Goodness-of-Fit for Length-Biased Survival Data with Right-Censoring

Jaime Younger

Thesis submitted to the Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements for the degree of Master of Science in
Biostatistics [1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

Cross-sectional surveys are often used in epidemiological studies to identify subjects with a disease. When estimating the survival function from onset of disease, this sampling mechanism introduces bias, which must be accounted for. If the onset times of the disease are assumed to be coming from a stationary Poisson process, this bias, which is caused by the sampling of prevalent rather than incident cases, is termed length-bias. A one-sample Kolomogorov-Smirnov type of goodness-of-fit test for right-censored length-biased data is proposed and investigated with Weibull, log-normal and log-logistic models. Algorithms detailing how to efficiently generate right-censored length-biased survival data of these parametric forms are given. Simulation is employed to assess the effects of sample size and censoring on the power of the test. Finally, the test is used to evaluate the goodness-of-fit using length-biased survival data of patients with dementia from the Canadian Study of Health and Aging.

Acknowledgements

Above all, I would like to express my sincere gratitude to my supervisor, Dr. Pierre-Jérôme Bergeron, without whom the success of this thesis would not be possible. His guidance, patience, and encouragement helped me greatly throughout this endeavour. I could not have imagined a better supervisor for my Master's thesis.

On a more personal note, I would to give a special thanks to my boyfriend Pat whose encouragement, advice and confidence in me played a crucial role in the success of this thesis. I would also like to thank my family for their continued support. Lastly, I would like to give credit to my friend Brad for Figures 2.2, 2.3 and 2.4.

The data reported in this article were collected as part of the Canadian Study of Health and Aging. The core study was funded by the Seniors' Independence Research Program, through the National Health Research and Development Program (NHRDP) of Health Canada (project no. 6606-3954-MC(S)). Additional funding was provided by Pfizer Canada Incorporated through the Medical Research Council/Pharmaceutical Manufacturers Association of Canada Health Activity Program, NHRDP (project no. 6603-1417-302(R)), Bayer Incorporated, and the British Columbia Health Research Foundation (projects no. 38 (93-2) and no. 34 (96-1)). The study was coordinated through the University of Ottawa and the Division of Aging and Seniors, Health Canada.

Contents

List of Figures **vi**

List of Tables **vii**

1 Introduction **1**

2 Background of Survival Analysis and Length-Biased Data **5**

2.1 Overview of Lifetime Data 5

2.1.1 Length-Biased Data 10

2.2 The Survival Function and Hazard Function 16

2.3 Likelihood Construction 19

2.4 Estimators of the Survival and Cumulative Hazard Functions . . 21

3 Length-Biased Likelihood, Parametric Survival Models, and Goodness-of-Fit Tests **25**

3.1 Likelihood for a Length-Biased Sample 25

3.2 Parametric Models in Survival Analysis 30

3.3 Tests for Goodness-of-Fit 34

4 Algorithms **41**

4.1 Simulating Length-Biased Samples 41

4.2 Nonparametric Estimation of Length-Biased Survival Data with Right-Censoring 49
4.3 Simulation and Estimation of p-values 51

5 Applications **54**
5.1 Simulation Studies 54
5.2 CSHA Data 65
5.3 Analysis 67

6 Conclusion **84**

List of Figures

2.1 Example of Survival Function 7
2.2 Incident study. 9
2.3 Prevalent study. 10
2.4 Common vs. random disease onset times. 12
2.5 Unbiased and length-biased densities. 13
2.6 Weibull biased and unbiased densities. 13
2.7 Unbiased and length-biased survival functions. 14

5.1 Biased and unbiased nonparametric estimates of $S(x)$. 68
5.2 Nonparametric and Weibull estimate of $S(x)$. 71
5.3 Nonparametric and log-normal estimate of $S(x)$. 72
5.4 Nonparametric and log-logistic estimates of $S(x)$. 73
5.5 Nonparametric and Weibull estimate of $S(x)$ (women). 76
5.6 Nonparametric and log-normal estimate of $S(x)$ (men). 77
5.7 Nonparametric and Weibull estimate of $S(x)$ for 0-10 years (women) . 78
5.8 Nonparametric and log-normal estimate of $S(x)$ for 0-10 years (men) . 79
5.9 Estimated hazard function for women (Weibull) and men (log-normal. 80
5.10 Impact of small survival time on Nonparametric (NP) estimate. . . . 82

List of Tables

2.1 Censoring Schemes and the Likelihood Function 21

5.1 Weibull critical values - varying sample size & amt. of censoring . . 56
5.2 Power when H_0 is log-normal and Weibull is true distribution 57
5.3 Power when H_0 is log-logistic and Weibull is true distribution 58
5.4 Log-normal critical values - varying sample size & amt. of censoring 59
5.5 Power when H_0 is Weibull and Log-normal is true distribution . . . 60
5.6 Power when H_0 is log-logistic and Log-normal is true distribution . . 61
5.7 Log-logistic critical values - varying sample size & amt. of censoring 62
5.8 Power when H_0 is Weibull and log-logistic is true distribution 63
5.9 Power when H_0 is log-normal and log-logistic is true distribution . . 64
5.10 CSHA parameter estimates 71
5.11 p-values estimated by simulation 74
5.12 CSHA Estimates - Women (Weibull) & Men (Log-normal) 75
5.13 p-values estimated by simulation - Women & Men 81

Chapter 1

Introduction

Due to an increase in life expectancy, the effects of the baby boom, as well as a decrease in the number of children per woman, seniors account for the fastest growing age group in Canada. Consequently, dementia is becoming one of the most important disorders in our aging society. Its prevalence in Canada has been estimated at 8% in persons over 65 years of age, and rises to approximately 35% among those aged 85 years and above (Lindsay et al. 2004). The Canadian Study of Health and Aging (CSHA) [1] was a study of the epidemiology of dementia and other health problems in the elderly in Canada, and the survival data collected therein serves as the motivation for this thesis. One of the goals of the CSHA is to estimate the survival distribution, from onset, of those with dementia. The CSHA was initiated in 1991, when prevalent cases with dementia were identified and recruited into the study.

The sampling of prevalent cases is common in epidemiological studies as logistical constraints often prevent the recruitment of incident cases. The ideal settings for survival data are studies in which individuals are observed at the initiation of the event, or immediately after. The individuals are then followed for the remainder of the study,

[1] The CSHA is detailed in the following 3 papers - Canadian Study of Health and Aging: study methods and prevalence of dementia (1994), The Canadian Study of Health and Aging: risk factors for Alzheimers disease in Canada (1994), The Canadian Study of Health and Aging: patterns of caring for people with dementia in Canada (1994).

at which point they will be recorded as either censored or not censored, depending on their status. These studies can be termed 'incident studies.' The Kaplan-Meier estimator is a nonparametric estimator of the survival function for incident cases that are subject to right-censoring, and relies on the important assumption that censoring is independent about survival time. Unfortunately, due to issues such as time and cost, it is not always possible to capture individuals in this manner. Oftentimes, the subjects have experienced the event prior to the initiation of the study, and these lifetimes are said to be left-truncated.

The sampling of subjects from a prevalent cohort is a form of biased sampling. When subjects are identified cross-sectionally, the onset of disease has already occurred. Under this type of sampling scheme, individuals who are longer-lived have a higher probability of being recruited into the study, and there is a possibility that shorter-lived individuals may go completely unnoticed. Here, the recruited sample is not representative of the incident population. If we are interested in estimating the survival distribution, from onset of disease, the fact that the lifetimes are left-truncated must be accounted for. The time from onset until recruitment into the study is contained in both the time from onset until death and the time from onset until censoring. In this case, the assumption that censoring is uninformative about survival time is violated, meaning that standard techniques that rely on this assumption do not apply. When there has been no epidemic of disease, the incidence rate of the disease can be assumed to be constant over time. Under this scenario, the probability of sampling a subject is directly proportional to its length, and the sampled lifetimes are said to be length-biased.

Though nonparametric and semiparametric models are widely used for statistical inference, one may wish to infer upon a data set using a parametric model. A one-sample goodness-of-fit test is one in which the discrepancy between a nonparametric estimator and some hypothesized parametric distribution is quantified. For length-biased survival data subject to right-censoring, a nonparametric estimator has been

developed by Vardi (1989). In order to test several specific parametric models for a particular set of data, we seek a versatile one-sample test, one that is adaptable to various distributions. In this thesis, we propose a one-sample goodness-of-fit test for length-biased right-censored data that can be applied to several parametric models.

We will use a Kolmogorov-Smirnov type test to evaluate the goodness-of-fit of three parametric models to the CHSA data - a set of purely length-biased survival data subject to right-censoring. Goodness-of-fit techniques aim to quantify the discrepancy between an observed set of data and a hypothesized distribution, in hopes of finding a model that accurately describes the population from which the data arose. Kolmogorov-Smirnov type statistics rely on the empirical distribution function of the underlying data, which is the nonparametric maximum likelihood estimator (NPMLE). In the case of full information (i.e no censoring or truncation), the empirical survival function is the NPMLE for a set of survival data.

However, full information in survival analysis is generally not the case, and consequently the empirical survival function is not an appropriate estimator. The Kaplan-Meier (Kaplan & Meier 1958) estimator has been shown to be the NPMLE for censored survival data, and has also been modified to handle the general problem of left-truncation (Turnbull 1976) when stationarity does not hold. This modified Kaplan-Meier estimator can be viewed as a conditional approach to estimation. However, when data are properly length-biased, the unconditional NPMLE derived by Vardi (1989) has been shown to be more efficient. As a result, this estimator will be used to carry out our Kolmogorov-Smirnov type goodness-of-fit testing.

This thesis is organized as follows: Chapter 2 will present an overview of survival data, including length-biased survival data and how it arises in epidemiology and medicine. Some fundamental quantities and estimators will also be reviewed. Chapter 3 will detail how to construct the likelihood function for length-biased data, review some common parametric models arising in survival analysis, as well as several goodness-of-fit statistics used in modeling. Chapter 4 will be devoted to the

algorithms necessary to carry out our goodness-fit-testing. In Chapter 5, the power of the one-sample test will be investigated, and the results of our analysis using survival data from the CSHA will be presented. Finally, Chapter 6 serves as a conclusion and highlights some routes for further research.

Chapter 2

Background of Survival Analysis and Length-Biased Data

We will begin this chapter with an overview of lifetime data, where it arises, and some common peculiarities involved with analysis, namely censoring and truncation. Next, we will define and describe length-biased data and how it arises in epidemiology and medicine. The next section will focus on two key functions relating to lifetime data - the survival function and the hazard function. Following this we will look at how to construct the likelihood function under several scenarios, and finally some common estimators in survival analysis will be defined.

2.1 Overview of Lifetime Data

Survival data, also referred to as lifetime data or time-to-event data, arises in a variety of areas including medicine, epidemiology, public health, biology, economics, and manufacturing. The main goal in survival analysis is to study the time until a particular event occurs. Let X be the time until a specific event occurs. X, the duration under study, could be the time from onset of disease until death, time for a

disease to recur, time for an electrical component to fail, or time for a rehabilitated drug user to relapse. Time may be measured, for example, in years, months or days, or alternatively time may refer to the age of the subject when the event occurred. An event is also often referred as a failure, although the event in question is not always a negative one. For example, the duration under study could be the time taken for completion of a university degree. In this case the *failure* is obtaining a diploma, which is of course a positive occurrence.

The survival function, $S(x)$, is a fundamental quantity used to describe time-to-event data. $S(x)$ is defined as the probability that an individual survives to time x, $P(X > x)$, and is the complement of $F(x)$, the cumulative distribution function (i.e. $S(x) = 1 - F(x)$). If X represents the time from onset of some disease until death, then $S(x)$ would represent the probability that an individual survives beyond time x. In survival analysis, especially in the medical and epidemiological fields, the probability of an individual surviving past a certain time point is generally of more interest, hence $S(x)$ is used instead of $F(x)$. For a continuous lifetime variable X, $S(x)$ is a continuous, monotone decreasing function with $S(0) = 1$ and $S(\infty) = \lim_{x\to\infty} S(x) = 0$. Figure 2.1 provides an example of the survival function.

The analysis of survival data is complicated by a few common factors. To begin with, the main variable under study is a positive random variable that is generally not normally distributed. A second feature that proposes difficulties during analysis is censoring. In order to obtain a complete data set, each individual would need to be followed from the initiation of the event until the event occurs. In this situation, both the start and end time of the event under study are known, in which case the entire lifetime is known. Due to issues such as time and cost, it is not always feasible to follow a subject from initiation until they experience the event under study, which means that the variable of interest may not be fully observed. Hence, information collected on some subjects is unavoidably incomplete. Sometimes the only information available is that the individual survived until at least a certain

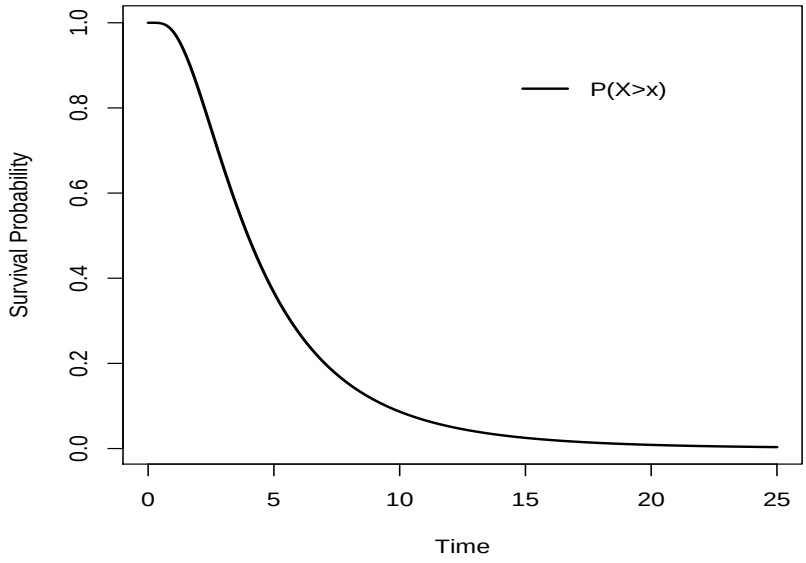

Figure 2.1: Example of Survival Function

point in time. This situation is termed *right-censoring*. An individual dropping out of the study, being lost to follow-up, or surviving past the conclusion of the study are all examples of right-censoring. Here, we know that the event of interest will occur to the right of our data point.

With right-censoring, data on each individual generally consists of a time under study, and a binary variable, δ, that indicates whether this time is an event time or a censoring time. Generally, δ=1 represents an event time, and δ=0 represents a censoring time. So for example, if an individual has the corresponding data point (x=62, δ=1), where x represents age, then we know that this individual experienced the event under study at age 62. For an individual with the data point (x=57, δ=0), we know that this individual was under study until age 57 and had not yet experienced the event.

Left censoring occurs when the event is known to have occurred before a certain time. In this case, the event of interest has already occurred at the time of observation,

but exactly when the event occurred is unknown. Consider studying the time until an individual develops a certain disease which commences without detectable symptoms. The exact onset of disease is unknown, and all that can be said is that the onset occurred prior to the first positive test. Here, the event of interest occurred to the left of our data point.

Interval censoring occurs when a failure time is known fall within two time points. Interval censoring is common in studies where the participants are not constantly monitored, but instead observed at certain time points during the study. For example, consider a cohort of individuals being followed until they contract malaria. When a subject first tests positive for malaria, the exact time of disease contraction may not be known, and the only conclusion we can draw about the event time is that it occurred between the time of the positive test and the previous testing date. Note that left-censoring is a form of interval censoring, since with left censoring we know that the event occurred sometime between time zero and the time of observation.

Another attribute of lifetime data is truncation, which should not be confused with censoring. Truncation refers to the idea that only individuals whose lifetimes fall within a certain interval can be selected for the sample. No information is available on individuals whose event times fall outside of this window, which is in contrast to censoring where we have at least partial information on each individual. If the survival time with a certain condition is being investigated, the participants must survive long enough to take part in the study. Those who die before the study begins do not have a chance of being included in the sample. For example, suppose we wish to study how long patients who have suffered a stroke survive at home after their initial hospitalization. Stroke patients who do not survive the initial hospitalization will not be included in the sample. This is known as left-truncation of the sample. Right-truncation occurs when an individual's event time must be less than some threshold in order to be included in the study.

The ideal setting for lifetime data is what is commonly referred to as an incident

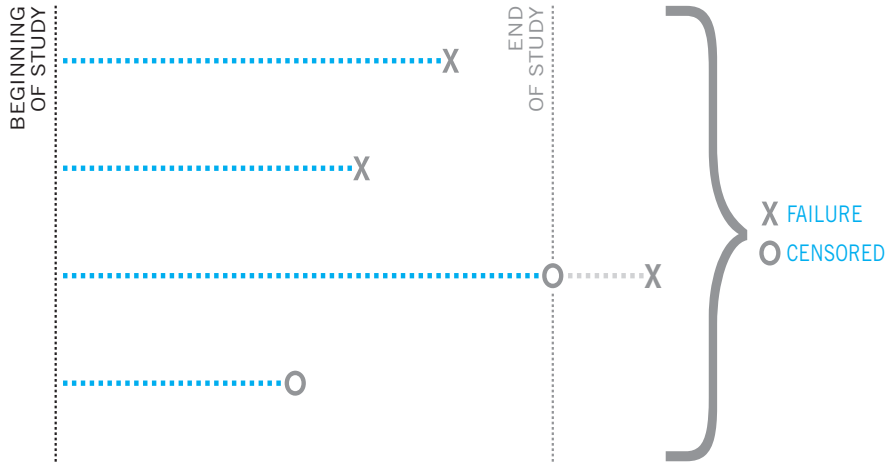

Figure 2.2: Incident study.

study. In this type of study, subjects are recruited before or at the time of initiation of the event in question. Figure 2.2 illustrates this scenario.

The subjects are followed until the event occurs or the study ends, whichever comes first. Those who have not experienced the event at conclusion of the study period are censored. For the censored individuals, it is known that they survived with the disease at least until the end of the study (note - individuals may also be censored if they drop out of the study or are lost to follow up, in which case they will be censored at the last time they were known to be alive). For analysis under this scenario, available methods (e.g. Kaplan Meier estimator) require the assumption of non-informative censoring. Non-informative censoring means that an individual's censoring time is independent of their failure time. In other words, non-informative censoring means that someone who has been censored will have the same risk of death as someone who experienced the event.

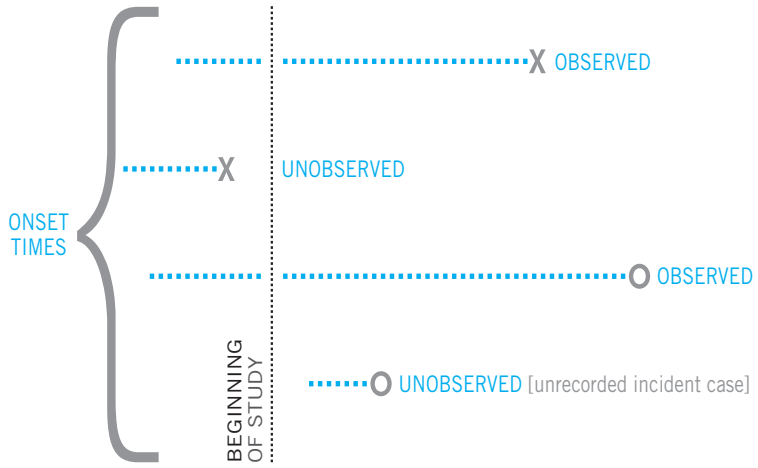

Figure 2.3: Prevalent study.

