

Note: This chapter is a draft excerpt from:
Mills, M. (forthcoming/2010) *Introducing Survival and Event History Analysis*.
London: Sage Publications. [The book is written for accessibility and is a practical
guide with a focus on using the computer program R.]

CHAPTER 1

The Fundamentals of Survival and Event History Analysis

Contents

1.1	INTRODUCTION: WHAT IS SURVIVAL AND EVENT HISTORY ANALYSIS?.....	2
1.2	KEY CONCEPTS AND TERMINOLOGY.....	2
1.3	CENSORING AND TRUNCATION.....	5
1.4	MATHEMATICAL EXPRESSION AND RELATION OF BASIC STATISTICAL FUNCTIONS..	8
1.5	HOW DO THE SURVIVOR, DENSITY AND HAZARD FUNCTION RELATE?	11
1.6	WHY USE SURVIVAL AND EVENT HISTORY ANALYSIS?	12
1.7	OVERVIEW OF SURVIVAL AND EVENT HISTORY MODELS.....	14

Objectives of this chapter:

After reading this chapter, the researcher should be able to:

1. Define, recognize and describe the **fundamental concepts and terminology** used in survival and event history analysis
2. Recognize and describe the **reason** why we use these methods and the **type of problem that can be solved**
3. Describe **censored data** and different types of censoring
4. Define and describe **truncated data**
5. Define, recognize and interpret a **survivor and hazard function**
6. Describe the **relationship between a survivor and hazard function**
7. Be able to argue **why it is necessary to use survival and event history models**
8. Recognize the **different types of survival and event history models and classes**

*****DRAFT VERSION - DO NOT DISTRIBUTE*****

1.1 Introduction: What is survival and event history analysis?

Survival and event history analysis is an umbrella term for a collection of statistical methods that examine the time it takes for an ‘event’ to occur. This method appeals to multiple scientific disciplines due to fact that many research questions involve timing and duration issues. This technique is often referred to as survival analysis in biostatistics, medical science and epidemiology, engineers call it reliability analysis; it is duration models for most economists and often termed event history analysis by sociologists, demographers and political scientists.

The goal of this book is introduce these methods in an accessible, practical and engaging manner starting from basic methods and ranging to the most cutting-edge techniques used in the field today. Written for accessibility, this book will appeal to students and researchers who want to understand the basics and apply these methods without getting entangled in the mathematical and theoretical technicalities. Readers are offered a blueprint for their entire research project from data preparation to model selection and diagnostics, allowing them to independently master these advanced methods. This book is written from the perspective of the ‘user’ with numerous examples and hands-on exercises, making it suitable as both a self-learning or textbook. Using the advanced, powerful and free computer program R, readers are provided with clear instructions on how to prepare data, run various types of models and enhance the expression of results with powerful graphics.

This chapter provides a basic description of the fundamental concepts and terminology of these techniques, which are the foundations for the rest of the book. The key concepts of censoring and truncation are discussed in detail with multiple examples. The mathematical expressions and relation of statistical functions are then presented in a manner that requires only a basic background in mathematics, statistics and data analysis. We then turn to the basic logic of why it is necessary to use these types of models with certain data and research problems. The final section provides a brief overview of the different types of survival and event history models, which simultaneously serves as an overview of this book. As with the entire book, the focus of this chapter will be example-based to allow the reader to quickly grasp the central concepts.

1.2 Key concepts and terminology

The goal of survival and event history analysis is to statistically model and analyze the time until an event occurs. As Figure 1.1 illustrates, the focus is therefore on the time that it takes for an event to occur, which is which we will model throughout this entire book. The classic example of an **event** is death. As

Table 1.1 illustrates, an event could also be infection, marriage, 'death' of a bank or the end of a U.N. peacekeeping mission. A key distinction of these methods is that they take **censoring** into account. As we discuss in more detail in the next section, this is the ability to examine failure time by also taking into account those who are still 'alive' or did not experience the event at the time of the survey or end of the clinical trail. A very common type is what is referred to as right-censoring, which is when the event that we are studying does not occur by the time of the last observation, such as the survey date or the last clinical observation point.

