3 Censoring and Truncation

3.1 Introduction

Time-to-event data present themselves in different ways which create special problems in analyzing such data. One peculiar feature, often present in time-to-event data, is known as censoring, which, broadly speaking, occurs when some lifetimes are known to have occurred only within certain intervals. The remainder of the lifetimes are known exactly. There are various categories of censoring, such as right censoring, left censoring, and interval censoring. Right censoring will be discussed in section 3.2. Left or interval censoring will be discussed in section 3.3. To deal adequately with censoring in the analysis, we must consider the design which was employed to obtain the survival data. There are several types of censoring schemes within both left and right censoring. Each type will lead to a different likelihood function which will be the basis for the inference. As we shall see in section 3.5, though the likelihood function is unique for each type of censoring, there is a common approach to be used in constructing it.

A second feature which may be present in some survival studies is that of *truncation*, discussed in section 3.4. Left truncation occurs when subjects enter a study at a particular age (not necessarily the origin for the event of interest) and are followed from this *delayed entry time* until the event occurs or until the subject is censored. Right truncation occurs when only individuals who have experienced the event of interest are observable. The main impact on the analysis, when data are truncated, is that the investigator must use a conditional distribution in constructing the likelihood, as shown in section 3.5, or employ a statistical method which uses a selective risk set to be explained in more detail in Chapter 4.

Sections 3.5 and 3.6 present an overview of some theoretical results needed to perform modern survival analysis. Section 3.5 shows the construction of likelihoods for censored and truncated data. These likelihoods are the basis of inference techniques for parametric models and, suitably modified, as partial likelihoods for semiparametric models. Section 3.6 gives a brief introduction to the theory of counting processes. This very general theory is used to develop most nonparametric techniques for censored and truncated data and is the basis for developing the statistical properties of both parametric and nonparametric methods in survival analysis.

3.2 Right Censoring

First, we will consider *Type I censoring* where the event is observed only if it occurs prior to some prespecified time. These censoring times may vary from individual to individual. A typical animal study or clinical trial starts with a fixed number of animals or patients to which a treatment (or treatments) is (are) applied. Because of time or cost considerations, the investigator will terminate the study or report the results before all subjects realize their events. In this instance, if there are no accidental losses or subject withdrawals, all censored observations have times equal to the length of the study period.

Generally, it is our convention that random variables are denoted by upper case letters and fixed quantities or realizations of random variables are denoted by lower case letters. With censoring, this convention will obviously present some difficulties in notation because, as we shall see, some censoring times are fixed and some are random. At the risk of causing some confusion we will stick to upper case letters for censoring times. The reader will be expected to determine from the context whether the censoring time is random or fixed.

In right censoring, it is convenient to use the following notation. For a specific individual under study, we assume that there is a lifetime X and a fixed censoring time, C_r (C_r for "right" censoring time). The X's are assumed to be independent and identically distributed with probability density function $f(x)$ and survival function $S(x)$. The exact lifetime X of an individual will be known if, and only if, X is less than or equal to C_r. If X is greater than C_r, the individual is a survivor, and his or her event time is censored at C_r. The data from this experiment can be conveniently represented by pairs of random variables (T, δ), where δ indicates whether the lifetime X corresponds to an event ($\delta = 1$) or is censored ($\delta = 0$), and T is equal to X, if the lifetime is observed, and to C_r if it is censored, i.e., $T = \min(X, C_r)$.

EXAMPLE 3.1 Consider a large scale animal experiment conducted at the National Center for Toxicological Research (NCTR) in which mice were fed a particular dose of a carcinogen. The goal of the experiment was to assess the effect of the carcinogen on survival. Toward this end, mice were followed from the beginning of the experiment until death or until a prespecified censoring time was reached, when all those still alive were sacrificed (censored). This example is illustrated in Figure 3.1.

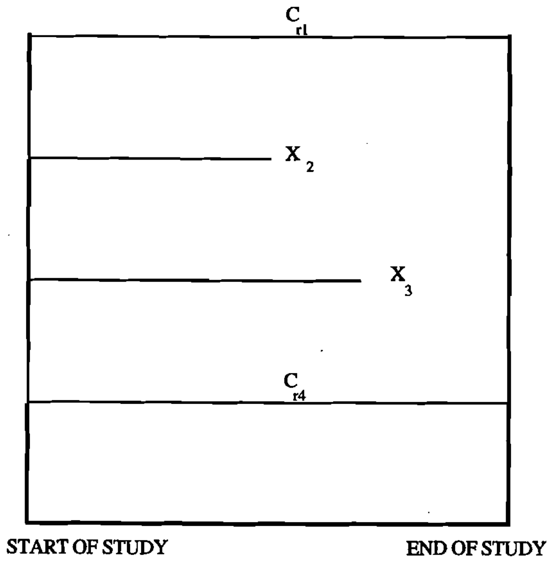

Figure 3.1 *Example of Type I censoring*

When animals have different, fixed-sacrifice (censoring) times, this form of Type I censoring is called *progressive Type I censoring*. An advantage of this censoring scheme is that the sacrificed animals give information on the natural history of nonlethal diseases. This type of censoring is illustrated in the following example.

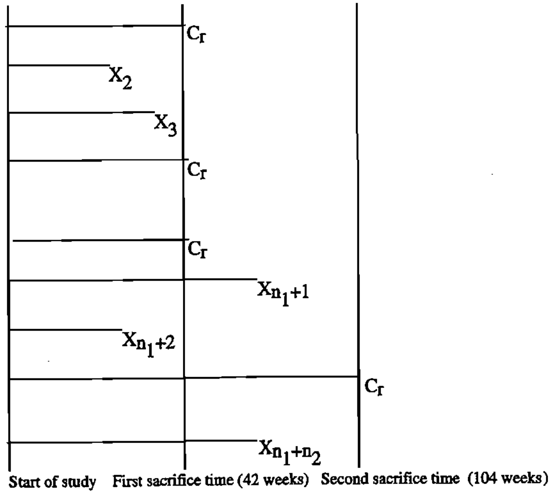
Cr
X2
X3
Cr
Cr
Xn1+1
Xn1+2
Cr
Xn1+n2
Start of study
First sacrifice time (42 weeks)
Second sacrifice time (104 weeks)

Figure 3.2 *Type I censoring with two different sacrifice times*

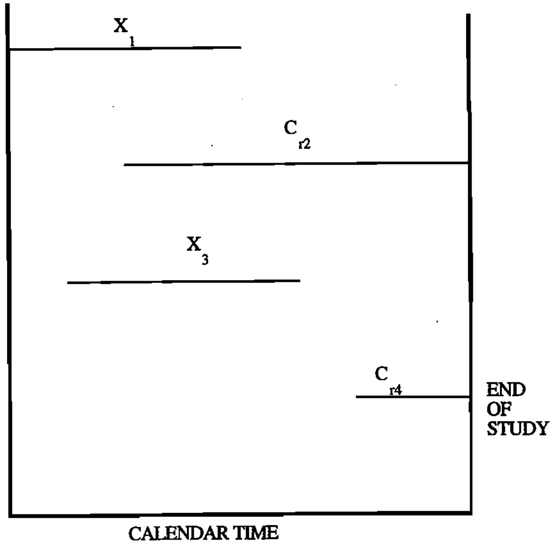

Figure 3.3 *Generalized Type I censoring when each individual has a different starting time*

EXAMPLE 3.2 Consider a mouse study where, for each sex, 200 mice were randomly divided into four dose-level groups and each mouse was followed until death or until a prespecified sacrifice time (42 or 104 weeks) was reached (see Figure 3.2 for a schematic of this trial for one gender and one dose level). The two sacrifice times were chosen to reduce the cost of maintaining the animals while allowing for limited information on the survival experience of longer lived mice.