2.1.1 Length-Biased Data

It is often the case with epidemiological studies that prevalent cases with a certain disease are identified through a cross-sectional study and followed forward in time, and this is an example of how length-bias may arise in survival analysis. A cross-sectional study provides a 'snapshot' of the characteristics of the population at a particular point in time. For the recruited individuals, onset of the disease occurred prior to the initiation of the study, and these lifetimes are said to be left-truncated. Figure 2.3 illustrates this scenario.

Cross-sectional sampling from a prevalent cohort may be implemented because the disease in question is quite rare, or simply due to certain time and cost constraints that prevent the recruitment of incident cases. When individuals selected for a study already have the disease in question, the situation is no longer an ideal one, and the fact that lifetimes ascertained in this fashion are left-truncated must be taken into account. Under this sampling scheme, it is more likely that longer-lived individuals will be recruited into the study, and shorter-lived individuals may go completely

unnoticed. As the probability of recruiting a longer-lived individual is higher than that of recruiting a shorter-lived individual, it follows that the prevalent population is not representative of the incident population. If one is interested in estimating the survival distribution from onset, using prevalent cases instead of incident cases will overestimate the survival probabilities of the true population, since the sample of prevalent cases is more likely to be made up of individuals who are longer-lived.

As an example, consider using data collected cross-sectionally on individuals with dementia. In order to be included in the sample, the individuals must be alive with dementia at the beginning of the study. Individuals with dementia who did not survive long enough to be included in the study, will not be accounted for. Since it is necessary to be alive at the time of recruitment, individuals with shorter survival times are more likely to be missed by the study. Hence, the sample will be made up those who survive longer with the disease. Using this sample to estimate the survival time of individuals with dementia in the population will overestimate the true survival, since those who die quickly with dementia do not have an impact on the estimation.

For left-truncated data to be properly length-biased, stationarity of the process generating the onset times is required. The incidence process of the disease can be assumed to be a segment of a stationary Poisson process if there has been no epidemic of the disease during the onset times of the subjects (Asgharian et al. 2002). When this assumption holds, the probability of an individual being included in the study is directly proportional to their lifetime, and the sampled lifetimes are said to be properly length-biased. For properly length-biased data, an unbiased density function, $f_U(x)$, has the corresponding length-biased density, $f_{LB}(x)$, given by (Cox 1969)

$$f_{LB}(x) = \frac{x f_U(x)}{\mu}, \tag{2.1.1}$$

where μ is the mean of $f_U(x)$.

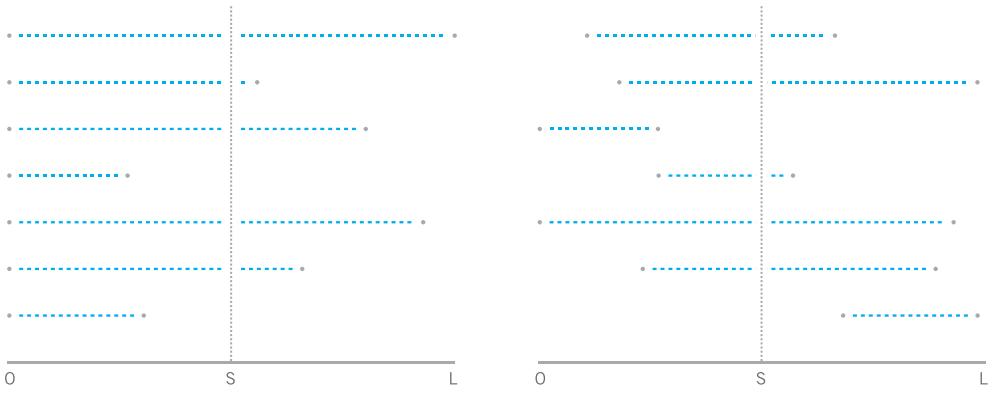

Figure 2.4: Common vs. random disease onset times.

To see the importance of the Poisson assumption, consider the diagrams in Figure 2.4. The dashed blue lines represent an individual's lifetime with a particular disease, from onset until death. In the first figure, all of the disease onset times are the same. S signifies the beginning of the study. In order for an individual to be included in the study, they must have a lifetime $X \geq S$. The corresponding density in this case is $\frac{f_U(x)}{S(s)}$. In the second figure, disease onset times appear to begin randomly. Here, the individual must be alive at the time of study recruitment, but shorter-lived individuals still have the possibility of being selected. When the onset times are generated according to a stationary Poisson process, the probability of selecting an individual is directly proportional to their lifetime with the disease.

Figure 2.5 shows an unbiased density along with its corresponding length-biased density. Notice that the length-biased density shifts the weight toward higher values of the variable. Note that how the length-biased density is shifted depends on the density parameters. Figure 2.6 shows two different Weibull densities along with their associated length-biased densities.

If the length-bias is not accounted for, the survival distribution will be overestimated. In terms of real applications, this overestimation can cause problems. For example, if we overestimate the survival probabilities of individuals with dementia,

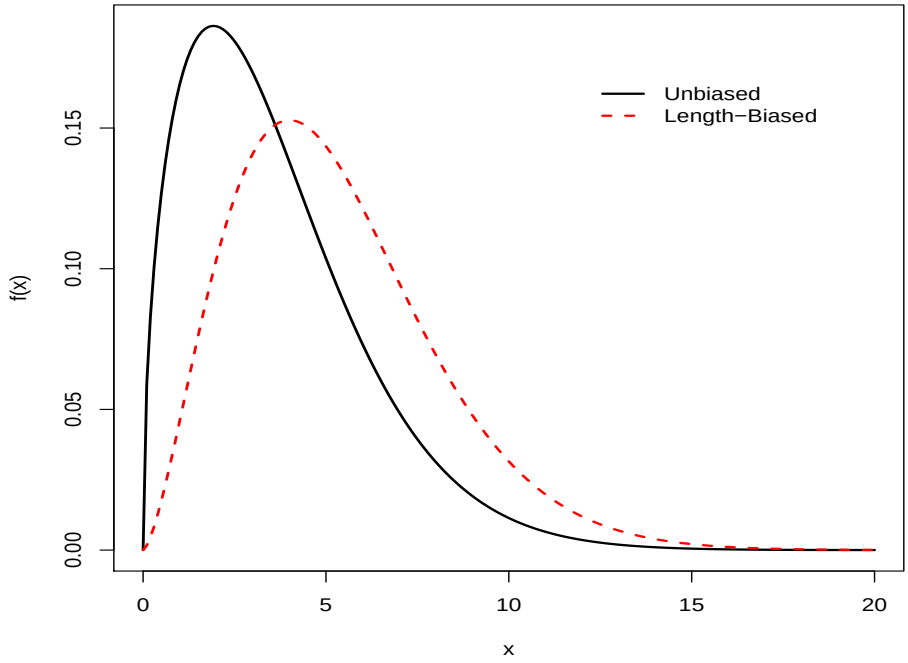

Figure 2.5: Unbiased and length-biased densities.

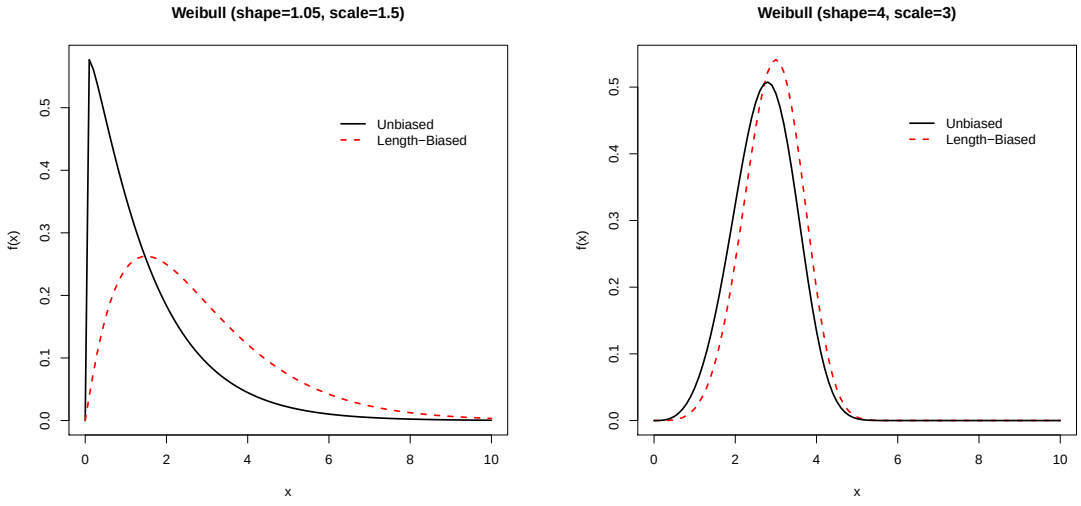

Figure 2.6: Weibull biased and unbiased densities.

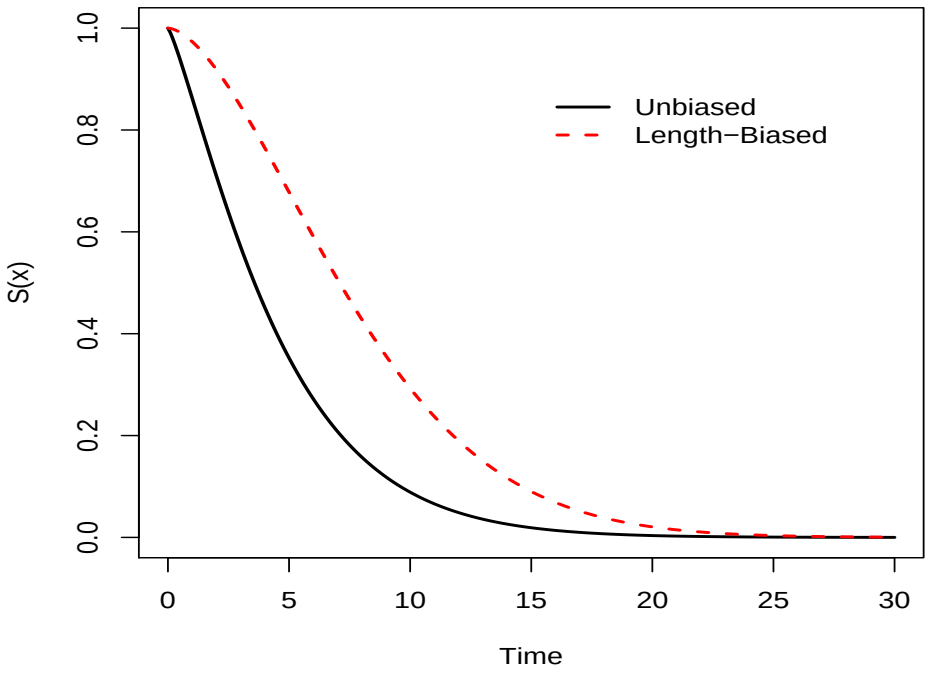

Figure 2.7: Unbiased and length-biased survival functions.

this translates to an overestimation of predicted survival once diagnosed with dementia, as well as an overestimation of the cost of dementia on the health care system. Figure 2.7 illustrates how ignoring the length-bias leads to an overestimation of the survival function.

Length-biased sampling has been studied for many years, with the first paper known to detail the problem mathematically written by Wicksell (1925). The phenomenon was further studied systematically by McFadden (1962), Blumenthal (1967), and Cox (1969). The literature has since been extensively developed, both in theory and application. In terms of advances of length-biased sampling in the medical and epidemiological fields, Zelen & Feinleib (1969) made significant contributions in the 60's and 70's. Much of Zelen's work concentrated on the early detection of chronic diseases, with a focus on breast cancer. The above papers are important in terms of

the development of the theory of length-biased sampling issues, but more importantly for this thesis is length-biased sampling when time is the main variable of interest. In medicine and epidemiology, length-bias arises most often through the sampling of subjects from a prevalent cohort. Lagakos et al. (1988) recognized the application of left-truncation in the study of AIDS from a prevalent cohort. Stern et al. (1997) analyzed data on patients with Alzheimer's disease from a prevalent cohort, but did not take into account the length-bias. This resulted in an overestimation of the survival function, which was validated by Wolfson et al. (2001), who showed that by correcting for length-bias, the median survival of patients diagnosed with Alzheimer's is much shorter. Gao & Hui (2000) used a two-phase sampling scheme to estimate incidence, again using data on patients with dementia.

When using a prevalent population to estimate the survival distribution from onset, the stationarity of the incidence process plays an important role in which estimation procedure should be used. When stationarity cannot be assumed, although bias still exists in the prevalent population compared to the incident population, the sampled lifetimes are not properly length-biased, and methods based on purely length-biased data cannot be used. In addition, the censoring times are informative, since the time from onset to recruitment is contained in both the full lifetime and censoring time, so methods used for incident cases will not apply to prevalent cases without adjustment. The bias stems from the lifetimes being left-truncated (since the event has already occurred) and this can be dealt with by conditioning upon the truncation times. This approach has been referred to as a 'conditional approach' as the estimates obtained are conditional upon the smallest truncation time. Turnbull (1976) developed an approach for dealing with left-truncation, which is an extension of work done by Kaplan & Meier (1958) who developed the nonparametric product limit estimator. Both the Kaplan-Meier estimator and the estimator modified for left-truncation will be discussed in further detail in section 2.4. The asymptotics of the product limit estimate with random truncation were derived by Wang et al. (1986),

and a nonparametric maximum likelihood estimation of the truncation distribution was derived by Wang (1991). An advantage of the conditional approach to estimation is that a truncating distribution need not be specified, however, when stationarity holds, the conditional approach does not offer the most efficient estimates for the survival function.

When stationarity can be assumed, an unconditional approach to estimation of left-truncated data can be used. This approach provides estimates of the survival distribution from onset, and hence has been termed an 'unconditional approach.' Vardi (1982, 1985) and Gill et al. (1988) developed an unconditional maximum likelihood approach which allows one to recover the unbiased survival distribution. Vardi (1989) derived the NPMLE for length-biased data with right-censoring; the asymptotics for the NPMLE were later given by Vardi & Zhang (1992). The algorithm for recovering the unbiased survival distribution developed by Vardi assumes that the number of censored and uncensored observations in known *a priori*, which is a very strong assumption. In a prevalent cohort study with follow-up, the number of censored and uncensored cases are not known until the end of the study. If this assumption is not fulfilled, though the likelihood remains the same, Vardi notes that the asymptotics will need to be derived separately. These asymptotic results for the NPMLE of both the length-biased and unbiased survival functions for length-biased, right-censored data were derived by Asgharian et al. (2002) and Asgharian & Wolfson (2005). An informal test for stationarity was developed by Asgharian et al. (2006) and the first formal test for stationarity was offered by Addona & Wolfson (2006).

2.2 The Survival Function and Hazard Function

Before going any further, it is necessary to define some quantities and relations used in the modeling of lifetime data. The notation used is adopted from Klein & Moeschberger (2003).

Let X represent the time until a specific event occurs. The survival function, the probability of an individual surviving past time x, is defined as

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

If X is a continuous, random variable, then $S(x)$ is a continuous, strictly decreasing function. Note that $S(x)$ is the complement of $F(x)$, the cumulative distribution function (*cdf*), and so

$$S(x) = P(X > x) = 1 - P(X \leq x) = 1 - F(x). \tag{2.2.2}$$

Using the fact that $\int_0^\infty f(x)d(x) = 1$ (X strictly positive),

$$S(x) = 1 - P(X \leq x) = 1 - \int_0^x f(t)d(t) = \int_x^\infty f(t)d(t). \tag{2.2.3}$$

We also have

$$f(x) = \frac{dF(x)}{dx} = \frac{d(1 - S(x))}{dx} = -\frac{dS(x)}{dx}. \tag{2.2.4}$$

An important property of the survival curve is that it is equal to one at time zero, and is equal to zero as time approaches infinity. It is always a monotone, nonincreasing function, and the rate of decline at time x is determined by the risk of experiencing the event at that instant. The survival curve is extremely useful in comparing mortality patterns, however determining the nature of a failure pattern by simply looking at the survival curve proves more difficult (Klein & Moeschberger 2003).

A second fundamental quantity in survival analysis is the hazard rate or hazard function. The hazard rate, $h(x)$, is defined by

$$h(x) = \lim_{\Delta x \to 0} \frac{P(x \leq X < x + \Delta x | X \geq x)}{\Delta x} \tag{2.2.5}$$

Looking at the numerator and denominator of the above expression separately may help with interpretation of the hazard rate. The numerator gives the probability

that the event will occur between time x and $x + \Delta x$, given that the event has not already occurred. When divided by the denominator, the width of the interval, a rate is obtained. By letting the width of the interval tend to zero, an instantaneous rate of failure is obtained. Unlike the survival function which is bounded between 0 and 1, the hazard rate can range from zero to infinity. From the equation above, the probability of an individual aged x failing in the next instant can be approximated by $h(x)\Delta x$.

When X is a continuous random variable, the hazard rate can be expressed as

$$h(x) = \frac{f(x)}{S(x)}. \tag{2.2.6}$$

The cumulative hazard rate, which can be thought of as the accumulation of hazard over time, is defined as

$$H(x) = \int_0^x h(t)d(t). \tag{2.2.7}$$

Note some useful relations between the survival function and the hazard function:

$$h(x) = \frac{f(x)}{S(x)} = \frac{-\frac{dS(x)}{dx}}{S(x)} = \frac{-d\ln(S(x))}{dx}, \tag{2.2.8}$$

$$H(x) = \int_0^x h(t)d(t) = \int_0^x \frac{-d\ln(S(t))}{dt} = -\ln[S(x)] \tag{2.2.9}$$

and

$$S(x) = \exp(-H(x)). \tag{2.2.10}$$

The hazard function is very useful in describing how the chance of an event occurring varies with time. There are numerous different shapes that the hazard function can take on, and some general shapes will be described, as well as scenarios where these shapes may occur.

An increasing hazard shape is when the hazard increases as time goes on, and this type of hazard is common. Natural aging and wear and tear are examples of increasing hazard. A decreasing hazard shape is when the hazard rate decreases with time, which

is not quite as common. A scenario where an individual may experience a decreasing hazard rate is following an organ transplant. When someone undergoes a transplant, their hazard rate may be very high immediately after the surgery, and slowly decline as they recover. A bath-tub shaped hazard is one that is initially decreasing, followed by remaining rather constant for a period of time, and finally increasing. This type of hazard is applicable to the human population followed from birth. Babies experience a higher hazard rate due to infant mortality in the early stages of life. After this stage the hazard rate decreases and remains relatively constant, and then the hazard rate will begin to increase as aging proceeds. Another interesting example that exhibits a bath-tub shaped hazard is an airplane flight. Typically, most accidents occur at, or close to take-off or landing. Lastly, there is the hump-shaped hazard, which sees the hazard rate initially increasing, followed by a decreasing hazard rate. An example where this hazard may be observed is after surgery. Initially, there is an increasing hazard rate due to the possibility of infection or related complications, and as the patient recovers the hazard rate gradually declines.

2.3 Likelihood Construction

As stated above, the incompleteness inevitable in lifetime data can pose certain difficulties during analysis, and it is necessary to pay careful attention to censoring and truncation when constructing the likelihood function. Whether an observation is censored, truncated, or an exact lifetime will have a different effect on the likelihood, and considering what type of information is provided by each observation will aid in a better understanding of the construction of the likelihood function.