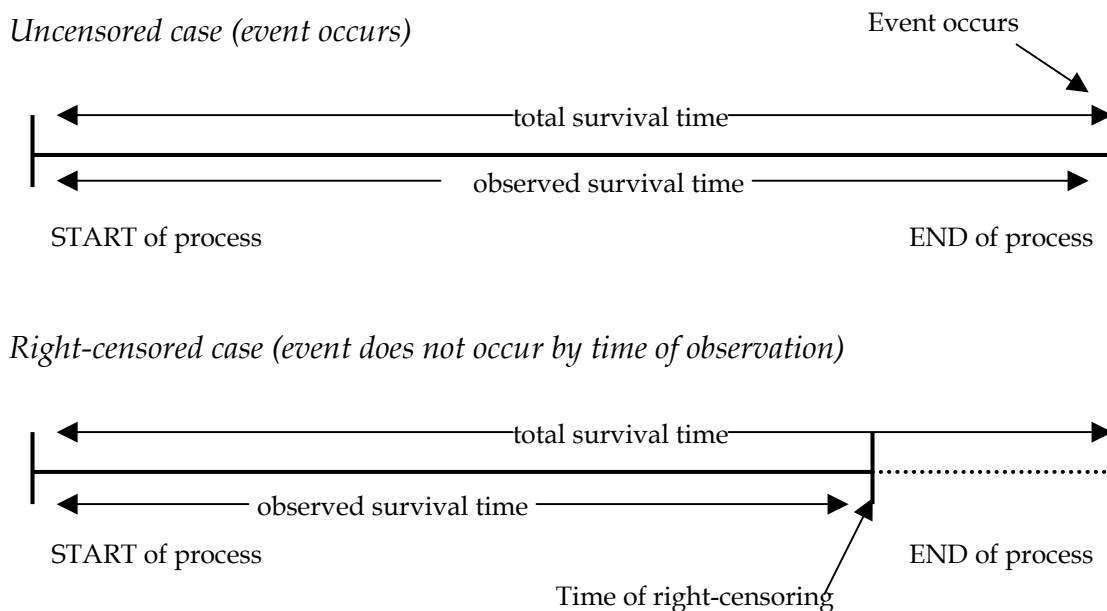



Figure 1.1 Understanding survival time in a single episode

Survival and event history analysis is used extensively in epidemiology and health sciences, such as the studies shown in Table 1.1 of leukemia, heart transplant survival and infections of kidney dialysis patients to other areas such as the study of primary biliary cirrhosis (Markus et al. 1989) or AIDS (Hammer et al., 1997). In the social sciences, there are also many applications such as the study of organizational change (Hannan and Carroll 1981), changes in hate crime law (Grattet, Jenness and Curry 1998), changes in social insurance legislation (Usui 1994), social movements and the evolution of right-wing movements (Koopmans and Olzak 2004), job mobility (Hachen 1992) and union decline (Western 1995). In this sense, these techniques are highly interdisciplinary and can be used to study a variety of research questions.

It is not only the occurrence of an event that is of interest to researchers, but the history or time that it took before that event happened. For this reason, **time**

is an essential aspect of these models and can be measured in diverse **units**, such as seconds, days, weeks, months, or years. This duration or time that it takes before an event occurs is what we refer to as **survival time**. Put in another way, it is the time that a person or other unit of analysis (e.g., bolt in a machine, bank, political regime) ‘survives’ the specified duration. The time that the unit under analysis spends in a specific discrete state is defined as a **spell**, or interchangeably often referred to as an **episode, interval, or risk period**.

Table 1.1 *Examples of survival analysis showing starting time and event status*

START	SURVIVAL TIME	EVENT
Patient with acute myelogenous leukemia enters remission (Miller 1997)		Death or ‘censored’ (i.e., still alive at last observation)
Patient joins waiting list for heart transplant (Crowley and Hu 1997)		Death or ‘censored’ (i.e., still alive at last observation)
Insertion of catheter in kidney dialysis patient (McGilchrist and Aisbett 1991)		Infection or ‘censored’ (i.e., no infection at last observation)
Woman in (non-marital) cohabiting relationship becomes pregnant (Blossfeld and Mills 2001)		Marriage or ‘censored’ (i.e., still cohabiting at last observation)
Commercial bank opens (Bergström, Engvall and Wallerstedt 1997)		Closure of bank or ‘censored’ (i.e., bank still alive at last observation)
Start of U.N. peacekeeping mission (Green, Kahl and Diehl 1998)		End of U.N. peacekeeping mission or ‘censored’ (i.e., mission still ongoing at last observation)

Another way to understand survival is in terms of **risk**. For example, given that an individual has remained in remission from cancer for three years, what is the risk that he or she eventually relapses? The occurrence of an event is also often referred to as a **failure**, usually attributed to fact that the event is death or disease. Even though we often use the term ‘failure’ or ‘failure time’, it might actually refer to a highly positive event such as birth of a child, cure from a disease or fall of an oppressive dictator.

These terms provide the basic foundations for simple models, but we will also explore analyses that are much more complex including competing risks, recurrent events and entire histories. The examples presented in Table 1.1 only

describe the occurrence of one event. Some events are not only one-dimensional, but may have several competing causes, often referred to as **competing risks** (Larson, 1984; Hachen, 1988; Tuma et al., 1979). For example, the event might not only be death, but multiple competing causes of death. Or, if a bank is facing serious problems, it not only has the option of closing but could merge with another financial partner.

The examples shown above also suggest that events can occur only once. However, it is also possible to examine **recurrent events** such as multiple relapses from remission, repeated heart attacks, repeated infections, multiple marriages, births or unemployment episodes. The topics of competing risks and recurrent events will be covered in more detail later chapters. The examples also show only one transition from one state to another. However, it is also possible to examine the concatenation (i.e., serial sequence of spells) that constitutes an event history in the form of **multistate models** and **sequence analysis**, which is examined in detail in the last chapter and is often not included in existing textbooks on this topic.

