for } j = 2, \\ 1, \text{ for } j = 3, \text{and} \\ 0, \text{ elsewhere.} \end{cases}$$

Note that the hazard rate is zero for a discrete random variable except at points where a failure could occur.

Practical Notes

1. Though the three survival functions in Figure 2.1 have the same basic shape, one can see that the three hazard functions shown in Figure 2.5 are dramatically different. In fact, the hazard function is usually more informative about the underlying mechanism of failure than the survival function. For this reason, consideration of the hazard function may be the dominant method for summarizing survival data.
2. The relationship between some function of the cumulative hazard function and some function of time has been exploited to develop hazard papers (Nelson, 1982), which will give the researcher an intuitive impression as to the desirability of the fit of specific models. For example, if X has a Weibull distribution, as in Example 2.1, then

its cumulative hazard rate is $H(x) = \lambda x^{\alpha}$, so a plot of ln $H(x)$ versus ln x is a straight line with slope α and y intercept ln λ. Using a nonparametric estimator of $H(x)$, developed in Chapter 4, this relationship can be exploited to provide a graphical check of the goodness of fit of the Weibull model to data (see section 12.5 for details and examples).

Theoretical Notes

1. For discrete lifetimes, we shall define the cumulative hazard function by

$$H(x) = \sum_{x_j \le x} h(x_j). \tag{2.3.8}$$

Notice that the relationship $S(x) = \exp\{-H(x)\}$ for this definition no longer holds true. Some authors (Cox and Oakes, 1984) prefer to define the cumulative hazard for discrete lifetimes as

$$H(x) = -\sum_{x_j \le x} \ln[1 - h(x_j)], \tag{2.3.9}$$

because the relationship for continuous lifetimes $S(x) = \exp[-H(x)]$ will be preserved for discrete lifetimes. If the $h(x_j)$ are small, (2.3.8) will be an approximation of (2.3.9). We prefer the use of (2.3.8) because it is directly estimable from a sample of censored or truncated lifetimes and the estimator has very desirable statistical properties. This estimator is discussed in Chapter 4.
2. For continuous lifetimes, the failure distribution is said to have an increasing failure-rate (IFR) property, if the hazard function $h(x)$ is nondecreasing for $x \ge 0$, and an increasing failure rate on the average (IFRA) if the ratio of the cumulative hazard function to time $H(x)/x$ is nondecreasing for $x > 0$.
3. For continuous lifetimes, the failure distribution is said to have a decreasing failure-rate (DFR) property if the hazard function $h(x)$ is nonincreasing for $x \ge 0$.

2.4 The Mean Residual Life Function and Median Life

The fourth basic parameter of interest in survival analyses is the *mean residual life* at time x. For individuals of age x, this parameter measures

their expected remaining lifetime. It is defined as $\text{mrl}(x) = E(X - x \mid X > x)$. It can be shown (see Theoretical Note 1) that the mean residual life is the area under the survival curve to the right of x divided by $S(x)$. Note that the mean life, $\mu = \text{mrl}(0)$, is the total area under the survival curve.

For a continuous random variable,

$$\text{mrl}(x) = \frac{\int_x^{\infty}(t-x)f(t)dt}{S(x)} = \frac{\int_x^{\infty} S(t)\,dt}{S(x)} \tag{2.4.1}$$

and

$$\mu = E(X) = \int_0^{\infty} tf(t)\,dt = \int_0^{\infty} S(t)\,dt. \tag{2.4.2}$$

Also the variance of X is related to the survival function by

$$\text{Var}(X) = 2\int_0^{\infty} tS(t)dt - \left[\int_0^{\infty} S(t)\,dt\right]^2. \tag{2.4.3}$$

The pth quantile (also referred to as the $100p$th percentile) of the distribution of X is the smallest x_p so that

$$S(x_p) \le 1 - p, \text{ i.e., } x_p = \inf\{t : S(t) \le 1 - p\}. \tag{2.4.4}$$

If X is a continuous random variable, then the pth quantile is found by solving the equation $S(x_p) = 1 - p$. The median lifetime is the 50th percentile $x_{0.5}$ of the distribution of X. It follows that the median lifetime for a continuous random variable X is the value $x_{0.5}$ so that

$$S(x_{0.5}) = 0.5. \tag{2.4.5}$$

EXAMPLE 2.4 The mean and median lifetimes for an exponential life distribution are $1/\lambda$ and $(\ln 2)/\lambda$ as determined from equations (2.4.2) and (2.4.5), respectively. Furthermore, the mean residual life for an exponential distribution is also $1/\lambda$ from equation (2.4.1). Distributions with this property are said to exhibit lack of memory. The exponential distribution is the unique continuous distribution possessing this characteristic.

EXAMPLE 2.1 *(continued)* For the Weibull distribution the $100p$th percentile is found by solving the equation $1 - p = \exp\{-\lambda x_p^{\alpha}\}$ so that $x_p = \{-\ln[1-p]/\lambda\}^{1/\alpha}$.

EXAMPLE 2.2 *(continued)* The median and other percentiles for the population mortality distribution of black men may be determined graphically by using

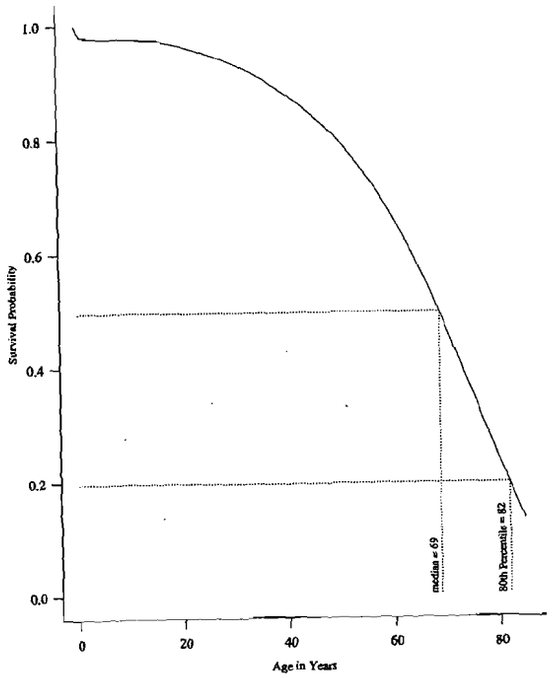

Figure 2.7 *Determination of the median lifetime and 80th percentile of lifetimes for black men in the US population in 1989*

the survival function plot depicted in Figure 2.2. First, find the appropriate survival probability, and, then, interpolate to the appropriate time. Determination of the median and 80th percentile, as illustrated in Figure 2.7, give values of about 69 and 82 years, respectively. More accurate values can be found by linear interpolation in Table 2.1. We see that $S(68) = 0.51679 > 0.5$ and $S(69) = 0.49520 < 0.5$, so the median lies between 68 and 69 years. By linear interpolation,

$$x_{0.5} = 68 + \frac{S(68) - 0.5}{S(68) - S(69)} = 68.78 \text{ years}.$$

Similar calculations yield $x_{0.8} = 81.81$ years.

Theoretical Notes

1. For a continuous random variable X,

$$E(X - x \mid X > x) = \frac{\int_x^\infty (t - x) f(t)\,dt}{S(x)}.$$

We integrate by parts to establish equation (2.4.1) using the fact that $f(t)dt = -dS(t)$, so that $E(X - x \mid X > x)S(x) = -(t - x)S(t)\,|_x^\infty + \int_x^\infty S(t)dt$. The first term on the right-hand side of the equation is 0 because $S(\infty)$ is 0. For a discrete, random variable, the result that the mean residual life is related to the area under the survival curve is obtained by using a partial summation formula.

2. Interrelationships between the various quantities discussed earlier, for a continuous lifetime X, may be summarized as

$$\begin{aligned}
S(x) &= \int_x^\infty f(t)\,dt \\
&= \exp\left[-\int_0^x h(u)\,du\right] \\
&= \exp[-H(x)] \\
&= \frac{\text{mrl}(0)}{\text{mrl}(x)} \exp\left[-\int_0^x \frac{du}{\text{mrl}(u)}\right].
\end{aligned}$$

$$\begin{aligned}
f(x) &= -\frac{d}{dx} S(x) \\
&= h(x)S(x) \\
&= \left(\frac{d}{dx}\text{mrl}(x) + 1\right)\left(\frac{\text{mrl}(0)}{\text{mrl}(x)^2}\right) \exp\left[-\int_0^x \frac{du}{\text{mrl}(u)}\right]
\end{aligned}$$

$$\begin{aligned}
h(x) &= -\frac{d}{dx} ln[S(x)] \\
&= \frac{f(x)}{S(x)} \\
&= \left(\frac{d}{dx}\text{mrl}(x) + 1\right) / \text{mrl}(x)
\end{aligned}$$

$$\begin{aligned}
\text{mrl}(x) &= \frac{\int_x^\infty S(u)\,du}{S(x)} \\
&= \frac{\int_x^\infty (u - x) f(u)\,du}{S(x)}.
\end{aligned}$$