When an exact lifetime is observed, information is gained on the chance of the event occurring at this exact time. This probability can be approximated by the corresponding probability density function (pdf) evaluated at the exact lifetime. When a right-censored observation occurs, the information given is that the individual sur-

vived past the censoring time. This probability can be approximated by the survival function evaluated at the censoring time. When a left-censored observation occurs, we know that the event has already taken place, and this probability is approximated by the corresponding *cdf* evaluated at the censoring time. For an interval-censored observation, it is known that the event occurred between two time points, and hence the information gained is the probability that event occurred during this interval (this can be calculated by using either $S(x)$ or $F(x)$). The idea is the same for truncated observations, but in this case a conditional probability is required. When left-truncation occurs, only the individuals who survive past the truncation time will be observed, and so the probability of the event occurring is conditional on the individual surviving at least until the truncation time. It follows that this quantity is approximated by the ratio of the *pdf* evaluated at the event time, and the survival function evaluated at the truncation time. Right-truncation is similar, except now we have the *cdf* evaluated at the truncation time in the denominator since this probability is conditional upon not experiencing the event by the end of the study period. The same idea follows for interval truncation. Table 2.1 summarizes how the likelihood function is affected under the different censoring and truncation schemes just described (Klein & Moeschberger 2003). Note that x represents an event time, C_r and C_l represent right and left censoring times respectively, and (Y_L, Y_R) represents the truncation window.

From here, the likelihood, L, for a data set is constructed by taking the product of each individual component. For example, consider a data set that consists of observed lifetimes and right-censored observations. The i^{th} observed lifetime and censoring time are denoted by x_i and r_i respectively. Let D and R represent the set of observed lifetimes and right-censoring times respectively. Then,

$$L \propto \prod_{i \in D} f(x_i) \prod_{i \in R} S(r_i). \tag{2.3.1}$$

The above likelihood can easily be adapted to the other censoring and truncation

Table 2.1: Censoring Schemes and the Likelihood Function

Likelihood Contribution	Censoring Scheme
$f(x)$	Observed Lifetime
$S(C_r)$	Right-Censoring
$1 - S(C_l)$	Left-Censoring
$S(C_r) - S(C_l)$	Interval-Censoring
$f(x)/[1 - S(Y_R)]$	Right-Truncation
$f(x)/S(Y_L)$	Left-Truncation
$f(x)/[S(Y_L) - S(Y_R)]$	Interval-truncation

schemes discussed above.

2.4 Estimators of the Survival and Cumulative Hazard Functions

Due to the incompleteness inherent in lifetime data, special techniques are required to properly draw inference about the survival distribution. The first estimator that will be described in this section is the Product-Limit estimator, also known as the Kaplan-Meier estimator, which gives an estimate for the survival function from onset. Next, the Nelson-Aalen estimator which estimates the hazard rate will be described. These estimates are appropriate for right-censored survival data, with non-informative censoring.

The Product-Limit estimator was proposed by Kaplan & Meier (1958) in order to estimate the proportion of the population whose lifetimes surpass time t, when there is right-censoring in the data. Kaplan and Meier refer to this quantity as $P(t)$, but it will be denoted as $\hat{S}(t)$ here. Suppose we have lifetime data on n individuals. It is possible that two events may occur at the same time (due to rounding), and it is

necessary to account for these ties. Let D be the number of distinct event times, and let t_i be the i^{th} distinct event time, such that $t_1 < t_2 < \cdots < t_D$. Let d_i represent the number of distinct events at each time t_i. Define Y_i to be a count of the number of individuals who are alive at time t_i, or who experience the event at time t_i (Y_i can be thought of as the number of subjects at risk of experiencing the event at time t_i). The Kaplan-Meier estimator is defined over the range of time where there is data, and is a decreasing step function with jumps at the event times. Note that the Kaplan-Meier estimator is not well defined for values of t greater than the event times in the data set (Kaplan & Meier 1958). The formula for the Kaplan-Meier estimator, $\hat{S}(t)$, as well as the estimator's variance, $\hat{V}[\hat{S}(t)]$, are given below.

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{if } t \geq t_1 \end{cases} \tag{2.4.1}$$

and

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}. \tag{2.4.2}$$

Due to the relationship between the survival function and cumulative hazard function shown above, the Kaplan-Meier estimator can also be used to estimate cumulative hazard function by $\hat{H}(t) = -\ln(\hat{S}(t))$.

As an alternative to using the Kaplan-Meier estimator to estimate the cumulative hazard function, there is a second estimator for $H(t)$ that performs better for smaller sample sizes. This estimator is due to work by Nelson (1972) and Aalen (1978), and is appropriately known as the Nelson-Aalen estimator. It is defined up until the largest observation time as

$$\hat{H}(t) = \begin{cases} 0 & \text{if } t < t_1 \\ \sum_{t_i \leq t} \frac{d_i}{Y_i} & \text{if } t \geq t_1. \end{cases} \tag{2.4.3}$$

The variance can be estimated by (Aalen 1978)

$$\sigma^2_{H_{NA}} = \sum_{t_i \leq t} \frac{d_i}{Y_i^2}. \tag{2.4.4}$$

Again, due to the relationship between the survival function and the hazard function, the Nelson-Aalen estimator can also be used to estimate the survival function.

The estimators described above are suitable for right-censored data. As left-truncation is common in survival analysis, it is necessary to have estimators that take into account both right-censoring and left-truncation. The estimators defined above can be slightly modified in order to accommodate this. Let L_j represent the age at which the j^{th} individual entered into the study, and T_j represent their corresponding time of death or censoring. Again, D is the number of distinct times, and t_i the i^{th} distinct event time with $t_1 < t_2 < \cdots < t_D$. Y_i, the number of individuals who are at risk of experiencing the event at time t_i, is slightly different when left-truncation is involved. Define Y_i as a count of individuals who entered into the study before time t_i and who have a study time of at least t_i. Equivalently, the risk set under left-truncation is the number of individuals with $L_j < t_i < T_j$. With this new definition of Y_i, the quantities above can be used for estimation when left-truncation is present (Klein & Moeschberger 2003).

However, it is important to note that now these estimates are conditional. For the Kaplan-Meier estimate, this means that the survival function for time t is an estimate of the probability of survival beyond t, given survival to at least the smallest truncation time. One caution when using the Kaplan-Meier estimator modified for left-truncation must be noted. If it happens that for some t_i, $Y_i = d_i$, then by definition the survival estimate will be zero for all values of t beyond this point. With left-truncated data, Y_i and d_i may be equal for small values of t, and even though prior to this point more survivors and deaths will be observed, the survival estimate will be zero, which is very uninformative. In circumstances such as these, the survival function is commonly estimated by only considering death times beyond this point

(Klein & Moeschberger 2003).

These modified estimators assume that the truncating distribution is unknown, and so conditioning upon the truncation times results in little information loss. When the data are properly length-biased, assumptions can be made about the truncating distribution, and in this scenario an unconditional analysis will be more informative. This unconditional approach, as well as common parametric models in survival analysis and goodness-of-fit tests will be discussed in detail in Chapter 3.

Chapter 3

Length-Biased Likelihood, Parametric Survival Models, and Goodness-of-Fit Tests

This chapter will begin with a look at how to write the likelihood function for a length-biased sample with right-censoring. Next we will talk about parametric models that are commonly used in survival analysis. And finally, some goodness-of-fits tests used in the modeling will be reviewed.

3.1 Likelihood for a Length-Biased Sample

In the case of full information (i.e. no censoring or truncation), the empirical survival function provides a nonparametric maximum likelihood estimate for the true survival distribution from onset. This is of little use, as full information in survival analysis is unrealistic in most real life applications, due to budget and time constraints. The Kaplan-Meier estimator is the nonparametric maximum likelihood estimate for life-time data with right-censoring, and reduces to the empirical survival function in the

case of no censoring. The Kaplan-Meier estimator is an efficient estimator for lifetime data in the presence of noninformative right-censoring, and can be modified to accommodate left-truncation, with the assumption that the truncating distribution is unknown. This modified Kaplan-Meier estimate conditions upon the truncation time, and can be referred to as a conditional approach to estimation. Conditioning upon the truncation time results in little information loss when the truncating distribution is unknown. However, if assumptions can be made about the truncating distribution, incorporating this information into the estimation procedure provides a more efficient estimate.

An unconditional approach to a nonparametric estimation of the survival function for length-biased survival data has been developed by Vardi (1989), but requires the assumption that the truncation times follow a uniform distribution. This is referred to as the assumption of stationarity and implies that the initiation times of a disease follow a stationary Poisson process (Wang 1991). A stationary Poisson process can be assumed so long as there has been no epidemic of the disease during the time period that covers the onset times of the subjects under study. An informal test for stationarity was investigated by Asgharian et al. (2006), and the first formal test for stationarity of the incidence rate in prevalent cohort studies was proposed by Addona & Wolfson (2006). If this assumption holds, lifetimes sampled from a prevalent cohort are considered to be length-biased. For length-biased lifetimes, the probability of selecting a subject from the target population is directly proportional to their lifetime. When estimating the survival distribution for a certain disease, the lifetime corresponds to the time from onset of the disease until death from the disease. When it is safe to assume stationarity, it has been shown that this unconditional estimate is more efficient than the modified Kaplan-Meier estimate (Asgharian et al. 2002).

Before describing Vardi's method of estimation, it is necessary to detail mathematically length-biased sampling. Suppose we have a random variable, Y, with

corresponding *cdf* $F_U(y)$. Y represents the true, unbiased event times. The length-biased distribution of Y, $F_{LB}(y)$, is defined as (Cox 1969)

$$F_{LB}(y) = \frac{1}{\mu}\int_0^y x dF_U(x) \tag{3.1.1}$$

where $\mu = \int_0^\infty x dF_U(x) < \infty$. The above distribution is the corresponding length-biased distribution of F_U. The length-biased distribution arises when a random variable, with *cdf* F_U, is observed with probability proportional to its length (Cox 1969). In the case where F_U has a density, f_U, with respect to Lebesgue measure, the length-biased distribution can be written as

$$F_{LB}(y) = \frac{1}{\mu}\int_0^y x f_U(x) dx \tag{3.1.2}$$

which implies that the length-biased density is (Correa & Wolfson 1999)

$$f_{LB}(y) = \frac{y f_U(y)}{\mu} \qquad y \geq 0. \tag{3.1.3}$$

Suppose we obtain a sample from F_{LB}, $\tilde{Y}_1, \ldots, \tilde{Y}_n$, that is, a sample from the length-biased distribution. The $\tilde{Y}_i$'s can be thought of as sampled prevalent cases, where stationarity of the disease onset times holds. When sampling from a prevalent cohort, the subjects already have the disease in question. They are selected into the study and, ideally, followed until they experience the event in question. We can split their lifetime into two segments: the time from disease onset until recruitment into the study, and the time from study recruitment until the event occurs. These time periods are respectively termed the backward and forward recurrence times. The backward recurrence time is also referred to as the truncation time, and will be denoted by T. The forward recurrence time is also known as the residual lifetime, and will be denoted by R. Note that each $\tilde{Y}_i$ in the above sample can be represented as the sum of T_i and R_i. However, as has been discussed, it is not always possible to follow every individual under study until the event occurs. Hence, the R_i's may be subject to censoring. Define C_i to be random residual censoring variables with *cdf*

$F_C(c)$. The observed residual lifetime will now be such that only the minimum of R_i and C_i is observed, and therefore the i^{th} observed or censored lifetime, X_i, can be represented as

$$X_i = T_i + R_i \wedge C_i. \tag{3.1.4}$$

We can refer to $A_i = T_i + C_i$ as a full censoring time, and $B_i = T_i + R_i$ as an observed lifetime. Complete lifetimes and complete censoring times both contain a common backward recurrence time, and therefore are not independent. However, the residual lifetimes and residual censoring times are independent in most practical situations. Further, we will assume that the C_i's are independent of the (T_i, R_i) pairs, then

$$\begin{aligned} Cov(A,B) &= Cov(T+R, T+C) \\ &= E[(T+R)(T+C)] - E[T+R]E[T+C] \\ &= E[T^2 + TR + TC + RC] - E^2[T] - E[T]E[R] - E[T]E[C] - E[R]E[C] \\ &= Var(T) + Cov(T,R) > 0. \end{aligned} \tag{3.1.5}$$

Recall that earlier δ was defined as the variable indicating whether a lifetime is censored or not, with $\delta = 1$ for an uncensored observation, and $\delta = 0$ for a censored observation. In terms of R_i and C_i, δ is defined as

$$\delta_i = \begin{cases} 1 & \text{if } R_i \leq C_i \\ 0 & \text{if } R_i > C_i. \end{cases} \tag{3.1.6}$$

Survival data on each individual can be represented as a pair, (X_i, δ_i), with X indicating the lifetime and δ indicating whether or not it is a censored observation. Note that censoring can occur either during the follow-up period, or at the end of the follow-up period.

Let the data described above be written as $W_i = (T_i, R_i \wedge C_i, \delta_i)$ as opposed to the above (X_i, δ_i), for $i = 1, \ldots, n$. With the assumption that the residual censoring times

are independent of the T_i's and R_i's, the independence of the W_i's follows (Bergeron 2006).

Consider two sets of observations, U_i's and Y_i's from some length-biased lifetime distribution F_{LB}. Let $V_i = H_i Y_i$, where the H_i follow a $U(0,1)$ distribution. This type of data distortion has been termed "multiplicative censoring", since the incompleteness of the V's comes from them being scaled down by the H's (Vardi 1989).

Suppose the goal is to nonparametrically estimate F_{LB}, using a set of full observations (U's) and multiplicatively censored observations (V's). The likelihood for this setting, derived by Vardi is,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} L_i(\theta) \\ &= \prod_{i=1}^{n} \Big(\frac{f_u(u_i;\theta)}{\mu(\theta)} \Big)^{\delta_i} \Big(\int_{w \geq v_i} \frac{f_u(w;\theta)}{\mu(\theta)} \Big)^{1-\delta_i}, \end{aligned} \tag{3.1.7}$$

where $\mu(\theta)$ is the mean of f_U. The multiplicatively censored times are $V = HY$, where $H \sim U(0,1)$, which is our truncation setting.

Vardi's problem A, which will be discussed in Chapter 4, is equivalent to following a fixed m individuals to the event, and throwing out a fixed n individuals at recruitment once their onset time is known. This is an unrealistic setting as the number of censored and uncensored observations are not known in advance, and so all individuals must be followed forward for the duration of the study. This type of situation can be likened to our truncation setting with length-biased data. Vardi states that under cross-sectional sampling, the likelihood will be proportional to $L(\theta)$, though the setup differs from multiplicative censoring. For cross-sectional sampling, the asymptotic properties of the MLE's obtained from $L(\theta)$ have been derived by Asgharian et al. (2002).

A detailed description of Vardi's problem and how the NPMLE can be obtained through an EM algorithm will be given in Chapter 4.

3.2 Parametric Models in Survival Analysis

Though nonparametric and semi-parametric models are extremely useful in analyzing survival data, parametric models are also widely used. When fitting a parametric model to a set of data, estimation is reduced to only a few parameters. Parametric survival models allow one to describe easily the behaviour of lifetimes, to compute values efficiently, as well as to predict model changes in parameters more intuitively. Also, parametric models are advantageous in building an understanding of the survival and hazard functions described in Chapter 2 (Klein & Moeschberger 2003). Some common models used in survival analysis will briefly be discussed including the exponential, Weibull, Pareto, log-normal, and log-logistic model, as well as the generalized Gamma family. Note that the class of densities considered here are used to model f_U.

The exponential model is a relatively simple model with some notable properties, namely, its constant hazard rate. Recall the *pdf* of the exponential distribution:

$$f(x) = \lambda \exp(-\lambda x). \tag{3.2.1}$$

From here the survival and hazard functions are easily derived as

$$S(x) = \int_x^\infty \lambda \exp(-\lambda t) dt = \exp(-\lambda x), \tag{3.2.2}$$

and

$$h(x) = \frac{\lambda \exp(-\lambda x)}{\exp(-\lambda x)} = \lambda. \tag{3.2.3}$$

It can be easily shown that the mean and standard deviation of the exponential distribution are both equal to $1/\lambda$.

Notice that for all values of x, the hazard rate will remain constant at λ. This means that the time until failure does not depend on any past history, which can prove to be quite restrictive in real-life applications. As Klein and Moeschbeger point out,

a constant hazard rate has been termed the "no-aging" property or "old-as-good-as-new" property. This property is likewise known as the memory-less property, which can be illustrated from

$$\begin{aligned} P(X > x+z \mid X > x) &= \frac{P(X > x+z \cap X > x)}{P(X > x)} \\ &= \frac{P(X > x+z)}{P(X > x)} \\ &= \frac{\exp(-\lambda(x+z))}{\exp(-\lambda x)} \\ &= \exp(-\lambda z) \\ &= P(X > z). \end{aligned} \tag{3.2.4}$$

It follows that the exponential model is more applicable to an industrial setting, for example, modeling the failure rate of electrical components. Since human beings age, and hazard increases, this model is not often applicable in the medical field.

The Weibull distribution is more widely used in survival analysis. It has relatively simple survival and hazard functions, but can accommodate an increasing, decreasing, or constant hazard rate. Recall the *pdf* of the Weibull distribution:

$$f(x) = \alpha\lambda x^{\alpha-1}\exp(-\lambda x^{\alpha}). \tag{3.2.5}$$

The survival function can be derived as

$$S(x) = \int_x^{\infty} \alpha\lambda t^{\alpha-1}\exp(-\lambda t^{\alpha})dt = \exp(-\lambda x^{\alpha}), \tag{3.2.6}$$

which implies that the hazard function is equal to

$$h(x) = \frac{\alpha\lambda x^{\alpha-1}\exp(-\lambda x^{\alpha})}{\exp(-\lambda x^{\alpha})} = \alpha\lambda x^{\alpha-1}. \tag{3.2.7}$$

The mean of the Weibull distribution is given by $\mu = \frac{\Gamma(1+1/\alpha)}{\lambda^{1/\alpha}}$. α is referred to as the shape parameter and λ is referred to as the rate parameter. Note that when $\alpha = 1$, the exponential distribution is obtained, and hence the exponential distribution is a

special case of the Weibull distribution. As mentioned above the hazard rate can be increasing, decreasing, or constant which corresponds to $\alpha > 1$, $\alpha < 1$, or $\alpha = 1$ (exponential).

The Pareto model is not as commonly used in survival analysis as say, the Weibull distribution, but it has closed forms for the survival function, density function and hazard rate, making it easy to work with. The density function, survival function, and hazard function are given as:

$$f(x) = \frac{\theta\lambda^\theta}{x^{\theta+1}}, \tag{3.2.8}$$

$$S(x) = \frac{\lambda^\theta}{x^\theta}, \tag{3.2.9}$$

$$h(x) = \frac{\theta}{x}, \tag{3.2.10}$$

where $\theta > 0$, $\lambda > 0$, and $x \geq \lambda$. The mean of the Pareto distribution is given by $\mu = \frac{\theta\lambda}{\theta-1}$ and exists only if $\theta > 1$. Due to its heavy skewness, the Pareto model is common in economics for modeling the distribution of incomes (Rice 1995).

A random variable X follows a log-normal distribution if $Y = \ln X$ follows a normal distribution. Some authors have observed that ages at the onset of certain diseases can be approximated by the log-normal distribution (Feinleib 1960, Horner 1987). The corresponding *pdf* and survival function are

$$f(x) = \frac{\exp[-\frac{1}{2}(\frac{\ln x-\mu}{\sigma})^2]}{x(2\pi)^{1/2}\sigma} \tag{3.2.11}$$

and

$$S(x) = 1 - \Phi\left[\frac{\ln x - \mu}{\sigma}\right] \tag{3.2.12}$$

where μ and σ are the mean and standard deviation of Y, and $\Phi(x)$ is the *cdf* for a standard normal variable.