In the chapters that follow, the **dependent or outcome variable** is therefore the time until an event occurs. The goal of survival analysis is not only to examine the time until an event occurs, but to also assess the relationship of survival time to explanatory variables. **Explanatory variables** (also known as covariates or independent variable) assess the impact of certain characteristics (e.g., receiving treatment, level of education) on the time until an event occurs. As we will explore in the chapters to come, these variables may be fixed in time (e.g., sex, place of birth) or be time-varying and change their values over time (e.g., age, labor force experience).

1.3 Censoring and Truncation

Censoring. Before going any further, it is essential to grasp the concepts of censoring and truncation. Censoring in particular is key for survival analysis. As Figure 1.2 shows, there are various types of censoring that can occur. A simple definition of censoring is that we have information about an individual's survival time, but do not know the exact survival time (Kleinbaum and Klein 2005). In some ways you could consider censoring as akin to missing data, but it is for several substantial reasons, quite different.

Figure 1.2 illustrates the various types of censoring often faced when undertaking survival analysis. **Uncensored cases** represent the information where we know both the starting and ending time of episodes, often the majority of the data. The most common type of censoring is in the form of a **right-censored case**, which occurs when the event under study is not experienced by the last observation. This is most often random censoring, which occurs very often in medical and social science studies. In a study of takeovers of Fortune 500

firms between 1980-1990, Davis et al. (1994) for example, had the dates of all takeover attempts for all firms. However, out of all of the firms that were at risk of takeover, only 30 percent of the firms experienced a takeover during the observation period. In other words, since a takeover was considered as the 'event', all firms that did not experience a takeover were specified as right censored.

Another common data collection technique is cross-sectional (i.e., data collected at one point in time) from retrospective surveys. Here questions are asked retrospectively about events such as medical, employment or fertility histories. Individuals would be questioned about for example, the birth of their first to last child or the start and end dates of jobs. If they have not (yet) progressed to the next event (e.g., death, birth of second child, end of first job), events are considered as right-censored by the interview date. For instance, if we were examining the transition from first to second birth, all of those individuals who had a first child, but had not a second a second child would be 'right-censored' by the interview date.

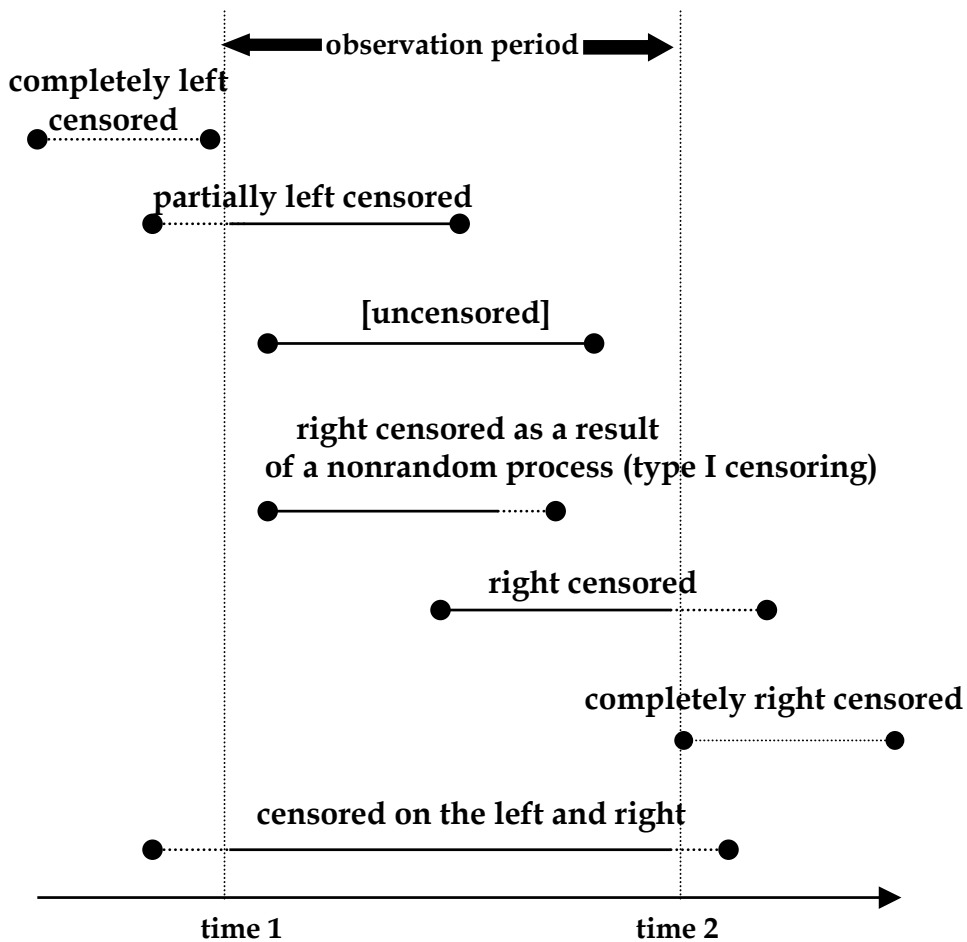


Figure 1.2 *Types of censoring*

In some medical studies and panel or longitudinal designs, the individual may not be present for follow-up (e.g., they have moved), drop out of the study (e.g., due to bad side effects or refusal to participate) or the study simply ends. The first two reasons are often considered as random whereas the latter is considered as non-random right-censored data. An essential point is that although there is no information about the *occurrence* of the event for right-censored cases, there is information about the survival or exposure time until the last point of observation. Right-censored data constitutes missing data insofar as we have information on the unit of analysis' event history and time at risk up to the last observation point. Therefore, we require the assumption that censoring is random and that the processes governing censoring and occurrence of events are independent of one another (Tsiastis, 1975).