Another instance, which gives rise to Type I censoring, is when individuals enter the study at different times and the terminal point of the study is predetermined by the investigator, so that the censoring times are known when an individual is entered into the study. In such studies (see Figure 3.3 for a hypothetical study with only four subjects), individuals have their own specific, fixed, censoring time. This form of censoring has been termed *generalized Type I censoring* (cf. David and Moeschberger, 1978). A convenient representation of such data is to shift each individual's starting time to 0 as depicted in Figure 3.4. Another method for representing such data is the Lexis diagram (Keiding, 1990). Here calendar time is on the horizontal axis, and life length is represented by a 45° line. The time an individual spends on study is represented by the height of the ray on the vertical axis. Figure 3.5 shows a Lexis diagram for the generalized Type I censoring scheme depicted in Figure 3.4. Here patients 1 and 3 experience the event of interest prior to the end of the study and are exact observations with $\delta = 1$. Patients 2 and 4, who experience the event after the end of the study, are only known to be alive at the end of the study and are censored observations ($\delta = 0$). Examples of studies with generalized Type I censoring are the breast-cancer trial in section 1.5, the acute leukemia trial in section 1.2, the study of psychiatric patients in section 1.15, and the study of weaning of newborns in section 1.14.

A second type of right censoring is *Type II censoring* in which the study continues until the failure of the first r individuals, where r is some predetermined integer ($r < n$). Experiments involving Type II censoring are often used in testing of equipment life. Here, all items are put on test at the same time, and the test is terminated when r of the n items have failed. Such an experiment may save time and money because it could take a very long time for all items to fail. It is also true that the statistical treatment of Type II censored data is simpler because the data consists of the r smallest lifetimes in a random sample of n lifetimes, so that the theory of order statistics is directly applicable to determining the likelihood and any inferential technique employed. Here, it should be noted that r the number of failures and $n - r$ the

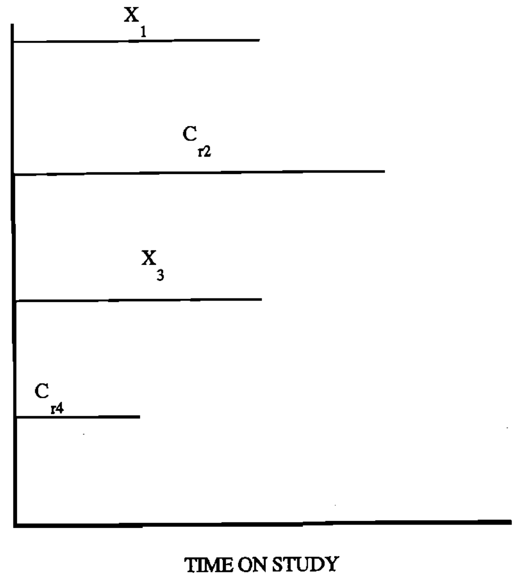

Figure 3.4 *Generalized Type I censoring for the four individuals in Figure 3.3 with each individuals starting time backed up to* 0. $T_1 = X_1$ *(death time for first individual)* ($\delta_1 = 1$); $T_2 = C_{r2}$ *(right censored time for second individual)* ($\delta_2 = 0$); $T_3 = X_3$ *(death time for third individual)* ($\delta_3 = 1$); $T_4 = C_{r4}$ *(right censored time for fourth individual)* ($\delta_4 = 0$).

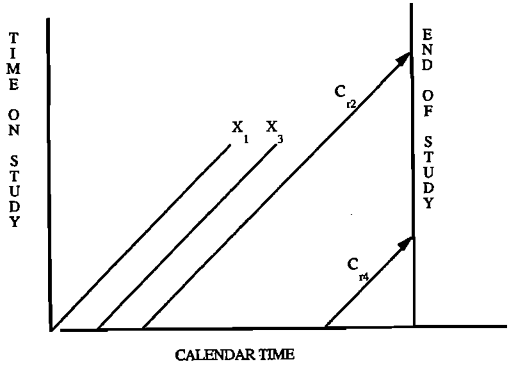

Figure 3.5 *Lexis diagram for generalized Type I censoring in Figure 3.3*

number of censored observations are fixed integers and the censoring time $T_{(r)}$, the rth ordered lifetime is random.

A generalization of Type II censoring, similar to the generalization of Type I censoring with different sacrifice times, is *progressive Type II censoring*. Here the first r_1 failures (an integer chosen prior to the start of the study) in a sample of n items (or animals) are noted and recorded. Then $n_1 - r_1$ of the remaining $n - r_1$ unfailed items (or animals) are removed (or sacrificed) from the experiment, leaving $n - n_1$ items (or animals) on study. When the next r_2 items (another integer chosen prior to the start of the study) fail, $n_2 - r_2$ of the unfailed items are removed (or animals sacrificed). This process continues until some predecided series of repetitions is completed. Again, r_i and n_i $(i = 1, 2)$ are fixed integers and the two censoring times, $T_{(n_1)}$ and $T_{(n_1+n_2)}$, are random.

A third type of right censoring is *competing risks* censoring. A special case of competing risks censoring is random censoring. This type of censoring arises when we are interested in estimation of the marginal distribution of some event but some individuals under study may experience some competing event which causes them to be removed from the study. In such cases, the event of interest is not observable for those who experience the competing event and these subjects are random right censored at that time. As shown in section 2.7, in the competing risk framework, to be able to identify the marginal distribution from competing risks data we need the event time and censoring times to be independent of each other. This relationship cannot be determined from the data alone. Typical examples of where the random censoring times may be thought to be independent of the main event time of interest are accidental deaths, migration of human populations, and so forth.

Whenever we encounter competing risks it is important to determine precisely what quantity we wish to estimate. We need to decide if we want to estimate a marginal (net), crude, or partial crude probability as discussed in section 2.7. If we wish to estimate a marginal probability, which is the chance of the event's occurring in a world where all other risks cannot occur, the other competing risks are random observations. Here we need an assumption of independence between the time to the event of interest and the competing events to make a meaningful inference. Techniques for estimation in this framework are discussed in sections 4.1–4.6. When interest centers on estimation of crude probabilities (that is, the probability of the event in the real world where a person can fail from any of the competing causes), then each competing risk is modeled by a cumulative incidence curve (see section 4.7) and no independence assumption is needed. For partial crude probabilities (that is, the probability of the event's occurring in a world where only a subset of competing risks are possible causes of failure) some of the competing risks are treated as random censored observations (those to be eliminated) and others are modeled by a cumulative incidence curve. In this case we require that those causes treated as random censored observations need to be independent of the other causes to obtain consistent estimates of the desired probabilities.

In many studies, the censoring scheme is a combination of random and Type I censoring. In such studies, some patients are randomly censored when, for example, they move from the study location for reasons unrelated to the event of interest, whereas others are Type I censored when the fixed study period ends.

Theoretical Note

1. In Type I progressive censoring, the sacrifice times are fixed (predetermined prior to the start of the study), whereas, in Type II progressive censoring, the sacrifice times are random times at which a predetermined number of deaths has occurred. This distinction is extremely important in constructing the likelihood function in section 3.5. An advantage of either type of censoring scheme is that the sacrificed animals give information on the natural history of nonlethal diseases.

3.3 Left or Interval Censoring

A lifetime X associated with a specific individual in a study is considered to be *left censored* if it is less than a censoring time C_l(C_l for "left" censoring time), that is, the event of interest has already occurred for the individual before that person is observed in the study at time C_l. For such individuals, we know that they have experienced the event sometime before time C_l, but their exact event time is unknown. The exact lifetime X will be known if, and only if, X is greater than or equal to C_l. The data from a left-censored sampling scheme can be represented by pairs of random variables (T, ε), as in the previous section, where T is equal to X if the lifetime is observed and ε indicates whether the exact lifetime X is observed ($\varepsilon = 1$) or not ($\varepsilon = 0$). Note that, for left censoring as contrasted with right censoring, $T = \max(X, C_l)$.

EXAMPLE 3.3 In a study to determine the distribution of the time until first marijuana use among high school boys in California, discussed in section 1.17, the question was asked, When did you you first use marijuana?" One of the responses was "I have used it but can not recall just when the first time was." A boy who chose this response is indicating that the event had occurred prior to the boy's age at interview but the exact age at

which he started using marijuana is unknown. This is an example of a left-censored event time.