The hazard function does not reduce to anything less complicated, so it will just be written as

$$h(x) = \frac{f(x)}{S(x)} = \frac{\frac{\exp[-\frac{1}{2}(\frac{\ln x-\mu}{\sigma})]}{x(2\pi)^{1/2}\sigma}}{1 - \Phi[\frac{\ln x-\mu}{\sigma}]}. \tag{3.2.13}$$

The mean of a log-normal variable is $\exp(\mu + \frac{1}{2}\sigma^2)$, again, where μ and σ are the mean and standard deviation of Y. The value of the hazard function is always 0 at time zero, and then it increases to a maximum, followed by a decrease, with the hazard rate approaching zero as x tends to infinity (i.e. - the hazard is hump-shaped). Decreasing hazard for larger values of x is not always optimal, but in situations where large values of x are not important, this model may be applicable (Klein & Moeschberger 2003).

A random variable X is said to follow a log-logistic distribution if $Y = \ln X$ follows a logistic distribution. The log-logistic distribution closely resembles the log-normal distribution, having heavier tails, and the advantage that the survival and hazard functions are more tractable. The *pdf* for Y is

$$f(y) = \frac{\exp(\frac{y-\mu}{\sigma})}{\sigma[1+\exp(\frac{y-\mu}{\sigma})]^2} \tag{3.2.14}$$

where μ and σ^2 are respectively the mean and scale parameter of Y. The corresponding *pdf*, survival function and hazard function for X are

$$f(x) = \frac{\alpha x^{\alpha-1}\theta}{[1+\theta x^{\alpha}]^2}, \tag{3.2.15}$$

$$S(x) = \frac{1}{1+\theta x^{\alpha}}, \tag{3.2.16}$$

$$h(x) = \frac{\alpha x^{\alpha-1}\theta}{1+\theta x^{\alpha}}, \tag{3.2.17}$$

where $\lambda = \exp(-\mu/\theta)$ and $\alpha = 1/\sigma > 0$. The mean of the log-logistic distribution is $\mu = \frac{\pi \csc(\frac{\pi}{\alpha})}{\alpha\theta^{1/\alpha}}$. The numerator of the hazard function resembles the Weibull hazard function, but behaves differently due to the denominator. Namely, the hazard is monotone decreasing for $\alpha \leq 1$ and for $\alpha > 1$ the hazard rate increases to a maximum at $(\frac{\alpha-1}{\theta})^{1/\alpha}$, followed by a decrease to 0 as x approaches infinity.

The gamma distribution is a little more complicated to work with, but has properties similar to the Weibull distribution. Its *pdf* is given by

$$f(x) = \frac{\lambda^{\beta} x^{\beta-1}\exp(-\lambda x)}{\Gamma(\beta)}, \tag{3.2.18}$$

where $\Gamma(\beta) = \int_0^\infty x^{\beta-1}\exp(-x)dx$. β is referred to as the shape parameter. When $\beta > 1$, the hazard function is monotone increasing with $h(0) = 0$ and $h(x) \to \lambda$ as $x \to \infty$. When $\beta < 1$, the hazard function is monotone decreasing with $h(0) \to \infty$ and $h(x) \to \lambda$ as $x \to \infty$. By adding an extra parameter, the generalized gamma distribution is obtained, allowing for additional flexibility in the hazard function. The generalized gamma distribution has *pdf*

$$f(x) = \frac{\alpha\lambda^{\alpha\beta}x^{\alpha\beta-1}\exp(-(\lambda x)^\alpha)}{\Gamma(\beta)}. \tag{3.2.19}$$

When both α and β are equal to 1, the exponential distribution is obtained. When $\beta = 1$ the Weibull distribution is obtained, and when $\alpha = 1$ the gamma distribution is obtained. The generalized gamma distribution will be useful when investigating how to generate length-biased samples, which will be discussed in Chapter 4.

In order to infer upon a set of data using a parametric model, it is necessary to obtain estimates of the model parameters. This is often done through maximum likelihood estimation. Estimates can be obtained by using the data to maximize the likelihood either analytically, when possible, or numerically otherwise. A popular numerical technique for likelihood maximization is the Newton-Raphson method. Some other techniques for likelihood maximization include steepest ascent, quasi-Newton, and conjugate gradient methods (Lawless 2003*b*). Once maximum likelihood estimates, $\hat{\boldsymbol{\theta}}_{ML}$, for the model parameters have been obtained, one can use the model to compute various quantities, such as the survival function, hazard function, mean or median.

3.3 Tests for Goodness-of-Fit

This section will focus on one-sample goodness-of-fit tests. One-sample goodness-of-fit tests compare a sample to a specified parametric distribution (either fully specified,

or up to unknown parameters). This is in contrast to two-sample goodness-of-fit tests where the goal is to determine whether two independent samples come from the same *unspecified* distribution. It should be noted that an asymptotic most efficient nonparametric test for the equality of survival distributions, for right-censored data with biased sampling, has recently been proposed by Ning et al. (2010).

Checking the adequacy of a model is an extremely important part of statistical inference. Suppose we have a random sample, $x_1, \ldots, x_n$, from some unknown distribution, and would like to know how well a particular model fits this set of observations. Goodness-of-fit statistics attempt to quantify the discrepancy between values expected under a specified model and the observed values from a random sample. Goodness-of-fit testing generally involves investigating this random sample to test the null hypothesis that it is, in fact, from some known, specified distribution. For example, the null hypothesis could specify some distribution $F^*(x)$, and then $x_1, \ldots, x_n$ will be compared to $F^*(x)$ to see if it is reasonable to conclude that $F^*(x)$ is the true underlying distribution of our random sample. Model checking needs vary depending on the model's strength and complexity of its assumptions (Lawless 2003*b*).

Graphical procedures, such as probability and residual plots, can be used as an informal first step for model checking. However, due to the high amount of variation inherent in graphical procedures, it can prove difficult to determine whether the features of the plot are due to natural variation. Even when data are generated from the assumed model, the eye will not always come to the correct conclusion. As a result, formal hypothesis tests are required.

A well-known goodness-of-fit statistic used for testing the hypothesis that X has some specified distribution is Pearson's χ^2 statistic. The null hypothesis that the distribution of the observed random variable is $F^*(x)$ is tested against the alternative hypothesis that the distribution of the observed random variable is different than

$F^*(x)$. Put another way,

$$H_0 \quad : F(x) = F^*(x) \tag{3.3.1}$$

$$H_1 \quad : F(x) \neq F^*(x). \tag{3.3.2}$$

Use of Pearson's χ^2 test requires that the observed values are independent and identically distributed. It is commonly used for categorical data, but can be extended to accommodate the continuous case. Consider a data set with n observations of a random variable X. These n observations can be grouped into c classes, and the frequency of each class recorded. Below is a table representing a data set organized in this fashion.

	C_1	C_2	C_3	...	C_c
Frequency	O_1	O_2	O_3	...	O_c

where O_i is the realized number of observations in i^{th} class.

Let p_i^* represent the probability of an observation being in class i under the assumption of the null hypothesis (i.e. under $F^*(x)$ being the true distribution of X). The expected number of observations in each class is then defined as

$$E_i = p_i^* n, \qquad i = 1, \ldots, c. \tag{3.3.3}$$

Pearson's χ^2 test statistic T is given by

$$T = \sum_{i=1}^{c} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{c} \frac{O_i^2}{E_i} - n. \tag{3.3.4}$$

If some of the E_i are small, the chi-square distribution may not be appropriate, and it may be necessary to combine cells with small E_i with another class in a suitable way. The general rule of thumb is that each E_i is at least 5 (Conover 1971).

For large values of n, T follows a chi-square distribution with $(c-1)$ degrees of freedom. The null hypothesis is rejected for values of T greater than $x_{(1-\alpha)}$, where

$x_{(1-\alpha)}$ corresponds to the $(1-\alpha)$ quantile of a chi-square distribution with $(c-1)$ degrees of freedom. If T exceeds $x_{(1-\alpha)}$, then H_0 is to be rejected. Note that if parameters are being estimated from the sample, a chi-square distribution with fewer degrees of freedom is required. One degree of freedom is subtracted for each estimated parameter. If k is number of parameters estimated, and c the number of classes, then the appropriate chi-square distribution is one with $(c-1-k)$ degrees of freedom (Conover 1971).

If Pearson's chi-squared test is to be used for continuous distributions, the observed data are grouped into discrete intervals. The expected frequency in each interval is the product of the probability associated with each interval and the number of observations, n. However, due to this discretization of the data, Pearson's chi-squared test is not that powerful for continuous data (D'Agostino & Stephens 1986). The choice of location and number of discrete intervals, and how the observations fall within them, has an effect on the performance of the test. For continuous data, using a test that is based on the empirical distribution is a more general and adequate approach.

There are several one-sample goodness-of-fit tests that involve comparing the empirical distribution function with a hypothesized distribution function, based on some measure of distance between the two. Recall that the empirical distribution function, $\hat{F}_n(x)$, is defined as the fraction of x_i's from a random sample of $x_1, \ldots, x_n$ that are less than or equal to x, or

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(x_i \leq x). \tag{3.3.5}$$

The empirical distribution is a useful estimator of the true unknown distribution function. In fact, in the case of complete information, the empirical distribution function is the nonparametric maximum likelihood estimate for $F(x)$. If a good agreement does not exist between the empirical distribution function and the hypothesized distribution function, then it seems reasonable to conclude that the true distribution

function is different than the hypothesized distribution function. Formal goodness-of-fit tests based on this idea have been studied in depth, and a few well-known tests will be discussed here. It is noted by Conover (1971) that statistics which are functions of the distance between the empirical distribution function and the hypothesized distribution function are known as Kolmogorov-type statistics, whereas statistics which are functions of the vertical distance between two empirical distribution functions are known as Smirnov-type statistics.

Let $F^*(x)$ be the hypothesized distribution. Define D^+ to be the largest vertical difference when $\hat{F}_n(x)$ is greater than $F^*(x)$, and D^- the largest vertical distance when $\hat{F}_n(x)$ is smaller than $F^*(x)$. Mathematically, D^+ and D^- are defined as

$$D^+ = \sup_x\{\hat{F}_n(x) - F^*(x)\}, \qquad D^- = \sup_x\{F^*(x) - \hat{F}_n(x)\}. \tag{3.3.6}$$

Perhaps one of the most well-known statistics is the Kolmogorov statistic, D, defined as

$$D = \sup_x |\hat{F}_n(x) - F^*(x)| = \max(D^+, D^-), \tag{3.3.7}$$

which was introduced by Kolmogorov in 1933. D'Agostino & Stephens (1986) refer to the above statistics as 'supremum statistics'.

A second class of statistics measuring discrepancies between the empirical and hypothesized distribution function is given by the Cramer-von Mises family and defined as

$$Q = n\int_{-\infty}^{\infty}\{\hat{F}_n(x) - F^*(x)\}^2\psi(x)dF^*(x), \tag{3.3.8}$$

where $\psi(x)$ is a weight function for the squared difference between $\hat{F}_n(x)$ and $F^*(x)$. If $\psi(x) = 1$, the Cramer-von Mises statistic, W^2, is obtained. If $\psi(x) = [F^*(x)(1 - F^*(x))]^{-1}$, the Anderson-Darling statistic (1952), A^2, is obtained. Large values of the above statistics provide evidence against the null hypothesis. Finite sample and asymptotic distributions are available, when testing a completely specified distribution.

Written in the above form, the expressions appear difficult to work with. In order to obtain more straight forward expressions, an important theorem, known as the *Probability Integral Transform* theorem, must be recalled:

Theorem 3.3.1 *Let X have continuous cdf $F_X(x)$ and define the random variable Y as $Y = F_X(x)$. Then Y is uniformly distributed on (0,1), that is, $P(Y \leq y) = y$, $0 < y < 1$.*

The proof for this theorem can be found in Casella & Berger (2001). Consequently, under the null hypothesis, the $F_X(x_i)$'s are the order statistics from a random sample distributed uniformly on the interval $(0, 1)$. So the distributions of D, W^2, and A^2 do not depend on the distribution $F^*(x)$, assuming the null hypothesis is true, which is clear from the alternate expressions (D'Agostino & Stephens 1986):

$$D = \max_{1 \leq i \leq n} \left[\frac{i}{n} - F^*(x_{(i)}), F^*(x_{(i)}) - \frac{i-1}{n} \right], \tag{3.3.9}$$

$$W^2 = \sum_{i=1}^{n} \left[F^*(x_{(i)}) - \frac{(i-.5)}{n} \right]^2 + \frac{1}{12n}, \tag{3.3.10}$$

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)[\ln F^*(x_{(i)}) + \ln\{1 - F^*(x_{(n+1-i)})\}]. \tag{3.3.11}$$

In the case of a fully specified $F^*(x)$, distribution theory for the above Kolmogorov-Smirnov type statistics is well developed and finite sample and asymptotic distributions are available.

Unfortunately, it is rare that one wishes to test a fully specified distribution, as parameters are often estimated from the data (Lawless 2003*b*). In this case, often the maximum likelihood estimates are used in place of the unknown parameters. The above statistics can be calculated in the same way, however now the distribution theory is more complicated. The distribution theory under a fully specified model does not apply in this situation, as the distributions of these statistics will depend on the distribution being tested, the parameters estimated and method of estimation,

as well as the sample size (D'Agostino & Stephens 1986). However, as will be done here, simulations often suffice in approximating p-values (Lawless 2003*b*). This idea will be discussed in more detail in the following chapter.

As stated earlier, three models will be investigated for goodness-of-fit to length-biased, right-censored survival data. Many of the goodness-of-fit statistics discussed above rely on the empirical distribution function, which is complicated by censoring and truncation. It follows that the nonparametric estimate for the survival function developed by Vardi (1989) will be used in place of the empirical survival function. Estimation in survival analysis uses $\hat{S}(x)$, and for the purpose of testing, $F(x)$ and $S(x)$ are essentially equivalent quantities, since $F(x) = 1 - S(x)$. Theoretically, all of the above tests could be extended using nonparametric estimates of the survival function.

Chapter 4

Algorithms

This chapter will focus on the algorithms necessary to test the goodness-of-fit of three parametric models to a set of length-biased survival data with right-censoring. First we will look at how to generate length-biased Weibull, log-normal and log-logistic samples with uniform left-truncation and right-censoring. Next, Vardi's algorithm for a nonparametric estimation of the survival function corrected for length-bias will be given. In the last section we will detail how simulation can be used to approximate the p-values for our goodness-of-fit tests.

4.1 Simulating Length-Biased Samples

This section will focus on how to generate length-biased data. One way to obtain a length-biased sample of size n is to generate a large amount of data, say $N \gg n$, from the appropriate distribution. We then sample n units from N, with probability proportional to size. The resulting sample of size n will be a length-biased sample from the given distribution. However, this is not an efficient way to generate data, especially when dealing with large sample sizes.

Here we will show that by making the appropriate transformations, certain

length-biased distributions have alternative representations, making the simulation of the data more efficient. Using this approach, we need only to generate n values from the corresponding distribution and make an appropriate transformation in order to obtain our length-biased sample. Specifically, upon transformation, the length-biased Weibull distribution can be obtained through a gamma distribution, the length-biased log-normal distribution can be obtained through a normal distribution, and the length-biased log-logistic distribution can be obtained through a beta distribution.

The length-biased distribution and density of a random variable Y were given as

$$F_Y(t) = \int_0^t \frac{xf(x)dx}{\mu}, \tag{4.1.1}$$

and

$$f_Y(t) = \frac{tf(t)}{\mu}. \tag{4.1.2}$$

Correa & Wolfson (1999) show how the generalized gamma distribution can be used to generate length-biased Weibull samples. Recall the generalized gamma distribution discussed in the previous chapter. The density of a $GG(k, \lambda, p)$ was given as

$$f(x) = \frac{\lambda p(\lambda x)^{pk-1} exp[-(\lambda x)^p]}{\Gamma(k)}, \tag{4.1.3}$$

where $\mu = \frac{\Gamma(k+\frac{1}{p})}{\lambda\Gamma(k)}$.

The length-biased distribution of a generalized gamma random variable is therefore given by

$$\begin{aligned} F_Y(t) &= \int_0^t \frac{x\lambda p(\lambda x)^{pk-1} \exp[-(\lambda x)^p]\lambda\Gamma(k)dx}{\Gamma(k)\Gamma(k+\frac{1}{p})} \\ &= \int_0^t \frac{\lambda p(\lambda x)^{p(k+\frac{1}{p})-1} \exp[-(\lambda x)^p]dx}{\Gamma(k+1/p)}, \end{aligned} \tag{4.1.4}$$

which is the distribution of a $GG(k+\frac{1}{p}, \lambda, p)$ random variable. Recall that the GG distribution reduces to an $exponential(\lambda)$ when $k = p = 1$, a $gamma(k, \lambda)$ when

$p = 1$, and a $Weibull(\lambda, p)$ when $k = 1$. From equation (4.1.4) we see that the form of the length-biased distribution of an $exponential(\lambda)$ is $gamma(2, \lambda)$, and the length-biased distribution of a $gamma(k, \lambda)$ is $gamma(k+1, \lambda)$. For a $Weibull(\lambda, p)$, the length-biased distribution is of the form $GG(1 + \frac{1}{p}, \lambda, p)$.

Correa & Wolfson (1999)) show that by letting $z = (\lambda x)^p$ in equation (4.1.4), we obtain

$$F_Y(t) = \int_0^{(\lambda t)^p} \frac{z^{(k+1/p)-1} e^{-z}}{\Gamma(k + 1/p)} dz \tag{4.1.5}$$

which shows that $F_Y(t)$ can be written as $F_Y(t) = F_Z(h(t))$, where $F_Z \sim gamma(k + 1/p, 1)$ and $h(t) = (\lambda x)^p$. $h(t)$ is non-negative, strictly increasing and continuous, and its inverse is given by

$$g(s) = \frac{s^{1/p}}{\lambda}. \tag{4.1.6}$$

Using the gamma distribution, which is standard for all statistical packages, we can easily generate survival times when data come from a length-biased Weibull distribution. This is done as follows:

Algorithm 1

- *Generate a* $gamma(1 + 1/p, 1)$ *random variable,* Z*, and*
- *Take* $W = g(Z)$ *where* g *is given in equation (4.1.6).*

The random variable W *follows a length-biased Weibull distribution with parameters* λ *and* p*.*

Recall the density for a log-normal random variable as

$$f(x) = \frac{\exp[-\frac{1}{2}(\frac{\ln x - \mu}{\sigma})^2]}{x(2\pi)^{1/2}\sigma} \tag{4.1.7}$$

with corresponding mean $\tilde{\mu} = \exp(\mu+\frac{1}{2}\sigma^2)$, where μ and σ^2 are the mean and variance of $\ln X$. The length-biased distribution is then given by

$$\begin{aligned} F_Y(t) &= \int_0^t \frac{x \exp[-\frac{1}{2}(\frac{\ln x-\mu}{\sigma})^2]dx}{x(2\pi)^{1/2}\sigma \exp(\mu+\frac{1}{2}\sigma^2)} \\ &= \int_0^t \frac{\exp(-\mu-\frac{1}{2}\sigma^2)\exp[-\frac{1}{2}(\frac{\ln x-\mu}{\sigma})^2]dx}{(2\pi)^{1/2}\sigma}. \end{aligned} \tag{4.1.8}$$

Letting $z = \ln x$, we obtain

$$\begin{aligned} F_Y(t) &= \int_0^{\ln t} \frac{\exp(z-\mu-\frac{1}{2}\sigma^2)\exp[-\frac{1}{2}(\frac{z-\mu}{\sigma})^2]dz}{(2\pi)^{1/2}\sigma} \\ &= \int_0^{\ln t} \frac{\exp[-\frac{1}{2}(\frac{z^2-2\mu z+\mu^2-2z\sigma^2+2\mu\sigma^2+\sigma^4}{\sigma^2})dz]}{(2\pi)^{1/2}\sigma} \\ &= \int_0^{\ln t} \frac{\exp[-\frac{1}{2}(\frac{z-(\mu+\sigma^2)}{\sigma})^2]dz}{(2\pi)^{1/2}\sigma} \end{aligned} \tag{4.1.9}$$

which shows that $F_Y(t)$ can be written as $F_Y(t) = F_Z(h(t))$, where $F_Z \sim normal(\mu+\sigma^2, \sigma^2)$ and $h(t) = \ln t$. $h(t)$ is non-negative, strictly increasing and continuous. Its inverse is given by

$$g(s) = e^s. \tag{4.1.10}$$

Note that this representation of length-biased log-normal data was given by Patil & Rao (1978).