If the event occurs before the observation and the length of time spent in the origin state is unknown, the observation is considered as **partially left-censored**. In a study of the mortality of 572 life insurance organizations in New York from the period of 1881-1931, for instance, Lehrman (1994) was only able to begin the investigation in 1881 due to the fact that reliable and complete records were available after this date. Thus instead of using the starting date of when the organizations were founded, the analysis started at a particular time point, which resulted in left-censoring (i.e., length of time since founding of organization was unknown).

A special condition is when an episode is **right censored as a result of a nonrandom process**, often referred to as Type I censoring. This type of censoring arises less in social science research and more in disciplines such as engineering. In engineering applications, such as the studies of Wallodi Weibull (e.g., 1939; 1951), which we will discuss later in this book, the focus was on the strength of materials, fatigue, rupture in solids, and bearings. In these studies, tubes, chips or bearings for instance, are all started for testing at a particular time and the time until event is recorded until their failure (i.e., breakdown). However, since some items may take a very long time to fail or breakdown, the experiment is prematurely terminated at a predetermined nonrandom time.

Another type of censoring, common for panel studies, is for cases that are both **left and right censored**. This is when the researcher knows that the unit of analysis is in a particular state at the first point of observation and maintains this same state at the second observation, but there is no information about the entrance or ending times of these states. For more detailed discussions of censored data, refer to sources such as Kalbfleisch and Prentice (1980: 39-41), Tuma and Hannan (1984: 118), Guo (1993), Yamaguchi (1991: 3-9) and Vermunt (1997: 117-130).

Truncation. There is often some confusion between whether observations are 'truncated' or 'censored' as it is a tricky distinction. Following Klein and Moeschberger (1997), truncation is a condition other than the event of interest

that is used to screen respondents or patients. This occurs when the individual experienced the event of interest prior to the survey or study. These individuals are not included in the study due to the fact that there is no information about them or in other words, the researcher is not aware of their existence. A common type of truncation is referred to as **left-truncation** (Guo, 1993). For example, consider a longitudinal panel survey that starts in a particular year and collects prospective data on unemployment. If there is no retrospective element in the survey, all of those respondents who experienced unemployment prior to the data collection date will have left-truncated observations. Or, consider the occurrence of exposure to a certain disease before a patient enters a study. The most common type of left-truncation is when subjects enter the study at a random age and are followed until the event of interest or the subject is right-censored. **Right-truncation** can also occur. Klein and Moeschberger (1997) provide the example of the examination of an episode from HIV infection until the development of AIDS. If the sample only includes those who have developed AIDS prior to the end of the study, those HIV infected individuals (who have not yet progressed to AIDS), are excluded from the sample.

1.4 Mathematical expression and relation of basic statistical functions

This section provides a brief outline the key statistical concepts of survival analysis. As described previously, the text is written at the level of a non-mathematician and primarily for ‘users’ of these types of the methods. That being said, understanding these expressions and how they are calculated are key for your general understanding and interpretations of these methods. Non-mathematicians can refer to Box 1.1 for a review of some of the notation used in the equations.

A core statistical concept in these models is the **hazard function**, which also goes by the name of the instantaneous transition or hazard rate. It is denoted by $h(t)$ which encompasses the *type* of change from one state to another, which defined in mathematical terms is the transition from state j to k and its *timing* (duration until the event occurs). It is formally specified as:

$$h(t)_{jk} = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < \Delta t | T \geq t)}{\Delta t} \quad [1.1]$$

Where T is the survival time or the duration between entering a particular origin state j (e.g., alive and under radiation treatment for lymphoma) to the destination

state k (e.g., death). The transition rate $h(t)_{jk}$ represents the instantaneous risk that the event occurs at time Δt , given that the event did not occur before time t .

Box 1.1 A review of notation for non-mathematicians

T	Random variable of survival time ($T \geq 0$)
$T \geq 0$	Means that T can be any number equal to or greater than zero (i.e., cannot have negative values)
t	Specific value for T
δ	(0,1) random variable = 1 if failure, = 0 if censored (Greek letter delta)
$S(t)$	Survivor function
$h(t)$	Hazard function
$f(t)$	Density function
∞	infinity
P	Probability
	given
Δt	Small time interval
lim	limit
$\lim_{\Delta t \rightarrow 0}$	Instantaneous potential

As Box 1.2 describes, the hazard is a *rate* and not a probability. The probability statement (often referred to as the conditional probability) refers to the part of the equation in the numerator of the equation that comes after the limit (lim) sign. Put in non-mathematical terms, it is the conditional probability that an individual's survival time (T) will lie in the time interval between t and $t + \Delta t$, given that the survival time is greater than or equal to t . The denominator of the equation is Δt , which refers to a small time interval. Since the equation is a ratio, the result is a probability per unit time, which is a rate that can vary between 0 and infinity $[0, \infty)$.