3. Interrelationships between the various quantities discussed earlier, for discrete lifetimes X, may be summarized as

$$S(x) = \sum_{x_j > x} p(x_j)$$

$$= \prod_{x_j \le x} [1 - h(x_j)],$$

$$p(x_j) = S(x_{j-1}) - S(x_j) = h(x_j)S(x_{j-1}), \quad j = 1, 2, \ldots,$$

$$h(x_j) = \frac{p(x_j)}{S(x_{j-1})},$$

$$\text{mrl}(x) = \frac{(x_{i+1} - x)S(x_i) + \sum_{j \ge i+1}(x_{j+1} - x_j)S(x_j)}{S(x)},$$

$$\text{for } x_i \le x < x_{i+1}.$$

4. If X is a positive random variable with a hazard rate $h(t)$, which is a sum of a continuous function $h_c(t)$ and a discrete function which has mass $h_d(x_j)$ at times $0 \le x_1 \le x_2 \le \cdots$, then the survival function is related to the hazard rate by the so called "product integral" of $[1 - h(t)]dt$ defined as follows:

$$S(x) = \prod_{x_j \le x} [1 - h_d(x_j)] \exp\left[-\int_0^x h_c(t)dt\right].$$

5. Sometimes (particularly, when the distribution is highly skewed), the median is preferred to the mean, in which case, the quantity median residual lifetime at time x, mdrl(x), is preferred to the mean residual lifetime at time x, mrl(x), as defined in (2.4.1). The median residual lifetime at time x is defined to be the median of the conditional distribution of $X - x \mid X > x$ and is determined using (2.4.4) except that the conditional distribution is used. It is the length of the interval from x to the time where one-half of the individuals alive at time x will still be alive. Note that the mdrl(0) is simply the median of the unconditional distribution.

2.5 Common Parametric Models for Survival Data

Although nonparametric or semiparametric models will be used extensively, though not exclusively, in this book, it is appropriate and necessary to discuss the more widely used parametric models. These models are chosen, not only because of their popularity among researchers who analyze survival data, but also because they offer insight into the nature of the various parameters and functions discussed in previous sections, particularly, the hazard rate. Some of the important models discussed include the exponential, Weibull, gamma, log normal, log logistic, normal, exponential power, Gompertz, inverse Gaussian, Pareto, and the generalized gamma distributions. Their survival functions, hazard rates, density functions, and expected lifetimes are summarized in Table 2.2.

First, because of its historical significance, mathematical simplicity, and important properties, we shall discuss the *exponential* distribution. Its survival function is $S(x) = \exp[-\lambda x], \lambda > 0, x > 0$. The density function is $f(x) = \lambda \exp[-\lambda x]$, and it is characterized by a constant hazard function $h(x) = \lambda$.

The exponential distribution has the following properties. The first, referred to as the lack of memory property, is given by

$$P(X \ge x + z \mid X \ge x) = P(X \ge z), \tag{2.5.1}$$

which allows for its mathematical tractability but also reduces its applicability to many realistic applied situations. Because of this distributional property, it follows that $E(X - x \mid X > x) = E(X) = 1/\lambda$; that is, the mean residual life is constant. Because the time until the future occurrence of an event does not depend upon past history, this property is sometimes called the "no-aging" property or the "old as good as new" property. This property is also reflected in the exponential distribution's constant hazard rate. Here, the conditional probability of failure at any time t, given that the event has not occurred prior to time t, does not depend upon t. Although the exponential distribution has been historically very popular, its constant hazard rate appears too restrictive in both health and industrial applications.

The mean and standard deviation of the distribution are $1/\lambda$ (thus, the coefficient of variation is unity) and the pth quantile is $x_p = -\ln(1-p)/\lambda$. Because the exponential distribution is a special case of both the Weibull and gamma distributions, considered in subsequent paragraphs, other properties will be implicit in the discussion of those distributions.

Though not the first to suggest the use of this next distribution, Rosen and Rammler (1933) used it to describe the "laws governing the fineness of powdered coal," and Weibull (1939, 1951) proposed the same distribution, to which his name later became affixed, for describing the life length of materials. Its survival function is $S(x) = \exp[-\lambda x^\alpha]$, for $x > 0$. Here $\lambda > 0$ is a scale parameter, and $\alpha > 0$ is a shape parameter. The two-parameter Weibull was previously introduced in Example 2.1. The exponential distribution is a special case when $\alpha = 1$. Figure 2.1, already presented, exhibits a variety of Weibull survival functions. Its

TABLE 2.2
Hazard Rates, Survival Functions, Probability Density Functions, and Expected Lifetimes for Some Common Parametric Distributions

Distribution	Hazard Rate $h(x)$	Survival Function $S(x)$	Probability Density Function $f(x)$	Mean $E(X)$
Exponential $\lambda > 0, x \geq 0$	λ	$\exp[-\lambda x]$	$\lambda \exp(-\lambda x)$	$\frac{1}{\lambda}$
Weibull $\alpha, \lambda > 0, x \geq 0$	$\alpha\lambda x^{\alpha-1}$	$\exp[-\lambda x^{\alpha}]$	$\alpha\lambda x^{\alpha-1}\exp(-\lambda x^{\alpha})$	$\frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$
Gamma $\beta, \lambda > 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - I(\lambda x, \beta)^*$	$\frac{\lambda^{\beta}x^{\beta-1}\exp(-\lambda x)}{\Gamma(\beta)}$	$\frac{\beta}{\lambda}$
Log normal $\sigma > 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - \Phi\left[\frac{\ln x - \mu}{\sigma}\right]$	$\frac{\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right]}{x(2\pi)^{1/2}\sigma}$	$\exp(\mu + 0.5\sigma^2)$
Log logistic $\alpha, \lambda > 0, x \geq 0$	$\frac{\alpha x^{\alpha-1}\lambda}{1+\lambda x^{\alpha}}$	$\frac{1}{1+\lambda x^{\alpha}}$	$\frac{\alpha x^{\alpha-1}\lambda}{[1+\lambda x^{\alpha}]^2}$	$\frac{\pi\mathrm{Csc}(\pi/\alpha)}{\alpha\lambda^{1/\alpha}}$ if $\alpha > 1$
Normal $\sigma > 0, -\infty < x < \infty$	$\frac{f(x)}{S(x)}$	$1 - \Phi\left[\frac{x-\mu}{\sigma}\right]$	$\frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]}{(2\pi)^{1/2}\sigma}$	μ
Exponential power $\alpha, \lambda > 0, x \geq 0$	$\alpha\lambda^{\alpha}x^{\alpha-1}\exp\{[\lambda x]^{\alpha}\}$	$\exp\{1 - \exp[(\lambda x)^{\alpha}]\}$	$\alpha e\lambda^{\alpha}x^{\alpha-1}\exp[(\lambda x)^{\alpha}] - \exp\{\exp[(\lambda x)^{\alpha}]\}$	$\int_0^{\infty} S(x)dx$
Gompertz $\theta, \alpha > 0, x \geq 0$	$\theta e^{\alpha x}$	$\exp\left[\frac{\theta}{\alpha}(1 - e^{\alpha x})\right]$	$\theta e^{\alpha x}\exp\left[\frac{\theta}{\alpha}(1 - e^{\alpha x})\right]$	$\int_0^{\infty} S(x)dx$
Inverse Gaussian $\lambda \geq 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$\Phi\left[\left(\frac{\lambda}{x}\right)^{1/2}\left(1 - \frac{x}{\mu}\right)\right] - e^{2\lambda/\mu}\Phi\left\{-\left[\frac{\lambda}{x}\right]^{1/2}\left(1 + \frac{x}{\mu}\right)\right\}$	$\left(\frac{\lambda}{2\pi x^3}\right)^{1/2}\exp\left[\frac{\lambda(x-\mu^2)}{2\mu^2 x}\right]$	μ
Pareto $\theta > 0, \lambda > 0$ $x \geq \lambda$	$\frac{\theta}{x}$	$\frac{\lambda^{\theta}}{x^{\theta}}$	$\frac{\theta\lambda^{\theta}}{x^{\theta+1}}$	$\frac{\theta\lambda}{\theta-1}$ if $\theta > 1$
Generalized gamma $\lambda > 0, \alpha > 0,$ $\beta > 0, x \geq 0$	$\frac{f(x)}{S(x)}$	$1 - I[\lambda x^{\alpha}, \beta]$	$\frac{\alpha\lambda^{\beta}x^{\alpha\beta-1}\exp(-\lambda x^{\alpha})}{\Gamma(\beta)}$	$\int_0^{\infty} S(x)dx$

* $I(t, \beta) = \int_0^t u^{\beta-1}\exp(-u)du/\Gamma(\beta)$.

hazard function has the fairly flexible form

$$h(x) = \lambda\alpha x^{\alpha-1}. \tag{2.5.2}$$

One can see from Figure 2.5 that the Weibull distribution is flexible enough to accommodate increasing ($\alpha > 1$), decreasing ($\alpha < 1$), or constant hazard rates ($\alpha = 1$). This fact, coupled with the model's relatively simple survival, hazard, and probability density functions, have made it a very popular parametric model. It is apparent that the shape of the Weibull distribution depends upon the value of α, thus, the reason for referring to this parameter as the "shape" parameter.