The algorithm for generating length-biased log-normal data is as follows:

Algorithm 2

- *Generate a normal$(\mu+\sigma^2, \sigma^2)$ random variable, Z, and*
- *Take $W = g(Z)$ where g is given in equation (4.1.10).*

The random variable W follows a length-biased log-normal distribution with parameters μ and σ.

Correa & Wolfson (1999) detail how to generate length-biased log-logistic samples, using a one-parameter log-logistic distribution. Note that the two-parameter log-logistic distribution is the reciprocal of the one-parameter log-logistic distribution, with an extra rate parameter. Algorithm 3, that follows, gives the procedure for generating length-biased samples from a two-parameter length-biased log-logistic function. This is an extension of the results from Correa & Wolfson (1999).

As mentioned earlier, the two-parameter log-logistic distribution, as given in Klein & Moeschberger (2003), has the following *pdf* and mean:

$$f(x) = \frac{\alpha\theta x^{\alpha-1}}{[1+\theta x^{\alpha}]^2} \tag{4.1.11}$$

$$\mu = \frac{\pi \csc(\frac{\pi}{\alpha})}{\alpha\theta^{1/\alpha}}. \tag{4.1.12}$$

Let $\theta = \lambda^{\alpha}$. The equations for $f(x)$ and μ become

$$f(x) = \frac{\alpha\lambda^{\alpha}x^{\alpha-1}}{[1+(\lambda x)^{\alpha}]^2} \tag{4.1.13}$$

$$\mu = \frac{\pi \csc(\frac{\pi}{\alpha})}{\alpha\lambda}. \tag{4.1.14}$$

The mean, μ, can be represented using the gamma function:

$$\mu = \frac{\pi \csc(\frac{\pi}{\alpha})}{\alpha\lambda} = \frac{\pi}{\alpha\lambda \sin(\frac{\pi}{\alpha})} = \frac{\frac{1}{\alpha}\Gamma(\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})}{\lambda} = \frac{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})}{\lambda}, \tag{4.1.15}$$

using the relationships $\csc(x) = \frac{1}{\sin(x)}$, $x\Gamma(x) = \Gamma(1+x)$ and $\frac{\pi}{\sin(\pi x)} = \Gamma(x)\Gamma(1-x)$.

The length-biased density for a log-logistic random variable is given by

$$\begin{aligned} F_Y(t) &= \int_0^t \frac{x\alpha\lambda^{\alpha}x^{\alpha-1}\lambda dx}{[1+(\lambda x)^{\alpha}]^2\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})} \\ &= \int_0^t \frac{\alpha x^{\alpha}\lambda^{\alpha+1}dx}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})(1+(\lambda x)^{\alpha})^2} \\ &= \int_0^t \frac{\alpha\lambda(\lambda x)^{\alpha}dx}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})(1+(\lambda x)^{\alpha})^2} \\ &= \int_0^t \frac{\alpha\lambda \exp[\alpha\log(\lambda x)]dx}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})(1+\exp[\alpha\log(\lambda x)])^2} \end{aligned} \tag{4.1.16}$$

Taking $z = \alpha \log(\lambda x)$, we obtain

$$\begin{aligned} F_Y(t) &= \int_{-\infty}^{\alpha log(\lambda t)} \frac{\alpha\lambda \exp(z) \exp(\frac{z}{\alpha}) dz}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})(1+\exp(z))^2 \alpha\lambda} \\ &= \int_{-\infty}^{\alpha log(\lambda t)} \frac{\exp(z(1+\frac{1}{\alpha}))dz}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})(1+\exp(z))^2} \end{aligned} \tag{4.1.17}$$

Letting $u = (1+e^z)^{-1}$ we obtain

$$\begin{aligned} F_Y(t) &= -\int_{1}^{(1+\exp[\alpha \log(\lambda t)])^{-1}} \frac{u^{-1/\alpha}(1-u)^{1/\alpha} du}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})} \\ &= \int_{(1+\exp[\alpha \log(\lambda t)])^{-1}}^{1} \frac{u^{-1/\alpha}(1-u)^{1/\alpha} du}{\Gamma(1+\frac{1}{\alpha})\Gamma(1-\frac{1}{\alpha})} \end{aligned} \tag{4.1.18}$$

So $F_Y(t)$ can be written as $F_Z(h(t))$, where F_Z is the distribution of a $Beta(1-\frac{1}{\alpha}, 1+\frac{1}{\alpha})$ random variable and $h(t) = (1+\exp[\alpha \log(\lambda t)])^{-1} = \frac{1}{1+(\lambda x)^\alpha}$. $h(t)$ is non-negative, strictly decreasing and continuous. Its inverse is given by

$$\tilde{g}(s) = \frac{\left(\frac{1-s}{s}\right)^{1/\alpha}}{\lambda}. \tag{4.1.19}$$

Under the original parametrization for the log-logistic distribution, the inverse function is given by

$$g(s) = \left(\frac{1-s}{\theta s}\right)^{1/\alpha}. \tag{4.1.20}$$

The algorithm for generating length-biased log-logistic data is as follows:

Algorithm 3

- *Generate a beta*$(1-\frac{1}{\alpha}, 1+\frac{1}{\alpha})$ *random variable,* Z*, and*
- *Take* $W = g(Z)$ *where* g *is given in equation (4.1.20).*

The random variable W *follows a length-biased log-logistic distribution with parameters* α *and* θ*.*

Using the three algorithms stated above, we can easily and efficiently generate length-biased Weibull, log-normal and log-logistic samples.

We wish to generate samples with uniform left-truncation and right-censoring. Once we have the generated length-biased times, the next step is to split the lifetimes into two parts; the truncation time, T, and the residual lifetime, R. The truncation time is the time from onset of disease until the beginning of the study, and the residual lifetime is the time from the beginning of the study until the event occurs. Given a generated time, x_i, the truncation time, t_i, is generated uniformly on the interval $(0, x_i)$. The residual lifetime is the difference between x_i and t_i, namely $r_i = x_i - t_i$. We also require a corresponding censoring indicator, δ_i, for each generated time. This is obtained by generating a random censoring variable, c_i, and comparing it to r_i such that

$$\delta_i = \begin{cases} 1 & \text{if } r_i \leq c_i \\ 0 & \text{if } r_i > c_i. \end{cases} \tag{4.1.21}$$

The censoring times can be generated in various ways, and three different approaches will be used here. We chose to employ three different censoring schemes to investigate if the type of censoring has any effect on the results of the goodness-of-fit tests. The first is a fixed censoring approach, which takes $c_i = c$ for $i = 1, \ldots, n$. In this case the censoring time point for every individual in the study is the same. This is a realistic approach as it is a common scenario that the end of the study period is the same for all individuals, and on this end date each individual is recorded as either censored or not censored, depending on their status.

A second approach to generate censoring times is to use some known (essentially positive) distribution. We chose to use a normal distribution with mean larger than 3σ to avoid negative values that would need to be resampled, and out of convenience. Exponential censoring is also a common choice for censoring distributions in simu-

lations. Here, the censoring times are not all the same, allowing for some variation. This is also a realistic approach to censoring. It is not always possible, especially in larger studies, to follow up with every individual on the same day.

The third censoring approach is dependent on having a real data set, which is not necessary for all simulations. Since R_i and C_i are assumed to be independent, we can obtain an estimate of the survival function of the residual censoring times, $S_C(c)$, through the usual Kaplan-Meier, with the δ_i's considered as the event times, as the event of interest here is censoring time. Then the $\hat{p}(C = c_i) = \hat{S}_C(c_i^-) - \hat{S}_C(c_i)$. If by chance, $\hat{S}_C(c)$ is undefined beyond some c_{max}, we can use $\hat{S}(c_{max}) = 0$ without loss of generality. We then sample from the residual censoring times from our data set, according to these probabilities. The sampled residual censoring times become the c_i's. By comparing each generated r_i to the sampled c_i, δ_i is obtained.

Now we have all the tools necessary to generate length-biased samples with left-truncation and right-censoring, and the following algorithm provides the steps to do so.

Algorithm 4

- *Use the data to estimate the appropriate parameters, depending on which model is being used. Fix n.*
- *Generate n length-biased times, $\boldsymbol{x}$, using the either Algorithm 1, 2 or 3 depending on the model being used.*
- *For each x_i, generate a truncation time t_i from a $U(0, x_i)$.*
- *Compute the residual lifetime, r_i, for each observation as $r_i = x_i - t_i$.*
- *Generate the censoring times, $\boldsymbol{c}$, using one of the three censoring approaches.*
- *Let $y_i = t_i + r_i \wedge c_i$ and $\delta_i = I[r_i < c_i]$.*

During analysis, it is not necessary to keep track of t_i and r_i, and so the generated data has the form (x_i, δ_i).

4.2 Nonparametric Estimation of Length-Biased Survival Data with Right-Censoring

In this section we will detail an algorithm introduced by Vardi (1989) to find a nonparametric estimate for the unbiased distribution of the survival function arising from properly length-biased data with right-censoring. In Vardi's 1989 paper he details Problem A, which is as follows: Consider a sample which has m independent complete observations and n independent incomplete observations. The m complete observations, $X_1, \ldots, X_m$ are drawn from a distribution G. Consider the random variable U which is independently selected from a uniform distribution on the interval $(0, 1)$. The n incomplete observations, $Z_1, \ldots, Z_n$ are such that $Z_i = Y_i U_i$, where Y_i has the same distribution, G, as the X's. The incompleteness of the Z's is a consequence of their being scaled down by the U's, and this form of censoring can be referred to as "multiplicative censoring" (Vardi 1989). Based on the sample of $X_1, \ldots, X_m$ and $Z_1, \ldots, Z_n$, the nonparametric maximum likelihood estimate of G will be derived. In other words, the goal is to estimate G, based on a set of complete and incomplete observations from G. Vardi gives the likelihood for the observations in Problem A as

$$L(G) = \prod_{i=1}^{m} G(dx_i) \prod_{i=1}^{n} \int_{y \geq z_i} \frac{1}{y} G(dy). \tag{4.2.1}$$

Note that the set of complete data can be considered as the uncensored observations and the set of incomplete data as the censored observations. The data we are considering can be thought of as the pairs $(t_i, 1)$ (the uncensored observations) and $(t_i, 0)$ (the censored observations),which correspond to the x_i's and z_i's respectively. Vardi shows that the above likelihood can be maximized nonparametrically

by assigning positive weights only on the observed x_i's and z_i's, since if mass were placed anywhere else this would only decrease the likelihood. Let $w_1 < w_2 <, \ldots, w_N$ represent the entire sorted data set, that is, both the complete and incomplete observations, in increasing order. If there are no ties in the data set then $N = m + n$, otherwise $N < n + m$. In theory ties will not exist in the data if the underlying distribution is continuous, but there may be ties in real applications.

Define ξ_j to be the number of complete observations at time w_j, and ζ_j to be the number of incomplete observations at w_j. Define $\mathbf{p} = (p_1, \ldots, p_N)'$ to be a probability vector where $p_j = p(w_j) = G(dw_j)$. Now the likelihood can be expressed as

$$L(\mathbf{p}) = \prod_{j=1}^{N} p_j^{\xi_j} \left(\sum_{k=j}^{N} \frac{1}{w_k} p_k \right)^{\zeta_j}. \tag{4.2.2}$$

Because the data set contains both complete and incomplete observations, it is considered to be incomplete. It is noted by Vardi that the EM algorithm is a natural solution for maximizing the likelihood in the above equation as our data set is incomplete and the form of the complete data is known. The algorithm works as follows: Given an estimate for $\mathbf{p}$, call it $\mathbf{p}^{old}$, the expected conditional number of complete observations at w_j, given the observed data and the previous probability vector $\mathbf{p}^{old}$, is assigned to each p_j^{new}. In order for $\mathbf{p}^{new}$ to be a probability vector, each p_j^{new} is divided by the total number of observations. Vardi's algorithm can be written in two steps:

Algorithm 5

- *Begin with an arbitrary probability vector, say* $\mathbf{p}^{old}$*, such that*

$$\sum_{j=1}^{N} p_j^{old} = 1$$

$$p_j^{old} > 0$$

for $j = 1, \ldots, N$.

- *Replace each* p_j^{old} *with*

$$p_j^{new} = \frac{1}{m+n} E(\sum_{i=1}^{m} I(x_i = w_j) + \sum_{i=1}^{n} (y_i = w_j) | (x_1, \ldots, x_m, z_1, \ldots, z_n), p^{old})$$
$$= \frac{1}{m+n} \left\{ \xi_j + \frac{p_j^{old}}{w_j} \sum_{k=1}^{j} \zeta_k \left(\sum_{i=k}^{N} \frac{p_i^{old}}{w_i} \right)^{-1} \right\}. \tag{4.2.3}$$

In the case of full information (i.e. - no censored observations), $\hat{\mathbf{p}}$ is fixed and no iterations will take place. Note that for our length-biased sampling setting, $\hat{\mathbf{p}}$ represents probabilities from the length-biased distribution (i.e. - G can be likened to F_{LB} discussed earlier). A consistent estimate for the length-biased distribution is obtained because 4.2.3 will converge to the unique maximizer of 4.2.2 by definition of the EM algorithm.

By using the inverse length-bias transformation to adjust the probabilities in $\hat{\mathbf{p}}$, the unbiased distribution can be found. $\hat{\mathbf{p}}$ is adjusted as follows:

$$\hat{p}_{U,j} = \frac{\frac{\hat{p}_j}{w_j}}{\sum_{i=1}^{N} \frac{\hat{p}_i}{w_i}} \tag{4.2.4}$$

The unbiased survival function is calculated as

$$\hat{S_{V}}_k = 1 - \sum_{i=1}^{k} \hat{p}_{U,i}. \tag{4.2.5}$$

4.3 Simulation and Estimation of p-values

In this section we will show how simulation can be used to obtain an approximate p-value for our goodness-of-fit tests. We will employ the a Kolmogorov-Smirnov type statistic in order to assess the goodness-of-fit of the Weibull, log-normal, and log-logistic models to our real set of length-biased, right-censored survival data. Recall that the survival distribution is equal to $1 - F(x)$. The Kolmogorov statistic was

defined in chapter 3 as

$$D = \max |\hat{F}_n(x) - F^*(x)|, \tag{4.3.1}$$

where $\hat{F}_n(x)$ is the empirical distribution, and $F^*(x)$ is the hypothesized distribution. Replacing $\hat{F}_n(x)$ with $1 - \hat{S}_n(x)$ and $F^*(x)$ with $1 - S^*(x)$, we obtain

$$\begin{aligned} D &= \max |(1 - \hat{S}_n(x)) - (1 - S^*(x))| \\ &= \max |S^*(x) - \hat{S}_n(x)|, \end{aligned} \tag{4.3.2}$$

and so D can be calculated using the survival function instead of the *cdf*, as will be done here.

Since we are dealing with length-biased, right-censored survival data, we will use Vardi's algorithm to compute the nonparametric maximum likelihood estimate, and this estimate will be used in place of the empirical distribution function when computing D. Let $\hat{S}_V(x)$ represent the estimate of the survival function corrected for length-bias obtained from the data using Vardi's algorithm.

When discussing some different goodness-of-fit techniques in Chapter 3, it was noted that it is rarely the case that one wishes to test a fully specified distribution. Often, we wish to test that a given data set is of some distributional form, say $S_{\boldsymbol{\theta}}(x)$, with unknown parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. For example, we may want to test that a data set is distributed according to a Weibull distribution, with unknown shape and scale parameters. It follows that it is necessary to first use the data to estimate the parameters, which can be done through maximum likelihood estimation. Let $\hat{\boldsymbol{\theta}}$ represent the vector of estimated parameters, and let $S^*_{\boldsymbol{\theta}}(x)$ represent the hypothesized survival distribution with the estimated parameters. The hypothesis being tested becomes:

$$\begin{aligned} H_0 &: S(x) = S^*_{\boldsymbol{\theta}}(x) \\ H_1 &: S(x) \neq S^*_{\boldsymbol{\theta}}(x), \end{aligned} \tag{4.3.3}$$

and so in order to compute the goodness-of-fit tests we simply replace $S^*(x)$ with $S^*_{\boldsymbol{\theta}}(x)$. However, now the test quantity obtained does not correspond to a specific

p-value, and simulation is required in order to roughly estimate the p-value for our test.

Once we have used the data to estimate the parameters of the hypothesized distribution, as well as the nonparametric maximum likelihood estimate according to Vardi's algorithm, the Kolmogorov-Smirnov type statistic can be calculated as

$$d_{data} = \max |\hat{S}_V(x) - S_{\hat{\boldsymbol{\theta}}}(x)|. \tag{4.3.4}$$

$P_{S_{\hat{\theta}}}(D \geq d_{data})$ can be used to roughly approximate the p-value (Ross 2006) The simulation is to be done as follows:

Algorithm 6

- *First use the data under study to estimate the parameters* $\boldsymbol{\theta}$ *by* $\hat{\boldsymbol{\theta}}$*, then compute* d_{data} *as shown in equation (4.3.4).*

- *Generate a sample of size n from* $\hat{S}_{\boldsymbol{\theta}}$*. Estimate the parameters of the simulated data by* $\hat{\boldsymbol{\theta}}_{sim}$*. Estimate the npmle of the simulated data,* $\hat{S}_{V_{sim}}$*, according to the Vardi algorithm. Compute*

$$d_{sim} = \max |\hat{S}_{V_{sim}}(x) - \hat{S}_{\boldsymbol{\theta}_{sim}}(x)|. \tag{4.3.5}$$

  Note whether $d_{sim_i} \geq d_{data}$*. Repeat many times, say* N*. The p-value is approximated by*

$$p = \sum_{i=1}^{N} \frac{I(d_{sim_i} \geq d_{data})}{N}. \tag{4.3.6}$$

Now we have all the tools necessary to carry out the goodness-of-fit analysis for a set of length-biased survival data subject to right-censoring. However, before we apply our goodness-of-fit test to a real set of data, we will use simulation techniques to approximate the distribution of D under the null hypothesis. The behaviour of D under different scenarios will be explored to get an idea of the power of the test, before finally applying this method to the CSHA data.