The transition rate is interrelated to two other fundamental concepts, which are the survivor function $S(t)$ and the density function $f(t)$. The **survivor function** is specified as:

$$S(t) = P(T > t) \quad [1.2]$$

and expresses the probability that an individual (or other unit of analysis) survives longer than some specified time t . In other words, $S(t)$ provides the probability of that the random variable T exceeds the specified time t . Some fields, such as engineering, also refer to this as the reliability function.

The **density function** $f(t)$ expresses the unconditional instantaneous probability that an event occurs in the time interval $(t, \Delta t)$ and is formally specified as:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < \Delta t)}{\Delta t - t} \quad [1.3]$$

The most common functions used in research are the hazard and survivor function. These functions are interrelated and different ones can be selected to define a model. The hazard function is often used in model specifications since by definition it is 'devoid of time.' The transition rate is the probability of occurrence of an event, conditional on having survived up to that time, whereas the density function is the unconditional probability. The rate is therefore computed by referring to those subjects that are still at risk of experiencing the event at time t . The **relationship** between the three functions is:

$$h(t) = \frac{f(t)}{S(t)} \quad [1.4]$$

One way to visualize the relationship between the hazard and survivor function is via the following diagram adapted from Kleinbaum and Klein (2005) and shown in Figure 1.3. Another way to remember the difference between them is to think about it in simple linguistic terms. The hazard focuses on failing whereas the survivor function focuses on surviving (i.e., not failing).

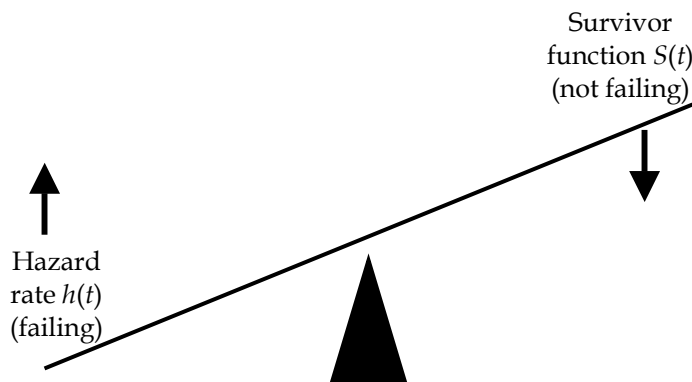


Figure 1.3 *Visualizing the relationship between the hazard and survivor function*

The equations for more advanced parametric regression models, also with explanatory variables, will be described in more detail in the upcoming chapters.

Box 1.2 Why is the hazard a rate and not a probability?

A common confusion is the distinction between a rate and probability. The hazard, which is the rate that is modeled in these techniques, can be distinguished from a probability by several factors. First, a rate can assume any values from 0 to infinity, or defined as $[0, \infty)$. Conversely, a probability can only take on numbers that range between 0 and 1, or $[0,1)$. A second difference, put in simple terms, lies in the denominator of the equation. For both, the frequency is the numerator, but the denominator for the probability is the number at risk and for the hazard rate is the amount of exposure time at risk.

$$\text{Probability} = \frac{\text{Frequency of Events}}{\text{Number at Risk}}$$

$$\text{Rate} = \frac{\text{Frequency of Events}}{\text{Amount of Exposure Time at Risk}}$$

To illustrate why the hazard is rate and not a probability, consider the following example from Tableman and Kim (2004: 6), where, $P = P(t \leq T < \Delta t | T \geq t) = 1/4$. This example shows that if a probability (P) is 1/4 and the time interval is a quarter of a day, the probability divided by the time interval is 1/4 divided by 1/3, which equals 0.75 per day.

P (Probability)	Δt (exposure time)	$\frac{P}{\Delta t} = \text{rate}$
$\frac{1}{4}$	$\frac{1}{3}$ day	$\frac{1/4}{1/3} = 0.75$ day
$\frac{1}{4}$	$\frac{1}{21}$ week	$\frac{1/4}{1/21} = 5.25$ week

1.5 How do the survivor, density and hazard function relate?

As we will see in the upcoming chapters, a common starting point in survival analysis is to estimate and plot the survivor, density and hazard function. This raises two common and interrelated questions. Are they interchangeable or can they differ? Should I estimate and plot the survivor, density or hazard function?

The answer, of course, is not always straightforward. While different survivor functions can have the same shape, their respective hazard functions can differ dramatically, illustrated in Figure 1.4. The figures demonstrate the differences between the hazard rates and their respective densities. Following Tableman and Kim (2004), consider the following three models: (a) an increasing hazard rate, such as processes related to aging, (b) a decreasing hazard rate with an elevated likelihood of early failure, such as death after an organ transplant; and, (c) a bath-tub shaped hazard such as the modeling of a population from birth to death. In general, the hazard function expresses more about the

underlying mechanism behind the failure or event in the data and is often the choice for summarizing the data.

