Alan B. Cantor

SAS Survival Analysis Techniques

for Medical Research

Second Edition

SAS Publishing

The correct bibliographic citation for this manual is as follows: Cantor, Alan B. 2003. *SAS® Survival Analysis Techniques for Medical Research, Second Edition*. Cary, NC: SAS Institute Inc.

SAS® Survival Analysis Techniques for Medical Research, Second Edition

ISBN 1-59047-135-0

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, January 2003

Table of Contents

Chapter 4 Proportional Hazards Regression

Chapter 5 Parametric Methods

Preface

Cox's proportional hazards regression model is one of the most popular methods used in survival analysis, and you will learn about it in Chapter 4, "Proportional Hazards Regression." Although Cox first described it in 1972, a search of Pub Med using the phrase "Cox Regression" finds the first article in medical literature using this method written in 1980. By the end of 1985, there were still only 69 such articles. During the 1990s, there were 1625 such articles. Of course, this simple search doesn't completely document the use of Cox regression during each of these time periods, but it does make the point that Cox regression, although an extremely useful tool for analyzing survival data, was not used much in medical literature until several years after it was published.

Now, you might think that this is a reflection of the slowness of biostatisticians to become aware of Cox regression. I don't think that this is the case. As a young assistant professor at a medical school in the late 1970s and early 1980s, I was aware of Cox regression but did not use it in a paper during those years. (I recall suggesting it once to a physician, but he rejected the idea as being too technical for a medical journal.) What augured against its use, by other biostatisticians and by me, was the difficulty of its implementation. Cox regression requires numerous matrix inversions and multiplications that make it impractical, in most real applications, to do by hand—even with a calculator. Computers were available in academic and research environments, but one would still have the task of writing a program in one of the popular languages of the time such as FORTRAN or BASIC.

With the implementation of Cox regression in commercial statistical software packages such as SAS software, and with the development of inexpensive desktop computers that could run this software, everything changed. Now, methods such as Cox regression, which would previously have required tedious calculation or computer programming, could be performed with a few lines of code on a computer sitting on the desk of a researcher or statistician. I'm convinced that this is why Cox regression became so much more widely used in the second half of the 1980s. Of course, this phenomenon has been repeated for a large number of statistical methods. As an experiment, you might want to note whether confidence bands for survival curves become more common in medical literature during the next few years. SAS is introducing them in Version 9.

While we are greatly indebted to those brilliant individuals who develop new methods that add to our armamentarium, we often fail to recognize the critical role of commercial software in allowing us to use these methods easily. Such software encourages the use of these methods by those who know about and understand them but who, in the absence of such software, would be reluctant to do the work required. But there is another, more insidious, effect as well. Such software enables those who do not understand the assumptions of a statistical method, or do not know how to interpret the results, to perform it anyway. This undoubtedly leads to a great deal of sloppy, and even misleading, reports in medical literature.

One of my goals in writing this book is to link the methods of survival analysis to the implementation of these methods in SAS software in a way that is accessible to those who are not professional statisticians. Hopefully, this will make some small contribution toward the remediation of the problem alluded to above. In doing so, I have not attempted to write "Survival Analysis for Dummies." I wouldn't know how to do that. The basic concepts of survival analysis cannot be understood without certain mathematical prerequisites. This includes some understanding of elementary differential and integral calculus. It also requires some understanding of statistics and probability. This includes the idea of likelihood-based estimation and inference.

In order to enable those without such a background to benefit from this book, I have included three appendixes. Appendix A provides the mathematical prerequisites for the material in this book. Appendix B presents the statistical background that the reader will need. Finally, because many readers will not have had experience with SAS software, Appendix C provides a brief introduction to SAS software. It is enough information to get started. I suggest that the reader glance through this material, and study those sections that are not already familiar. An instructor, using this book as a textbook, might want to look at the material in the appendixes and review, in the first part of the course, the subjects that the class needs.

A few other comments about this book are in order. First of all, SAS software is a moving target, so all I can do is provide information on the most up-to-date version. In mid-2002, when this book was nearing completion, that was an early release of Version 9. I am grateful to SAS for providing me with an early release of Version 9, and I've incorporated those new features that are related to survival analysis. Those readers who are using earlier versions will, of course, not be able to use them. Beginning with Version 7, SAS allows variable names with more than eight characters. Prior to Version 7, the limit was eight characters. In order to allow variable names to be more easily interpreted, I have used longer variable names in the macros, programs, and data sets that I included. To allow you to use the book's code without excessive typing, the code is provided on the SAS website. The details of how to download code are on the inside front cover. Users of earlier versions of SAS will have to find such variable names and shorten them. Readers of this book are strongly encouraged to download the programs and data sets and try them for themselves. You are also encouraged to make modifications and improvements. In addition to Base SAS and SAS/STAT, this book also makes extensive use SAS/GRAPH and SAS/IML software, so you should have those products installed. Finally, I make extensive use of the SAS macro language as well as of the IML procedure, two facilities that many readers will not be familiar with. That will not prevent you from using my programs, but you might want to spend some time studying their documentation to try to see how they work. It would be an additional benefit of this book if it leads to your learning these important skills.

