

Chapter 8 Survival Analysis

8.1 Introduction

We've all been faced with midterm exams in college (maybe you just took one today). Besides your knowledge of the material on the test, what other challenges did you face when you sat there at your desk? Chances are, on at least one of the midterms you've taken during your educational career, you faced the pressure of completing your test in the time allotted. In fact, it's common for some students not to complete the exam in the hour (or two) provided. However, it's also common for some students to finish early. The point is that there is variability from student to student in the time required to finish an exam, and there is the real possibility that some students will not finish.

Now consider the collection of lengths of times students take to complete their exams. This collection of time measurements is an example of **survival data** or **time-to-event data**. So suppose that you take a statistics midterm along with the other 9 students in your class, and you have 60 minutes to finish. Here are some hypothetical times (in minutes and ordered from smallest to largest) for 10 students to complete the exam:

52 54 54 55 55 56 58 60 60 60

You'll notice that the last 3 times are all 60 minutes, the maximum allowed time for the exam. The three students with times of 60 minutes didn't finish the exam in the time provided. Hence, we have 7 recorded values of the time taken to finish the test, and we have 3 individuals with times that are greater than 60 minutes.

What problem do we face if we want to compute descriptive statistics for the time-to-event random variable, e.g. how do we compute the average time to complete the test? The fact that 3 students have **incomplete** times poses some challenges.

A variety of techniques and methods are available to analyze time-to-event data with incomplete observations, and these fall under the broad topic of **survival analysis**. In survival analysis studies, the response variable is the time until a target event (event-of-interest) occurs. Specific questions that can be addressed using survival analysis methods:

- When are individuals more likely to experience the event of interest, such as finishing a 60-minute exam?
- Who or what particular group of individuals is at the greatest risk of experiencing the event of interest during the time of the study or observation period? For example, are students of a particular major more likely to finish the 60-minute exam than students with another major?

In survival analysis studies, the response variable of interest, denoted by T , is the time until the event of interest occurs (also called the **failure time**, **survival time**, or **time-to-event** random variable).

Consider some specific examples of time-to-event random variables are:

- Time taken to complete a task, e.g. a midterm.
- Time-to-relapse for recently treated drug addicts (Public Health)
- Time until graduation for students entering a post-secondary institution. (Education)
- Time until death due to AIDS for those individuals diagnosed with HIV (Medicine)
- Age at first consideration of suicide (Sociology or Psychology)

Notice that in all the examples, the variable of interest is the duration of time until an event-of-interest occurs. **Time-to-event** data are fundamentally different from time series data where measurements on the same observational units are collected at different points in time. **Time-to-event** data are time durations until the observational unit experiences the target event.

Studies that use survival analysis techniques should possess the following features:

- A well-defined *event of interest* whose occurrence is being explored with individuals (or objects) “at risk” of experiencing the event (e.g. death due to AIDS, graduation from college).
- A clearly defined *beginning-of-time*: starting point (or time zero) when no object or individual under study has yet to experience the event (e.g. the date an individual is diagnosed with HIV, date student enters a post-secondary institution)
- A *meaningful scale for measuring time* (e.g. seconds, minutes, days, months, years, etc.)

Exercise 1: Midterm Exams. Suppose you wanted to analyze the times taken by students to complete their midterm exams. Then:

- 1) Describe the event-of-interest.
- 2) Define the time-to-event random variable, T .
- 3) Define a clear beginning of time
- 4) Provide a meaningful scale for measuring time for this study.

The survival time random variable can be described with certain numerical quantities and functions. Some quantities will already be familiar to you. For example, the **median lifetime** provides the time at which 50% of the subjects have experienced the event of interest. Other functions that are specific to the field of survival analysis, such as the survivorship function, hazard function, and cumulative hazard function, and the Cox regression model will probably be new to you.

Topics in survival analysis will be covered in two sections. Survival Analysis: Part A focuses on the features of time-to-event data and the **survival function**, the primary method for describing survival data. Survival Analysis: Part B discusses additional models and descriptive measures for analyzing survival data including the hazard

function, mean and median lifetimes, and the Cox regression model. Inference procedures for survival probabilities are also covered. Since, for most students, eating chocolate tends to be more enjoyable than taking exams, the data for the first lab will focus on conducting survival analysis on chocolate chips.

What type of chocolate chips melts faster- white chocolate or milk chocolate? When the instructor gives approval, place either a white or milk chocolate chip into your mouth and record the time until it melts. Flip a coin to determine which type of chip should be eaten first.