The rth moment of the Weibull distribution is $[\Gamma(1 + r/\alpha)]\lambda^{-r/\alpha}$. The mean and variance are $[\Gamma(1 + 1/\alpha)]\lambda^{-1/\alpha}$ and $\{\Gamma(1 + 2/\alpha) - [\Gamma(1 + 1/\alpha)]^2\}\lambda^{-2/\alpha}$, respectively, where $\Gamma[\alpha] = \int_0^{\infty} u^{\alpha-1}e^{-u}du$ is the well-known gamma function. $\Gamma[\alpha] = (\alpha - 1)!$ when α is an integer and is tabulated in Beyer (1968) when α is not an integer. The pth quantile of the Weibull distribution is expressed by

$$x_p = \{-[\ln(1 - p)]/\lambda\}^{1/\alpha}.$$

It is sometimes useful to work with the logarithm of the lifetimes. If we take $Y = \ln X$, where X follows a Weibull distribution, then, Y has the density function

$$\alpha\exp\{\alpha[y - (-(\ln\lambda)/\alpha)] - \exp\{\alpha[y - (-(\ln\lambda)/\alpha)]\}\}, -\infty < y < \infty. \tag{2.5.3}$$

Writing the model in a general linear model format, $Y = \mu + \sigma E$, where $\mu = (-\ln\lambda)/\alpha, \sigma = \alpha^{-1}$ and E has the standard extreme value distribution with density function

$$\exp(w - e^{w}), -\infty < w < \infty. \tag{2.5.4}$$

A random variable (more familiar to the traditional linear model audience) X is said to follow the *log normal* distribution if its logarithm $Y = \ln X$, follows the normal distribution. For time-to-event data, this distribution has been popularized because of its relationship to the *normal* distribution (a distribution which we assume is commonly known from elementary statistics courses and whose hazard rate, survival function, density function and mean are reported in Table 2.2 for completeness) and because some authors have observed that the log normal distribution approximates survival times or ages at the onset of certain diseases (Feinleib, 1960 and Horner, 1987).

Like the normal distribution, the log normal distribution is completely specified by two parameters μ and σ, the mean and variance of Y. Its density function is expressed by

$$f(x) = \frac{\exp\left[-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right]}{x(2\pi)^{1/2}\sigma} = \phi\left(\frac{\ln x - u}{\sigma}\right)/x \tag{2.5.5}$$

and its survival function is given by

$$S(x) = 1 - \Phi\left[\frac{\ln x - \mu}{\sigma}\right], \tag{2.5.6}$$

where $\Phi(\phi)$ is the cumulative distribution function (density function) of a standard normal variable.

The hazard rate of the log normal is hump-shaped, that is, its value at 0 is zero, and it increases to a maximum and, then, decreases to 0 as x approaches infinity (see Figure 2.8). This model has been criticized as a lifetime distribution because the hazard function is decreasing for large x which seems implausible in many situations. The model may fit certain cases where large values of x are not of interest.

For the log normal distribution the mean lifetime is given by $\exp(\mu + \sigma^2/2)$ and the variance by $[\exp(\sigma^2)-1]\exp(2\mu+\sigma^2)$. The pth percentile,

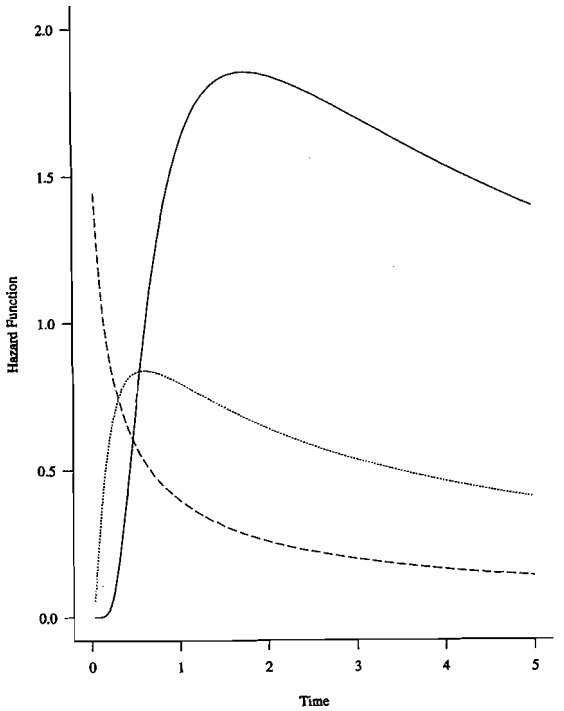

Figure 2.8 *Log normal hazard rates.* $\mu = 0, \sigma = 0.5$ (———); $\mu = 0, \sigma = 0.1$ (——); $\mu = 0, \sigma = 2.0$ (— — —)

x_p is expressed as $\exp(\mu + \sigma z_p)$, where z_p is the pth percentile of a standard normal distribution.

A variable X is said to follow the *log logistic* distribution if its logarithm $Y = \ln X$ follows the logistic distribution, a distribution closely resembling the normal distribution, but the survival function is mathematically more tractable. The density function for Y is expressed by

$$\frac{\exp(\frac{y-\mu}{\sigma})}{\sigma[1 + \exp(\frac{y-\mu}{\sigma})]^2}, -\infty < y < \infty, \tag{2.5.7}$$

where μ and σ^2 are, respectively, the mean and scale parameter of Y. Again, we can cast this distribution in the linear model format by taking $Y = \mu + \sigma W$, where W is the standardized logistic distribution with $\mu = 0$ and $\sigma = 1$.

The hazard rate and survival function, respectively, for the log logistic distribution may be written as relatively simple expressions:

$$h(x) = \frac{\alpha\lambda x^{\alpha-1}}{1 + \lambda x^{\alpha}}, \tag{2.5.8}$$

and

$$S(x) = \frac{1}{1 + \lambda x^{\alpha}}, \tag{2.5.9}$$

where $\alpha = 1/\sigma > 0$ and $\lambda = \exp(-\mu/\sigma)$.
The numerator of the hazard function is the same as the Weibull hazard, but the denominator causes the hazard to take on the following characteristics: monotone decreasing for $\alpha \leq 1$. For $\alpha > 1$, the hazard rate increases initially to a maximum at time $[(\alpha - 1)/\lambda]^{1/\alpha}$ and then decreases to zero as time approaches infinity, as shown in Figure 2.9. The mean and variance of X are given by $E[X] = \pi\csc(\pi/\alpha)/(\alpha\lambda^{1/\alpha})$, if $\alpha > 1$, and $\mathrm{Var}(X) = 2\pi\csc(2\pi/\alpha)/(\alpha\lambda^{2/\alpha}) - E[X]^2$, if $\alpha > 2$. The pth percentile is $x_p = \{p/[\lambda(1-p)]\}^{1/\alpha}$.

This distribution is similar to the Weibull and exponential models because of the simple expressions for $h(x)$ and $S(x)$ above. Its hazard rate is similar to the log normal, except in the extreme tail of the distribution, but its advantage is its simpler hazard function $h(x)$ and survival function $S(x)$.

The *gamma* distribution has properties similar to the Weibull distribution, although it is not as mathematically tractable. Its density function is given by

$$f(x) = \lambda^{\beta} x^{\beta-1}\exp(-\lambda x)/\Gamma(\beta), \tag{2.5.10}$$

where $\lambda > 0$, $\beta > 0$, $x > 0$, and $\Gamma(\beta)$ is the gamma function. For reasons similar to those of the Weibull distribution, λ is a scale parameter and β is called the shape parameter. This distribution, like the Weibull, includes the exponential as a special case ($\beta = 1$), approaches

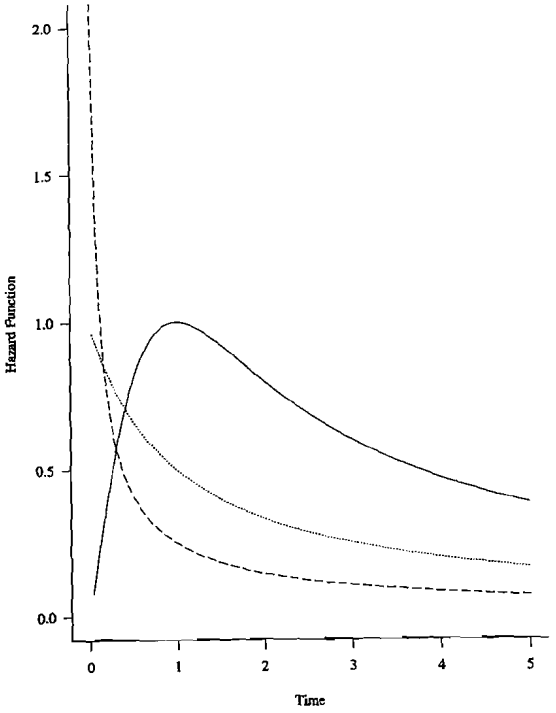

Figure 2.9 *Log logistic hazard rates.* $\lambda = 1, \sigma = 0.5$ (———); $\lambda = 1, \sigma = 1.0$ (·····); $\lambda = 1, \sigma = 2.0$ (——)

a normal distribution as $\beta \to \infty$, and gives the chi-square distribution with ν degrees of freedom when $\nu = 2\beta$ (β, an integer) and $\lambda = 1/2$. The mean and variance of the gamma distribution are β/λ and β/λ^2, respectively.