Acknowledgments

In a book's preface, it is customary to thank those who have helped the author. Probably no author ever had more help from more people. I have been very fortunate to have had excellent instructors when I was a student. In addition, throughout my career I have learned a great deal from my colleagues. In this regard, some special words need to be said about Al Gross, with whom I worked from 1978 to 1983. Al, together with Virginia Clark, wrote the first widely used textbook on survival analysis. By doing so, they transformed the subject from a set of isolated articles in journals to a subject to be taught to students in graduate schools. In addition to his brilliance as a statistician, Al was known as a generous mentor to students and junior colleagues. His recent death was a source of great sadness to all who knew him.

I have also benefited greatly from my associations with biomedical researchers. For the past eight years, this has been at the H. Lee Moffitt Cancer Center and Research Institute. My colleagues probably thought that our primary relationship was that I was helping them. In fact, by providing me with challenging problems and constantly requiring me to enhance my skills, they helped me as much as I helped them.

The Survival Analysis class in the Department of Epidemiology at the University of South Florida in the Spring of 2001 deserves my special thanks as well. They served, without providing informed consent, as guinea pigs for much of the material in this book.

This book was vastly improved as a result of the critiques and suggestions of several reviewers, both external to and from within SAS. Their efforts are appreciated. As with the first edition, my thanks also go out to the folks at SAS, especially those in the BBU program. Their assistance has been vital.

And finally, my heartfelt thanks goes out to my dear wife Bootsie, for her support, encouragement, and love. Without that, this book, and much more in my life that gives me joy and satisfaction, would not be possible.

Chapter 1 What Survival Analysis Is About

1.1 The Nature of Survival Data

Survival data are special and, thus, they require special methods for their analyses. Before going into what makes these data special and how they are analyzed, let's establish some terminology and explain what is meant by survival data.

Although you might naturally think of survival data as dealing with the time until death, actually the methods that will be discussed in this book are used for data that deal with the time until the occurrence of *any* well-defined event. In addition to death, that event can be, for example:

- Relapse of a patient in whom disease had been in remission
- Death from a specific cause
- Development of a disease in someone at high risk
- Resumption of smoking by someone who had quit
- Commission of a crime by someone after serving a prison term
- Cancellation of service by a credit card (or phone service) customer
- Recovery of platelet count after bone marrow transplantation to some predefined level
- Relief from symptoms such as headache, rash, and nausea.

Note that for the first six examples, longer times until the event occurs are better; while for the last two, shorter times are better. Nevertheless, the methods to be described in this book can be applied to any of them. Note also that the methods we will be discussing can be used in a variety of settings, some nonmedical. For simplicity, words like "survival" and "death" are used in describing these methods, but you should be aware of the broader areas of applicability.

1.1.1 Cause- and Non-cause-specific Death

This might be a good place for a few words about cause-specific death. When analyzing survival of patients with some form of cancer, you might want to focus on death caused by cancer, particularly in an older population in which we expect deaths from other causes as well. You would then count only those deaths that are caused by cancer as "events." A death from any other cause would be treated the same as if the patient had suddenly moved out of state and could no longer be followed. Of course, this requires that you establish rules and a mechanism for distinguishing between cause-specific and non-cause-specific deaths.

As an alternative, there are methods of dealing with death from one cause, in the presence of competing risks. We will not be discussing such methods in this book. In the New York Health Insurance Plan study designed to assess the efficacy of mammography (Shapiro, et al., 1988), women were randomized to a group who received annual mammography or to a group that did not. Since the study's planners realized there would be considerable mortality not related to breast cancer, they took as their endpoint death caused by breast cancer. A committee was created to determine whether the death of a woman in the study was due to breast cancer. This committee, which was blinded with respect to the woman's group assignment, followed a detailed algorithm described in the study protocol. It's interesting to note that a recent *Lancet* article (Olsen and Gotzsche, 2001) called into question this study and others that are the basis for the widespread use of screening mammography. One of the authors' contentions was that the determination of cause of death was biased. We won't attempt to deal with this controversy here. In fact, the issue of what should be the endpoint for a study of a screening modality is a difficult one. In fact, because of the problems in assessing cause-specific mortality, some have advocated overall mortality as a more appropriate endpoint for such studies.

What makes analyses of these types of data distinctive is that often there will be many subjects for whom the event has not occurred during the time that the patient has been followed. This can happen for several reasons. Here are some examples:

- The event of interest is death, but at the time of analysis the patient is still alive.
- A patient was lost to follow-up without having experienced the event of interest.
- A competing event that precludes the event of interest has occurred. For example, in a study designed to compare two treatments for prostate cancer, the event of interest might be death caused by the cancer. A patient may die of an unrelated cause, such as an automobile accident.
- A patient is dropped from the study without having experienced the event of interest because of a major protocol violation or for reasons specified by the protocol.

In all of these situations, you don't know the time until the event occurs. Without knowledge of the methods to be described in this book, a researcher might simply exclude such cases. But clearly this throws out a great deal of useful information. In all of these cases we know that the time to the event exceeds some known number. For example, a subject who was known to be alive three years into a study and then moved to another state and could no longer be followed is known to have a survival time of at least three years. This subject's time is said to be right censored. A subject's observed time, t, is right censored if, at time t, he or she is known to still be alive. Thus you know that this subject's survival time is at least t. A survival time might also be left censored. This happens if all that is known about the time to death is that it is less than or equal to some value. A death is interval censored if it is known only that it occurred during some time interval. Although there is a great deal of current research on ways to deal with left- and interval-censored data, most survival analytic methods deal only with right-censored data, since this is the type of censoring most commonly seen. Of the three SAS procedures that deal explicitly with survival data, two deal only with right censoring. This is the type of censoring most commonly seen in medical research. The third, PROC

LIFEREG, discussed in Chapter 5, deals with left and interval censoring as well. Except for that chapter, and a brief discussion in Chapter 2, this book will not consider left- or interval-censored times and the term "censored" will always mean "right censored" unless another form of censoring is specified.