EXAMPLE 3.4 In early childhood learning centers, interest often focuses upon testing children to determine when a child learns to accomplish certain specified tasks. The age at which a child learns the task would be considered the time-to-event. Often, some children can already perform the task when they start in the study. Such event times are considered left censored.

Often, if left censoring occurs in a study, right censoring may also occur, and the lifetimes are considered *doubly censored* (cf. Turnbull, 1974). Again, the data can be represented by a pair of variables (T, δ), where $T = \max[\min(X, C_r), C_l]$ is the on study time; δ is 1 if T is a death time, 0 if T is a right-censored time, and -1 if T is a left-censored time. Here C_l is the time before which some individuals experience the event and C_r is the time after which some individuals experience the event. X will be known exactly if it is less than or equal to C_r and greater than or equal to C_l.

EXAMPLE 3.3 *(continued)* An additional possible response to the question "When did you first use marijuana?" was "I never used it" which indicates a right-censored observation. In the study described in section 1.17, both left-censored observations and right-censored observations were present, in addition to knowing the exact age of first use of marijuana (uncensored observations) for some boys. Thus, this is a doubly censored sampling scheme.

EXAMPLE 3.4 *(continued)* Some children undergoing testing, as described in Example 3.4, may not learn the task during the entire study period, in which case such children would be right-censored. Coupled with the left-censored observations discussed earlier, this sample would also contain doubly censored data.

A more general type of censoring occurs when the lifetime is only known to occur within an interval. Such *interval censoring* occurs when patients in a clinical trial or longitudinal study have periodic follow-up and the patient's event time is only known to fall in an interval $(L_i, R_i]$ (L for left endpoint and R for right endpoint of the censoring interval). This type of censoring may also occur in industrial experiments where there is periodic inspection for proper functioning of equipment items. Animal tumorigenicity experiments may also have this characteristic.

EXAMPLE 3.5 In the Framingham Heart Study, the ages at which subjects first developed coronary heart disease (CHD) are usually known exactly. However, the ages of first occurrence of the subcategory angina pectoris may be known only to be between two clinical examinations, approximately two years apart (Odell et al., 1992). Such observations would be interval-censored.

EXAMPLE 3.6 In section 1.18, the data from a retrospective study to compare the cosmetic effects of radiotherapy alone versus radiotherapy and adjuvant chemotherapy on women with early breast cancer are reported. Patients were observed initially every 4–6 months but, as their recovery progressed, the interval between visits lengthened. The event of interest was the first appearance of moderate or severe breast retraction, a cosmetic deterioration of the breast. The exact time of retraction was known to fall only in the interval between visits (interval-censored) or after the last time the patient was seen (right-censored).

In view of the last two examples, it is apparent that any combination of left, right, or interval censoring may occur in a study. Of course, interval censoring is a generalization of left and right censoring because, when the left end point is 0 and the right end point is C_l we have left censoring and, when the left end point is C_r and the right end point is infinite, we have right censoring.

The main impact on the analysis, when data are truncated, is that the investigator must use a conditional distribution in constructing the likelihood, as shown in section 3.5, or employ a statistical method which uses a selective risk set, explained in more detail in section 4.6.

3.4 Truncation

A second feature of many survival studies, sometimes confused with censoring, is *truncation*. Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed. An individual whose event time is not in this interval is not observed and no information on this subject is available to the investigator. This is in contrast to censoring where there is at least partial information on each subject. Because we are only aware of individuals with event times in the observational window, the inference for truncated data is restricted to conditional estimation.

When Y_R is infinite then we have *left truncation*. Here we only observe those individuals whose event time X exceeds the truncation

time Y_L. That is we observe X if and only if $Y_L < X$. A common example of left truncation is the problem of estimating the distribution of the diameters of microscopic particles. The only particles big enough to be seen based on the resolution of the microscope are observed and smaller particles do not come to the attention of the investigator. In survival studies the truncation event may be exposure to some disease, diagnosis of a disease, entry into a retirement home, occurrence of some intermediate event such as graft-versus-host disease after a bone marrow transplantation, etc. In this type of truncation any subjects who experience the event of interest prior to the truncation time are not observed. The truncation time is often called a *delayed entry time* since we only observe subjects from this time until they die or are censored. Note that, as opposed to left censoring where we have partial information on individuals who experience the event of interest prior to age at entry, for left truncation these individuals were never considered for inclusion into the study.

EXAMPLE 3.7 In section 1.16, a survival study of residents of the Channing House retirement center located in California is described. Ages at death (in months) are recorded, as well as ages at which individuals entered the retirement community (the truncation event). Since an individual must survive to a sufficient age to enter the retirement center, all individuals who died earlier will not enter the center and thus are out of the investigator's cognizance; i.e., such individuals have no chance to be in the study and are considered left truncated. A survival analysis of this data set needs to account for this feature.

Right truncation occurs when Y_L is equal to zero. That is, we observe the survival time X only when $X \leq Y_R$. Right truncation arises, for example, in estimating the distribution of stars from the earth in that stars too far away are not visible and are right truncated. A second example of a right-truncated sample is a mortality study based on death records. Right-censored data is particularly relevant to studies of AIDS.

EXAMPLE 3.8 Consider the AIDS study described in section 1.19. Here cases of patients with transfusion-induced AIDS were sampled. Retrospective determination of the transfusion times were used to estimate the waiting time from infection at transfusion to clinical onset of AIDS. The registry was sampled on June 30, 1986, so only those whose waiting time from transfusion to AIDS was less than the time from transfusion to June 30, 1986, were available for observation. Patients transfused prior to June 30, 1986, who developed AIDS after June 30, 1986, were not observed and are right truncated.

The main impact on the analysis when data are truncated is that the investigator must use a conditional distribution in constructing the likelihood, as shown in section 3.5, or employ a statistical method which uses a selective risk set, which will be explained in more detail in section 4.6.

3.5 Likelihood Construction for Censored and Truncated Data

As stated previously, the design of survival experiments involving censoring and truncation needs to be carefully considered when constructing likelihood functions. A critical assumption is that the lifetimes and censoring times are independent. If they are not independent, then specialized techniques must be invoked. In constructing a likelihood function for censored or truncated data we need to consider carefully what information each observation gives us. An observation corresponding to an exact event time provides information on the probability that the event's occurring at this time, which is approximately equal to the density function of X at this time. For a right-censored observation all we know is that the event time is larger than this time, so the information is the survival function evaluated at the on study time. Similarly for a left-censored observation, all we know is that the event has already occurred, so the contribution to the likelihood is the cumulative distribution function evaluated at the on study time. Finally, for interval-censored data we know only that the event occurred within the interval, so the information is the probability that the event time is in this interval. For truncated data these probabilities are replaced by the appropriate conditional probabilities.

More specifically, the likelihoods for various types of censoring schemes may all be written by incorporating the following components:

exact lifetimes	-	$f(x)$
right-censored observations	-	$S(C_r)$
left-censored observations	-	$1 - S(C_l)$
interval-censored observations	-	$[S(L) - S(R)]$
left-truncated observations	-	$f(x)/S(Y_L)$
right-truncated observations	-	$f(x)/[1 - S(Y_R)]$
interval-truncated observations	-	$f(x)/[S(Y_L) - S(Y_R)]$

The likelihood function may be constructed by putting together the component parts as

$$L \propto \prod_{i\varepsilon D} f(x_i) \prod_{i\varepsilon R} S(C_r) \prod_{i\varepsilon L} (1 - S(C_l)) \prod_{i\varepsilon I} [S(L_i) - S(R_i)], \qquad (3.5.1)$$

where D is the set of death times, R the set of right-censored observations, L the set of left-censored observations, and I the set of interval-censored observations. For left-truncated data, with truncation interval (Y_{Li}, Y_{Ri}) independent from the jth death time, we replace $f(X_i)$ by

$f(x_i)/[S(Y_{Li}) - S(Y_{Ri})]$ and $S(C_i)$ by $S(C_i)/[S(Y_{Li}) - S(Y_{Ri})]$ in (3.5.1).