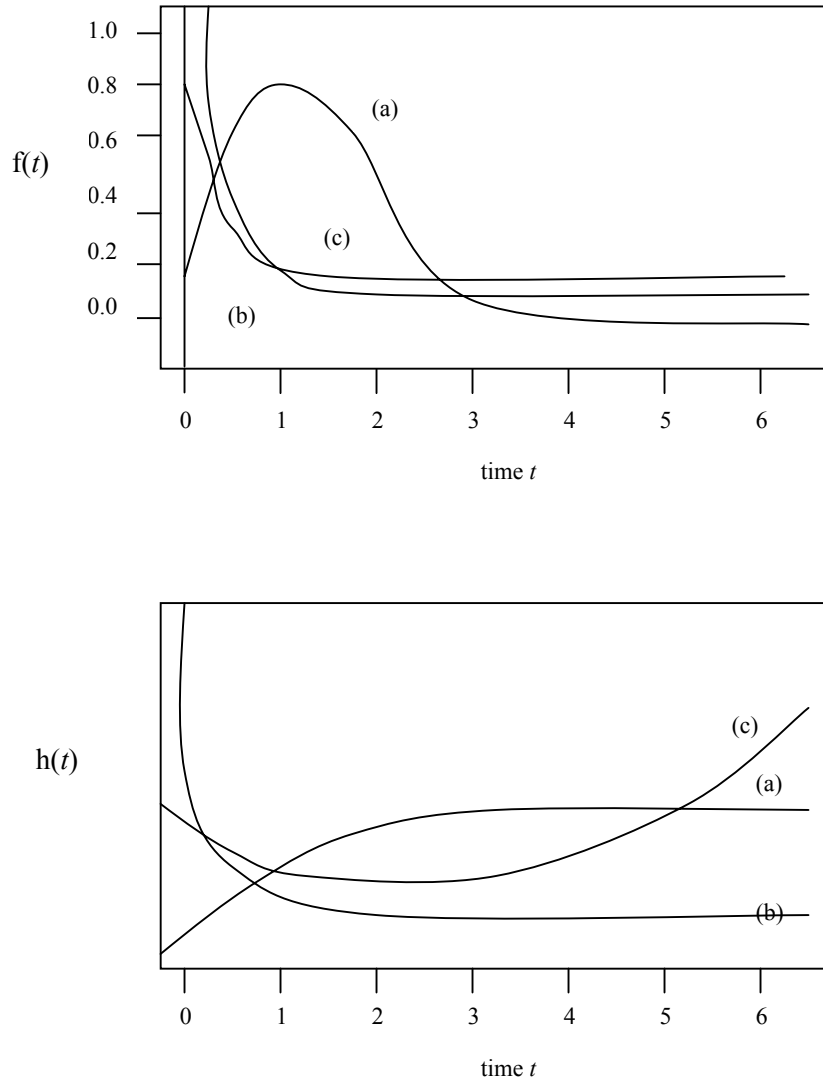


Figure 1.3 *Types of related hazard and density rates*

1.6 Why use survival and event history analysis?

Now that the basic concepts have been outlined, we can now adequately address the very fundamental questions of: Why is it necessary to engage in survival and event history analysis? What does it offer that ordinary regression models do not?

Why use survival and event history analysis? Anyone working with censored data should seriously consider using these types of methods. A simple descriptive example, adapted from Tableman and Kim (2004) provides a useful case to understand why we need survival analysis to deal with data that has censored information. The example draws from the Embury et al. (1977) study shown in Table 1.1, which contains the results from a clinical trial that examines the efficacy of maintenance chemotherapy for individuals with acute myelogenous leukemia (AML). The goal of the study was to see if maintenance chemotherapy extended the time until relapse. Once patients entered into remission after treatment by chemotherapy, they were randomly assigned into two groups of a 'maintained' (continued to receive maintenance chemotherapy) and a 'nonmaintained' (control group who received no chemotherapy).

Figure 1.2 illustrates how the simple mean duration of the time until the event (death or censoring) would be calculated if we ignored censoring (i.e., removed censored observations) versus accounting for censoring. Using the first approach, the difference between the two groups appears to be negligible (also between the medians that are 23.0 for both groups). We might conclude from the analysis that the survival time for the group that received maintenance chemotherapy was only slightly more skewed to the right or in other words, had only a marginally higher survival time than the control group. When we calculate the mean properly by also accounting for censored data, we see a markedly larger gap between mean in the maintained versus the nonmaintained group. Here the median also differs from 31.0 and 23.0 for the maintained versus the nonmaintained respectively. In fact, the actual distribution of the maintained group is far more right-skewed and the survival difference between the two groups is very large.

What does survival analysis offer that ordinary regression models do not? These techniques have several unique factors that distinguish them from other types of methods. A common question is: Why can't I just use OLS (ordinary least squares) or logistic regression? What is the added value of these models? First and foremost, survival analysis **adds information about timing**. This in turn makes it possible to account for 'censoring'. Unlike other techniques, it also takes a different approach by not only focusing on the outcome, but also analyzing the time to an event. This enables us to compare the survival between two or more groups and assess the relationship between explanatory variables and survival time. Another unique feature is the ability to include time-varying covariates (i.e., explanatory variables that change their values over time such as age or education), which is not possible in OLS or logistic regression. For this reason, these models are often referred to as dynamic.