The hazard function for the gamma distribution is monotone increasing for $\beta > 1$, with $h(0) = 0$ and $h(x) \to \lambda$ as $x \to \infty$, and monotone decreasing for $\beta < 1$, with $h(0) = \infty$ and $h(x) \to \lambda$ as $x \to \infty$. When $\beta > 1$, the mode is at $x = (\beta - 1)/\lambda$. A plot of the gamma hazard function is presented in Figure 2.10.

The survival function of the gamma distribution is expressed as

$$S(x) = \left[\int_x^\infty \lambda(\lambda t)^{\beta-1} \exp(-\lambda t)dt\right] / \Gamma(\beta) \qquad (2.5.11)$$

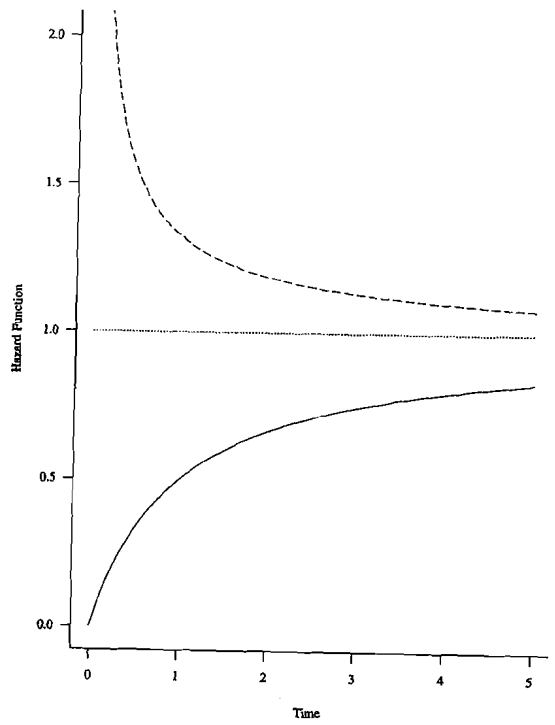

Figure 2.10 *Gamma hazard rates.* $\lambda = 1, \beta = 2.0$ (———); $\lambda = 1, \beta = 1.0$ (——); $\lambda = 1, \beta = 0.5$ (— — —)

$$= 1 - \left[\int_0^{\lambda x} u^{\beta-1} \exp(-u)du\right] / \Gamma(\beta)$$
$$= 1 - I(\lambda x, \beta),$$

where I is the incomplete gamma function.

For $\beta = n$, an integer, we obtain the Erlangian distribution whose survival function and hazard function, respectively, calculated from (2.2.2) and (2.3.2) simplify to

$$S(x) = \exp(-\lambda x) \sum_{k=0}^{n-1} (\lambda x)^k / k! \qquad (2.5.12)$$

and

$$h(x) = \lambda(\lambda x)^{n-1}\left[(n-1)!\sum_{k=0}^{n-1}(\lambda x)^k/k!\right]^{-1}.$$

Practical Notes

1. A relationship for the exponential distribution is $H(x) = -\ln S(x) = \lambda x$. This provides an empirical check for an exponential fit to data by plotting $H(x)$ vs x. The resulting plot should be a straight line through the origin with slope λ.
2. An empirical check of the Weibull distribution is accomplished by plotting $\ln[H(x)]$ vs. $\ln(x)$ (utilizing the relationship in the continuation of Example 2.1). The plot should result in a straight line with slope α and y intercept $\ln(\lambda)$. Later, in Chapter 12, we shall use this technique to give crude estimates of the parameters.
3. The *generalized gamma* distribution introduces an additional parameter α allowing additional flexibility in selecting a hazard function. This model has density function

$$f(x) = \frac{\alpha\lambda^{\beta}x^{\alpha\beta-1}\exp\{-\lambda x^{\alpha}\}}{\Gamma(\beta)} \qquad (2.5.13)$$

and survival function

$$S(x) = 1 - I(\lambda x^{\alpha}, \beta).$$

This distribution reduces to the exponential when $\alpha = \beta = 1$, to the Weibull when $\beta = 1$, to the gamma when $\alpha = 1$, and approaches the log normal as $\beta \to \infty$. It is a useful distribution for model checking.
4. Occasionally, the event of interest may not occur until a threshold time ϕ is attained, in which case, $S(x) < 1$ only for $x > \phi$. In reliability theory, ϕ is called the "guarantee time." For example, in this instance, the Weibull survival function may be modified as follows:

$$S(x) = \exp[-\lambda(x-\phi)^{\alpha}], \lambda > 0, \alpha > 0 \text{ and } x > \phi.$$

Similar modifications may be made to the other distributions discussed in this section to accommodate the notion of a threshold parameter.
5. A model that has a hazard function capable of being bathtub-shaped, i.e., decreasing initially and, then, increasing as time increases, is the *exponential power* distribution with $\alpha < 1$ (Smith-Bain, 1975).
6. A distribution with a rich history in describing mortality curves is one introduced by Gompertz (1825) and later modified by Makeham (1860) by adding a constant to the hazard function (see Chiang, 1968, pp. 61–62). Again, the hazard function, survival function, density function, and mean of the *Gompertz* distribution are summarized in Table 2.2.

7. Other distributions which have received some attention in the literature are the *inverse Gaussian* and the *Pareto* distributions. These distributions are tabulated in Table 2.2 along with their survival, hazard, and probability density functions.

Theoretical Notes

1. The exponential distribution is summarized, with references, by Galambos (1982), Galambos and Kotz (1978), and Johnson and Kotz (1970). It is known to have been studied as early as the nineteenth century by Clausius (1858) in connection with the kinetic theory of gases. More recently, in studies of manufactured items (Davis, 1952; Epstein and Sobel, 1954; Epstein, 1958) and, to a lesser extent, in health studies (Feigl and Zelen, 1965; Sheps, 1966), the exponential distribution has historically been used in describing time to failure. As has been already noted, its constant hazard rate and lack of memory property greatly limit its applicability to modern survival analyses.
2. The Weibull distribution has been widely used in both industrial and biomedical applications. Lieblein and Zelen (1956), Berretoni (1964), and Nelson (1972) used it to describe the life length of ball bearings, electron tubes, manufactured items, and electrical insulation, respectively. Pike (1966) and Peto and Lee (1973) have given a theoretical motivation for its consideration in representing time to appearance of tumor or until death in animals which were subjected to carcinogenic insults over time (the multi-hit theory). Lee and Thompson (1974) argued, in a similar vein, that, within the class of proportional hazard rate distributions, the Weibull appears to be the most appropriate choice in describing lifetimes. Other authors (Lee and O'Neill, 1971; Doll, 1971) claim that the Weibull model fits data involving time to appearance of tumors in animals and humans quite well.
3. The Weibull distribution is also called the first asymptotic distribution of extreme values (see Gumbel, 1958, who popularized its use). The Weibull distribution arises as the limiting distribution of the minimum of a sample from a continuous distribution. For this reason, the Weibull distribution has been suggested as the appropriate distribution in certain circumstances.

2.6 Regression Models for Survival Data

Until this point, we have dealt exclusively with modeling the survival experience of a homogeneous population. However, a problem fre-

quently encountered in analyzing survival data is that of adjusting the survival function to account for concomitant information (sometimes referred to as covariates, explanatory variables or independent variables). Populations which exhibit such heterogeneity are prevalent whether the study involves a clinical trial, a cohort study, or an observational study.

Consider a failure time $X > 0$, as has been discussed in the previous sections, and a vector $\mathbf{Z}^t = (Z_1, \ldots, Z_p)$ of explanatory variables associated with the failure time X. $\mathbf{Z}^t$ may include quantitative variables (such as blood pressure, temperature, age, and weight), qualitative variables (such as gender, race, treatment, and disease status) and/or time-dependent variables, in which case $\mathbf{Z}^t(x) = [Z_1(x), \ldots, Z_p(x)]$. Typical time-dependent variables include whether some intermediate event has or has not occurred by time x, the amount of time which has passed since the same intermediate event, serial measurements of covariates taken since a treatment commenced or special covariates created to test the validity of given model. Previously, we have stressed the importance of modeling the survival function, hazard function, or some other parameter associated with the failure-time distribution. Often a matter of greater interest is to ascertain the relationship between the failure time X and one or more of the explanatory variables. This would be the case if one were comparing the survival functions for two or more treatments, wanting to determine the prognosis of a patient presenting with various characteristics, or identifying pertinent risk factors for a particular disease, controlling for relevant confounders.