- 5) What is the target event (event-of-interest)?
- 6) What is an appropriate beginning of time?
- 7) What is an appropriate metric for time?
- 8) Record the number of seconds required for each chip to completely melt in your mouth or if you ran out of time. A brief discussion about the definition of "completely melts" should take place to ensure consistency in the recorded times.

What makes time-to-event data different from other types of data, e.g. heights, weights, income? Some time-to-event observations will be *complete* (we can observe the time until the event occurs), while others will only be partially known, or *incomplete*. In this chocolate chip activity and the midterm data presented at the beginning of the chapter contains censored data. **Censored survival times** occur when individuals may not experience the target event during a study's data collection time.

8.1 Censoring and Truncation

Two primary mechanisms responsible for incomplete times are called **censoring** and **truncation**. **Censoring** occurs when an individual's event time (time until the event of interest occurs) is not observed. Event times can be censored when the target event occurs outside the period of time that a study is conducted. **Truncation** occurs when a subject is observed only if his/her event time falls within a certain observational window of time, i.e. between (L, R) , where L = left truncation time, and R = right truncation time.

The fundamental difference between censoring and truncation mechanisms is that censoring pertains to when subjects *leave* a study (e.g. they die, they finish a midterm, etc.), while truncation refers to when subjects can *enter* a study. In addition, truncation is a selection process that applies to all individuals in the study, while censoring applies to particular individuals in the study.

An example of *right censoring*: In early childhood learning centers, interest often focuses on testing children to determine when a child learns to accomplish certain tasks, e.g. tie their shoe laces. The age at which a child learns the task is considered the time-to-event.

An example of *left censoring*: Consider investigating the lifetimes of computer batteries, i.e. the time until the battery burns out. A new fully charged battery is placed

in each of 10 computers and the computers are left running while you take an hour lunch break. When you return from lunch you start keeping track of the time until the power runs out.

An example of **left truncation**: Consider a survival study of residents in a retirement center. Ages at death are recorded, as well the ages at which individuals entered the retirement community. Since an individual must survive to a certain age, for example 65 years, to enter the retirement center, all individuals who died earlier will not enter the center, and have no chance to be in the study. In other words, individuals must survive to at least age $L=65$ years in order to be in the study, All other individuals are left truncated. Left truncation occurs when individuals do not come under observation until after a particular time, say L (for left truncated time). All times must exceed this fixed value L , and those subjects whose times do not exceed that value are excluded from analysis.

An example of **right truncation**: A study investigated the time until the infection and induction times for adults who were infected with the HIV virus through a contaminated blood transfusion. The data consist of the time in years measured from April 1, 1978 when adults were infected by the virus from a contaminated blood transfusion, and the waiting time to development of AIDS, measured from the date of infection. The waiting times from infection at transfusion to clinical onset of AIDS were estimated by retrospectively determining the transfusion times by inspecting a registry on June 30, 1986. Since the registry was sampled on June 30, 1986, infected individuals who developed AIDS after this data were not included in the study. These individuals are right truncated.

Survival data can be represented by pairs of random variables (T, C) . T is the time-to-event random variable representing the actual survival time of an individual. The censoring status indicator variable C takes on values of 0 to indicate that the event-of-interest was not observed and a value of 1 if the subject experienced the event of interest.

- 9) What type of censoring/truncation took place in the chocolate chip lab activity? For the chocolate chip melting times, write out the observed pairs of data (T, C) for the first 10 students in the study.

In addition, we will assume throughout that the censoring mechanism operates independently of event occurrence and the risk of event occurrence, known as noninformative censoring. The censoring is **noninformative** when it is caused by planned termination of follow-up or by a subject leaving a study for reasons unrelated to the risk of the event.

8.2 The Survival Function

Now that we've discussed particular features of time-to-event data and introduced some notation and terminology, we introduce techniques for drawing inferences about the distribution of the time-to-event random variable T .

The primary function used to characterize the distribution of a time-to-event random variable T is the **survival function** (or **survivorship function**) $S(t)$ given by:

$$S(t) = P(T > t).$$

$S(t)$ provides the probability of surviving (not experiencing the event-of-interest) beyond time t . Another interpretation of $S(t)$ is that it provides the proportion of subjects in a population who have yet to experience the event of interest by time t .