Chapter 5

Applications

5.1 Simulation Studies

This section will provide the results of a number of simulations assessing the behaviour of D under different sample sizes and amounts of censoring. Namely, we are interested in investigating the power of the test to reject a false null hypothesis. We do this by first simulating from a particular known distribution, say, Weibull. We re-fit the simulated data using a Weibull model, as well as fit the data nonparametrically according to Vardi's algorithm discussed in the previous chapter, and then we can obtain d_i for each simulated set of data. Next, we calculate the 90th, 95th, and 99th percentiles of d_i for $i = 1, \ldots, n$ ($n = 1000$ was used). These percentiles become the critical values at which we would reject the null hypothesis at the 10%, 5%, and 1% significance levels, respectively. Using the same simulated sets of Weibull data, we re-fit using a log-normal model and calculate the d_i values under this model. We then calculate the proportion of d_i's that are above the critical values, and these proportions provide an idea of the probability that a log-normal null hypothesis would be rejected, given the data is Weibull, at the given significance level. We do the same with the log-logistic model. This gives us insight into how often a log-normal or

log-logistic null hypothesis would be rejected, if the data was, indeed, distributed according to a Weibull distribution. We do this for various sample sizes and amounts of censoring. Next, we repeat the above procedure starting with the log-normal model, and finally the log-logistic model. A fixed censoring approach will be used since, as we will see later, the different censoring distributions have negligible effects on the results of the goodness-of-fit tests.

Sample sizes of 30, 100, 250, 500 and 1000 were considered. Amounts of censoring of $\approx 5\%$, $\approx 20\%$, and $\approx 50\%$ were considered as light, medium, and heavy, respectively. The following tables give the results of the simulation studies.

Weibull Data		Critical Values for D		
Sample Size	Censoring	$\alpha_{.10}$	$\alpha_{.05}$	$\alpha_{.01}$
30	$\approx 5\%$	0.285	0.329	0.478
	$\approx 20\%$	0.290	0.345	0.479
	$\approx 50\%$	0.277	0.327	0.478
100	$\approx 5\%$	0.193	0.227	0.355
	$\approx 20\%$	0.199	0.227	0.369
	$\approx 50\%$	0.191	0.223	0.303
250	$\approx 5\%$	0.139	0.163	0.226
	$\approx 20\%$	0.136	0.162	0.261
	$\approx 50\%$	0.134	0.160	0.235
500	$\approx 5\%$	0.102	0.122	0.174
	$\approx 20\%$	0.106	0.125	0.203
	$\approx 50\%$	0.100	0.123	0.204
1000	$\approx 5\%$	0.078	0.090	0.153
	$\approx 20\%$	0.080	0.092	0.130
	$\approx 50\%$	0.078	0.092	0.134

Table 5.1: Weibull critical values - varying sample size & amt. of censoring

Log-Normal H_0		Power		
Sample Size	Censoring	$1-\beta_{.10}$	$1-\beta_{.05}$	$1-\beta_{.01}$
30	$\approx 5\%$	0.129	0.076	0.015
	$\approx 20\%$	0.116	0.067	0.018
	$\approx 50\%$	0.128	0.061	0.017
100	$\approx 5\%$	0.221	0.111	0.019
	$\approx 20\%$	0.198	0.113	0.014
	$\approx 50\%$	0.201	0.108	0.028
250	$\approx 5\%$	0.459	0.250	0.055
	$\approx 20\%$	0.443	0.241	0.024
	$\approx 50\%$	0.409	0.214	0.038
500	$\approx 5\%$	0.820	0.601	0.114
	$\approx 20\%$	0.731	0.474	0.045
	$\approx 50\%$	0.734	0.456	0.036
1000	$\approx 5\%$	0.987	0.934	0.150
	$\approx 20\%$	0.983	0.991	0.336
	$\approx 50\%$	0.970	0.849	0.208

Table 5.2: Power when H_0 is log-normal and Weibull is true distribution

Log-Logistic H_0		Power		
Sample Size	Censoring	$1-\beta_{.10}$	$1-\beta_{.05}$	$1-\beta_{.01}$
30	$\approx 5\%$	0.222	0.148	0.027
	$\approx 20\%$	0.200	0.121	0.033
	$\approx 50\%$	0.214	0.113	0.028
100	$\approx 5\%$	0.487	0.287	0.045
	$\approx 20\%$	0.454	0.274	0.034
	$\approx 50\%$	0.416	0.246	0.067
250	$\approx 5\%$	0.877	0.676	0.198
	$\approx 20\%$	0.860	0.643	0.075
	$\approx 50\%$	0.821	0.597	0.125
500	$\approx 5\%$	0.993	0.973	0.598
	$\approx 20\%$	0.990	0.939	0.224
	$\approx 50\%$	0.997	0.943	0.174
1000	$\approx 5\%$	1.000	1.000	0.838
	$\approx 20\%$	1.000	1.000	0.965
	$\approx 50\%$	1.000	1.000	0.921

Table 5.3: Power when H_0 is log-logistic and Weibull is true distribution

Log-Normal Data		Critical Values for D		
Sample Size	Censoring	$\alpha_{.10}$	$\alpha_{.05}$	$\alpha_{.01}$
30	$\approx 5\%$	0.193	0.214	0.240
	$\approx 20\%$	0.186	0.203	0.237
	$\approx 50\%$	0.202	0.220	0.261
100	$\approx 5\%$	0.112	0.121	0.149
	$\approx 20\%$	0.117	0.130	0.151
	$\approx 50\%$	0.115	0.123	0.147
250	$\approx 5\%$	0.074	0.080	0.094
	$\approx 20\%$	0.074	0.081	0.096
	$\approx 50\%$	0.075	0.082	0.095
500	$\approx 5\%$	0.054	0.060	0.073
	$\approx 20\%$	0.052	0.058	0.066
	$\approx 50\%$	0.054	0.060	0.070
1000	$\approx 5\%$	0.038	0.042	0.053
	$\approx 20\%$	0.038	0.042	0.050
	$\approx 50\%$	0.037	0.040	0.048

Table 5.4: Log-normal critical values - varying sample size & amt. of censoring

Weibull H_0		Power		
Sample Size	Censoring	$1-\beta_{.10}$	$1-\beta_{.05}$	$1-\beta_{.01}$
30	$\approx 5\%$	0.692	0.616	0.516
	$\approx 20\%$	0.675	0.608	0.462
	$\approx 50\%$	0.564	0.487	0.323
100	$\approx 5\%$	0.981	0.974	0.938
	$\approx 20\%$	0.966	0.953	0.901
	$\approx 50\%$	0.949	0.927	0.856
250	$\approx 5\%$	1.000	1.000	1.000
	$\approx 20\%$	1.000	1.000	0.998
	$\approx 50\%$	0.998	0.997	0.994
500	$\approx 5\%$	1.000	1.000	1.000
	$\approx 20\%$	1.000	1.000	1.000
	$\approx 50\%$	1.000	1.000	1.000
1000	$\approx 5\%$	1.000	1.000	1.000
	$\approx 20\%$	1.000	1.000	1.000
	$\approx 50\%$	1.000	1.000	1 .000

Table 5.5: Power when H_0 is Weibull and Log-normal is true distribution

Log-Logistic H_0		Power		
Sample Size	Censoring	$1-\beta_{.10}$	$1-\beta_{.05}$	$1-\beta_{.01}$
30	$\approx 5\%$	0.176	0.101	0.051
	$\approx 20\%$	0.179	0.122	0.048
	$\approx 50\%$	0.148	0.080	0.027
100	$\approx 5\%$	0.335	0.243	0.075
	$\approx 20\%$	0.253	0.170	0.067
	$\approx 50\%$	0.281	0.193	0.076
250	$\approx 5\%$	0.593	0.487	0.264
	$\approx 20\%$	0.571	0.464	0.212
	$\approx 50\%$	0.520	0.393	0.195
500	$\approx 5\%$	0.869	0.727	0.437
	$\approx 20\%$	0.848	0.753	0.569
	$\approx 50\%$	0.789	0.652	0.396
1000	$\approx 5\%$	0.991	0.967	0.856
	$\approx 20\%$	0.986	0.957	0.849
	$\approx 50\%$	0.985	0.969	0.839

Table 5.6: Power when H_0 is log-logistic and Log-normal is true distribution

Log-Logistic Data		Critical Values for D		
Sample Size	Censoring	$\alpha_{.10}$	$\alpha_{.05}$	$\alpha_{.01}$
30	$\approx 5\%$	0.200	0.228	0.282
	$\approx 20\%$	0.198	0.220	0.269
	$\approx 50\%$	0.214	0.239	0.287
100	$\approx 5\%$	0.119	0.134	0.168
	$\approx 20\%$	0.121	0.135	0.170
	$\approx 50\%$	0.117	0.133	0.171
250	$\approx 5\%$	0.079	0.088	0.107
	$\approx 20\%$	0.077	0.087	0.107
	$\approx 50\%$	0.079	0.088	0.112
500	$\approx 5\%$	0.057	0.062	0.073
	$\approx 20\%$	0.056	0.063	0.079
	$\approx 50\%$	0.056	0.062	0.077
1000	$\approx 5\%$	0.040	0.045	0.055
	$\approx 20\%$	0.040	0.044	0.055
	$\approx 50\%$	0.039	0.043	0.053

Table 5.7: Log-logistic critical values - varying sample size & amt. of censoring

Weibull H_0		Power		
Sample Size	Censoring	$1-\beta_{.10}$	$1-\beta_{.05}$	$1-\beta_{.01}$
30	$\approx 5\%$	0.901	0.848	0.708
	$\approx 20\%$	0.866	0.822	0.680
	$\approx 50\%$	0.761	0.687	0.565
100	$\approx 5\%$	1.000	0.999	0.996
	$\approx 20\%$	0.998	0.996	0.994
	$\approx 50\%$	0.997	0.988	0.963
250	$\approx 5\%$	1.000	1.000	1.000
	$\approx 20\%$	1.000	1.000	1.000
	$\approx 50\%$	1.000	1.000	1.000
500	$\approx 5\%$	1.000	1.000	1.000
	$\approx 20\%$	1.000	1.000	1.000
	$\approx 50\%$	1.000	1.000	1.000
1000	$\approx 5\%$	1.000	1.000	1.000
	$\approx 20\%$	1.000	1.000	1.000
	$\approx 50\%$	1.000	1.000	1.000

Table 5.8: Power when H_0 is Weibull and log-logistic is true distribution

Log-Normal H_0		Power		
Sample Size	Censoring	$1-\beta_{.10}$	$1-\beta_{.05}$	$1-\beta_{.01}$
30	$\approx 5\%$	0.236	0.156	0.052
	$\approx 20\%$	0.234	0.150	0.068
	$\approx 50\%$	0.202	0.117	0.047
100	$\approx 5\%$	0.513	0.362	0.139
	$\approx 20\%$	0.447	0.327	0.143
	$\approx 50\%$	0.462	0.314	0.108
250	$\approx 5\%$	0.783	0.679	0.423
	$\approx 20\%$	0.780	0.647	0.383
	$\approx 50\%$	0.737	0.602	0.297
500	$\approx 5\%$	0.973	0.938	0.824
	$\approx 20\%$	0.960	0.902	0.666
	$\approx 50\%$	0.937	0.882	0.658
1000	$\approx 5\%$	0.998	0.994	0.978
	$\approx 20\%$	0.999	0.996	0.979
	$\approx 50\%$	0.999	0.994	0.965

Table 5.9: Power when H_0 is log-normal and log-logistic is true distribution

The simulation results generally are as expected. That is, we see an increased power as we increase the sample size. The effect of censoring does not seem to have a discernible impact on the power of the test. For data distributed according to a Weibull distribution, the power to reject a false log-normal or log-logistic hypothesis is quite high for very large samples (even at the 1% level for log-logistic). For samples of size 500, the power to reject a false log-logistic hypothesis is high at the 5% level. In the log-normal case, the power to reject a false Weibull hypothesis is high for all sample sizes, except very small samples of size $n = 30$. The power to reject a false log-logistic hypothesis when the data are log-normal is high for very large samples. And lastly, when the data are distributed according to a log-logistic distribution, the power to reject a false Weibull null hypothesis is very high for samples greater than

$n = 100$, but under a false log-normal hypothesis it is very high only for sample sizes of above $n = 500$.

Since our test appears to be powerful enough at sufficiently large sample sizes, and acts as expected, we can now apply it to the CSHA data. Note that using the above definitions for sample size and censoring, the CSHA data set is considered as large in terms of sample size and medium in terms of amount of censoring. A detailed description of the CSHA data will be given in the next section, and the results of the CSHA goodness-of-fit tests will be given in section 5.3. A closer look at the length-biased likelihood for each model, as well as how to carry out the testing step-by-step, will also be given in section 5.3.

5.2 CSHA Data

The Canadian Study of Health and Aging (CSHA) was a longitudinal study of the epidemiology of dementia and other health problems affecting the elderly across Canada. Diseases that affect the elderly are becoming more common due to our aging population and longer lifespans. Affecting the aging population, dementia is a devastating disease which involves memory deterioration as well as a decline in cognitive, emotional and intellectual functioning. The CSHA has many aims, including estimating the prevalence and incidence of dementia among the elderly, investigating the risk factors for Alzheimer's Disease, as well as estimating the survival distribution of those with dementia. The study has undergone three phases, with the first phase beginning in 1991 (CSHA-1), the second phase in 1996 (CSHA-2), and the third phase in 2001 (CSHA-3).

During the first phase, 10,263 individuals aged 65 and above were recruited cross-sectionally. Canada was split into five geographic regions (British Columbia, the Prairie provinces, Ontario, Quebec, and the Atlantic region) and equal sized samples were drawn from each region. In the final sample there were 9,008 study participants

from the community and 1,255 study participants from long-term care facilities.

It is estimated that approximately 8% of Canadians 65 years of age and over suffer from dementia. Splitting into increasing age groups, a dramatic rise in prevalence is observed with rates of 2% for those aged 65-74, 11% for those aged 75-84, and 35% for individuals 85 years and over (Lindsay et al. 2004). Considering the increase in prevalence with age, the random samples drawn from each area were stratified by age. Those aged 75-84 were sampled at a rate twice than those aged 65-74, and individuals aged 85 and above were sampled at a rate 2.5 times those aged 65-74. The Modified Mini-Mental State (3MS) test was used to screen for cognitive impairment of individuals from the community. Depending on their score, a full clinical assessment and final diagnosis by a committee of health professionals was completed. All sampled individuals living in institutions were given a full clinical assessment as the rate of dementia is much higher among these subjects.

Those who screened positive for dementia were included in the final sample. Three dementia categories were used; "probable Alzheimer's Disease", "possible Alzheimer's Disease", and "vascular dementia." The individuals were classified into one of the categories and then followed until the second phase of the study in 1996. The final sample included 816 possibly censored survival times. The second phase of the study began with a follow-up of the individuals from phase one who were still alive.

For this thesis, the phase one portion of the CSHA will be used. The approximate date of onset of dementia, date of death or censoring, as well as the death indicator will be used in estimating the survival function. Gender will also be included following the initial analysis. As is common to epidemiological studies, the data were subject to both right-censoring and left-truncation. Subjects were censored if they were still alive at the end of the study or if they were lost to follow-up, although loss to follow-up occurred in only a small percentage of subjects. Left-truncation follows from the fact that prevalent cases, instead of incident cases, were selected. Only individuals who were alive with dementia at the beginning of the study had the possibility of being

included in the sample, and so individuals with longer survival times were favoured.

Both Asgharian et al. (2006) and Addona & Wolfson (2006) proposed tests to assess stationarity of incidence rates, and could not reject the stationarity assumption for the CSHA data with their methods. This, coupled with the fact that dementia rates should have remained relatively constant over the period that covers the CSHA data dementia onset times, allows us to assume that the random left-truncation times are uniformly distributed. Under the assumption of stationarity discussed earlier, it follows that the data are length-biased. When the length-bias of the data is considered, the median survival of demented patients is estimated to be much shorter than previous studies have shown. Wolfson et al. (2001) reported an adjusted median survival of 3.3 years when correcting for length-bias, contrasted with median survival times varying from 5 to 9.3 years as suggested by previous studies.

5.3 Analysis

Before moving on to the results of the goodness-of-fit testing, Figure 5.1 shows both the unbiased and length-biased nonparametric survival estimates, using Vardi's algorithm. The importance of correcting for length-bias is readily observed.

Using length-biased Weibull (1), log-normal (2), and log-logistic (3) models, goodness-of-fit tests will be performed and compared using the length-biased, right-censored survival data of dementia patients from the CSHA. The Kolmogorov-Smirnov type statistic, D, discussed in Chapter 4, will be used as the measure of goodness-of-fit. The following steps will be used for all three models. The first step is to use the data to estimate the parameters of the survival function for the length-biased model. This will be done by maximum likelihood estimation using the non-linear minimization (nlm) function in the statistical package R. The length-biased likelihood discussed

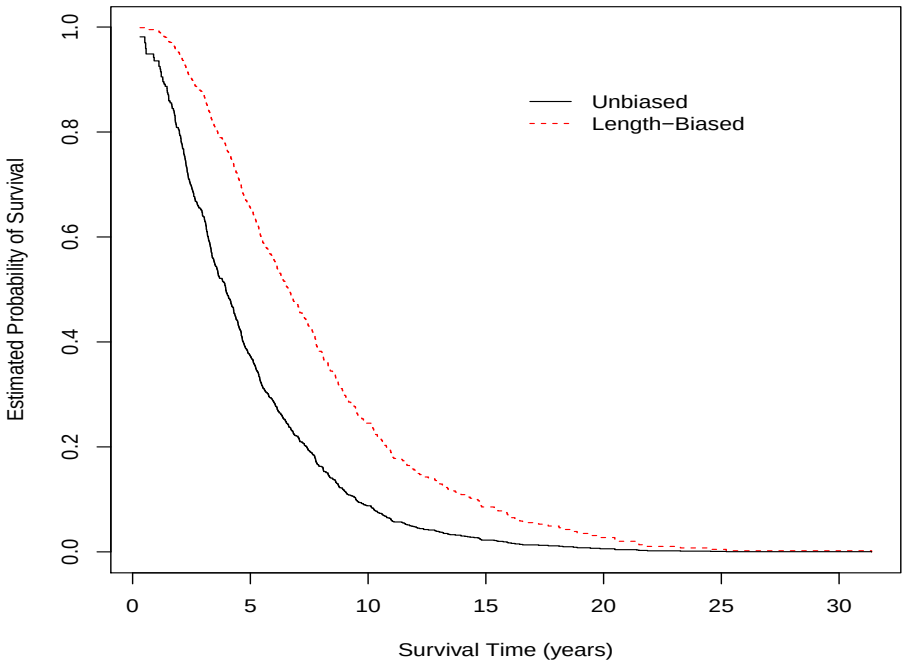

Figure 5.1: Biased and unbiased nonparametric estimates of $S(x)$.

in Chapter 3 can be written as:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \frac{f_{\boldsymbol{\theta}}(x_i)^{\delta_i} S_{\boldsymbol{\theta}}(x_i)^{1-\delta_i}}{\mu(\theta)}. \tag{5.3.1}$$

As maximizing the log-likelihood is equivalent to maximizing the likelihood (and often simpler), the following are the log-likelihoods of the length-biased Weibull, log-normal and log-logistic models; respectively,

$$L_1(\boldsymbol{\theta_1}) = \sum_{i=1}^{n} \Big[(p\delta_i+1)\log(\lambda)+\delta_i \log(p)+\delta_i(p-1)\log(x_i-(\lambda x_i)^p-\log(1+\frac{1}{p})\Big] \tag{5.3.2}$$

$$\begin{aligned} L_2(\boldsymbol{\theta_2}) = \sum_{i=1}^{n} \Big[\delta_i[-\frac{1}{2}(\frac{\log(x_i)-\mu}{\sigma})^2] + (1-\delta_i)\log[1-\Phi(\frac{\log(x_i)-\mu}{\sigma})] \\ - \delta_i \log[x_i(2\pi)^{1/2}\sigma] - \mu - \frac{1}{2}\sigma^2\Big] \end{aligned} \tag{5.3.3}$$

$$\begin{aligned} L_3(\boldsymbol{\theta_3}) = \sum_{i=1}^{n} \Big[(\delta_i+1)\log(\alpha) + \delta_i(\alpha-1)\log(x_i) + (\delta_i+\frac{1}{\alpha})\log(\theta) \\ + \log[\sin(\frac{\pi}{\alpha})] - (\delta_i+1)\log(1+\theta x_i^{\alpha}) - \log(\pi)\Big] \end{aligned} \tag{5.3.4}$$

Once the likelihoods have been maximized, maximum likelihood estimates for the parameters of each model are obtained. Specifically, for the Weibull model we have $\hat{\lambda}$ and $\hat{p}$, $\hat{\mu}$ and $\hat{\sigma}$ for the log-normal model, and $\hat{\alpha}$ and $\hat{\theta}$ for the log-logistic model. Using the regular survival function (i.e. unbiased) with the obtained estimates, we have a parametric model which estimates the survival distribution corrected for length-bias. For $i = 1, 2, 3$, let $\hat{S}_i(x)$ represent the estimated unbiased parametric survival curve.