For right-truncated data, only deaths are observed, so that the likelihood is of the form

$$L \propto \prod_i f(Y_i)/[1 - S(Y_i)].$$

If each individual has a different failure distribution, as might be the case when regression techniques are used,

$$L = \prod_{i\varepsilon D} f_i(x_i) \prod_{i\varepsilon R} S_i(C_r) \prod_{i\varepsilon L} [1 - S_i(C_l)] \prod_{i\varepsilon I} [S_i(L_i) - S_i(R_i)]. \qquad (3.5.2)$$

We will proceed with explicit details in constructing the likelihood function for various types of censoring and show how they all basically lead to equation (3.5.1).

Data from experiments involving right censoring can be conveniently represented by pairs of random variables (T, δ), where δ indicates whether the lifetime X is observed ($\delta = 1$) or not ($\delta = 0$), and T is equal to X if the lifetime is observed and to C_τ if it is right-censored, i.e., $T = \min(X, C_\tau)$.

Details of constructing the likelihood function for *Type I censoring* are as follows. For $\delta = 0$, it can be seen that

$$\begin{aligned} Pr[T, \delta = 0] &= Pr[T = C_\tau \mid \delta = 0] Pr[\delta = 0] = Pr(\delta = 0) \\ &= Pr(X > C_\tau) = S(C_\tau). \end{aligned}$$

Also, for $\delta = 1$,

$$\begin{aligned} Pr(T, \delta = 1) &= Pr(T = X \mid \delta = 1) Pr(\delta = 1), \\ &= Pr(X = T \mid X \leq C_\tau) Pr(X \leq C_\tau) \\ &= \left[\frac{f(t)}{1 - S(C_\tau)}\right] [1 - S(C_\tau)] = f(t). \end{aligned}$$

These expressions can be combined into the single expression

$$Pr(t, \delta) = [f(t)]^{\delta} [S(t)]^{1-\delta}.$$

If we have a random sample of pairs (T_i, δ_i), $i = 1, \ldots, n$, the likelihood function is

$$L = \prod_{i=1}^{n} Pr[t_i, \delta_i] = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \qquad (3.5.3)$$

which is of the same form as (3.5.1). Because we can write $f(t_i) = h(t_i)S(t_i)$ we can write this likelihood as

$$L = \prod_{i=1}^{n}[h(t_i)]^{\delta_i}\exp[-H(t_i)]$$

EXAMPLE 3.9 Assume $f(x) = \lambda e^{-\lambda x}$.

Then, the likelihood function is

$$\begin{aligned} L_I &= \prod_{i=1}^{n}(\lambda e^{-\lambda t_i})^{\delta_i}\exp[-\lambda t_i(1-\delta_i)] \qquad (3.5.4)\\ &= \lambda^r \exp[-\lambda S_T], \end{aligned}$$

where $r = \sum \delta_i$ is the observed number of events and S_T is the total time on test for all n individuals under study.

EXAMPLE 3.10 A simple random censoring process encountered frequently is one in which each subject has a lifetime X and a censoring time C_r, X and C_r being independent random variables with the usual notation for the probability density and survival function of X as in Type I censoring and the p.d.f. and survival function of C_r denoted by $g(c_r)$ and $G(c_r)$, respectively. Furthermore, let $T = \min(X, C_r)$ and δ indicates whether the lifetime X is censored ($\delta = 0$) or not ($\delta = 1$). The data from a sample of n subjects consist of the pairs (t_i, δ_i), $i = 1, \ldots, n$. The density function of this pair may be obtained from the joint density function of X and C_r, $f(x, c_r)$, as

$$\begin{aligned} Pr(T_i = t, \delta = 0) &= Pr(C_{r,i} = t, X_i > C_{r,i})\\ &= \frac{d}{dt}\int_0^t \int_v^\infty f(u, v)\,du\,dv. \qquad (3.5.5) \end{aligned}$$

When X and C_r are independent with marginal densities f and g, respectively, (3.5.5) becomes

$$\begin{aligned} &= \frac{d}{dt}\int_0^t \int_v^\infty f(u)g(v)\,du\,dv\\ &= \frac{d}{dt}\int_0^t S(v)g(v)\,dv\\ &= S(t)g(t) \end{aligned}$$

and, similarly,

$$Pr(T_i = t, \delta = 1) = Pr(X_i = t, X_i < C_r, i) = f(t)G(t).$$

So,

$$\begin{aligned} L &= \prod_{i=1}^{n}[f(t_i)G(t_i)]^{\delta_i}[g(t_i)S(t_i)]^{1-\delta_i}\\ &= \left\{\prod_{i=1}^{n} G(t_i)^{\delta_i} g(t_i)^{1-\delta_i}\right\}\left\{\prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}\right\}. \end{aligned}$$

If the distribution of the censoring times, as alluded to earlier, does not depend upon the parameters of interest, then, the first term will be a constant with respect to the parameters of interest and the likelihood function takes the form of (3.5.1)

$$L \propto \prod_{i=1}^{n}[f(t_i)]^{\delta_i}[S(t_i)]^{1-\delta_i}. \qquad (3.5.6)$$

Practical Notes

1. The likelihoods constructed in this section are used primarily for analyzing parametric models, as discussed in Chapter 12. They also serve as a basis for determining the partial likelihoods used in the semiparametric regression methods discussed in Chapters 8 and 9.
2. Even though, in most applications of analyzing survival data, the likelihoods constructed in this section will not be explicitly used, the rationale underlying their construction has value in understanding the contribution of the individual data components depicted in (3.5.1).

Theoretical Notes

1. For Type II censoring, the data consist of the rth smallest lifetimes $X_{(1)} \le X_{(2)} \le \cdots \le X_{(r)}$ out of a random sample of n lifetimes $X_1, \ldots, X_n$ from the assumed life distribution. Assuming $X_1, \ldots, X_n$ are i.i.d. and have a continuous distribution with p.d.f. $f(x)$ and survival function $S(x)$, it follows that the joint p.d.f. of $X_{(1)}, \ldots, X_{(r)}$ is (cf. David, 1981)

$$L_{II,1} = \frac{n!}{(n-r)!}\left[\prod_{i=1}^{r} f(x_{(i)})\right] [S(x_{(r)})]^{n-r}. \qquad (3.5.7)$$

2. For simplicity, in the progressive Type II censoring case, assume that the censoring (or serial sacrifice) has just two repetitions. Here we observe the r_1 ordered failures $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(r_1)}$, then, n_1 items are removed from the study and sacrificed. Of the remaining $(n - r_1 - n_1)$ items we observe the next r_2 ordered failures $X^*_{(1)} \leq X*_{(2)} \leq \cdots \leq X^*_{(r_2)}$ after which the study stops with the remaining $n - n_1 - r_1 - r_2$ items being censored at $X^*_{r_2}$. The likelihood for this type of data may be written as

$$Pr(X_{(1)}, \ldots, X_{(r_1)}, X^*_{(1)}, \ldots, X^*_{(r_2)},)$$
$$= P_1(X_{(1)}, \ldots,\ X_{(r_1)}) P_2(X^*_{(1)}, \ldots, X^*_{(r_2)} \mid X_{(1)}, \ldots, X_{(r_1)}).$$

By equation (3.5.7), the first term above becomes

$$\frac{n!}{(n - r_1)!} \prod_{i=1}^{r_1} f(t_{(i)})[S(t_{(r_1)})]^{n-r_1}$$

and, by a theorem in order statistics (David, 1981), the second term above becomes

$$\frac{(n - r_1 - n_1)!}{(n - r_1 - n_1 - r_2)!} \prod_{i=1}^{r_2} f^*(x^*_{(i)})[S^*(x^*_{(r_2)})]^{n-r_1-n_1-r_2}$$

where $f^*(x) = \frac{f(x)}{S(x_{r_1})}$, $x \geq x_{(r_i)}$ is the truncated p.d.f. and $S^*(x) = \frac{S(x)}{S(x_{r_1})}$, $x \geq x_{(r_i)}$ is the truncated survival function so that

$$L_{\text{II},2} = \frac{n!(n - r_1 - n_1)!}{(n - r_1)!(n - r_1 - n_1 - r_2)!} \prod_{i=1}^{r_1} f(x_{(i)})[S(t_{(r_1)})]^{n-r_1}$$
$$x \frac{\prod_{i=1}^{r_2} f(t^*_{(i)})}{[S(t_{(r_1)})]^{r_2}} \left[\frac{S(t^*_{(r_2)})}{S(t_{(r_1)})}\right]^{n-r_1-n_1-r_2}$$

so that

$$L_{\text{II},2} \propto \prod_{i=1}^{r_1} f(x_{(i)}) \left[S(x_{(r_i)})\right]^{n_1} \prod_{i=1}^{r_2} f(x^*_{(i)}) \left[S(x^*_{(r_2)})\right]^{n-r_1-n_1-r_2}$$

which, again, can be written in the form of (3.5.1).