1.1.2 Random Variables

Survival data, therefore, are described by values of a pair of random variables, say (T, D). They can be interpreted as follows:

- T represents the time that the subject was observed on study.
- D is an indicator of the fact that the event in question either occurred or did not occur at the end of time T. The values of D might be 0 to indicate that the event did not occur and 1 to indicate that it did. Then D = 0 means that the corresponding T is a censored time. Of course, other values are possible.

The SAS survival analysis procedures, as well as the macros presented in this book, allow you to specify any set of values that indicate that a time is censored. This is convenient when you have data in which a variable indicating a subject's final status can have several values that indicate censoring. Subscripts will be used to distinguish the subjects. Thus, if there are n subjects on a study, their survival data might be represented by the n pairs $(t_1, d_1), (t_2, d_2), \ldots (t_n, d_n)$.

Sometimes, in textbooks or in journal articles, survival data are reported with only the time variable. Adding a plus sign to the time indicates censoring. For example, reporting survival data as 2.6, 3.7+, 4.5, 7.2, 9.8+ would mean that the second and the fifth observations are censored and the others are not. You can also store information on both the survival time and censoring value with only one variable. Making the time negative indicates censoring. Using this convention, these data would be 2.6, –3.7, 4.5, 7.2, –9.8. A SAS DATA step can easily be written to convert such a data set to the desired form. This is illustrated by the following example and output:

```
proc print data=original;
title 'Original Data Set';
data; set original;
d=1;
if time<0 then do;
        d=0;
        time=-time;
        end;
proc print;
title 'Modified Data Set';
run;
```

Output 1.1

Original Data Set

OBS	TIME
1	2.6
2	-3.7
3	4.5
4	7.2
5	-9.8

Output 1.2

Modified Data Set

OBS	TIME	D
1	2.6	1
2	3.7	0
3	4.5	1
4	7.2	1
5	9.8	0

There is another way of thinking about the random variables *T* and *D* described earlier. Each patient in the study is really subject to two random variables: the time until death (or the event of interest) and the time until censoring. Once one of these happens you can observe that time, but not the other. The value, *t*, of the random variable *T*, which you can observe, can be thought of as the minimum of the time until death and the time until censoring. The value, *d*, of the random variable *D* indicates whether that minimum is the time until death (d = 1) or the time until censoring (d = 0). An important assumption in all of what follows is that time until death and the time until censoring are independent. This would generally be true, for example, if censoring were due to the ending of the follow-up period.

On the other hand, suppose you are analyzing the data from a study in which patients with some sort of cardiac disease are randomized to drug treatment or surgery. In some cases it might later be decided that a patient who has been randomized to drug treatment now needs to have surgery. You might be tempted to take the patient off study with a censored survival time equal to the time until surgery. However, if the decision for surgery were based on the patient's deteriorating condition, to do so would create bias in favor of the drug treatment. That's because such a patient's death would not be counted as a death since he had previously been censored. A better approach might be to anticipate this possibility when planning the study. You might plan the study as a comparison of two treatment strategies: immediate surgery vs. initial drug treatment with surgery under certain conditions that are established in advance.

Another feature of survival data that distinguishes them from other types of data is the importance to estimation and inference of the distinction between nonparametric and parametric approaches. You will recall that methods such as *t* tests and ANOVAs, which are based on normally distributed random variables, tend to be valid even when data are not normally distributed if the sample sizes are reasonable large. When we are concerned about the appropriateness of the normality assumption we can replace the *t* tests and ANOVAs with nonparametric methods, such as the Wilcoxon Rank Sum Test and the Kruskal-Wallis Test, that don't require that assumption. With survival data, there are also parametric and nonparametric methods and often we need to choose between them. However, in this context, we generally don't even think about the data being normally distributed. Other distributions, some of which will be discussed later, are sometimes used, but the correctness of the distributional assumption that we make can be much more critical. To avoid making this choice, we often prefer nonparametric methods. Of course, the cost of using nonparametric methods is typically a less precise analysis. We will return to this issue later.

1.2 Exercises

1.2.1 Think of three other potential applications for survival analysis. At least one should be nonmedical.

1.2.2 Why do you think the planners of the NY HIP mammography study decided to use cause-specific mortality as the major endpoint instead of simply considering the cases diagnosed and their stage distribution? The latter approach would have made the study much shorter and cheaper. What do you think about the idea of using overall mortality, instead of breast-cancer-specific mortality as an endpoint?