The next step is to use the data to estimate the unbiased survival curve non-parametrically using the Vardi Algorithm discussed in Chapter 4. A step-function will be obtained with jumps at each time point from the data. Let $\hat{S}_V(x)$ represent the estimated nonparametric unbiased survival curve.

Our Kolmgorov-Smirnov type statistic for the i^{th} model is calculated as

$$d_i = \max|\hat{S}_V(x) - \hat{S}_i(x)|, \tag{5.3.5}$$

and d_i is the largest vertical distance between the two curves. If our hypothesized distribution was a fully specified model, d_i would correspond to a p-value for our test. Since this is not the case, simulation is the next step in assessing the goodness-of-fit for each model. As stated earlier, the p-value is approximated by $P(D \geq d_i)$. This p-value will be estimated by following the simulation steps detailed in Algorithm 6. The hypotheses being tested are

$$\begin{aligned} H_0 &: S(x) = S_i*(x) \\ H_1 &: S(x) \neq S_i^*(x). \end{aligned} \tag{5.3.6}$$

A small p-value corresponds to the conclusion that the model in question is not a good fit for the data (i.e. rejection of the null hypothesis). Initially, we will fix $n = 816$, the sample size of the CSHA data set. N, the number of iterations in our simulation, is required to be large, and a value of 10,000 will be used. For each model, the simulation will be performed five times using the three censoring schemes described in Chapter 4. Two simulations will use the fixed censoring scheme with values of $c_1 = 5.2$ and $c_2 = 5.8$ years. These values represent a constant follow-up at 5.2 and 5.8 years after the study began. The next two simulations will use a random normal censoring scheme with a mean of $\mu_1 = 5.2$, $\sigma_1 = 0.3$ and $\mu_2 = 5.8$, $\sigma_2 = 0.6$ respectively. Here, the follow-up times are not all the same, which is realistic as it is not always possible to follow-up with every subject at the exact same time. The values of 5.2 and 5.8 years were chosen by considering the follow-up time in the actual study. Lastly, a censoring scheme that samples values from the real residual censoring times in the CSHA data set will be used. After the initial simulations using a sample size of $n = 816$ are carried out, the women and men will be analyzed separately.

Table 5.10 contains the d values, estimated model parameters, along with their standard deviation and confidence intervals, for each model, followed by their respective graphs in Figures 5.2, 5.3 and 5.4.

Model	d_i	Estimate	Std. Dev.	C. I.
Weibull	0.096	$\lambda = 0.207$	0.009	(0.189, 0.225)
		$p = 1.215$	0.046	(1.125, 1.305)
Log-Normal	0.065	$\mu = 1.378$	0.033	(1.313, 1.443)
		$\sigma = 0.679$	0.018	(0.644, 0.715)
Log-Logistic	0.124	$\alpha = 2.808$	0.075	(2.660, 2.955)
		$\lambda = 0.018$	0.003	(0.012, 0.023)

Table 5.10: CSHA parameter estimates

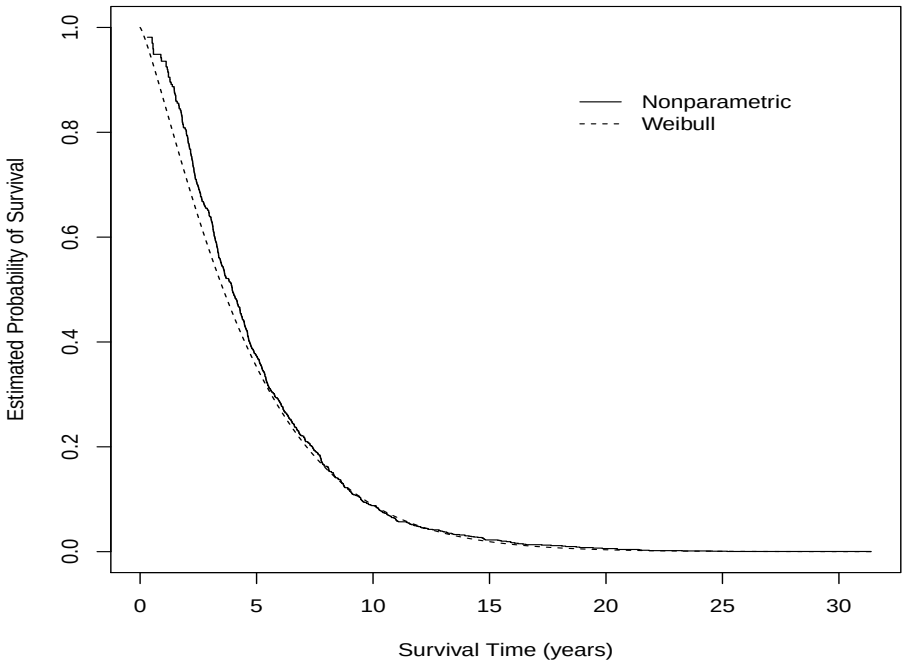

Figure 5.2: Nonparametric and Weibull estimate of $S(x)$.

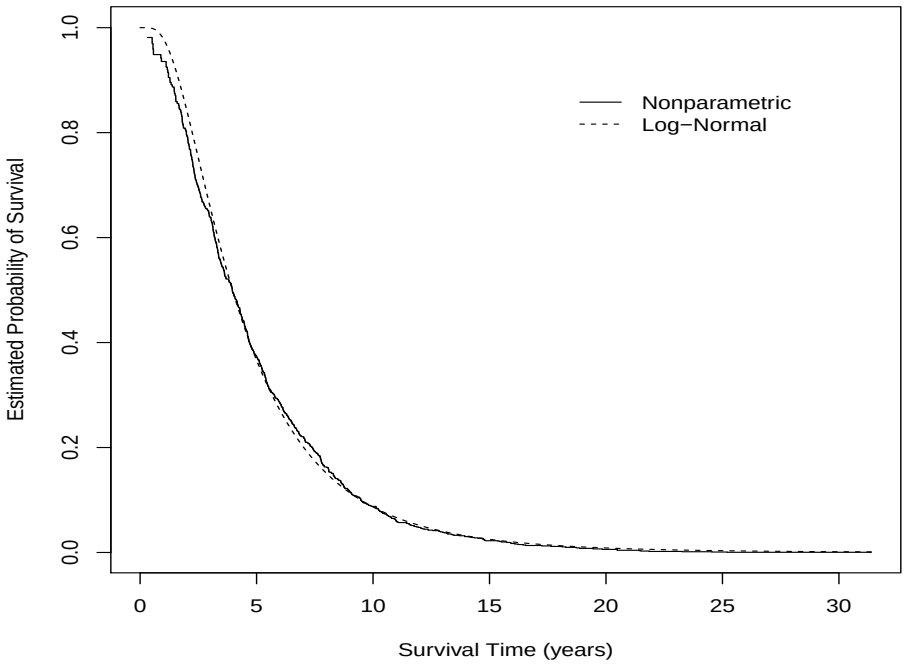

Figure 5.3: Nonparametric and log-normal estimate of $S(x)$.

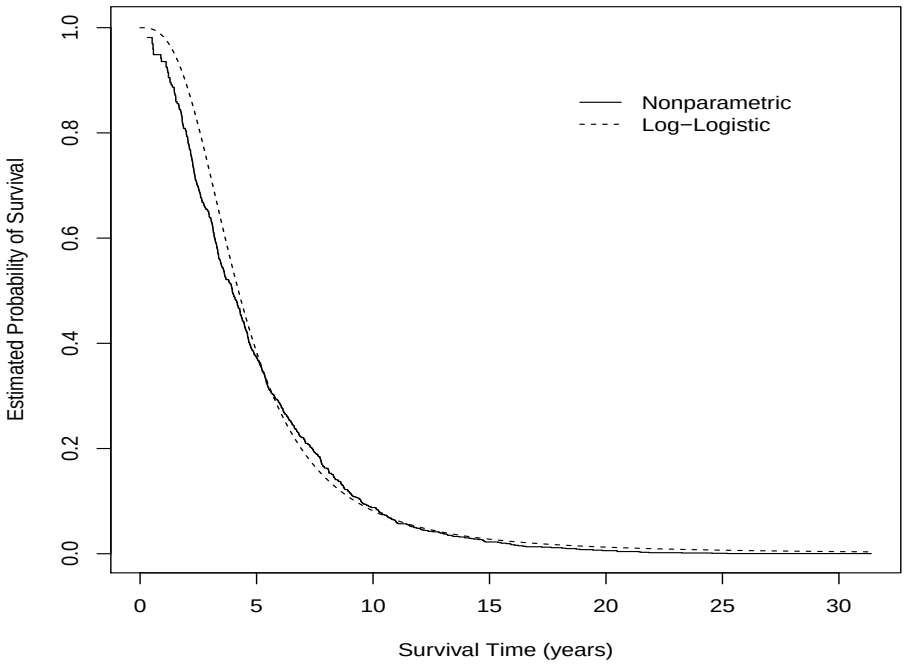

Figure 5.4: Nonparametric and log-logistic estimates of $S(x)$.

Upon first observation of the graphs, it appears that the log-normal distribution is the best fit. The overall shape of the log-normal model seems to accurately mimic the nonparametric estimate. However, a good visual fit is generally not sufficient. The log-normal model would also be best if we based our inference on the size of the D statistic. If we began with fully specified models, we would conclude that the log-normal model is best, since with fully specified distributions we can rely on the probability integral transform to obtain the p-value regardless of the form of S_θ. But since the data was used to estimate parameters, we 'lose degrees of freedom,' and can no longer rely on the smallest d to choose the best fit. Hence, simulations are required to obtain the approximated p-values for each model. The simulation results are given in Table 5.11.

Censoring Approach	Weibull	Log-Normal	Log-Logistic
KM	0.0550	0.0007	0.0000
Fixed C=5.2	0.0556	0.0015	0.0000
Fixed C=5.8	0.0576	0.0008	0.0000
Normal (5.2,0.3)	0.0550	0.0011	0.0000
Normal (5.2,0.3)	0.0544	0.0010	0.0000

Table 5.11: p-values estimated by simulation

From the results in Table 5.11, we reject the hypothesis that the data come from a log-normal or log-logistic population. At a 5% significance level we cannot reject the hypothesis that the data come from a Weibull population, although this is very close. It is also clear from the simulation results that the different approaches to censoring have negligible effects.

The next step taken was to stratify the CSHA data by gender, as was done by Cook & Bergeron (2011). Also, conventional epidemiological wisdom suggests that men and women should be analyzed separately. Earlier research has shown that

women have a higher risk of Alzheimer's disease than men (Seeman 1997). In the CHSA, a higher proportion of men suffered from vascular dementia, and a higher proportion of women suffered from Alzheimer's disease. Thus, men and women may suffer from different causes of dementia, and perhaps stratifying by gender may provide us with better insight.

The data was split by gender and d_i was calculated for men and women, for each model. The log-logistic was not a good fit for either gender, and so a simulation was not performed for this model. The Weibull model appeared to be a good fit for the women, and the log-normal model a good fit for the men. The d values, estimated model parameters and their standard deviations and confidence intervals are given in Table 5.12, and the respective graphs are given in Figure 5.5 and Figure 5.6.

Model	D	Estimate	Std. Dev.	C. I.
Weibull (women)	0.049	$\lambda = 0.187$	0.009	(0.169, 0.204)
		$p = 1.312$	0.059	(1.196, 1.429)
Log-Normal (men)	0.033	$\mu = 1.244$	0.063	(1.120, 1.369)
		$\sigma = 0.692$	0.034	(0.626, 0.758)

Table 5.12: CSHA Estimates - Women (Weibull) & Men (Log-normal)

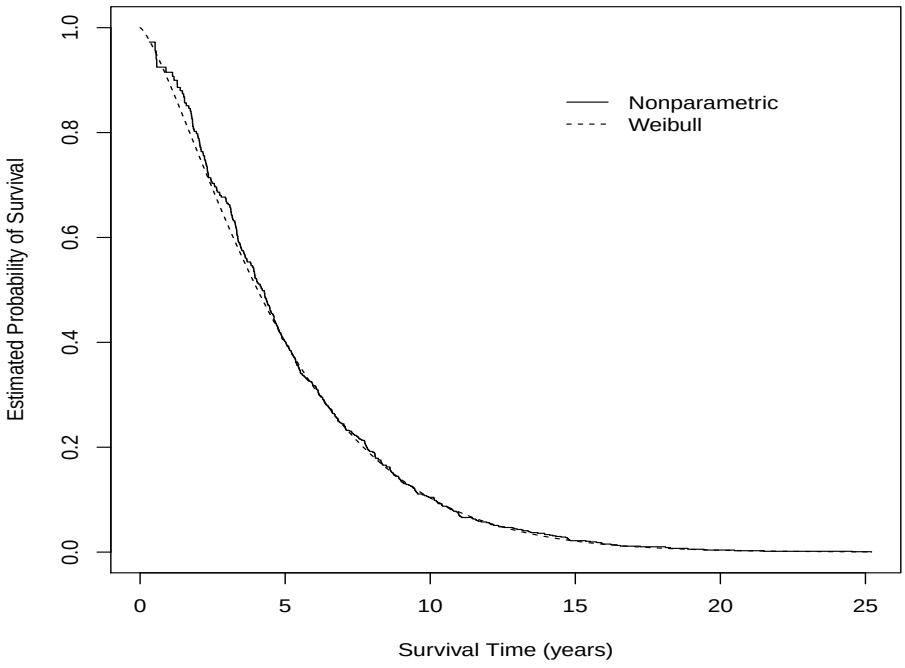

Figure 5.5: Nonparametric and Weibull estimate of $S(x)$ (women).

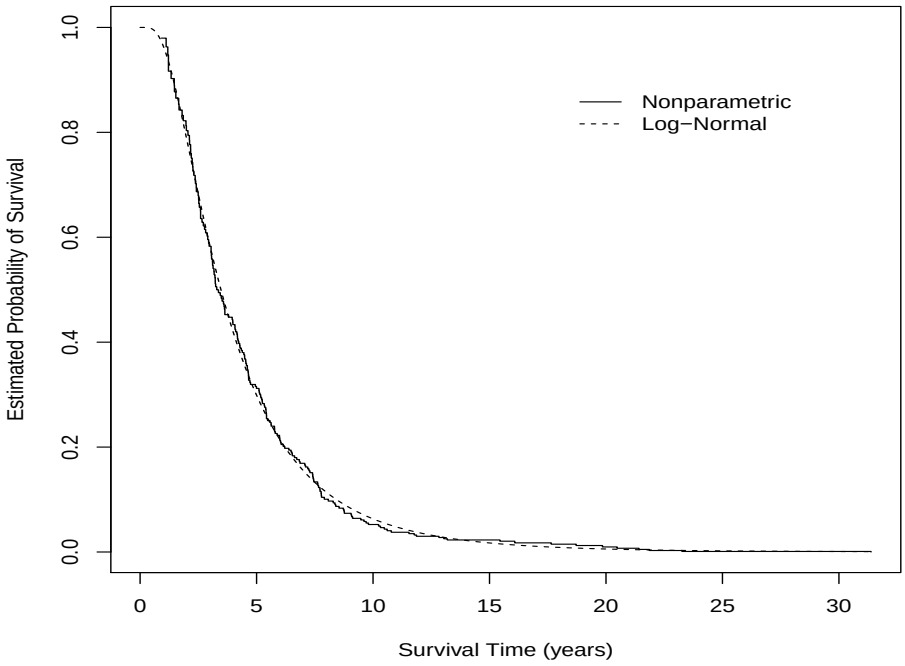

Figure 5.6: Nonparametric and log-normal estimate of $S(x)$ (men).

Graphs of the estimated survival for up to 10 years are shown in Figures 5.7 and 5.8. These are included since surviving for 25 years or above with dementia is quite rare, and shorter survival times may be more realistic.

The estimated hazard functions for women and men are shown in Figure 5.9. A monotone increasing hazard is seen for the women, which is expected as the estimated Weibull shape parameter is greater than 1. An increasing hazard function suggests that the risk of death with dementia increases as time goes on. For the men, a hump-shaped hazard is seen, which is an inherent property of the log-normal model. The hazard function peaks at approximately 5 years, which is not surprising given the data (i.e. - number of events becomes more sparse past 5 year mark, and hazard appears to decrease).

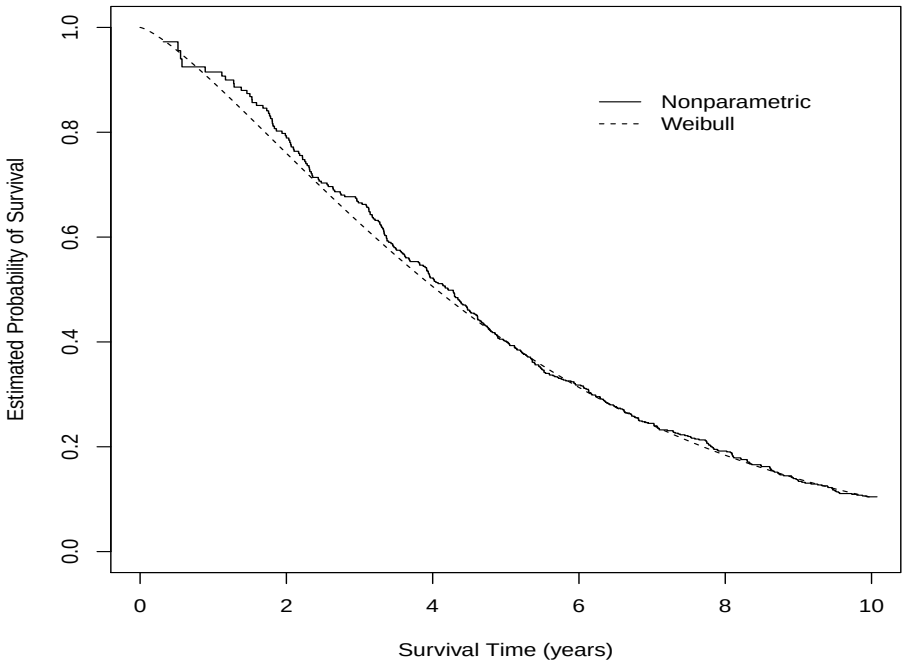

Figure 5.7: Nonparametric and Weibull estimate of $S(x)$ for 0-10 years (women) .