3. For random censoring, when X and C_r are not independent, the likelihood given by (3.5.6) is not correct. If the joint survival function of X and C_r is $S(x, c)$, then, the likelihood is of the form

$$L_{\text{III}} \propto \prod_{i=1}^{n} \{[-\partial S(x, t_i)/\partial x]_{x=t_i}\}^{\delta_i} \{[-\partial S(t_i, c)/\partial c]_{c=t_i}\}^{1-\delta_i},$$

which may be appreciably different from (3.5.6).

3.6 Counting Processes

In the previous section, we discussed the construction of classical likelihoods for censored and truncated data. These likelihoods can be used to develop some of the methods described in the remainder of this book. An alternative approach to developing inference procedures for censored and truncated data is by using counting process methodology. This approach was first developed by Aalen (1975) who combined elements of stochastic integration, continuous time martingale theory and counting process theory into a methodology which quite easily allows for development of inference techniques for survival quantities based on censored and truncated data. These methods allow relatively simple development of the large sample properties of such statistics. Although complete exposition of this theory is beyond the scope of this book, we will give a brief survey in this section. For a more rigorous survey of this area, the reader is referred to books by Andersen et al. (1993) and Fleming and Harrington (1991).

We start by defining a counting process $N(t), t \geq 0$, as a stochastic process with the properties that $N(0)$ is zero; $N(t) < \infty$, with probability one; and the sample paths of $N(t)$ are right-continuous and piecewise constant with jumps of size +1. Given a right-censored sample, the processes, $N_i(t) = I[T_i \leq t, \delta_i = 1]$, which are zero until individual i dies and then jumps to one, are counting processes. The process $N(t) = \sum_{i=1}^{n} N_i(t) = \sum_{t_i \leq t} \delta_i$ is also a counting process. This process simply counts the number of deaths in the sample at or prior to time t.

The counting process gives us information about when events occur. In addition to knowing this information, we have additional information on the study subjects at a time t. For right censored data, this information at time t includes knowledge of who has been censored prior to time t and who died at or prior to time t. In some problems, our information may include values for a set of fixed time covariates, such as age, sex, treatment at time 0 and possibly the values of time-dependent covariates, at all times prior to t. This accumulated knowledge about what has happened to patients up to time t is called the *history* or *filtration* of the counting process at time t and is denoted by $\mathbf{F}_t$. As time progresses, we learn more and more about the sample so that a natural requirement is that $\mathbf{F}_s \subset \mathbf{F}_t$ for $s \leq t$. In the case of right-censored data, the history at time t, $\mathbf{F}_t$, consists of knowledge of the pairs (T_i, δ_i) provided $T_i \leq t$ and the knowledge that $T_i > t$ for those individuals still under study at time t. We shall denote the history at an instant just prior to time t by $\mathbf{F}_{t-}$. The history $\{\mathbf{F}_t, t \geq 0\}$ for a given problem depends on the observer of the counting process.

For right-censored data, if the death times X_i and censoring times C_i are independent, then, the chance of an event at time t, given the history just prior to t, is given by

$$
\begin{aligned}
&Pr[t \leq T_i \leq t + dt, \delta_i = 1|F_{t-}] \qquad (3.6.1)\\
&= \begin{cases} Pr[t \leq X_i \leq t + dt, C_i > t + dt_i|X_i \geq t, C_i \geq t] = h(t)dt & \text{if } T_i \geq t, \\ 0 & \text{if } T_i < t \end{cases}
\end{aligned}
$$

For a given counting process, we define $dN(t)$ to be the change in the process $N(t)$ over a short time interval $[t, t + dt)$. That is $dN(t) = N[(t + dt)^-] - N(t^-)$ (Here t^- is a time just prior to t). In the right-censored data example (assuming no ties), $dN(t)$ is one if a death occurred at t or 0, otherwise. If we define the process $Y(t)$ as the number of individuals with a study time $T_i \geq t$, then, using (3.6.1),

$$
\begin{aligned}
E[dN(t)|\mathbf{F}_{t^-}] &= E[\text{Number of observations with}\\
&\qquad t \leq X_i \leq t + dt, C_i > t + dt_i \mid \mathbf{F}_{t^-}]\\
&= Y(t)h(t)dt. \qquad (3.6.2)
\end{aligned}
$$

The process $\lambda(t) = Y(t)h(t)$ is called the *intensity process* of the counting process. $\lambda(t)$ is itself a stochastic process that depends on the information contained in the history process, $\mathbf{F}_t$ through $Y(t)$.

The stochastic process $Y(t)$ is the process which provides us with the number of individuals at risk at a given time and, along with $N(t)$, is a fundamental quantity in the methods presented in the sequel. Notice that, if we had left truncated data and right-censored data, the intensity process would be the same as in (3.6.2) with the obvious modification to $Y(t)$ as the number of individuals with a truncation time less than t still at risk at time t.

We define the process $\Lambda(t)$ by $\int_0^t \lambda(s)ds, t \geq 0$. This process, called the *cumulative intensity process*, has the property that $E[N(t)|\mathbf{F}_{t^-}] = E[\Lambda(t) \mid \mathbf{F}_{t^-}] = \Lambda(t)$. The last equality follows because, once we know the history just prior to t, the value of $Y(t)$ is fixed and, hence, $\Lambda(t)$ is nonrandom. The stochastic process $M(t) = N(t) - \Lambda(t)$ is called *the counting process martingale*. This process has the property that increments of this process have an expected value, given the strict past, $\mathbf{F}_{t^-}$, that are zero. To see this,

$$
\begin{aligned}
E(dM(t) \mid \mathbf{F}_{t^-}) &= E[dN(t) - d\Lambda(t) \mid \mathbf{F}_{t^-}]\\
&= E[dN(t) \mid \mathbf{F}_{t^-}] - E[\lambda(t)dt \mid \mathbf{F}_{t^-}]\\
&= 0.
\end{aligned}
$$

The last inequality follows because $\lambda(t)$ has a fixed value, given $\mathbf{F}_{t^-}$.

A stochastic process with the property that its expected value at time t, given its history at time $s < t$, is equal to its value at time s is called a *martingale*, that is, $M(t)$ is a martingale if

$$
E[M(t) \mid \mathbf{F}_s] = M(s), \text{for all } s < t. \qquad (3.6.3)
$$

To see that this basic definition is equivalent to having $E[dM(t) \mid \mathbf{F}_{t^-}] = 0$ for all t, note that, if $E[dM(t) \mid \mathbf{F}_{t^-}] = 0$, then,

$$
\begin{aligned}
E(M(t) \mid \mathbf{F}_s] - M(s) &= E[M(t) - M(s) \mid \mathbf{F}_s]\\
&= E\left[\int_s^t dM(u) \mid \mathbf{F}_s\right]\\
&= \int_s^t E[E[dM(u) \mid \mathbf{F}_{u-}] \mid \mathbf{F}_s]\\
&= 0.
\end{aligned}
$$

Thus the counting process martingale is indeed a martingale.

