Group comparisons in the presence of length-biased data

Vittorio Addona

Department of Mathematics and Statistics, McGill University,
805 rue Sherbrooke Ouest, Montréal, Québec, Canada H3A 2K6

August 9, 2001

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements of the degree of Master of Science.

Abstract

The effects of length-bias and left-truncation in survival data have been well studied in the statistical literature. To a lesser extent, the phenomena of length-bias and left-truncation have also been investigated when group comparisons are of interest. This literature examines various biases that may occur under different scenarios, and also, on occasion, proposes procedures for the estimation of covariate effects when using prevalent data. In this thesis, we review the literature concerned with the analysis of length-biased and left-truncated data, paying particular attention to the issue of group comparisons. Some shortcomings of the methods developed in the literature are pointed out. We also assess the effects of failure to recognize the presence of length-bias when performing group comparisons in natural history of disease studies. To our knowledge, this issue has not yet been addressed in the literature.

Résumé

Les effets de biais de longueur et de troncation à gauche dans les données de survies ont été bien étudiés dans la littérature statistique. À un moindre degré, les phénomènes de biais de longueur et troncation à gauche ont aussi été examinés quand les comparaisons de groupes sont d'intérêt. Cette littérature examine divers biais qui peuvent se produirent selon différents scénarios, et aussi, à l'occasion, propose des procédures pour l'estimation des effets covariés lors de l'utilisation de donné prédominantes. Dans cette thèse, nous passons en revue la littérature concernant l'analyse de données à biais de longueur et à troncation à gauche, avec une attention particulière au sujet de la comparaison de groupes. Certains points faibles des méthodes développées dans la littérature sont indiqués. Nous évaluons également les effets du manque de reconnaissance de biais de longueur quand des comparaisons de groupes sont effectuées dans des études de l'histoire naturelles de maladies. À notre connaissance, ce problème n'a encore pas été adressé dans la littérature.

Acknowledgements

This work would not have been possible without the help and support of many people. I would like to thank my thesis supervisor, Dr. David Wolfson. I greatly appreciate all of the guidance and advice he has provided for this thesis and over the last three years. It is clear that he truly cares about his students, and I consider him a friend more than a supervisor.

I would also like to thank my mom and the rest of my family for all of the sacrifices they made to help me get to where I am now. In everyday life these words often get lost, but they should always remember that I love them very much. To my girlfriend Isadora, I send a gentle kiss. Thank you for being there through all the difficult moments when I may have been discouraged and was probably not the easiest person with whom to speak. I love you sweetheart. To my four best friends in the world, Dario, Mike, Paul and Marco, I also say thank you. I will always be there for you guys whenever you need anything.

Finally, I would like to thank Ben Marlin for the preparation of this thesis, and Pierre-Jérôme Bergeron for the translation of the abstract.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 General medical setting:

In the study of the natural history of a disease, one often wants to make a statement about the survival from onset of an individual who acquires the disease. For example, one might want to estimate the mean or median survival time with this disease, or to estimate the probability of surviving longer than a certain amount of time with the disease. Whatever the case may be, a sample of diseased individuals is necessary in order to make some kind of inference about the condition. One way in which this sample can be obtained is by assembling a cohort of individuals and following them forward until some of these subjects acquire the disease under study. These incident cases are then followed for a further fixed time period and their survival times noted. This is termed an incident follow-up study as new cases are identified from onset as they occur (see Figure 1.1). However, this method of obtaining a sample often leads to practical difficulties. Firstly, a very large cohort may be needed to ensure a reasonable number of occurrences of the disease. That is, the cohort may need to be followed for a long period of time in order for a substantial number to develop the condition. Secondly, further long follow-up may be needed to ensure that a substantial number of these cases have progressed to "failure". Hence, a different sampling scheme may be necessary whereby individuals who already have the disease of interest are identified at a certain point in time,