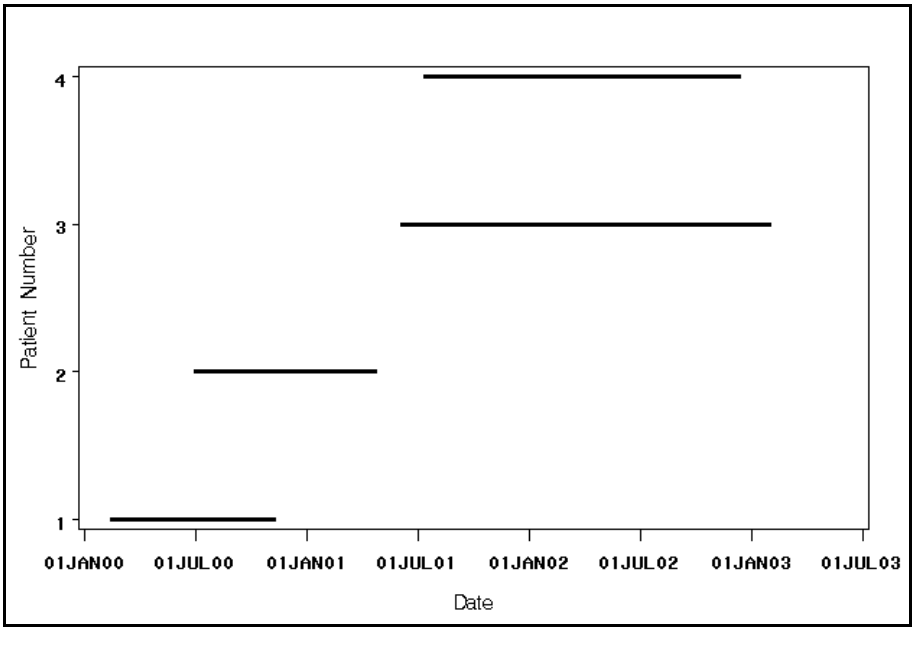
1.3 Calendar Time and Study Time

Another concept we need to discuss is how time is defined in survival studies. In most survival studies patients do not all begin their participation at the same time. Instead, they are accrued over a period of time. Often they are followed for a period of time after accrual has ended. Consider a study that starts accrual on February 1, 2000, and accrues for 24 months until January 31, 2002, with an additional 12 months of follow-up ending January 31, 2003. In other words, no more patients are entered on study after January 31, 2002, and those accrued are followed until January 31, 2003. Now consider the following patients:

- Patient #1: Enters on February 15, 2000, and dies on November 8, 2000.
- Patient #2: Enters on July 2, 2000, and is censored (lost to follow-up) on April 23, 2001.
- Patient #3: Enters on June 5, 2001, and is still alive and censored at the end of the follow-up period.
- Patient #4: Enters on July 13, 2001, and dies on December 12, 2002.

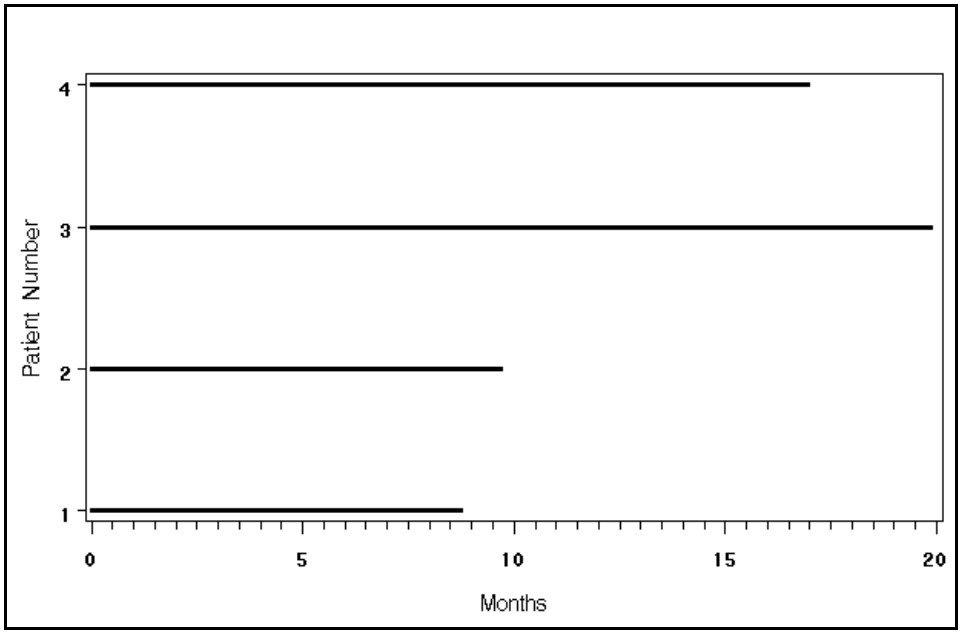
Their experiences are shown graphically in Figure 1.1.

Figure 1.1



In survival analyses, all patients are thought of as starting at time 0. Thus their survival experience can be represented as in Figure 1.2. When reference is made to the number surviving or the number at risk at some time, the time referred to is the time from each patient's study entry—not the time since the study started. For example, of the four patients we just described, two of them (#3 and #4) are still at risk at 12 months. None are still at risk at 24 months. Both of these facts are seen in Figure 1.2. If, at some later date, you speak of those at risk at t = 6 months, that has nothing to do with the situation on July 31, 2001, six months from the start of the study. Rather you mean those who, as of the last date that the data were updated, had been on study for at least 6 months without dying or being censored.