The counting process martingale, $M(t) = N(t) - \Lambda(t)$ is made up of two parts. The first is the process $N(t)$, which is a nondecreasing step function. The second part $\Lambda(t)$ is a smooth process which is predictable in that its value at time t is fixed just prior to time t. This random function is called a *compensator* of the counting process. The martingale can be considered as mean zero noise which arises when we subtract the smoothly varying compensator from the counting process

To illustrate these concepts, a sample of 100 observations was generated from an exponential population with hazard rate $h_X(t) = 0.2$. Censoring times were generated from an independent exponential distribution with hazard rate $h_C(t) = 0.05$. Figure 3.6 shows the processes $N(t)$ and the compensator of $N(t)$, $\Lambda(t) = \int_0^t h(u)Y(u)du$, of a single sample drawn from these distributions. Note that $N(t)$ is an increasing step function with jumps at the observed death times, $Y(t)$ is a decreasing step function with steps of size one at each death or censoring time, and $\Lambda(t)$ is an increasing continuous function that is quite close to $N(t)$.

Figure 3.7 depicts the values of $M(t)$ for 10 samples generated from this population. The sample in Figure 3.6 is the solid line on this figure. We can see in this figure that the sample paths of $M(t)$ look like a sample of random, mean 0, noise.

An additional quantity needed in this theory is the notion of the *predictable variation process* of $M(t)$, denoted by $\langle M\rangle(t)$. This quantity is defined as the compensator of the process $M^2(t)$. Although $M(t)$ reflects the noise left after subtracting the compensator, $M^2(t)$ tends to increase with time. Here, $\langle M\rangle(t)$ is the systematic part of this increase and is the predictable process needed to be subtracted from $M^2(t)$ to produce a martingale. The name, predictable variation process, comes from the fact that, for a martingale $M(t)$, $\text{var}(dM(t) \mid \mathbf{F}_{t^-}) = d\langle M\rangle(t)$. To see this, recall that, by definition, $E[dM(t)] = 0$. Now,

$$
\begin{aligned}
dM^2(t) &= M[(t + dt)^-]^2 - M(t^-)^2\\
&= [M(t^-) + dM(t)]^2 - M(t^-)^2\\
&= [dM(t)]^2 + 2M(t^-)dM(t).
\end{aligned}
$$

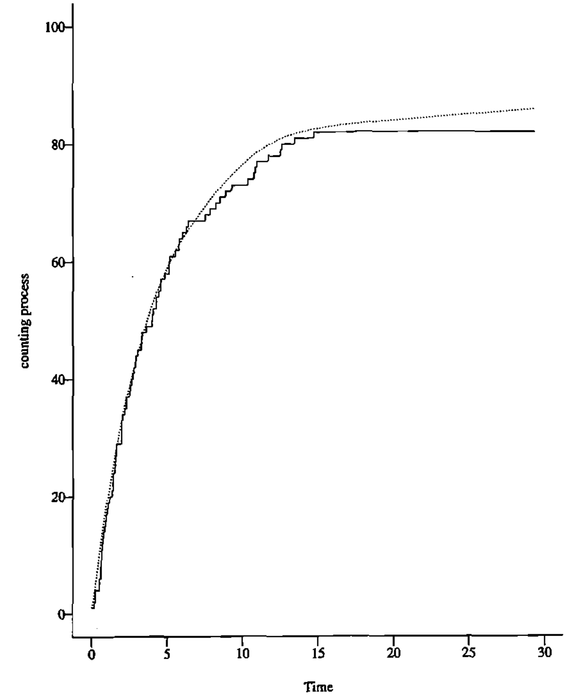

Figure 3.6 *Example of a counting process, N(t) (solid line) and its compensator, Λ(t), (dashed line) for a sample of 100 individuals*

So,

$$\begin{aligned}\mathrm{Var}[dM(t) \mid \mathbf{F}_{t^-}] &= E[(dM(t))^2 \mid \mathbf{F}_{t^-}] \\ &= E[(dM^2(t)) \mid \mathbf{F}_{t^-}] - 2E[M(t^-)dM(t) \mid \mathbf{F}_{t^-}] \\ &= d\langle M\rangle(t) - 2M(t^-)E[dM(t) \mid \mathbf{F}_{t^-}] = d\langle M\rangle(t)\end{aligned}$$

because once $\mathbf{F}_{t-}$ is known, $M(t^-)$ is a fixed quantity and $E[dM(t) \mid \mathbf{F}_{t^-}] = 0$.

To find $\mathrm{Var}[dM(t) \mid \mathbf{F}_{t^-}]$ recall that $dN(t)$ is a zero-one random variable with a probability, given the history, of $\lambda(t)$ of having a jump of size one at time t. The variance of such a random variable is $\lambda(t)[1-\lambda(t)]$. If there are no ties in the censored data case, $\lambda(t)^2$ is close to zero so that $\mathrm{Var}[dM(t) \mid \mathbf{F}_{t^-}] \cong \lambda(t) = Y(t)h(t)$. In this case, notice that the conditional mean and variance of the counting process $N(t)$ are the same and one can show that locally, conditional on the past history, the counting

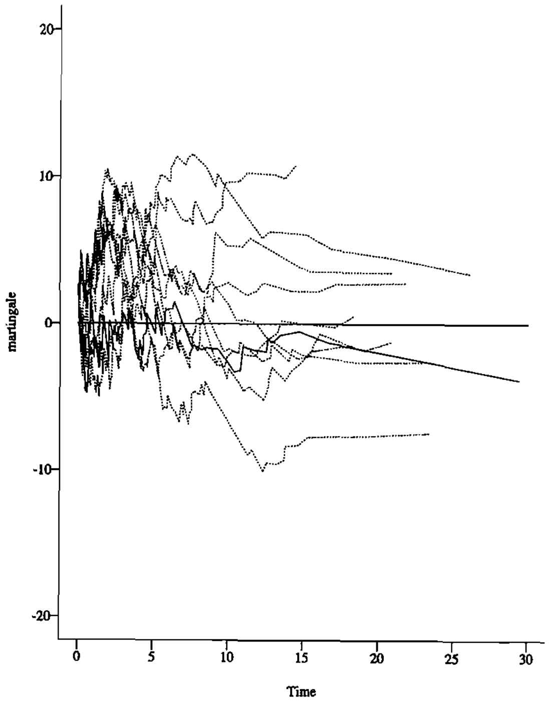

Figure 3.7 *Sample of 10 martingales. The compensated process in Figure 3.6 is the solid line.*

process behaves like a Poisson process with rate $\lambda(t)$. When there are ties in the data, the Bernoulli variance is used. Of course, in either case, these variances are conditional variances in that they depend on the history at time t^- through $Y(t)$. In many applications these conditional variances serve as our estimator of the variance of $dM(t)$.

Many of the statistics in later sections are *stochastic integrals* of the basic martingale discussed above. Here, we let $K(t)$ be a *predictable process*. That is $K(t)$ is a stochastic process whose value is known, given the history just prior to time t, $\mathbf{F}_{t^-}$. An example of a predictable process is the process $Y(t)$. Over the interval 0 to t, the stochastic integral of such a process, with respect to a martingale, is denoted by $\int_0^t K(u)dM(u)$. Such stochastic integrals have the property that they themselves are martingales as a function of t and their predictable variation process can be found from the predictable variation process of the original martingale by

$$\left\langle \int_0^t K(u)dM(u) \right\rangle = \int_0^t K(u)^2 d\langle M\rangle(u). \qquad (3.6.4)$$

To illustrate how these tools can be used to derive nonparametric estimators of parameters of interest, we shall derive a nonparametric estimator of the cumulative hazard rate $H(t)$ based on right-censored data, the so-called Nelson–Aalen estimator. Recall that we can write $dN(t) = Y(t)h(t)dt + dM(t)$. If $Y(t)$ is nonzero, then,

$$\frac{dN(t)}{Y(t)} = h(t)dt + \frac{dM(t)}{Y(t)}. \qquad (3.6.5)$$

If $dM(t)$ is noise, then, so is $dM(t)/Y(t)$. Because the value of $Y(t)$ is fixed just prior to time t,

$$E\left[\frac{dM(t)}{Y(t)} \mid \mathbf{F}_{t^-}\right] = \frac{E[dM(t) \mid \mathbf{F}_{t^-}]}{Y(t)} = 0.$$

Also, the conditional variance of the noise can be found as

$$\text{Var}\left[\frac{dM(t)}{Y(t)} \mid \mathbf{F}_{t^-}\right] = \frac{\text{Var}[dM(t) \mid \mathbf{F}_{t^-})]}{Y(t)^2} = \frac{d\langle M\rangle(t)}{Y(t)^2}.$$

If we let $J(t)$ be the indicator of whether $Y(t)$ is positive and we define $0/0 = 0$, then, integrating both sides of equation (3.6.5),

$$\int_0^t \frac{J(u)}{Y(u)} dN(u) = \int_0^t J(u)h(u)du + \int_0^t \frac{J(u)}{Y(u)} dM(u).$$

The integral $\int_0^t \frac{J(u)}{Y(u)} dN(u) = \hat{H}(t)$ is the Nelson–Aalen estimator of $H(t)$. The integral, $W(t) = \int_0^t \frac{J(u)}{Y(u)} dM(u)$, is the stochastic integral of the predictable process $\frac{J(u)}{Y(u)}$ with respect to a martingale and, hence, is also a martingale. Again, we can think of this integral as random noise or the statistical uncertainty in our estimate. The random quantity $H^*(t) = \int_0^t J(u)h(u)du$, for right-censored data is equal to $H(t)$ in the range where we have data, and, ignoring the statistical uncertainty in $W(t)$, the statistic $\hat{H}(t)$ is a nonparametric estimator of the random quantity $H^*(t)$.