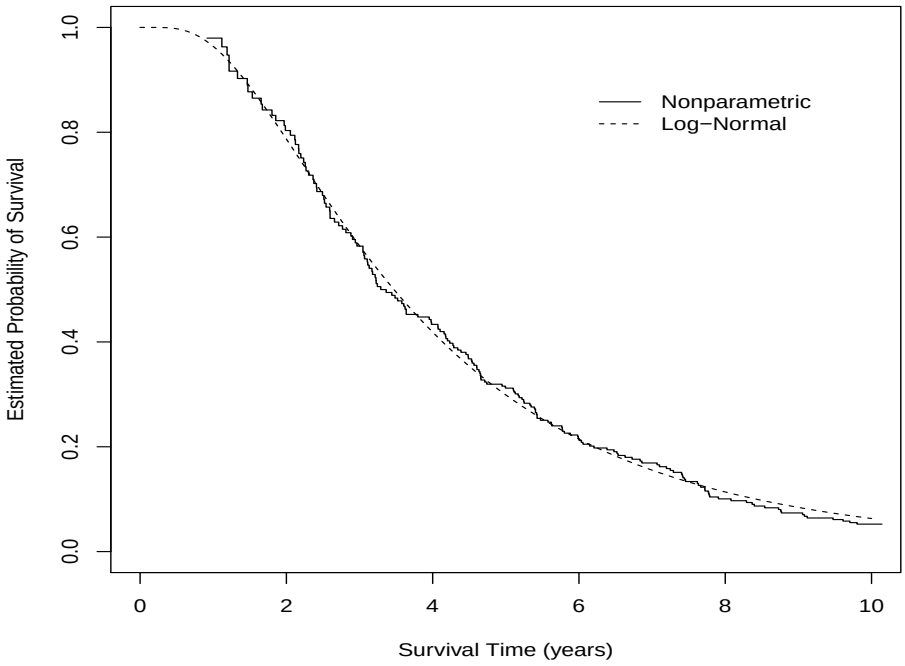

Figure 5.8: Nonparametric and log-normal estimate of $S(x)$ for 0-10 years (men) .

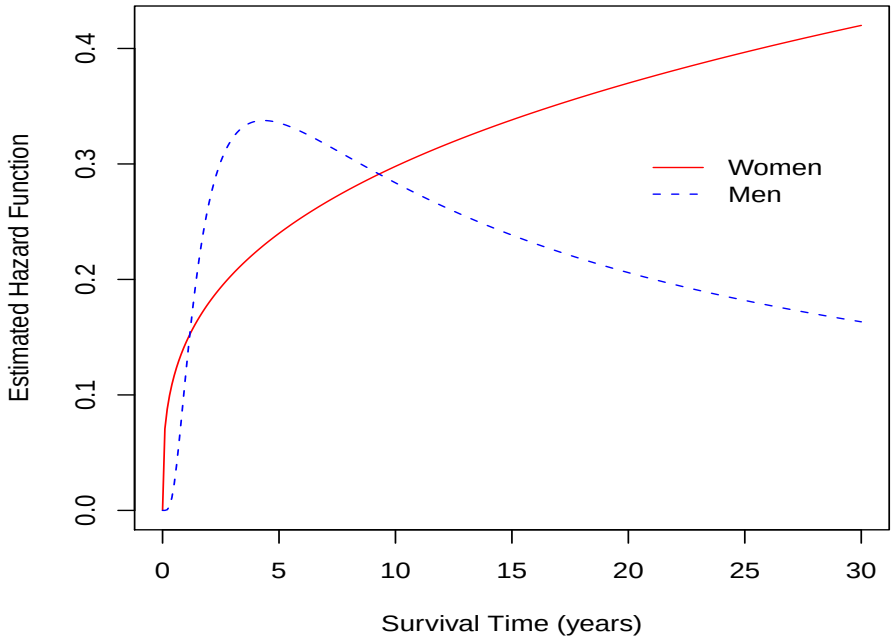

Figure 5.9: Estimated hazard function for women (Weibull) and men (log-normal.

Simulations were performed for each group, and the results are given in Table 5.13.

Censoring Approach	Women - Weibull	Men - Log-Normal
KM	0.5756	0.9753
Fixed C=5.2	0.5786	0.9753
Fixed C=5.8	0.5750	0.9749
Normal (5.2,0.3)	0.5723	0.9775
Normal (5.2,0.3)	0.5719	0.9767

Table 5.13: p-values estimated by simulation - Women & Men

The results when the subjects are split into men and women differ greatly from those when analyzing the group as a whole. The log-normal model seems to be good fit for the men and the Weibull model a good fit for the women. This suggests that the men and women come from two different underlying populations, as conventional epidemiological wisdom suggests. For future analyses, gender may be considered as an important variable. However, when split by gender, the men are a sample of size $n = 237$ and the women are a sample size of $n = 579$. As was seen earlier, for samples comparable in size to that of the men, sufficient power is lacking. This test may not be useful for smaller samples, in the sense that it may not detect a difference between parametric models when n is small.

When generating the length-biased samples, occasionally a very small time (e.g. $t \leq 0.005$) was generated, although this was rare. A survival time this small severely affects the unbiased nonparametric estimate of the survival curve, and so only times greater than t=0.01 were generated for the simulations. In real applications, times smaller than 0.01 are generally not observed, making this constraint inconsequential. Figure 5.10 shows three curves: the unbiased parametric estimate of $S(x)$, the unbiased nonparametric estimate of $S(x)$ for a simulated data set containing a very small time point (t=0.0049, generated from length-biased Weibull distribution in R), and

the unbiased nonparametric estimate when this small time point is removed. Only one parametric estimate is shown as the difference between the parametric estimates with and without the small time point is very small and and unobservable from the graph.

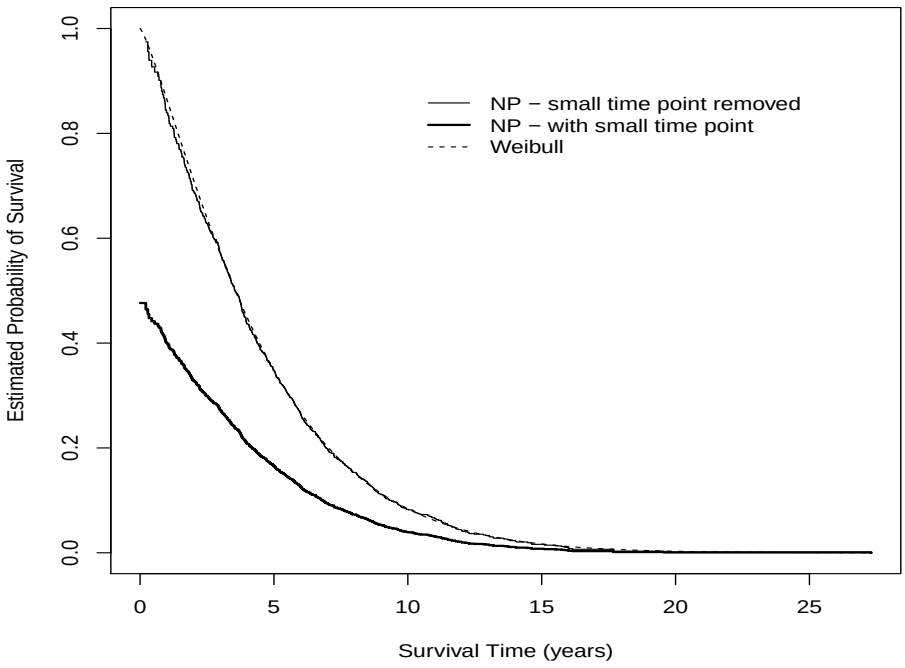

Figure 5.10: Impact of small survival time on Nonparametric (NP) estimate.

As can be seen, when the small data point is included in the data set, the nonparametric estimate does not perform well. It should be noted that 'small' in this case is a relative term. The relative size of the smallest survival time compared to the second smallest time (as well as the rest of the data set) is the cause of this problem (in Figure 5.10, $t_{(1)}$ is approximately 43 times smaller than $t_{(2)}$). This is due to the inverse length-bias transformation formula used to correct the final probability vector, $\hat{\mathbf{p}}^*$, to obtain the unbiased probability vector. Since survival times < 0.01 are rarely observed, it was not considered a problem to add this constraint when generating the

data sets for the simulation, and we can be ensured that the nonparametric estimate will perform as expected.

Chapter 6

Conclusion

The goal of this thesis was to develop a one-sample goodness-of-fit test for length-biased data subject to right-censoring. Using Weibull, log-normal and log-logistic models, the proposed test was applied to length-biased survival data collected from the Canadian Study of Health and Aging.

When potentially censored survival data on individuals is collected by means of an incident study, the standard tool to estimate the survival function, from onset, is the Kaplan-Meier estimator (Kaplan & Meier 1958). If, instead, a prevalent cohort is used to identify individuals, the survival times are then subject to left-truncation. By modifying the Kaplan-Meier approach, an estimate conditional upon truncation time can be used to estimate the survival function from onset. When onset times of a disease can be assumed to come from a segment of a stationary Poisson process, the survival times are said to be properly length-biased (Wang 1991), and in this case an unconditional approach to estimation of the survival function from onset, pioneered by Vardi (1989), will provide a more efficient estimate (Asgharian et al. 2002). These ideas were discussed in Chapter 2.

In Chapter 3, the length-biased likelihood for right-censored survival data was reviewed. This likelihood can be maximized both parametrically and nonparamet-

rically. Some common parametric models used in survival analysis were discussed, namely, the Weibull, log-normal, and log-logistic models. For each of these models, we use the length-biased likelihood to obtain estimates for the biased survival function from onset, and then we can obtain a parametric estimate for the unbiased survival function. Using the nonparametric maximum likelihood for the unbiased survival function, we can quantify the discrepancy of the hypothesized parametric models using an appropriate goodness-of-fit statistic. The Kolmogorov-Smirnov statistic, as well as other goodness-of-fit statistics used for lifetime data, were also discussed in Chapter 3. It should be noted that a two-sample nonparametric test for the equality of survival distributions, under length-biased sampling, has recently been developed (Ning et al. 2010).

We did not begin with a fully specified model, but instead used the data to estimate the parameters for our hypothesized distributions, and hence simulation is required in order to approximate a p-value for our goodness-of-fit tests. Chapter 4 was devoted to the algorithms necessary to carry out our goodness-of-fit testing. These algorithms include Vardi's algorithm for a nonparametric estimation of the survival function from onset, the simulation of length-biased samples of the appropriate parametric form, as well as an algorithm for p-value approximation for our tests.

Before applying our test, it was necessary to investigate its behaviour under different scenarios. Once sufficient power for the test was established, the next step was to apply these concepts to a real set of right-censored length-biased data; these results were discussed in Chapter 5. Individuals included in the CSHA were recruited cross-sectionally and then screened for dementia. Those who screened positively were included in the final sample, which consisted of 816 individuals, who were followed forward for 5 years, at which point they were recorded as either censored or not censored. The stationarity of the onset times has been previously verified for the CSHA data (Addona & Wolfson 2006), and the onset of dementia for those included in the sample occurred prior to recruitment, hence, the CSHA data is properly length-

biased.

The unbiased survival function, from onset, was estimated using a Weibull, log-normal, and log-logistic model and each parametric estimate curve was compared to the unbiased unconditional nonparametric survival estimate. Although visually the log-normal model appeared to be the best fit to the data, it was determined using simulation that this was not the case, hence the need to develop adequate one-sample goodness-of-fit tests. Both the log-normal and log-logistic models had p-values very close to 0, and therefore the hypothesis that the data come from a log-normal or log-logistic population was rejected. The p-value for the Weibull model was slightly above 0.05, but we chose to reject the null hypothesis in the case. The next step was to split the data by gender and re-perform the analysis. In doing so, it was concluded that the log-normal model was a good fit for the men, and the Weibull model was a good fit for the women. These results imply that the men and women come from two different populations, which lines up with conventional epidemiological wisdom which suggests that men and women should be analyzed separately. However, since this was a preliminary analysis, an approach which includes covariates could be tested. For example, an accelerated failure time model or cox proportional hazard model could be used to estimate the survival function from onset.

In terms of future research ideas, a few things should be mentioned. Instead of only considering the maximum distance between the nonparametric and parametric curves, as the Kolmogorov-Smirnov statistic does, we could employ a different goodness-of-fit statistic, such as the Cramer-von Mises or Anderson-Darling statistic discussed in Chapter 3, to evaluate the fit of our parametric models. Using the maximum distance was a first step, and employing a different way of quantifying the discrepancy between the curves is a natural extension. Also, a possible extension of the Shapiro-Wilk test could be considered for the log-normal model (Shapiro & Wilk 1965).

The data collected during the CSHA is complex in nature. Older individuals were

sampled at a higher rate than younger individuals, and equal samples were selected from each geographic region, despite the different population sizes. An extension to estimating the survival function in the presence of length-biasedness would be to incorporate the appropriate sampling weights. For incident cases arising from complex surveys, a weighted Kaplan-Meier curve has been developed (Lawless 2003*a*). This Kaplan-Meier approach uses an estimated population proportion surviving through each time interval, instead of only the proportion surviving in the data set. As far as we know, sampling weights have yet to be incorporated in the estimation of the survival function in the presence of length-bias.

Bibliography

Aalen, O. (1978), 'Nonparametric inference for a family of counting processes', *The Annals of Statistics* pp. 701–726.

Addona, V. & Wolfson, D. (2006), 'A formal test for the stationarity of the incidence rate using data from a prevalent cohort study with follow-up', *Lifetime Data Analysis* **12**(3), 267–284.

Asgharian, M., M'Lan, C. & Wolfson, D. (2002), 'Length-biased sampling with right censoring', *Journal of the American Statistical Association* **97**(457), 201–209.

Asgharian, M. & Wolfson, D. (2005), 'Asymptotic behaviour of the NPMLE of the survivor function when the data are length-biased and subject to right censoring', *Annals of Statistics* **33**, 2109–2131.

Asgharian, M., Wolfson, D. & Zhang, X. (2006), 'Checking stationarity of the incidence rate using prevalent cohort survival data', *Statistics in medicine* **25**(10), 1751–1767.

Bergeron, P. (2006), Covariates and length-biased sampling: Is there more than meets the eye?, PhD thesis, McGill University.

Blumenthal, S. (1967), 'Proportional sampling in life length studies', *Technometrics* pp. 205–218.

Canadian Study of Health and Aging Working Group (1994*a*), 'The Canadian study of health and aging: risk factors for Alzheimer's disease in Canada', *Neurology* **44**, 2073–2080.

Canadian Study of Health and Aging Working Group (1994*b*), 'Canadian study of health and aging: study methods and prevalence of dementia', *Canadian Medial Association Journal* **150**, 899–913.

Canadian Study of Health and Aging Working Group (1994*c*), 'Patterns of caring for people with dementia in Canada', *Canadian Journal on Aging* **13**, 470–487.

Casella, G. & Berger, R. (2001), *Statistical inference*, Duxbury Press.

Conover, W. (1971), *Practical nonparametric statistics*, Wiley.

Cook, R. & Bergeron, P. (2011), 'Information in the sample covariate distribution in prevalent cohorts', *Statistics in Medicine* **30**(12), 1397–1409.

Correa, J. & Wolfson, D. (1999), 'Length-bias: some characterizations and applications', *Journal of statistical computation and simulation* **64**(3), 209–219.

Cox, D. (1969), Some sampling problems in technology, *in* 'New developments in survey sampling', Wiley.

D'Agostino, R. & Stephens, M. (1986), *Goodness-of-fit techniques*, Marcel Dekker, Inc.

Feinleib, M. (1960), 'A method of analyzing log-normally distributed survival data with incomplete follow-up', *Journal of the American Statistical Association* pp. 534–545.

Gao, S. & Hui, S. (2000), 'Estimating the incidence of dementia from two-phase sampling with non-ignorable missing data', *Statistics in medicine* **19**(11-12), 1545–1554.

Gill, R., Vardi, Y. & Wellner, J. (1988), 'Large sample theory of empirical distributions in biased sampling models', *The Annals of Statistics* pp. 1069–1112.

Horner, R. (1987), 'Age at onset of Alzheimer's disease: clue to the relative importance of etiologic factors?', *American journal of epidemiology* **126**(3), 409.

Kaplan, E. & Meier, P. (1958), 'Nonparametric estimation from incomplete observations', *Journal of the American statistical association* pp. 457–481.

Klein, J. & Moeschberger, M. (2003), *Survival analysis: techniques for censored and truncated data*, Springer.

Lagakos, S., Barraj, L. & Gruttola, V. (1988), 'Nonparametric analysis of truncated survival data, with application to AIDS', *Biometrika* **75**(3), 515.

Lawless, J. (2003*a*), Censoring and weighting in survival estimation from survey data, *in* 'Proceedings of the Survey Methods Section', pp. 31–36.

Lawless, J. (2003*b*), *Statistical methods and methods for lifetime data*, Wiley.

Lindsay, J., Sykes, E., McDowell, I., Verreault, R. & Laurin, D. (2004), 'More than the epidemiology of Alzheimer's disease: contributions of the Canadian study of health and aging', *Canadian journal of psychiatry* **49**(2), 83–91.

McFadden, J. (1962), 'On the lengths of intervals in a stationary point process', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 364–382.

Nelson, W. (1972), 'Theory and applications of hazard plotting for censored failure data', *Technometrics* pp. 945–966.

Ning, J., Qin, J. & Shen, Y. (2010), 'Non-parametric tests for right-censored data with biased sampling', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* .

Patil, G. & Rao, C. (1978), 'Weighted distributions and size-biased sampling with applications to wildlife populations and human families', *Biometrics* pp. 179–189.

Rice, J. (1995), *Mathematical statistics and data analysis*, Duxbury Press.

Ross, S. (2006), *Simulation*, Elsevier Academic Press.

Seeman, M. (1997), 'Psychopathology in women and men: focus on female hormones', *American Journal of Psychiatry* **154**(12), 1641.

Shapiro, S. & Wilk, M. (1965), 'An analysis of variance test for normality (complete samples)', *Biometrika* **52**(3/4), 591–611.

Stern, Y., Tang, M., Albert, M., Brandt, J., Jacobs, D., Bell, K., Marder, K., Sano, M., Devanand, D., Albert, S. et al. (1997), 'Predicting time to nursing home care and death in individuals with Alzheimer disease', *JAMA: the journal of the American Medical Association* **277**(10), 806.

Turnbull, B. (1976), 'The empirical distribution function with arbitrarily grouped, censored and truncated data', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 290–295.

Vardi, Y. (1982), 'Nonparametric estimation in the presence of length bias', *The Annals of Statistics* pp. 616–620.

Vardi, Y. (1985), 'Empirical distributions in selection bias models', *The Annals of Statistics* pp. 178–203.

Vardi, Y. (1989), 'Multiplicative censoring, renewal processes, deconvolution and decreasing density: nonparametric estimation', *Biometrika* **76**(4), 751.

Vardi, Y. & Zhang, C. (1992), 'Large sample study of empirical distributions in a random-multiplicative censoring model', *The Annals of Statistics* pp. 1022–1039.

Wang, M. (1991), 'Nonparametric estimation from cross-sectional survival data', *Journal of the American Statistical Association* pp. 130–143.

Wang, M., Jewell, N. & Tsai, W. (1986), 'Asymptotic properties of the product limit estimate under random truncation', *the Annals of Statistics* pp. 1597–1605.

Wicksell, S. (1925), 'The corpuscle problem: a mathematical study of a biometric problem', *Biometrika* **17**(1/2), 84–99.

Wolfson, C., Wolfson, D., Asgharian, M., M'Lan, C., Østbye, T., Rockwood, K. & Hogan, D. (2001), 'A reevaluation of the duration of survival after the onset of dementia', *New England Journal of Medicine* **344**(15), 1111–1116.

Zelen, M. & Feinleib, M. (1969), 'On the theory of screening for chronic diseases', *Biometrika* **56**(3), 601–614.