Figure 1.2



1.4 Exercise

For the data in the previous example, how many patients are at risk at 5 months? at 10 months? at 15 months? at 20 months?

1.5 Example

Sometimes, in studies involving follow-up of patients, there is interest in more than one time variable—for example, in an oncology study, time until death, which will be called survival time, and also time until death or relapse, which will be called disease-free survival time. The database will then have to contain information on both endpoints. Since SAS handles dates internally as numeric constants (the number of days before or after January 1, 1960) it is often convenient for the data sets to contain the dates of interest and to include in a SAS DATA step the statements to calculate the time values needed. As an example, consider a sample of patients treated for malignant melanoma. Presumably they are rendered disease-free surgically. Suppose that, in addition, they are treated with either treatment A or B, which are thought to inhibit relapse and improve survival. We might want to consider both survival and disease-free survival of these patients and how they are affected by treatment, tumor thickness, stage of disease, and tumor site. The first three records in the database might look like this:

Ptid	Date_of_surg	Date_of_relapse	Date_of_death	Date_of_last	Treatment	Site	Stage	Thickness
13725	10/5/95	11/6/96	1/5/97	1/5/97	A	1	III	1.23
25422	3/7/97	.	2/6/99	2/6/99	B	3	II	1.13
34721	9/6/94	.	.	3/18/2002	B	2	III	2.15

Note the inclusion of a unique patient identifying number, Ptid. While this number will play no role in the analyses of this data set, it is a good idea to associate such a number with each patient on a trial. This will facilitate merging with other data sets to add other variables of interest. Names are usually not good for this purpose because of the risk of spelling variations and errors. Also, we might want to exclude patient names in order to protect patient confidentiality. Note that Treatment, Site, and Stage are represented by codes or brief symbolic names. For obvious reasons we should avoid having long words or phrases for variables such as disease site or tumor histology. Treatment is a dichotomous variable. Although numbers are used for the possible sites, Site is categorical. The numbers used do not imply any ordering. Stage is ordinal with stages I, II, III, and IV representing successively more extensive disease. Finally Thickness is a continuous variable that is measured in millimeters. In later chapters, methods for dealing with all of these types of variables with SAS procedures will be discussed. In this case, missing values for date variables are used to indicate that the event did not occur. In order to analyze survival time and disease-free survival time, the following variables are needed:

Dfsevent — has the value 1 if the patient died or relapsed, 0 otherwise.

Dfstime — is the time, in months, from surgery to death or relapse if either occurred. Otherwise, it is the time that the patient was observed after surgery.

Survevent — has the value 1 if the patient died, 0 otherwise.

Survtime — is the time, in months, from surgery to death if the patient died. Otherwise, it is the time that the patient was observed after surgery.

The statements to add the variables needed to analyze survival and disease-free survival to the data set might look like this:

```
data melanoma; set melanoma;
  /*  Defining dfs time and event variables  */
dfsevent = 1 - (date_of_relapse EQ .)*(date_of_death EQ .);
  /*  Divide by 30.4 to convert from days to months */
if dfsevent = 0 then dfstime = (date_of_last -  date_of_surg)/30.4;
if date_of_relapse NE . then dfstime=(date_of_relapse -
date_of_surg)/30.4;
if date_of_relapse EQ . and date_of_death NE . then
dfstime=(date_of_death - date_of_surg)/30.4;

  /* Defining survival time and event variables */
survevent = (date_of_death ne .);
if survevent = 0 then survtime = (date_of_last - date_of_surg)/30.4;
else survtime = (date_of_death-date_of_surg)/30.4;
```

The divisions by 30.4 are simply to convert time from days to months, a more convenient time unit. Note that 30.4 is approximately 365/12. Also, when statements such as `(date_of_relapse EQ .)` or `(date_of_death EQ .)` are used in an arithmetic expression, they have the value 0 if false and 1 if true. The previous statements create the variables `Dfstime` and `Dfsevent` to be used in analyses of disease-free survival and the variables `Survtime` and `Survevent` to be used in analyses of survival. The first three observations of the resultant data set would look like this:

Output 1.3

PTID	DATE_OF_SURG	DATE_OF_RELAPSE	DATE_OF_DEATH	DATE_OF_LAST	TREATMENT	SITE	STAGE
13725	10/05/95	11/06/96	01/05/97	01/05/97	A	1	III
25422	03/07/97	.	02/06/99	02/06/99	B	3	II
34721	09/06/94	.	.	03/18/02	B	2	III

PTID	THICKNESS	DFSEVENT	DFSTIME	SURVEVENT	SURVTIME
13725	1.23	1	13.0592	1	15.0658
25422	1.13	1	23.0592	1	23.0592
34721	2.15	0	90.4605	0	90.4605

Now that these variables have been defined, there are several questions you might want to address. For example, you might want to estimate the survival and disease-free survival probabilities over time for the overall cohort and for subgroups defined by treatment, stage, site, and so on. Standard errors and confidence intervals for those estimates might also be desirable. This will be discussed in Chapter 2. You might also want to perform statistical tests to assess the evidence for the superiority of one treatment over the other. This can also be done. Methods will be discussed in Chapters 3 and 4. Now it might happen that the patients treated with treatment A were of worse prognosis (as seen by their stages, perhaps) then those treated with treatment B. If the treatment assignment was not randomized, this might happen if the treating physicians preferred treatment A for more advanced tumors. Even if the treatment assignment were randomized, it could happen by chance that one of the treatment groups had a higher proportion of patients with more advanced disease. You will learn how, using methods to be discussed in Chapters 3 and 4, to compare the two treatments after adjusting for stage. In addition, you will be able, if you make certain assumptions, to create a model that will produce estimated survival and disease-free survival probabilities for patients with specified values of three variables. Techniques for doing this are presented in Chapters 4 and 5. For example, you will learn how to estimate the probability that a patient with a stage II tumor of thickness 1.5 mm at site 1 treated by treatment A will survive for at least three years.