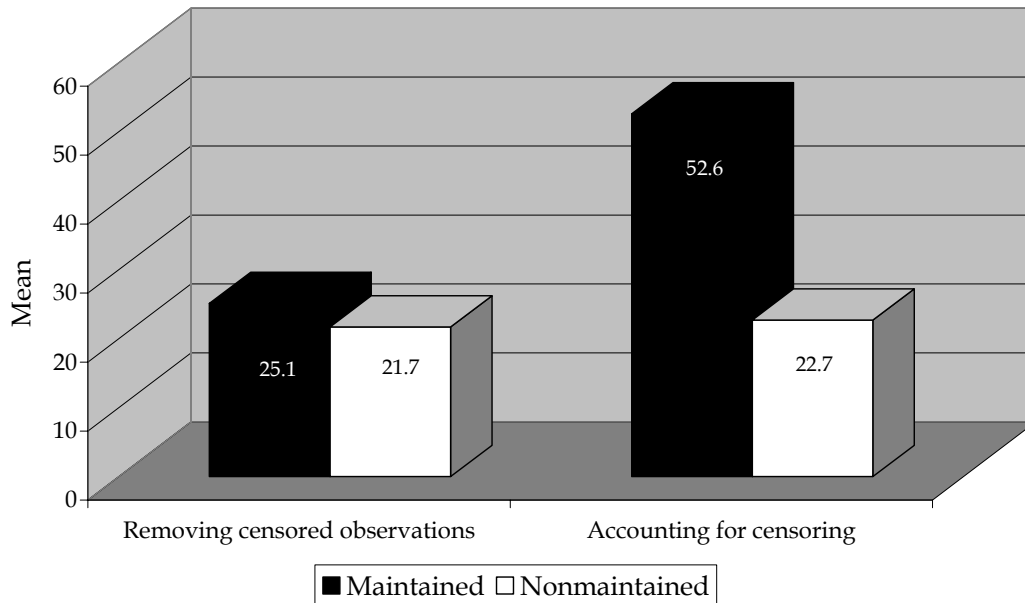


Figure 1.2 Difference in Mean in AML study by removing versus accounting for censoring

1.7 Overview of survival and event history models

This book will introduce you to various types of survival and event history models, which are summarized in Table 1.2. Do not feel disheartened if you do not understand some of the terminology or subtle differences between the models at this point. As you continue to read each chapter, the key aspects and advantages and disadvantages of each class of models should become clear. You can also refer to the detailed glossary at the end of this book for any unfamiliar terms. The table is not intended to be exhaustive, but to provide you with a brief overview.

Survival and event history analysis often begins with **non-parametric models**, which in this book focuses on the Kaplan-Meier or product-limit estimator. These analyses are often the first descriptive step in exploring and understanding your data and very useful as they make no assumption about the shape of the hazard or how the covariates can influence this shape. A central disadvantage of these models is that it is not possible to include or control for multiple covariates. The second class of models is often referred to as **semi-parametric models**, dominated by the widely used Cox model. This model has the advantage in that it is very flexible and does not make any restrictive assumptions about the shape of the hazard. It allows the inclusion of multiple covariates, but lives by the central assumption of the proportional hazards (i.e., there is a proportional hazard between groups over time), which we will explore

and test later in this book in detail. It is more flexible and less precise than parametric models and is not appropriate when you would like to test time-dependence or in other words, how the hazard may vary over time.

For **parametric models**, the researcher needs to decide in advance the shape of the hazard function and how covariates affect this function, which may be difficult if you are unfamiliar with the process under study, which is why researchers often start with the more flexible Cox model. When specified correctly, parametric models (e.g., Weibull, exponential), provide more precise parameter estimates. If incorrectly specified, however, the estimates can be seriously biased. Additional models include **event count models** such as the Poisson model, which examines event counts instead of the duration.

Models that are less often covered in survival and event history books, are models that examine not only one transition, but those that also analyze the entire state space. **Multistate models** are actually classic models and stem from applications in the early 1970s in demography (Rogers, 1973; 1975; Schoen, 1975; Willemkens, 1987). These models are what Willemkens (1987: 127) refers to as “a technique for describing the structure of life.” Using the theory of stochastic processes (i.e., consideration of random or probabilistic behavior) and the formulation of the model as a Markov-chain are essential components. Drawing from Hougaard (2000), the final chapter demonstrates how to estimate various types of these models. Later in this book you will also learn about a variant of these models, **competing risks models**, which are models that allow for the episode to end in two or more different outcomes. These allow a comparison of hazard functions across competing risks (e.g., competing causes of death). We will also examine **recurrent event models**, sometimes also referred to as multilevel or random effect models. These models acknowledge that the subject may experience the event more than once (e.g., multiple infections, jobs) and focus on understanding how the effect of covariates changes across episodes and consider the potential that due to some unmeasured cause (often referred to as unobserved heterogeneity), some subjects may be more likely to experience more repeated events than others.