The survivorship function is treated as a parameter since it characterizes the population distribution of the random variable T . Hence, we refer to $S(t)$ as the **population survivorship function**, since it describes the survival experiences of all subjects in the population. If we know the exact distribution of the time to event random variable T , then (theoretically) we can calculate an expression for the survival function $S(t)$ using some calculus and additional topics from intermediate mathematical statistics.

When we do not know the exact probability distribution for T (which is most often the case), we will need to find an estimator for $S(t)$ based on a sample of survival data. Loosely speaking, methods that are based on a distributional assumption about T are called *parametric*, while methods that do not make assumptions about the distribution of T are called *nonparametric*.

- 10) Sketch a plot of the population survivorship function the lifetimes of dark chocolate chips might look like. What is $S(t)$ at time 0? Is $S(t)$ increasing or decreasing? Would you expect $S(t)$ to be continuous or discrete? What would you estimate to be the median lifetime?

8.3 Estimators of $S(t)$

More than likely, the exact distribution for T will be unknown so a nonparametric estimator of $S(t)$ based on a given sample of survival data will be needed. We will discuss two methods for estimating survival probabilities: the *empirical survival function* and the *Kaplan-Meier estimator*. The first method is only appropriate when there is no censoring in the data, while the latter method incorporates the censoring into the procedure.

When no censoring is present, we can estimate $S(t)$ using the **empirical survival function** given by:

$$\begin{aligned}\hat{S}(t) &= \text{Number of individuals alive at time } t \div \text{Total number of individuals in the study} \\ &= 1 - \frac{\text{Number of individuals failed at time } t}{\text{Total number of individuals in the study}}\end{aligned}$$

When censoring is present in the data, careful consideration is needed when estimating survival probabilities. We illustrate a procedure to estimate survival probabilities with the chocolate chip melting times. Suppose we have the following fictitious melting times (in seconds) of milk chocolate chips for 7 students where the maximum time allowed for the experiment was 60 seconds:

Student	1	2	3	4	5	6	7
Time	35+	39	60+	45	25	55	30+

where the “+” denotes right censored observations. Let's assume that Students 1 and 7 with respective times 35+ and 30+ withdrew from the “study”, while Student 3 with time 60+ had not experienced a melted chip by the end of the experiment.

- 11) First consider approaches to computing an estimated survival probability that ignore the censoring. In particular, use the following two approaches to calculate $\hat{S}(45)$, the estimated probability that it takes at least 45 seconds for a chocolate chip to melt.
 - a) Treat all the censored times as complete times, and use $\hat{S}(t)$ above.
 - b) Eliminate all times that are censored and use the remaining *complete* observations, and use $\hat{S}(t)$ above.

These estimates are, respectively, overly optimistic and pessimistic because they ignore the censoring in the data. Next we'll consider a way to estimate the survival probabilities taking into account the censoring status of subjects.

Developing the Kaplan-Meier Estimator of $S(t)$: In preparation for estimating the survival probabilities using the Kaplan-Meier estimator, we will first need to create a set of time intervals based on the observed chocolate chip melting times.

We will denote the number of *complete* event times by m . Next, order the complete event times, T , from smallest to largest. The smallest *complete* time will be labeled as t_1 , the second smallest as t_2 , and so on. The largest *complete* time is denoted t_m .

- 12) Consider the fake chocolate chip melting time data. What is m ? List t_1 through t_m for the fake chip data.

These complete times, t_1 through t_m , are used to define time intervals beginning at one complete event time, and ending just prior to the next complete event time with some minor modifications for the first and last intervals as outlined below:

- By convention, the *first interval* begins at time 0, t_0 , and ends just prior to the time when the first event occurs, time t_1 . This interval is denoted by $[0, t_1)$.

- If the largest event time is complete the **last interval** consists of just one point, i.e. the interval is given by $[t_m, t_m]$. If the largest event time is censored, then the interval extends to this censored time, but the interval is open on the right, i.e. the interval is given by $[t_m, \text{Largest Event Time})$.
- The **intervals between the first and last** are created using the complete event times as endpoints, i.e. $[t_i, t_{i+1})$, for $i = 1, 2, \dots, m-1$. If there are no ties in the data, (i.e. no two complete times are identical), then each interval will contain exactly one event.

The time intervals for the fake chocolate chip data is given in Table 1 below.