1.6 Functions That Describe Survival

1.6.1 The Distribution Function, Survival Function, and Density

The survival time of a subject being followed on a clinical study will be thought of as a random variable, *T*. As with random variables in other areas of statistics, this random variable can be characterized by its cumulative distribution function which, you will recall (see Appendix B) is defined by

$$F(t) = Pr[T \leq t], t \geq 0 \tag{1.1}$$

That is, for any nonnegative value of *t*, F(t) is the probability that survival time will be less than or equal to *t*. Of course, you could just as well describe the random variable, *T*, in terms of the probability that survival time will be greater than *t*. This function is called the survival function and will be denoted S(t). We then have

$$S(t) = 1 - F(t) = Pr[T > t], t \geq 0 \tag{1.2}$$

By convention, S(t) is usually used in survival analysis, although F(t) is more commonly used in other areas of statistics.

The density function, denoted f(t), is also used to describe a random variable that represents survival time. Recall that it is the derivative of the distribution function. Thus $f(t) = F'(t) = -S'(t)$.

1.6.2 The Hazard Function

Another very useful way of characterizing survival is by a function called the hazard function, which we will usually denote by h(t). It is the instantaneous rate of change of the death probability at time *t*, on the condition that the patient survived to time *t*. The formula for the hazard is

$$h(t) = \frac{f(t)}{S(t)} = \frac{-S'(t)}{S(t)}, t \geq 0 \tag{1.3}$$

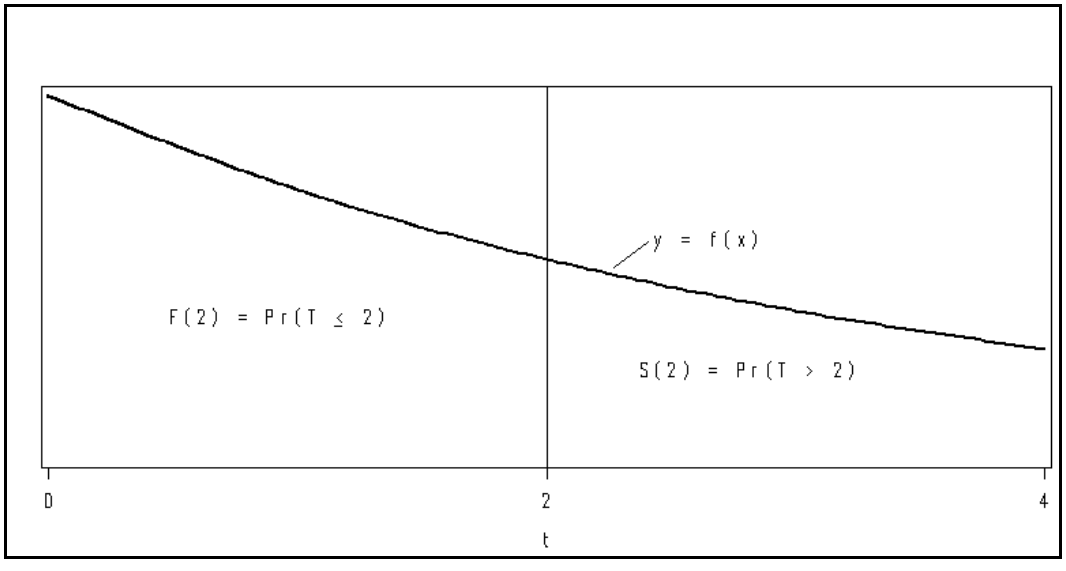
Although the hazard at time *t* conveys information about the risk of death at that time for a patient who has survived for that long, you should not think of it as a probability. In fact, it may exceed 1.0. A way to associate the hazard, h(t), at time *t* with a probability is to note that from (1.3) and the definition of the density, f(t), as given in Appendix B, we have the approximation, when Δt is near 0 of

$$h(t)\Delta t \approx \frac{F(t + \Delta t) - F(t)}{S(t)} \tag{1.4}$$

The numerator in (1.4) is the probability that the patient dies by time $t + \Delta t$ minus the probability that he or she dies by time *t*; that is, the probability that the patient dies at time between t and $t + \Delta t$. As noted earlier, dividing by S(t) conditions on surviving to time *t*. Thus the hazard at time *t* multiplied by a small increment of time approximates the probability of dying within that increment of time after *t* for a patient who survived to time *t*. This is a handy approximation that will be used later.

By the Fundamental Theorem of Calculus, if we plot the graph of the function y = f(t), then for any value, t_0, of *t*, $F(t_0)$ is the area above the horizontal axis, under the curve, and to the left of a vertical line at t_0. $S(t_0)$ is the area to the right of t_0. Figure 1.3 illustrates this property for $t_0 = 2$ and an arbitrary density function f(t).