Two approaches to the modeling of covariate effects on survival have become popular in the statistical literature. The first approach is analogous to the classical linear regression approach. In this approach, the natural logarithm of the survival time $Y = \ln(X)$ is modeled. This is the natural transformation made in linear models to convert positive variables to observations on the entire real line. A linear model is assumed for Y, namely,

$$Y = \mu + \boldsymbol{\gamma}^t\mathbf{Z} + \sigma W, \tag{2.6.1}$$

where $\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_p)$ is a vector of regression coefficients and W is the error distribution. Common choices for the error distribution include the standard normal distribution which yields a log normal regression model, the extreme value distribution (2.5.4), which yields a Weibull regression model, or a logistic distribution (2.5.7), which yields a log logistic regression model. Estimation of regression coefficients, which is discussed in detail in Chapter 12, is performed using maximum likelihood methods and is readily available in most statistical packages.

This model is called the accelerated failure-time model. To see why this is so, let $S_o(x)$ denote the survival function of $X = e^Y$ when $\mathbf{Z}$ is zero, that is, $S_o(x)$ is the survival function of $\exp(\mu + \sigma W)$.

Now,

$$\begin{aligned} Pr[X > x \mid \mathbf{Z}] &= Pr[Y > \ln x \mid \mathbf{Z}] \\ &= Pr[\mu + \sigma W > \ln x - \boldsymbol{\gamma}^t\mathbf{Z} \mid \mathbf{Z}] \\ &= Pr[e^{\mu+\sigma W} > x\exp(-\boldsymbol{\gamma}^t\mathbf{Z}) \mid \mathbf{Z}] \\ &= S_o[x\exp(-\boldsymbol{\gamma}^t\mathbf{Z})]. \end{aligned}$$

Notice that the effect of the explanatory variables in the original time scale is to change the time scale by a factor $\exp(-\boldsymbol{\gamma}^t\mathbf{Z})$. Depending on the sign of $\boldsymbol{\gamma}^t\mathbf{Z}$, the time is either accelerated by a constant factor or degraded by a constant factor. Note that the hazard rate of an individual with a covariate value $\mathbf{Z}$ for this class of models is related to a baseline hazard rate h_o by

$$h(x \mid \mathbf{Z}) = h_o[x\exp(-\boldsymbol{\gamma}^t\mathbf{Z})]\exp(-\boldsymbol{\gamma}^t\mathbf{Z}). \tag{2.6.2}$$

EXAMPLE 2.5

Suppose that the survival time X follows a Weibull distribution with parameters λ and α. Recall that in section 2.5 we saw that the natural logarithm of X, $Y = \ln(X)$, can be written as a linear model, $Y = \mu + \sigma W$, where $\mu = (-\ln(\lambda)/\alpha)$, $\sigma = \alpha^{-1}$, and W has a standard extreme value distribution with density function $f(w) = \exp\{w - e^w\}$, $-\infty < w < \infty$. Suppose that we also have a set of $p-1$ covariates, $\{Z_2, \ldots, Z_p\}$ which can explain some of the patient to patient variability observed for the lifetimes under study. We shall define the covariate $Z_1 = 1$ to allow for an intercept term in our log linear model and $\mathbf{Z}^t = (Z_1, \ldots, Z_p)$. Let $\boldsymbol{\gamma}^t = (\gamma_1, \ldots, \gamma_p)$ be a vector of regression coefficients. The natural log linear model for Y is given by

$$Y = \boldsymbol{\gamma}^t\mathbf{Z} + \sigma W.$$

With this model, the survival function for Y is expressed as

$$S_Y(y \mid \mathbf{Z}) = \exp\left[-\exp\left(\frac{y - \boldsymbol{\gamma}^t\mathbf{Z}}{\sigma}\right)\right].$$

On the original time scale the survival function for X is given by

$$\begin{aligned} S_X(x \mid Z) &= \exp\left[-x^{1/\sigma}\exp\left(\frac{-\boldsymbol{\gamma}^t\mathbf{Z}}{\sigma}\right)\right] = \exp\{-[x\exp(-\boldsymbol{\gamma}^t\mathbf{Z})]^\alpha\} \\ &= S_o(x\exp\{-\boldsymbol{\gamma}^t\mathbf{Z}\}), \end{aligned}$$

where $S_o(x) = \exp(-x^\alpha)$ is a Weibull survival function.

Although the accelerated failure-time model provides a direct extension of the classical linear model's construction for explanatory variables

for conventional data, for survival data, its use is restricted by the error distributions one can assume. As we have seen earlier in this chapter, the easiest survival parameter to model is the hazard rate which tells us how quickly individuals of a certain age are experiencing the event of interest. The major approach to modeling the effects of covariates on survival is to model the conditional hazard rate as a function of the covariates. Two general classes of models have been used to relate covariate effects to survival, the family of multiplicative hazard models and the family of additive hazard rate models.

For the family of multiplicative hazard rate models the conditional hazard rate of an individual with covariate vector $\mathbf{z}$ is a product of a baseline hazard rate $h_o(x)$ and a non-negative function of the covariates, $c(\boldsymbol{\beta}'\mathbf{z})$, that is,

$$h(x \mid \mathbf{z}) = h_o(x)c(\boldsymbol{\beta}'\mathbf{z}). \tag{2.6.3}$$

In applications of the model, $h_o(x)$ may have a specified parametric form or it may be left as an arbitrary nonnegative function. Any nonnegative function can be used for the link function $c()$. Most applications use the Cox (1972) model with $c(\boldsymbol{\beta}'\mathbf{z}) = \exp(\boldsymbol{\beta}'\mathbf{z})$ which is chosen for its simplicity and for the fact it is positive for any value of $\boldsymbol{\beta}'\mathbf{z}$.

A key feature of multiplicative hazards models is that, when all the covariates are fixed at time 0, the hazard rates of two individuals with distinct values of $\mathbf{z}$ are proportional. To see this consider two individuals with covariate values $\mathbf{z}_1$ and $\mathbf{z}_2$. We have

$$\frac{h(x \mid \mathbf{z}_1)}{h(x \mid \mathbf{z}_2)} = \frac{h_0(x)c(\boldsymbol{\beta}'\mathbf{z}_1)}{h_0(x)c(\boldsymbol{\beta}'\mathbf{z}_2)} = \frac{c(\boldsymbol{\beta}'\mathbf{z}_1)}{c(\boldsymbol{\beta}'\mathbf{z}_2)},$$

which is a constant independent of time.

Using (2.6.3), we see that the conditional survival function of an individual with covariate vector $\mathbf{z}$ can be expressed in terms of a baseline survival function $S_o(x)$ as

$$S(x \mid \mathbf{z}) = S_o(x)^{c(\boldsymbol{\beta}'\mathbf{z})}. \tag{2.6.4}$$

This relationship is also found in nonparametric statistics and is called a "Lehmann Alternative."

Multiplicative hazard models are used for modeling relative survival in section 6.3 and form the basis for modeling covariate effects in Chapters 8 and 9.

EXAMPLE 2.5 *(continued)* The multiplicative hazard model for the Weibull distribution with baseline hazard rate $h_o(x) = \alpha\lambda x^{\alpha-1}$ is $h(x \mid \mathbf{z}) = \alpha\lambda x^{\alpha-1}c(\boldsymbol{\beta}'\mathbf{z})$. When the Cox model is used for the link function, $h(x \mid z) = \alpha\lambda x^{\alpha-1}\exp(\boldsymbol{\beta}'\mathbf{z})$. Here the conditional survival function is given by $S(x \mid \mathbf{z}) = \exp[-\lambda x^{\alpha}]^{\exp[\boldsymbol{\beta}'\mathbf{z}]} = \exp[-\lambda x^{\alpha}\exp(\boldsymbol{\beta}'\mathbf{z})] = \exp[-\lambda(x\exp[\boldsymbol{\beta}'\mathbf{z}/\alpha])^{\alpha}]$, which is of the form of an accelerated failure-time model (2.6.2). The Weibull is the only continuous distribution which has the property of being both an accelerated failure-time model and a multiplicative hazards model.

A second class of models for the hazard rate is the family of additive hazard rate models. Here, we model the conditional hazard function by

$$h(x \mid \mathbf{z}) = h_o(x) + \sum_{j=1}^{p} z_j(x)\beta_j(x). \tag{2.6.5}$$

The regression coefficients for these models are functions of time so that the effect of a given covariate on survival is allowed to vary over time. The p regression functions may be positive or negative, but their values are constrained because (2.6.5) must be positive.

Estimation for additive models is typically made by nonparametric (weighted) least-squares methods. Additive models are used in section 6.3 to model excess mortality and, in Chapter 10, to model regression effects.

Practical Notes

1. From Theoretical Note 1 of section 2.4,

$$S(x \mid \mathbf{z}) = \exp\left[-\int_0^x h(t \mid \mathbf{z})dt\right] \tag{2.6.6}$$

and, in conjunction with (2.6.4),

$$\begin{aligned} S(x \mid \mathbf{z}) &= \exp\left[-\int_0^x h_o(t)\exp[\boldsymbol{\beta}'\mathbf{z}]dt\right] \\ &= \left\{\exp\left[-\int_0^x h_o(t)dt\right]\right\}^{\exp[\boldsymbol{\beta}'\mathbf{z}]} \\ &= [S_o(x)]^{\exp[\boldsymbol{\beta}'\mathbf{z}]} \end{aligned}$$

which implies that

$$\ln[-\ln S(x \mid \mathbf{z})] = \boldsymbol{\beta}'\mathbf{z} + \ln[-\ln S_o(x)]. \tag{2.6.7}$$

So the logarithms of the negative logarithm of the survival functions of X, given different regressor variables $\mathbf{z}_i$, are parallel. This relationship will serve as a check on the proportional hazards assumption discussed further in Chapter 11.