These time intervals are used to estimate \hat{p}_i , the conditional probability of experiencing the event in the i^{th} time interval, $[t_{i-1}, t_i)$, given that event has not occurred by the start of the interval.

$$\hat{p}_i = \frac{d_i}{n_i}$$

where d_i is the number of subjects who experienced the target event in interval i , and n_i is the total number of subjects (with complete and censored times) who are eligible (at risk) to experience the target event at the **beginning** of time period i .

Finally, the conditional probabilities can be used to calculate $\hat{S}(t)_{KM}$, the **Kaplan-Meier estimator** (or **product limit estimator**), for the population survivorship function $S(t)$. $\hat{S}(t)_{KM}$ is the estimate of the unconditional probability of surviving beyond time t . The Kaplan-Meier estimator considers survival to any point in the time as a series of steps by the observed survival and censored times. The estimator attempts to recover the true survivor function that would have been observed if no censoring occurred. Each step depends on successful completion of the previous step.

The following provides the formal mathematical definition for the Kaplan-Meier estimator. Suppose we have n individuals with observed event times t_1, t_2, \dots, t_n which include both uncensored and censored times. Let

- m = number of distinct uncensored event times, where $m \leq n$.
- t_1, t_2, \dots, t_m denote the **ordered complete** times, i.e. those times when the event-of-interest actually occurred ordered from smallest to largest.
- n_i = number at risk of experiencing the event at time t_i .
- d_i = number experiencing the event at time t_i .

Then the Kaplan-Meier estimator of the survivorship function is given by:

$$\hat{S}(t)_{KM} = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i} \right)$$

where \prod is the symbol for taking the products of terms $\left(1 - \frac{d_i}{n_i}\right)$ for all i such that the complete event times t_i are less than or equal to the time of interest t . Also $\hat{S}(t)_{KM} = 1$ if $t < t_1$.

[Mathematical Note:] The Kaplan-Meier estimator is derived using the Multiplication Rule from introductory probability. Let $A_i \equiv$ Event occurs after interval i . Then

$$\hat{S}(25)_{KM} = 1 - \hat{p}_1 = P(A_1) = P(A_1 \cap A_0) = P(A_1 | A_0)P(A_0) = (1 - \hat{p}_1)(1 - \hat{p}_0)$$

and

$$\hat{S}(30)_{KM} = P(A_2) = P(A_2 \cap A_1 \cap A_0) = P(A_2 | A_1 \cap A_0)P(A_1 \cap A_0) = (1 - \hat{p}_2)(1 - \hat{p}_1)(1 - \hat{p}_0)$$

13) Using the information derived above, complete Table 1.

Interval i	Time Interval	Number at Risk (n_i)	Number Censored	Number of Failures (d_i)	\hat{p}_i	$1 - \hat{p}_i$	$\hat{S}(t)_{KM}$
0	[0,25)	7	0	0	0/7	1	1
1	[25, 30)	7	0	1	1/7	6/7	6/7
2	[30, 45)	6	2	1	1/6	5/6	5/7
3	[45, 55)						
4	[55, 60)						

Table 1: Counts and estimated probabilities for fake chocolate chip data.

Notice that the value of $\hat{S}(t)_{KM}$ is constant for all t in each interval. From Table 1, we see that for example $\hat{S}(30) = 5/7$, i.e. the proportion of chips in the sample that have not melted after 30 seconds is $5/7$ or 71%.

- 14) Using the values $\hat{S}(t)_{KM}$ from Table 1, answer the following questions:
- What is the estimate for $S(45)$, i.e. what proportion of chips in the sample have not melted (survived) after 45 seconds?
 - How does your estimate $\hat{S}(45)_{KM}$ compare to the first two estimates computed in 11)?
 - Estimate the proportion of chips that have melted by 35 seconds.
 - Estimate the proportion of chips that have not melted after 50 seconds.

Once the survival probabilities have been estimated, a graph of the **Kaplan-Meier curve** can be constructed. The complete event times t_1, t_2, \dots, t_m are plotted against the values of $\hat{S}(t)_{KM}$. Figure 1 shows the Kaplan-Meier curve for the fake chocolate chip melting times was constructed Minitab.