Figure 1.3



If we take the definition of h(t) in (1.3) and integrate both sides, we get

$$\int_0^t h(u)du = -\int_0^t \frac{S'(u)}{S(u)}du = -\log[S(t)].$$

So that (1.5)

$$S(t) = e^{-\int_0^t h(u)du}.$$

The integral, $\int_0^t h(u)du$, in (1.5) is called the cumulative hazard at time *t* and it plays a critical role in long-term survival. If this integral increases without bound as $t \to \infty$, then S(t) approaches 0 as $t \to \infty$. In other words, there are no long-term survivors or "cures." If, however, the integral approaches a limit, $c < \infty$, as $t \to \infty$, then S(t) approaches exp(–c) as $t \to \infty$. In this case, we can think of exp(–c) as the "cure rate." Estimation of a cure rate is one of the most important and challenging problems of survival analysis. An approach to this problem will be presented in Chapter 5.

1.7 Exercises

1.7.1 Starting with a hazard function, $h(t) = \lambda t + \gamma$ for $\lambda > 0$ and $\gamma > 0$, find the associated survival function, S(t) and density, f(t).
1.7.2 Starting with a hazard function, $h(t) = \alpha\exp(-\beta t)$ for $\alpha > 0$ and $\beta \neq 0$, find the survival function, S(t), and density, f(t). What is the limit of S(t) as $t \rightarrow \infty$ if $\beta > 0$?
What is the limit of S(t) as $t \rightarrow \infty$ if $\beta < 0$?

1.8 Some Commonly Used Survival Functions

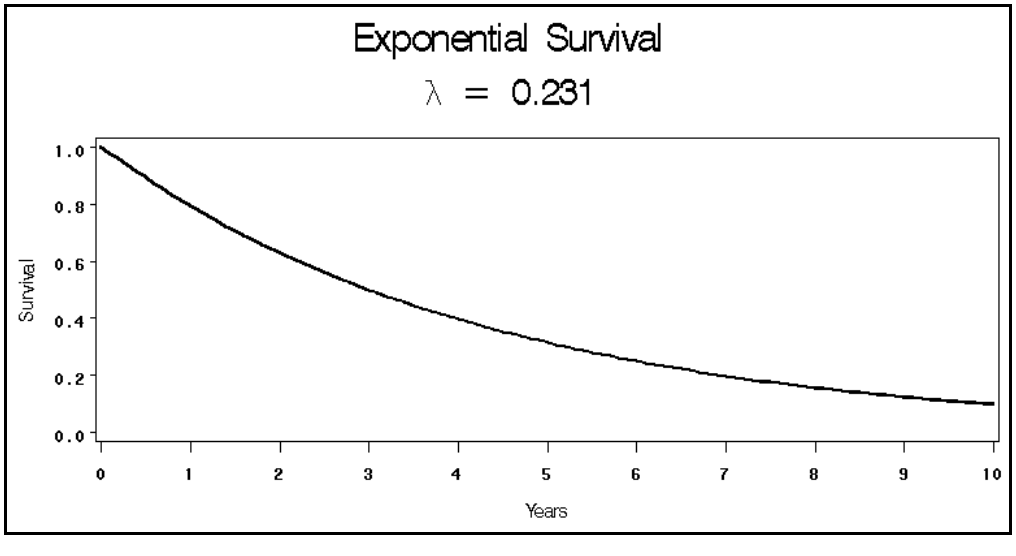
1.8.1 The Exponential Function

The simplest function that we might use to describe survival is the exponential function given by

$$S(t) = \exp(-\lambda t), \quad \lambda > 0, \; t \geq 0. \tag{1.6}$$

This survival function has only one parameter, the constant hazard, λ. The median survival time, defined as the solution of $S(t) = 0.5$, is easily seen to be $t = -\log(0.5)/\lambda$. Also, if we assume a probability of p of surviving for time t, then λ is determined by $\lambda = -\log(p)/t$. Because of its simplicity, the assumption that survival data are exponentially distributed is very popular, although its validity is sometimes questionable. A graph of an exponential survival function is given as Figure 1.4.