2.7 Models for Competing Risks

In the previous sections of this chapter we have examined parameters which can be used to describe the failure time, T, of a randomly selected individual. Here T may be the time to death (see, for example, sections 1.5, 1.7, 1.8, 1.11), the time to treatment failure (see, for example, sections 1.9, 1.10), time to infection (see sections 1.6, 1.12), time to weaning (section 1.14), etc.

In some medical experiments we have the problem of competing risks. Here each subject may fail due to one of K ($K \geq 2$) causes, called competing risks. An example of competing risks is found in section 1.3. Here the competing risks for treatment failure are relapse and death in remission. Occurrence of one of these events precludes us from observing the other event on this patient. Another classical example of competing risks is cause-specific mortality, such as death from heart disease, death from cancer, death from other causes, etc.

To discuss parameters for the competing-risks problem we shall formulate the model in terms of a latent failure time approach. Other formulations, as discussed in Kalbfleisch and Prentice (1980), give similar representations. Here we let X_i, $i = 1, \ldots, K$ be the potential unobservable time to occurrence of the ith competing risk. What we observe for each patient is the time at which the subject fails from any cause, $T = \text{Min}(X_1, \ldots, X_p)$ and an indicator δ which tells which of the K risks caused the patient to fail, that is, $\delta = i$ if $T = X_i$.

The basic competing risks parameter is the *cause-specific hazard rate for risk i* defined by

$$h_i(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t, \delta = i \mid T \geq t]}{\Delta t} \tag{2.7.1}$$
$$= \lim_{\Delta t \to 0} \frac{P[t \leq X_i < t + \Delta t, \delta = i \mid X_j \geq t, j = 1, \ldots, K]}{\Delta t}$$

Here $h_i(t)$ tells us the rate at which subjects who have yet to experience any of the competing risks are experiencing the ith competing cause of failure. The overall hazard rate of the time to failure, T, given by (2.3.1) is the sum of these K cause-specific hazard rates; that is

$$h_T(t) = \sum_{i=1}^{K} h_i(t).$$

The cause-specific hazard rate can be derived from the joint survival function of the K competing risks. Let $S(t_1, \ldots, t_K) = \Pr[X_1 > t_1, \ldots, X_K > t_K]$. The cause specific hazard rate is given by

$$h_i(t) = \frac{-\partial S(t_1, \ldots, t_K)/\partial t_i\big|_{t_1 = \cdots t_K = t}}{S(t, \ldots, t)} \tag{2.7.2}$$

EXAMPLE 2.6 Suppose that we have K competing risks and that the potential failure times are independent with survival functions $S_i(t)$ for $i = 1, 2, \ldots, K$. Then the joint survival function is $S(t_1, \ldots, t_K) = \prod_{i=1}^{K} S_i(t_i)$, and by (2.7.2) we have

$$h_i(t) = \frac{-\partial \prod_{j=1}^{K} S_j(t_j)/\partial t_i\big|_{t_1 = \cdots t_K = t}}{\prod_{j=1}^{K} S_j(t)} = \frac{-\partial S_i(t_i)/\partial t_i\big|_{t_i = t}}{S_i(t)},$$

which is precisely the hazard rate of X_i.

Example 2.6 shows that for independent competing risks the marginal and cause-specific hazard rates are identical. This need not be the case when the risks are dependent as we see in the following example.

EXAMPLE 2.7 Suppose we have two competing risks and the joint survival function is $S(t_1, t_2) = [1 + \theta(\lambda_1 t_1 + \lambda_2 t_2)]^{-1/\theta}$, $\theta \geq 0, \lambda_1, \lambda_2 \geq 0$. Here the two potential failure times are correlated with a Kendall's τ of $(\theta/(\theta + 2))$ (see section 13.3 for a discussion and derivation of this model). By (2.7.2) we have

$$h_i(t) = \frac{-\partial[1 + \theta(\lambda_1 t_1 + \lambda_2 t_2)]^{-1/\theta}/\partial t_i\big|_{t_1 = t_2 = t}}{1 + \theta t(\lambda_1 + \lambda_2)]^{-1/\theta}}$$
$$= \frac{\lambda_i}{1 + \theta t(\lambda_1 + \lambda_2)}, \; i = 1, 2.$$

Here the survival function of the time to failure, $T = \min(X_1, X_2)$ is $S(t, t) = S_T(t) = [1 + \theta t(\lambda_1 + \lambda_2)]^{-1/\theta}$ and its hazard rate is $(\lambda_1 + \lambda_2)/[1 + \theta t(\lambda_1 + \lambda_2)]^{-1/\theta}$. Note that the marginal survival function for X_1 is given by $S(t_1, 0) = [1 + \theta t \lambda_1]^{-1/\theta}$ and the marginal hazard rate is, from (2.3.2), $\lambda_1/(1 + \theta \lambda_1 t)$, which is not the same as the crude hazard rate.

In competing-risks modeling we often need to make some assumptions about the dependence structure between the potential failure times. Given that we can only observe the failure time and cause and not the potential failure times these assumptions are not testable with only competing risks data. This is called the *identifiability dilemma.*

We can see the problem clearly by careful examination of Example 2.7. Suppose we had two independent competing risks with hazard rates $\lambda_1/[1 + \theta t(\lambda_1 + \lambda_2)]$ and $\lambda_2/[1 + \theta t(\lambda_1 + \lambda_2)]$, respectively. By Example 2.6 the cause-specific hazard rates and the marginal hazard rates are identical when we have independent competing risks. So the crude hazard rates for this set of independent competing risks are identical to the set of dependent competing risks in Example 2.7. This means that given what we actually see, (T, δ), we can never distinguish a pair of dependent competing risks from a pair of independent competing risks.

In competing risks problems we are often interested not in the hazard rate but rather in some probability which summarizes our knowledge about the likelihood of the occurrence of a particular competing risk. Three probabilities are computed, each with their own interpretation. These are the *crude*, *net*, and *partial crude* probabilities. The crude probability is the probability of death from a particular cause in the real world where all other risks are acting on the individual. For example, if the competing risk is death from heart disease, then an example of a crude probability is the chance a man will die from heart disease prior to age 50. The net probability is the probability of death in a hypothetical world where the specific risk is the only risk acting on the population. In the potential failure time model this is a marginal probability for the specified risk. For example, a net probability is the chance that a man will die from heart disease in the counterfactual world where men can only die from heart disease. Partial crude probabilities are the probability of death in a hypothetical world where some risks of death have been eliminated. For example, a partial crude probability would be the chance a man dies from heart disease in a world where cancer has been cured.

Crude probabilities are typically expressed by the cause-specific sub-distribution function. This function, also known as the *cumulative incidence function*, is defined as $F_i(t) = P[T \le t, \delta = i]$. The cumulative incidence function can be computed directly from the joint density function of the potential failure times or it can be computed from the cause specific hazard rates. That is,

$$F_i(t) = P[T \le t, \delta = i] = \int_0^t h_i(u) \exp\{-H_T(u)\}\, du. \tag{2.7.3}$$

Here $H_T(t) = \sum_{j=1}^K \int_0^t h_j(u)\, du$ is the cumulative hazard rate of T. Note that the value of $F_i(t)$ depends on the rate at which all the competing risks occur, not simply on the rate at which the specific cause of interest is occurring. Also, since $h_i(t)$ can be estimated directly from the observed data, $F_i(t)$ is directly estimable without making any assumptions

about the joint distribution of the potential failure times (see section 4.7). $F_i(t)$ is not a true distribution function since $F_i(\infty) = P[\delta = i]$. It has the property that it is non-decreasing with $F_i(0) = 0$ and $F_i(\infty) < 1$. Such a function is called a "sub-distribution" function.

The net survival function, $S_i(t)$, is the marginal survival function found from the joint survival function by taking $t_j = 0$ for all $j \ne i$. When the competing risks are independent then the net survival function is related to the crude probabilities by

$$S_i(t) = \exp\left\{-\int_0^t \frac{dF_i(u)}{S_T(u)}\right\}.$$

This relationship is used in Chapter 4 to allow us to estimate probabilities when there is a single independent competing risk which is regarded as random censoring (see section 3.2 for a discussion of random censoring).

When the risks are dependent, Peterson (1976) shows that net survival probabilities can be bounded by the crude probabilities. He shows that

$$S_T(t) \le S_i(t) \le 1 - F_i(t).$$

The lower (upper) bounds correspond to perfect positive (negative) correlation between the risks. These bounds may be quite wide in practice. Klein and Moeschberger (1988) and Zheng and Klein (1994) show that these bounds can be tightened by assuming a family of dependence structures for the joint distribution of the competing risks.