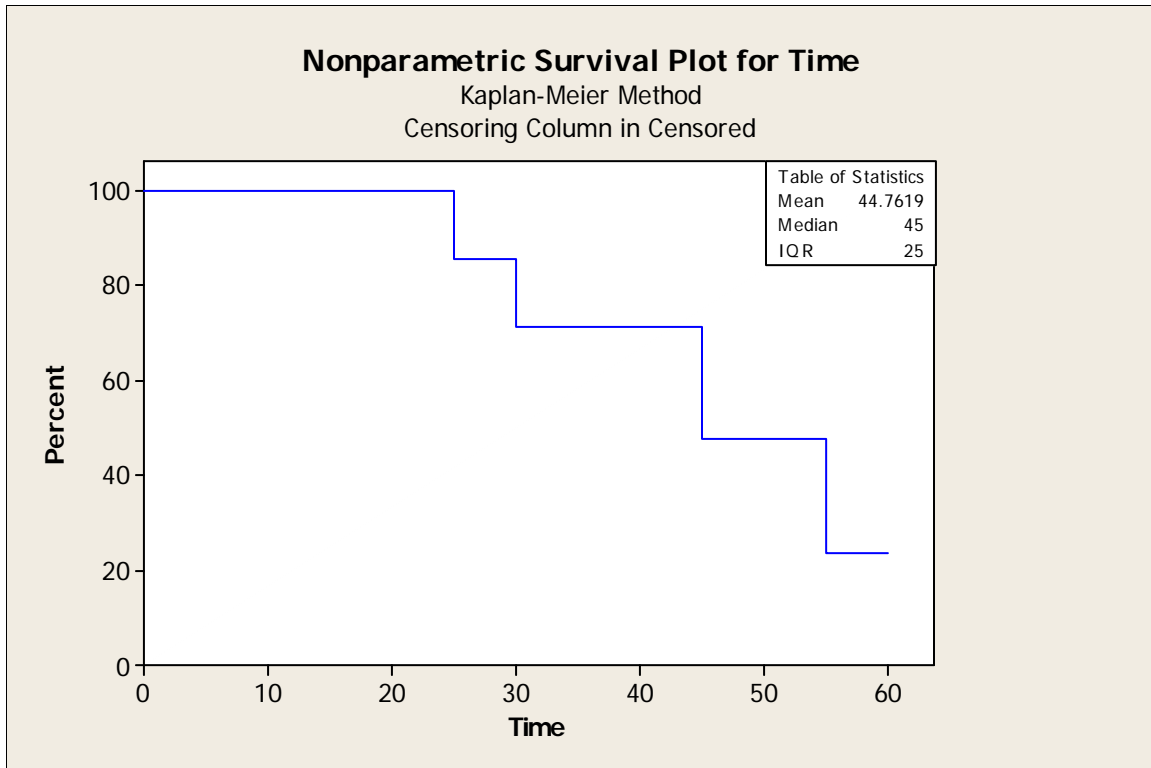


Figure 1: Kaplan-Meier estimated proportions (estimated survival probabilities) of chocolate chips.

From Figure 1 we can observe that the graph of the Kaplan-Meier curve is actually a decreasing series of steps with drops occurring at each complete event time t_i . This type of plot is called a **step function** because it looks like a series of steps.

The estimated probabilities remain constant within each time interval, and the height of each step corresponds to the value of $\hat{S}(t)_{KM}$ for t inside $[t_i, t_{i+1})$, $i = 1, \dots, m-1$.

A feature of Figure 1 is that the proportion of chips that remain unmelted after 60 seconds is nonzero, and so the last step of the Kaplan-Meier curve extends to the right up to 60 seconds. How would the curve change if the largest observed melting time were not censored?

- 15) From the in-class chocolate chip melting time experiment, construct a table similar to the one for the fake melting times. Then create the Kaplan-Meier curves for both the white and dark chips. If the largest observed melting time was uncensored, then repeat both parts of this exercise as if the largest time were censored. If the largest observed melting time was censored, then repeat both parts of this exercise as if the largest time were uncensored. Compare and contrast the two types of chips. Was there a period of time that experienced a rapid decrease in the (estimated) survival rates?
- 16) Repeat question 15) for the midterm exam times from the Survival Analysis Introduction.

8.4 Minitab instructions for calculating survival functions

Create 2 columns of data: Time and Censored. Use 1 to represent a censored item and 0 to represent an uncensored item.

In Minitab select Stat >> Reliability/Survival >> Distribution Analysis (Right Censoring) >> Nonparametric Distribution Analysis

In Variables: Time

Click Censor: Select Use censoring columns: Censored
Censoring Value: 1

Click Graphs: Survival Plot

Click Storage: Select Survival probabilities and Times for survival plots.