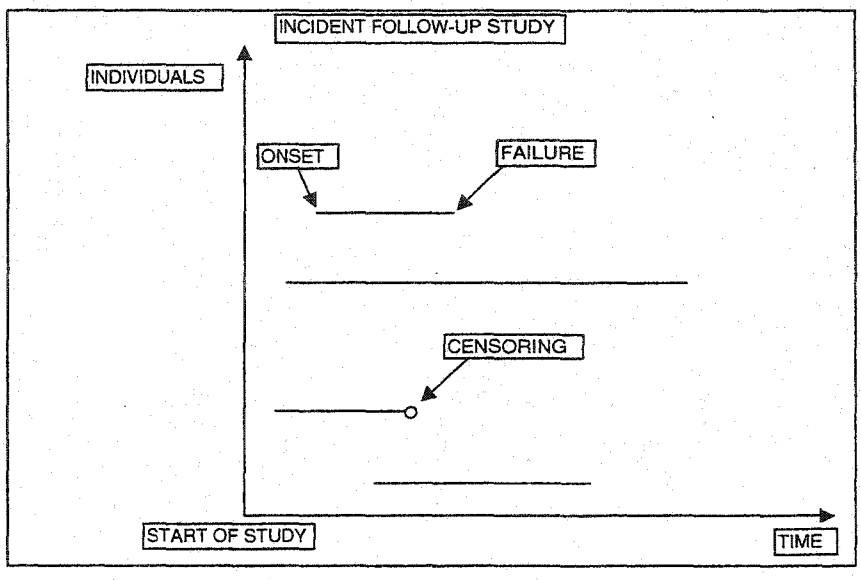


Figure 1.1: Incident follow-up study

and at this time their dates of onset ascertained, if this is feasible. These individuals can then be followed until death (or another end point of interest) or censoring. This is known as a prevalent follow-up study since the cases are initially identified as prevalent cases (see Figure 1.2). This second method of sampling alleviates the main practical difficulties associated with the incident sampling scheme. However, there are statistical difficulties induced by the prevalent sampling scheme which must be addressed. The survival times of individuals who are sampled in a prevalent cohort study are said to be length-biased. That is, the individuals included under this sampling scheme tend to have longer survival than those that would normally be included in an equivalent incident study. In a manner of speaking, the subjects who are sampled in a prevalent cohort study must survive long enough to be included in the study. Hence, observed survival from this sample will tend to be longer than would be obtained in an equivalent incident study (see Figure 1.3). One might term the observed survival times as length-biased since they describe a length-biased survivor function. As interest will always be in the unbiased survivor function adjustment for this bias must be made.