Figure 1.4



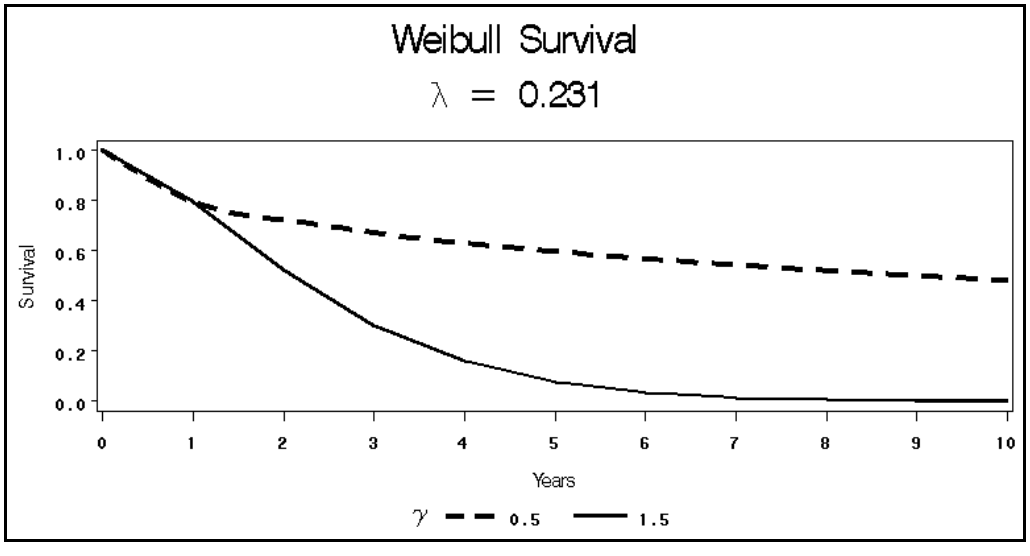
1.8.2 The Weibull Function

A more complex, but often more realistic, model for survival is given by the Weibull function

$$S(t) = \exp(-\lambda t^{\gamma}),\ t \geq 0, \lambda > 0, \gamma > 0 \qquad (1.7)$$

Note that the exponential survival function is a special case of the Weibull with $\gamma = 1$. The hazard function is given by $h(t) = \lambda\gamma t^{\gamma-1}$. It increases as t increases if $\gamma > 1$ and decreases as t increases if $0 < \gamma < 1$. Graphs of survival functions of each type are shown in Figure 1.5.

Figure 1.5



Other functions, such as the lognormal, gamma, and Rayleigh, are also sometimes used to describe survival, but will not be discussed in this chapter.

1.9 Exercises

1.9.1 Show that if $S(t) = e^{-\lambda t}$, the hazard function is the constant, λ, the density is given by $\lambda e^{-\lambda t}$, and the mean survival time is $1/\lambda$.

1.9.2 Find the value of λ and the median survival time for an exponential survival function if $S(3) = .4$.

1.9.3 Show that, for an exponential survival distribution, the probability of surviving past time $t_0 + t_1$ given that an individual has survived past t_0 equals the unconditional probability of surviving past t_1. In other words, the probability of surviving t_1 more units of time is the same at the beginning as it is after surviving t_0 units of time. This is often referred to as the "memoryless" property of the exponential distribution. Hint: In symbols, this means that $S(t_0 + t_1)/S(t_0) = S(t_0)$.

1.9.4 Show that the hazard function associated with a Weibull survival function is $\lambda\gamma t^{\gamma-1}$ and find the density function.

1.10 Functions That Allow for Cure

1.10.1 The Idea of "Cure Models"

The previously discussed survival functions are all based on proper distribution functions, that is $F(t) \to 1$ as $t \to \infty$. Of course this means that $S(t) \to 0$ as $t \to \infty$. Often, however, a model, to be realistic, must allow for a nonzero probability of indefinite survival—that is, a nonzero probability of cure. Suppose you were analyzing survival data for a cohort of children who had Hodgkin's Disease. You might find that a considerable number of patients were alive, apparently free of disease, and still being followed after ten years and that no deaths had occurred after four years. It would be reasonable to surmise that a nonzero proportion had been cured in this case. A survival function that goes to zero with increasing time would not be a good model for such data.

1.10.2 Mixture Models

One way to model such data is to assume that the population being studied is a mixture of two subpopulations. A proportion, π, is cured and the remaining proportion, $1 - \pi$, has a survival function as in Section 1.10.1. If, for example, the survival function of the non-cured patients is exponential, the survival of the entire population might be given by

$$S(t) = \pi + (1 - \pi)\exp(-\lambda t), \; t \geq 0 \tag{1.8}$$

The graph of such a survival function approaches a plateau at $S(t) = \pi$ as $t \to \infty$. Goldman (1984) and Sposto and Sather (1985) have studied this model. Of course, the exponential function in (1.8) can be replaced by any survival function. For example, Gamel et al. (1994) have considered such a model based on a lognormal survival function for the noncured patients.

1.10.3 The Stepwise Exponential Model

Another model that can allow for cure is the piecewise exponential model as described by Shuster (1992). This model assumes that the hazard is constant over intervals, but can be different for different intervals. For example, we might have $h(t) = \lambda$ for $0 \leq t < t_0$ and $h(t) = 0$ for $t \geq t_0$. For this model the survival function is given by

$$\begin{aligned} S(t) &= \exp(-\lambda t) \quad \text{for } 0 \leq t < t_0 \\ S(t) &= \exp(-\lambda t_0) \quad \text{for } t \geq t_0 \end{aligned} \tag{1.9}$$

1.10.4 The Gompertz Model

Still another model for survival that allows for cure is given by the Gompertz function defined by

$$S(t) = \exp\{-\frac{\gamma}{\theta}[\exp(\theta t) - 1]\} \quad \gamma > 0, \; t \geq 0 \tag{1.10}$$

Although this function appears to be rather complicated, it follows by (1.7) from the assumption that $h(t)$ is increasing or decreasing exponentially with rate θ as *t* increases. In fact, this function was first used by Gompertz (1825) to describe mortality in an aging male population in which he observed an exponentially increasing hazard. With $\theta < 0$ it's not hard to see that $S(t) \to \exp(\gamma/\theta)$ as $t \to \infty$. This function was first used to describe survival of patients by Haybittle (1959). It has also been studied in this context by others (Gehan and Siddiqui, 1973; Cantor and Shuster, 1992; Garg, Rao, and