For partial crude probabilities we let $\mathbf{J}$ be the set of causes that an individual can fail from and $\mathbf{J}^C$ the set of causes which are eliminated from consideration. Let $T^J = \min(X_i, i \in \mathbf{J})$ then we can define the partial crude sub-distribution function by $F_i^J(t) = \Pr[T^J \le t, \delta = i], i \in \mathbf{J}$. Here the ith partial crude probability is the chance of dying from cause i in a hypothetical patient who can only experience one of the causes of death in the set $\mathbf{J}$. One can also define a partial crude hazard rate by

$$\lambda_i^J(t) = \frac{-\partial S(t_1, \ldots, t_K)/\partial t_i\big|_{t_j = t, t_j \in \mathbf{J}, t_j = 0, t_j \in \mathbf{J}^C}}{S(t_1, \ldots, t_p)\big|_{t_j = t, t_j \in \mathbf{J}, t_j = 0, t_j \in \mathbf{J}^C}}. \tag{2.7.4}$$

As in the case of the crude partial incidence function we can express the partial crude sub-distribution function as

$$F_i^J(t) = \Pr[T^J \le t, \delta = i] = \int_0^t \lambda_i^J(x) \exp\left\{-\sum_{j \in \mathbf{J}} \int_0^t \lambda_j^J(u)\, du\right\} dx. \tag{2.7.5}$$

When the risks are independent then the partial crude hazard rate can be expressed in terms of the crude probabilities as

$$\lambda_i^J(t) = \frac{dF_i(t)/dt}{S_T(t)}. \tag{2.7.6}$$

EXAMPLE 2.8 Suppose we have three independent exponential competing risks with hazard rates $\lambda_1, \lambda_2, \lambda_3$, respectively. In this case, as seen in Example 2.6, the net and crude hazard rates for the first competing risk are equal to λ_1. The hazard rate of T is $h_T(t) = \lambda_1 + \lambda_2 + \lambda_3$. Equation (2.7.3), the crude sub-distribution function for the first competing risk is

$$F_1(t) = \int_0^t \lambda_1 \exp(-u(\lambda_1 + \lambda_2 + \lambda_3)\, du$$

$$= \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}\{1 - \exp\{-t(\lambda_1 + \lambda_2 + \lambda_3)\}\}.$$

Note that the crude probability of death from cause 1 in the interval $[0, t]$ is not the same as the net (marginal) probability of death in this interval given by $1 - \exp\{-\lambda_1 t\}$. Also $F_i(\infty) = \lambda_1/(\lambda_1 + \lambda_2 + \lambda_3)$, which is the probability that the first competing risk occurs first. If we consider a hypothetical world where only the first two competing risks are operating ($\mathbf{J} = \{1, 2\}$), the partial crude hazard rates are $\lambda_i^J(t) = \lambda_i$, $i = 1, 2$, and the partial crude sub-distribution function is given by

$$F_1^J(t) = \int_0^t \lambda_1 \exp(-u(\lambda_1 + \lambda_2)\, du = \frac{\lambda_1}{\lambda_1 + \lambda_2}\{1 - \exp\{-t(\lambda_1 + \lambda_2)\}\}.$$

EXAMPLE 2.7 *(continued)* Suppose we have two competing risks with joint survival function $S(t_1, t_2) = [1 + \theta(\lambda_1 t_1 + \lambda_2 t_2)]^{-1/\theta}$, $\theta \geq 0, \lambda_1, \lambda_2 \geq 0$. Here the crude hazard rates are given by $\lambda_i/[1 + \theta t(\lambda_1 + \lambda_2)]$, for $i = 1, 2$. The cause-specific cumulative incidence function for the ith risk is

$$F_i(t) = \int_0^t \frac{\lambda_i}{[1 + \theta x(\lambda_1 + \lambda_2)]} \exp\left\{-\int_0^x \frac{\lambda_1 + \lambda_2}{[1 + \theta u(\lambda_1 + \lambda_2)]}\, du\right\} dx$$

$$= \frac{\lambda_i}{\lambda_1 + \lambda_2}\left\{1 - [1 + \theta t(\lambda_1 + \lambda_2)]^{-1/\theta}\right\}.$$

In Figure 2.11 we plot the cumulative incidence function and the net probability for cause 1 when $\lambda_1 = 1$, $\lambda_2 = 2$, and $\theta = 2$. Here we see clearly that the cumulative incidence function levels off at one-third the probability that the first competing risk fails first. Also we see quite clearly that the crude probability is always less than the net probability.

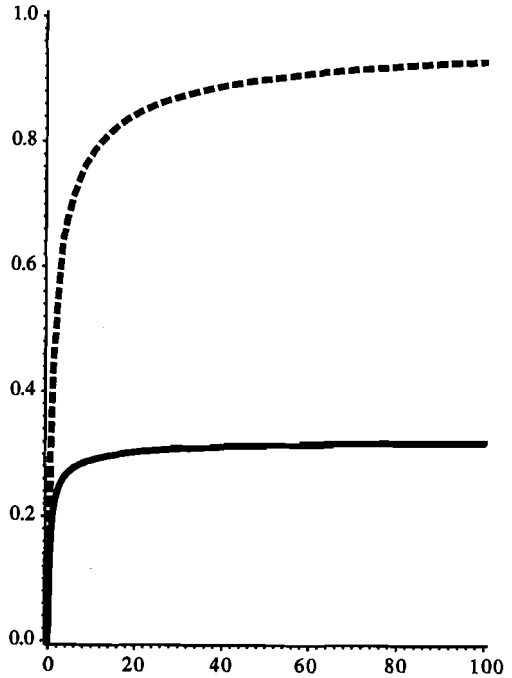

Figure 2.11 *Cumulative incidence function (solid line) and net probability for the first competing risk in Example 2.7.*

Practical Notes

1. Competing risk theory has an intriguing history going back to a memoir read in 1760 by Daniel Bernoulli before the French Academy of Sciences and published in 1765. It was motivated by a controversy on the merits of smallpox inoculation. Using Halley's Breslau life table of 1693, Bernoulli constructed a hypothetical lifetable, which reflected the mortality structure at different ages if smallpox was eliminated. A key assumption was, as Bernoulli recognized, that the hypothetical lifetimes of individuals saved from smallpox were independent of lifetimes associated with the other causes of death. Bernoulli's question "What would be the effect on mortality if the occurrence of one

or more causes of death were changed?" and the untestable assumption of independence of causes of death are still very much with us today.

2. For simplicity, we shall only assume one competing risk, whose event time will be denoted by Y (although all results may be generalized to many competing risks). In the competing-risks framework, as we have seen, we can only observe T = minimum (X, Y) and $\delta = I(X < Y)$, an indicator function which indicates whether or not the main event of interest has occurred. The early observation by Cox (1959, 1962) that there was a difficulty in the interpretation of bivariate data in the competing risk context was elucidated and clarified by later authors. Berman (1963) showed explicitly that the distribution of (T, δ) determined that of X, if X and Y are assumed to be independent. Tsiatis (1975) proved a nonidentifiability theorem which concluded that a dependent-risk model is indistinguishable from some independent risk model and that any analysis of such data should include a careful analysis of biological circumstances. Peterson (1976) argued that serious errors can be made in estimating the survival function in the competing risk problem because one can never know from the data whether X and Y are independent or not.
3. Heckman and Honore (1989) show, under certain regularity conditions, for both proportional hazards and accelerated failure time models that if there is an explanatory covariate, $\mathbf{Z}$, whose support is the entire real line then the joint distribution of (X, Y) is identifiable from $(T, \delta, \mathbf{Z})$. Slud (1992), in a slightly different vein, shows how the marginal distribution of the survival time X can be nonparametrically identifiable when only the data $(T, \delta, \mathbf{Z})$ are observed, where $\mathbf{Z}$ is an observed covariate such that the competing risk event time, Y, and $\mathbf{Z}$ are conditionally independent given X.

Theoretical Notes

1. Slud and Rubinstein (1983) have obtained tighter bounds on $S(x)$ than the Peterson bounds described earlier, in this framework, by utilizing some additional information. Their method requires the investigator to bound the function

$$\rho(t) = \frac{\{[f_i(t)/q_i(t)] - 1\}}{\{[S_i(t)/S_T(t)] - 1\}}$$

where

$$f_i(t) = -\frac{dS_i(t)}{dt},$$

and

$$q_i(t) = \frac{d}{dt}F_i(t).$$

Knowledge of the function $\rho(t)$ and the observable information, (T, δ), is sufficient to determine uniquely the marginal distribution of X. The resulting estimators $\hat{S}_\rho(x)$ are decreasing functions of $\rho(\cdot)$. These resulting bounds are obtained by the investigator's specification of two functions, $\rho_i(t)[\rho_1(t) < \rho_2(t)]$, so that if the true $\rho(t)$ function is in the interval $[\rho_1(t) < \rho_2(t)]$, for all t, then $\hat{S}_{\rho_2}(t) \leq S(t) \leq \hat{S}_{\rho_1}(t)$.