Because $W(t) = \hat{H}(t) - H^*(t)$ is a martingale, $E[\hat{H}(t)] = E[H^*(t)]$. Note that $H^*(t)$ is a random quantity and its expectation is not, in general, equal to $H(t)$. The predictable variation process of $W(t)$ is found quite simply, using (3.6.4), as

$$\begin{aligned}\langle W\rangle(t) &= \int_0^t \left[\frac{J(u)}{Y(u)}\right]^2 d\langle M\rangle(u) = \int_0^t \left[\frac{J(u)}{Y(u)}\right]^2 Y(u)h(u)du \\ &= \int_0^t \left[\frac{J(u)}{Y(u)}\right] h(u)du.\end{aligned}$$

A final strength of the counting process approach is the martingale central limit theorem. Recall that $Y(t)/n$ and $N(t)/n$ are sample averages and that, for a large sample, the random variation in both should be small. For large n, suppose that $Y(t)/n$ is close to a deterministic function $y(t)$. Let $Z^{(n)}(t) = \sqrt{n}W(t) = \sqrt{n}[\hat{H}(t) - H^*(t)]$. This process is almost equal to $\sqrt{n}[\hat{H}(t) - H(t)]$, because for large samples $H^*(t)$ is very close to $H(t)$. Given the history, the conditional variance of the jumps in $Z^{(n)}(t)$ are found to converge to $h(t)/y(t)$. To see this

$$\begin{aligned}\text{Var}[dZ^{(n)}(t) \mid \mathbf{F}_{t^-}] &= n\text{Var}[dW(t) \mid \mathbf{F}_{t^-}] \\ &= n\text{Var}[\frac{dM(t)}{Y(t)} \mid \mathbf{F}_{t^-}] \\ &= n\frac{d\langle M(t)\rangle}{Y(t)^2} \\ &= n\frac{\lambda(t)dt}{Y(t)^2} \\ &= n\frac{Y(t)h(t)dt}{Y(t)^2} = \frac{h(t)dt}{Y(t)/n},\end{aligned}$$

which converges to $h(t)dt/y(t)$ for large samples. Also, for large samples, $Z^{(n)}$ will have many jumps, but all of these jumps will be small and of order $1/\sqrt{n}$.

The above heuristics tell us that, for large samples, $Z^{(n)}$ has almost continuous sample paths and a predictable variation process very close to

$$\langle Z^{(n)}\rangle \approx \int_0^t \frac{h(u)du}{y(u)}. \qquad (3.6.6)$$

It turns out that there is one and only one limiting process, $Z^{(\infty)}$ which is a martingale with continuous sample paths and a deterministic predictable variation $\langle Z^{(\infty)}\rangle$ exactly equal to (3.6.6). This limiting process has independent increments and normally distributed finite-dimensional distributions. A process has independent increments if, for any set of nonoverlapping intervals (t_{i-1}, t_i), $i = 1, \ldots, k$ the random variables $Z^{(\infty)}(t_i) - Z^{(\infty)}(t_{i-1})$ are independent. The limiting process has normally distributed finite-dimensional distributions if the joint distribution of $[Z^{(\infty)}(t_1), \ldots, Z^{(\infty)}(t_k)]$ is multivariate normal for any value of k. For the process $Z^{(\infty)}$, $[Z^{(\infty)}(t_1), \ldots, Z^{(\infty)}(t_k)]$ has a k-variate normal distribution with mean 0 and a covariance matrix with entries

$$\text{cov}[Z^{(\infty)}(t), Z^{(\infty)}(s)] = \int_0^{\min(s,t)} \frac{h(u)du}{y(u)}.$$

This basic convergence allows us to find confidence intervals for the cumulative hazard rate at a fixed time because $\sqrt{n}[\hat{H}(t) - H^*(t)]$ will have an approximate normal distribution with mean 0 and variance

$$\sigma[Z^{(\infty)}] = \int_0^t \frac{h(u)du}{y(u)}.$$

An estimate of the variance can be obtained from

$$n \int_0^t \frac{dN(u)}{Y(u)^2}$$

because we can estimate $y(t)$ by $Y(t)/n$ and $h(t)$ by $dN(t)/Y(t)$. The fact that, as a process, $Z^{(n)}$ is approximated by a continuous process with normal margins also allows us to compute confidence bands for the cumulative hazard rate (see section 4.4).

To estimate the survival function, recall that, for a continuous random variable, $S(t) = \exp[-H(t)]$ and, for a discrete, random variable, $S(t) = \prod_{s=0}^{t}[1 - d\hat{H}(s)]$. Here, we say that $S(t)$ is the *product integral* of $1 - d\hat{H}(t)$. To obtain an estimator of the survival function, we take the product integral of $1 - d\hat{H}(t)$ to obtain

$$\hat{S}(t) = \prod_{s=0}^{t}[1 - d\hat{H}(t)] = \prod_{s=0}^{t}\left[1 - \frac{dN(s)}{Y(s)}\right].$$

This is the Kaplan–Meier estimator (see section 4.2) which is a step function with steps at the death times where $dN(t) > 0$. It turns out that $\hat{S}(t)/S(t) - 1$ is a stochastic integral with respect to the basic martingale M and is also a martingale. Thus confidence intervals and confidence bands for the survival function can be found using the martingale central limit theorem discussed above (see sections 4.3 and 4.4).

Counting processes methods can be used to construct likelihoods for survival data in a natural way. To derive a likelihood function based on $N(t)$ consider a separate counting process, $N_j(t)$, for each individual in the study. Given the history up to time t, $dN_j(t)$ has an approximate Bernoulli distribution with probability $\lambda_j(t)dt$ of having $dN_j(t) = 1$. The contribution to the likelihood at a given time is, then, proportional to

$$\lambda_j(t)^{dN_j(t)}[1 - \lambda_j(t)dt]^{1-dN_j(t)}.$$

Integrating this quantity over the range $[0, \tau]$ gives a contribution to the likelihood of

$$\lambda_j(t)^{dN_j(t)} \exp\left[-\int_0^{\tau} \lambda_j(u)du\right].$$

The full likelihood for all n observations based on information up to time τ is, then, proportional to

$$L = \left[\prod_{j=1}^{n} \lambda_j(t)^{dN_j(t)}\right] \exp\left[-\sum_{j=1}^{n} \int_0^{\tau} \lambda_j(u)du\right].$$

For right-censored data, where $\lambda_j(t) = Y_j(t)h(t)$, with $Y_j(t) = 1$ if $t \le t_j, 0$ if $t > t_j$, so

$$L \propto \left[\prod_{j=1}^{n} h(t_j)^{\delta_j}\right] \exp\left(-\sum_{j=1}^{n} H(t_j)\right),$$

which is exactly the same form as (3.5.1). This heuristic argument is precisely stated in Chapter 2 of Andersen et al. (1993).