Another type of advanced analysis we will examine is the technique of **sequence analysis**. Sequence analysis is commonly used by geneticists to examine the sequences of DNA, but can easily be used with various types of event history data to examine the entire sequence of a disease or employment history over time in order to define and compare sequences. Discrete Markov models and optimal-matching-based clustering are described with practical examples in the last chapter.

Table 1.2 Summary of survival and event history models

CLASS AND TYPE OF MODEL	DESCRIPTION	ADVANTAGES	DISADVANTAGES
NON-PARAMETRIC			
<ul style="list-style-type: none"> - life table estimates - Kaplan-Meier (Product-Limit) estimator 	<ul style="list-style-type: none"> - makes no assumption about shape of hazard - makes no assumption about how covariates affect shape of hazard - effects of covariates shown by stratifying data into groups 	<ul style="list-style-type: none"> - good method to understand basics and produce descriptive results 	<ul style="list-style-type: none"> - can only compare limited number of groups - cannot handle continuous data - does not allow inclusion of multiple covariates and multivariate controls
SEMI-PARAMETRIC			
<ul style="list-style-type: none"> - Cox model (most prominent) - piecewise constant exponential model 	<ul style="list-style-type: none"> - makes no assumption about shape of hazard - makes strong assumption about how covariates affect shape of hazard by assuming proportional hazard between groups over time 	<ul style="list-style-type: none"> - flexible model, often initial exploratory choice in analyses - allows inclusion of multiple covariates, multivariate analysis - results often similar to parametric models, but without (often) restrictive assumptions 	<ul style="list-style-type: none"> - not appropriate for testing hypotheses about time-dependence (i.e., how hazard varies over time) - less precise than parametric models - sometimes called 'overfitted'
PARAMETRIC MODELS			
<ul style="list-style-type: none"> - Exponential, Weibull, logistic, Gamma, Gaussian, complementary log-log, log-logistic, log-normal, Gompertz, Makeham, extreme value, Rayleigh and others 	<ul style="list-style-type: none"> - researcher needs to decide in advance shape of the hazard function and how covariates impact the hazard function - maximum likelihood methods - preferred when researcher wants to study the nature of time dependence and when time is meaningful in an independent variable - continuous and discrete-time models 	<ul style="list-style-type: none"> - more precise parameter estimates (if correct model assumptions) - allows multivariate analysis - allows analysis of discrete and continuous explanatory variables - specifies the shape of the hazard function, allowing for predictive modeling 	<ul style="list-style-type: none"> - if the hazard shape is incorrectly specified, parameter estimates can be seriously biased - needs preliminary work to first define shape of hazard function and understand how covariates affect the hazard function
EVENT COUNT MODELS			
<ul style="list-style-type: none"> - Poisson - Logit, probit, logistic, negative binomial regression (NBR models) 	<ul style="list-style-type: none"> - analyzes the number of events since a defined starting time - often for analysis of rare events (Poisson) 	<ul style="list-style-type: none"> - useful for analysis of rare events (Poisson, NBR) when number of zeros (i.e., number of trials without any events) is large 	<ul style="list-style-type: none"> - does not examine duration, but rather event-counts

Table 1.2 Summary of survival and event history models, continued

CLASS AND TYPE OF MODEL	DESCRIPTION	ADVANTAGES	DISADVANTAGES
MULTISTATE MODELS			
- Multistate models (also include competing risk, recurrent event and alternating state models)	- model for a stochastic process, which at any time point occupies one set of discrete states - specify state structure and form of hazard function for each transition	- appropriate for event-related dependence	- considers states, not events (problem for recurrent events) - all data considered longitudinal; less useful for repeated measurements
COMPETING RISK MODELS			
- Competing risk and multiple destination models: use one of the models described above (e.g., Cox) and make adjustments to risk group depending on whether risks are independent of one another	- episode can end in two or more different outcomes - central assumption is conditional independence of the risks under analysis	- considers more complex destination states - treat different reasons as different events, allowing comparison of hazard functions across competing risks	- problem if competing risks are not properly identified - hard to cope with assumption of conditional independence of the risks under analysis
RECURRENT EVENT MODELS			
- Recurrent event or multiple episode models - Frailty models, conditional frailty models (sometimes also referred to as multilevel models, random effect models)	- some subjects more likely to experience repeated event due to unmeasured cause (unobserved heterogeneity) - understanding how covariate effects change across episodes - Frailty: model as random effect - Conditional frailty: modifies frailty model to adjust for event dependence, stratify cases by event number	- goes beyond single-episode models that only compare effects between covariates to examine how covariate effects change across episodes - by estimating frailty as cause of unobserved heterogeneity as a random effect, coefficients for measured variables are less biased	- Frailty models may be badly biased if frailty is correlated with the covariates or the wrong distribution is assumed
SEQUENCE ANALYSIS			
- Discrete Markov models and optimal-matching-based clustering	- Obtain a matrix of proximities between sequences via optimal matching (or other metric) and cluster sequences via multidimensional scaling methods	- Provide a highly holistic view of entire event history - Derive prominent characteristics of complete trajectories	- Remains highly descriptive