2. Pepe (1991) and Pepe and Mori (1993) interpret the cumulative incidence function as a "marginal probability." Note that this function is not a true marginal distribution as discussed earlier but rather is the chance that the event of interest will occur prior to time t in a system where an individual is exposed to both risks. Pepe and Mori suggest as an alternative to the cumulative incidence function the "conditional probability" of X, defined by

$$P(\{X \leq t, X < Y\} \mid \{Y < t, X > Y\}^c) = \frac{F_i(t)}{F_i^c(t)},$$

which they interpret as the probability of X's occurring in $[0, t)$, given nonoccurrence of Y in $[0, t)$, where F^c denotes the complement of F.

2.8 Exercises

2.1 The lifetime of light bulbs follows an exponential distribution with a hazard rate of 0.001 failures per hour of use.

(a) Find the mean lifetime of a randomly selected light bulb.

(b) Find the median lifetime of a randomly selected light bulb.

(c) What is the probability a light bulb will still function after 2,000 hours of use?

2.2 The time in days to development of a tumor for rats exposed to a carcinogen follows a Weibull distribution with $\alpha = 2$ and $\lambda = 0.001$.

(a) What is the probability a rat will be tumor free at 30 days? 45 days? 60 days?

(b) What is the mean time to tumor? (Hint $\Gamma(0.5) = \sqrt{\pi}$.)

(c) Find the hazard rate of the time to tumor appearance at 30 days, 45 days, and 60 days.

(d) Find the median time to tumor.

2.3 The time to death (in days) following a kidney transplant follows a log logistic distribution with $\alpha = 1.5$ and $\lambda = 0.01$.

(a) Find the 50, 100, and 150 day survival probabilities for kidney transplantation in patients.

(b) Find the median time to death following a kidney transplant.

(c) Show that the hazard rate is initially increasing and, then, decreasing over time. Find the time at which the hazard rate changes from increasing to decreasing.

(d) Find the mean time to death.

2.4 A model for lifetimes, with a bathtub-shaped hazard rate, is the exponential power distribution with survival function $S(x) = \exp\{1 - \exp[(\lambda x)^{\alpha}]\}$.

(a) If $\alpha = 0.5$, show that the hazard rate has a bathtub shape and find the time at which the hazard rate changes from decreasing to increasing.

(b) If $\alpha = 2$, show that the hazard rate of x is monotone increasing.

2.5 The time to death (in days) after an autologous bone marrow transplant, follows a log normal distribution with $\mu = 3.177$ and $\sigma = 2.084$. Find

(a) the mean and median times to death;

(b) the probability an individual survives 100, 200, and 300 days following a transplant; and

(c) plot the hazard rate of the time to death and interpret the shape of this function.

2.6 The Gompertz distribution is commonly used by biologists who believe that an exponential hazard rate should occur in nature. Suppose that the time to death in months for a mouse exposed to a high dose of radiation follows a Gompertz distribution with $\theta = 0.01$ and $\alpha = 0.25$. Find

(a) the probability that a randomly chosen mouse will live at least one year,

(b) the probability that a randomly chosen mouse will die within the first six months, and

(c) the median time to death.

2.7 The time to death, in months, for a species of rats follows a gamma distribution with $\beta = 3$ and $\lambda = 0.2$. Find

(a) the probability that a rat will survive beyond age 18 months,

(b) the probability that a rat will die in its first year of life, and

(c) the mean lifetime for this species of rats.

2.8 The battery life of an internal pacemaker, in years, follows a Pareto distribution with $\theta = 4$ and $\lambda = 5$.

(a) What is the probability the battery will survive for at least 10 years?

(b) What is the mean time to battery failure?

(c) If the battery is scheduled to be replaced at the time t_o, at which 99% of all batteries have yet to fail (that is, at t_o so that $Pr(X > t_o) = .99$), find t_o.

2.9 The time to relapse, in months, for patients on two treatments for lung cancer is compared using the following log normal regression model:

$$Y = \text{Ln}(X) = 2 + 0.5Z + 2W,$$

where W has a standard normal distribution and $Z = 1$ if treatment A and 0 if treatment B.

(a) Compare the survival probabilities of the two treatments at 1, 2, and 5 years.

(b) Repeat the calculations if W has a standard logistic distribution. Compare your results with part (a).

2.10 A model used in the construction of life tables is a piecewise, constant hazard rate model. Here the time axis is divided into k intervals, $[\tau_{i-1}, \tau_i)$, $i = 1, \ldots, k$, with $\tau_o = 0$ and $\tau_k = \infty$. The hazard rate on the ith interval is a constant value, θ_i; that is

$$h(x) = \begin{cases} \theta_1 & 0 \le x < \tau_1 \\ \theta_2 & \tau_1 \le x < \tau_2 \\ \vdots & \\ \theta_{k-1} & \tau_{k-2} \le x < \tau_{k-1} \\ \theta_k & x \ge \tau_{k-1} \end{cases}$$

(a) Find the survival function for this model.

(b) Find the mean residual-life function.

(c) Find the median residual-life function.

2.11 In some applications, a third parameter, called a guarantee time, is included in the models discussed in this chapter. This parameter ϕ is the smallest time at which a failure could occur. The survival function of the three-parameter Weibull distribution is given by

$$S(x) = \begin{cases} 1 & \text{if } x < \phi \\ \exp[-\lambda(x - \phi)^{\alpha}] & \text{if } x \ge \phi. \end{cases}$$

(a) Find the hazard rate and the density function of the three- parameter Weibull distribution.

(b) Suppose that the survival time X follows a three-parameter Weibull distribution with $\alpha = 1$, $\lambda = 0.0075$ and $\phi = 100$. Find the mean and median lifetimes.

2.12 Let X have a uniform distribution on the interval 0 to θ with density function

$$f(x) = \begin{cases} 1/\theta, & \text{for } 0 \le x \le \theta \\ 0, & \text{otherwise.} \end{cases}$$

(a) Find the survival function of X.

(b) Find the hazard rate of X.

(c) Find the mean residual-life function.

2.13 Suppose that X has a geometric distribution with probability mass function

$$p(x) = p(1-p)^{x-1}, x = 1, 2, \ldots$$

(a) Find the survival function of X. (Hint: Recall that for $0 < \theta < 1$, $\sum_{j=k}^{\infty} \theta^j = \theta^k/(1-\theta)$.

(b) Find the hazard rate of X. Compare this rate to the hazard rate of an exponential distribution.

2.14 Suppose that a given individual in a population has a survival time which is exponential with a hazard rate θ. Each individual's hazard rate θ is potentially different and is sampled from a gamma distribution with density function

$$f(\theta) = \frac{\lambda^\beta \theta^{\beta-1} e^{-\lambda\theta}}{\Gamma(\beta)}$$

Let X be the life length of a randomly chosen member of this population.

(a) Find the survival function of X.

(Hint: Find $S(x) = E_\theta[e^{-\theta x}]$.)

(b) Find the hazard rate of X. What is the shape of the hazard rate?

2.15 Suppose that the hazard rate of X is a linear function $h(x) = \alpha + \beta x$, with α and $\beta > 0$. Find the survival function and density function of x.

2.16 Given a covariate Z, suppose that the log survival time Y follows a linear model with a logistic error distribution, that is,

$$Y = \ln(X) = \mu + \beta Z + \sigma W \text{ where the pdf of } W \text{ is given by}$$

$$f(w) = \frac{e^w}{(1+e^w)^2}, -\infty < w < \infty.$$

(a) For an individual with covariate Z, find the conditional survival function of the survival time X, given Z, namely, $S(x \mid Z)$.

(b) The odds that an individual will die prior to time x is expressed by $[1 - S(x \mid Z)]/S(x \mid Z)$. Compute the odds of death prior to time x for this model.

(c) Consider two individuals with different covariate values. Show that, for any time x, the ratio of their odds of death is independent of x. The log logistic regression model is the only model with this property.

2.17 Suppose that the mean residual life of a continuous survival time X is given by $\text{MRL}(x) = x + 10$.

(a) Find the mean of X.

(b) Find $h(x)$.

(c) Find $S(x)$.

2.18 Let X have a uniform distribution on 0 to 100 days with probability density function

$$f(x) = 1/100 \text{ for } 0 < x < 100,$$
$$= 0, \text{ elsewhere.}$$

(a) Find the survival function at 25, 50, and 75 days.

(b) Find the mean residual lifetime at 25, 50, and 75 days.

(c) Find the median residual lifetime at 25, 50, and 75 days.

2.19 Suppose that the joint survival function of the latent failure times for two competing risks, X and Y, is

$$S(x, y) = (1-x)(1-y)(1+.5xy), \quad 0 < x < 1, \quad 0 < y < 1.$$

(a) Find the marginal survival function for x.

(b) Find the cumulative incidence of T_1.

2.20 Let X and Y be two competing risks with joint survival function

$$S(x, y) = \exp\{-x - y - .5xy\}, 0 < x, y.$$

(a) Find the marginal cumulative distribution function of X.

(b) Find the cumulative incidence function of X.