The counting process techniques illustrated in this section can be used to derive a wide variety of statistical techniques for censored and truncated survival data. They are particularly useful in developing nonparametric statistical methods. In particular, they are the basis of the univariate estimators of the survival function and hazard rate discussed in Chapter 4, the smoothed estimator of the hazard rate and the models for excess and relative mortality discussed in Chapter 6, most of the k-sample nonparametric tests discussed in Chapter 7, and the regression methods discussed in Chapters 8, 9, and 10. A check of the martingale property is used to test model assumptions for regression models, as discussed in Chapter 11. Most of the statistics developed in the sequel can be shown to be stochastic integrals of some martingale, so large sample properties of the statistics can be found by using the predictable variation process and the martingale central limit theorem. In the theoretical notes, we shall point out where these methods can be used and provide references to the theoretical development of the methods. The books by Andersen et al. (1993) or Fleming and Harrington (1991) provide a sound reference for these methods.

3.7 Exercises

3.1 Describe, in detail, the types of censoring which are present in the following studies.

(a) The example dealing with remission duration in a clinical trial for acute leukemia described in section 1.2.

(b) The example studying the time to death for breast cancer patients described in section 1.5.

3.2 A large number of disease-free individuals were enrolled in a study beginning January 1, 1970, and were followed for 30 years to assess the age at which they developed breast cancer. Individuals had clinical exams every 3 years after enrollment. For four selected individuals described below, discuss in detail, the types of censoring and truncation that are represented.

(a) A healthy individual, enrolled in the study at age 30, never developed breast cancer during the study.

(b) A healthy individual, enrolled in the study at age 40, was diagnosed with breast cancer at the fifth exam after enrollment (i.e., the disease started sometime between 12 and 15 years after enrollment).

(c) A healthy individual, enrolled in the study at age 50, died from a cause unrelated to the disease (i.e., not diagnosed with breast cancer at any time during the study) at age 61.

(d) An individual, enrolled in the study at age 42, moved away from the community at age 55 and was never diagnosed with breast cancer during the period of observation.

(e) Confining your attention to the four individuals described above, write down the likelihood for this portion of the study.

3.3 An investigator, performing an animal study designed to evaluate the effects of vegetable and vegetable-fiber diets on mammary carcinogenesis risk, randomly assigned female Sprague-Dawley rats to five dietary groups (control diet, control diet plus vegetable mixture, 1; control diet plus vegetable mixture, 2; control diet plus vegetable-fiber mixture, 1; and control diet plus vegetable-fiber mixture, 2). Mammary tumors were induced by a single oral dose (5 mg dissolved in 1.0 ml. corn oil) of 7,12-dimethylbenz(α)anthracene (DMBA) administered by intragastric intubation, i.e., the starting point for this study is when DMBA was given.

Starting 6 weeks after DMBA administration, each rat was examined once weekly for 14 weeks (post DMBA administration) and the time (in days) until onset of the first palpable tumor was recorded. We wish to make an inference about the marginal distribution of the time until a tumor is detected. Describe, in detail, the types of censoring that are represented by the following rats.

(a) A rat who had a palpable tumor at the first examination at 6 weeks after intubation with DMBA.

(b) A rat that survived the study without having any tumors.

(c) A rat which did not have a tumor at week 12 but which had a tumor at week 13 after inturbation with DMBA.

(d) A rat which died (without tumor present and death was unrelated to the occurrence of cancer) at day 37 after intubation with DMBA.

(e) Confining our attention to the four rats described above, write down the likelihood for this portion of the study.

3.4 In section 1.2, a clinical trial for acute leukemia is discussed. In this trial, the event of interest is the time from treatment to leukemia relapse. Using the data for the 6-MP group and assuming that the time to relapse distribution is exponential with hazard rate λ, construct the likelihood function. Using this likelihood function, find the maximum likelihood estimator of λ by finding the value of λ which maximizes this likelihood.

3.5 Suppose that the time to death has a log logistic distribution with parameters λ and α. Based on the following left-censored sample, construct the likelihood function.

DATA: 0.5, 1, 0.75, 0.25-, 1.25-, where - denotes a left- censored observation.

3.6 The following data consists of the times to relapse and the times to death following relapse of 10 bone marrow transplant patients. In the sample patients 4 and 6 were alive in relapse at the end of the study and patients 7–10 were alive, free of relapse at the end of the study. Suppose the time to relapse had an exponential distribution with hazard rate λ and the time to death in relapse had a Weibull distribution with parameters θ and α.

Patient	Relapse Time (months)	Death Time (months)
1	5	11
2	8	12
3	12	15
4	24	33^+
5	32	45
6	17	28^+
7	16^+	16^+
8	17^+	17^+
9	19^+	19^+
10	30^+	30^+

$^+$ Censored observation

(a) Construct the likelihood for the relapse rate λ.

(b) Construct a likelihood for the parameters θ and α.

(c) Suppose we were only allowed to observe a patients death time if the patient relapsed. Construct the likelihood for θ and α based on this truncated sample, and compare it to the results in (b).

3.7 To estimate the distribution of the ages at which postmenopausal woman develop breast cancer, a sample of eight 50-year-old women were given yearly mammograms for a period of 10 years. At each exam, the presence or absence of a tumor was recorded. In the study, no tumors were detected by the women by self-examination between the scheduled yearly exams, so all that is known about the onset time of breast cancer is that it occurs between examinations. For four of the eight women, breast cancer was not detected during the 10 year study period. The age at onset of breast cancer for the eight subjects was in the following intervals:

$$(55, 56], (58, 59], (52, 53], (59, 60], \geq 60, \geq 60, \geq 60, \geq 60.$$

(a) What type of censoring or truncation is represented in this sample?

(b) Assuming that the age at which breast cancer develops follows a Weibull distribution with parameters λ and α, construct the likelihood function.

3.8 Suppose that the time to death X has an exponential distribution with hazard rate λ and that the right-censoring time C is exponential with hazard rate θ. Let $T = \min(X, C)$ and $\delta = 1$ if $X \leq C; 0$, if $X > C$. Assume that X and C are independent.

(a) Find $P(\delta = 1)$

(b) Find the distribution of T.

(c) Show that δ and T are independent.

(d) Let $(T_1, \delta_1), \ldots, (T_n, \delta_n)$ be a random sample from this model. Show that the maximum likelihood estimator of λ is $\sum_{i=1}^{n} \delta_i / \sum_{i=1}^{n} T_i$. Use parts a–c to find the mean and variance of $\hat{\lambda}$.

3.9 An example of a counting process is a Poisson process $N(t)$ with rate λ. Such a process is defined by the following three properties:

(a) $N(0) = 0$ with probability 1.

(b) $N(t) - N(s)$ has a Poisson distribution with parameter $\lambda(t - s)$ for any $0 \leq s \leq t$.

(c) $N(t)$ has independent increments, that is, for $0 \leq t_1 < t_2 < t_3 < t_4$, $N(t_2) - N(t_1)$ is independent of $N(t_4) - N(t_3)$.

Let $\mathbf{F}_s$ be the σ-algebra defined by $N(s)$. Define the process $M(t) = N(t) - \lambda t$.

i. Show that $E|M(t)| < \infty$.

ii. Show that $E[M(t) \mid N(s)] = M(s)$ for $s < t$, and conclude that $M(t)$ is a martingale and that λt is the compensator of $N(t)$. (Hint: Write $M(t) = M(t) - M(s) + M(s)$.)

4

Nonparametric Estimation of Basic Quantities for Right-Censored and Left-Truncated Data

4.1 Introduction

In this chapter we shall examine techniques for drawing an inference about the distribution of the time to some event X, based on a sample of right-censored survival data. A typical data point consists of a time on study and an indicator of whether this time is an event time or a censoring time for each of the n individuals in the study. We assume throughout this chapter that the potential censoring time is unrelated to the potential event time. The methods are appropriate for Type I, Type II, progressive or random censoring discussed in section 3.2.

To allow for possible ties in the data, suppose that the events occur at D distinct times $t_1 < t_2 < \cdots < t_D$, and that at time t_i there are d_i events (sometimes simply referred to as deaths). Let Y_i be the number of individuals who are at risk at time t_i. Note that Y_i is a count of the number