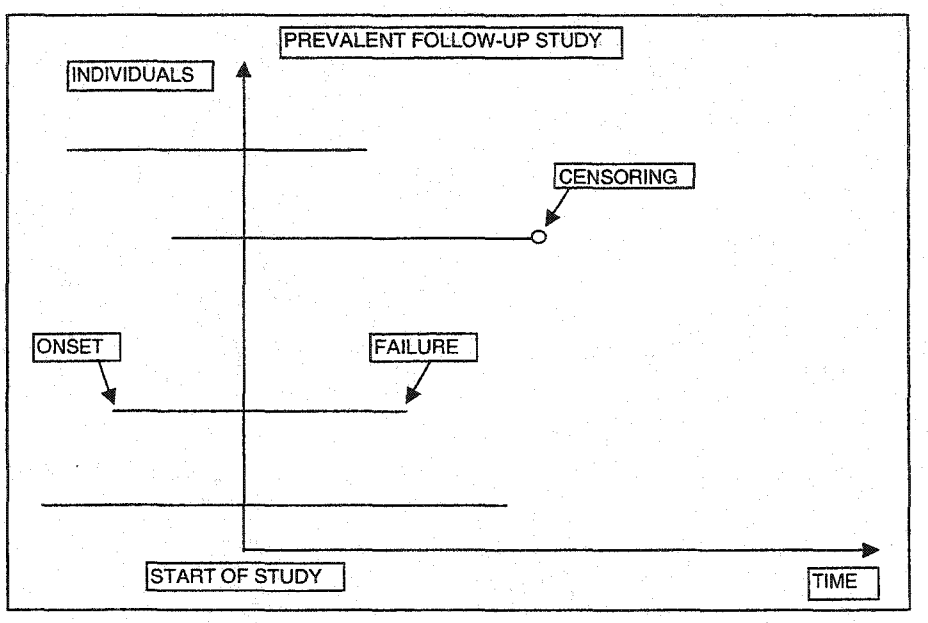


Figure 1.2: Prevalent follow-up study

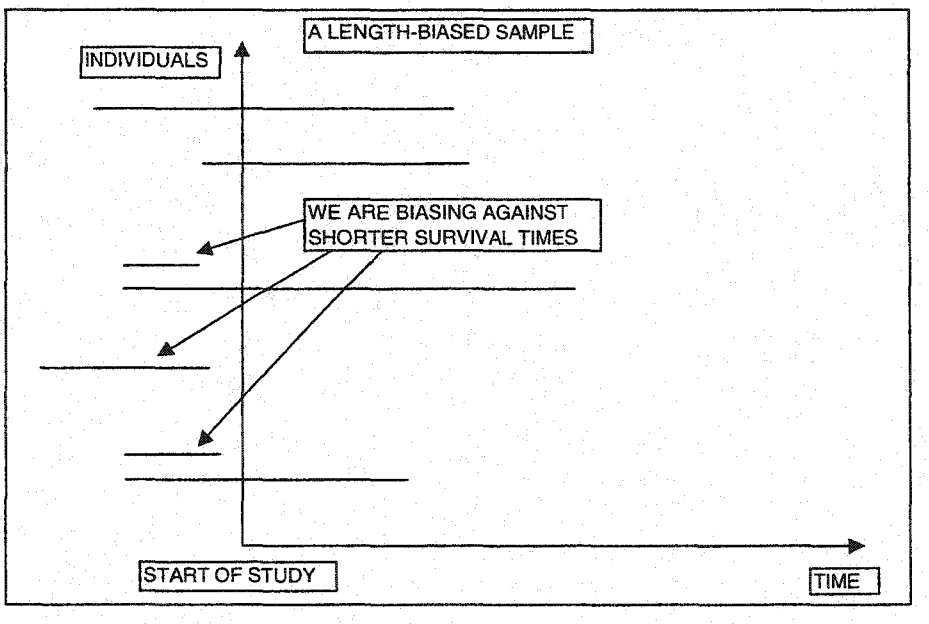


Figure 1.3: A length-biased sample

1.2 Historical work in the one sample problem:

The phenomenon of length-bias has been widely studied in the statistical literature. The manifestation of this phenomenon is not restricted to the medical field. It arises in economics, for example, when it is desired to estimate the mean duration, say, of welfare aid. Individuals who are currently on welfare at a given point in time are sampled and followed forward until they stop receiving aid. This is analogous to the situation described previously in that individuals who are already receiving welfare aid tend to be those who receive aid for longer periods of time. One of the original papers in this area, by Cox [7], discussed length-bias in the sampling of textile fibres. Cox mentioned that at the time much effort had been spent on issues other than the manner in which a sample was selected. This was because these methods were, for the most part, not general ones. However, he maintained that in order to obtain dependable and consistent results, clearly defined and well studied sampling techniques were necessary.

Now, the distribution of parallel fibre lengths in a piece of material can be seen to be analogous to the distribution of the survival times from onset of a disease of individuals in a population. The left end of the fibre would correspond to the onset of disease (initiating event) and the right end would represent the desired terminating event. Clearly, if all fibres, short and long, had equal chance of selection, this would give rise to an unbiased sample. However, one sampling method grips the material at a certain point, much in the way researchers might enter the population of diseased individuals at a certain point in time. All fibres which are not gripped are not selected, as are those people who have reached the terminating event before the study is commenced. Thus, fibres which are selected for the sample do not accurately represent the population of all fibres in the material. The probability of selecting any one fibre is proportional to its length, yielding a length-biased sample. Denoting $f(x)$ as the unbiased density of

fibre length in the population, and $g(x)$ as the length-biased density, the following relationship holds where μ is the mean of $f(x)$:

$$g(x) = \frac{xf(x)}{\mu} \tag{1.1}$$

Straight forward evaluation of the k^{th} moment of $g(x)$ yields (1.2) where μ_k is the k^{th} moment of $f(x)$:

$$E_g(X^k) = \frac{\mu_{k+1}}{\mu} \tag{1.2}$$

Setting $k = 1$ in (1.2) gives the mean of the length-biased density $g(x)$:

$$E_g(X) = \mu + \frac{\sigma^2}{\mu} \tag{1.3}$$

This is clearly larger than the mean μ of the unbiased density $f(x)$, which agrees with our intuition.

Cox developed some estimates of quantities relating to the unbiased distribution from the length-biased data and compared them to the analogous estimates that would be obtained from an unbiased sample. Even though Cox considered both parametric and nonparametric methods, the vital issue of censoring was not addressed. This is presumably because it was not an issue in the sampling of textile fibres. That is, the right end point of a fibre, which represents the "failure" or terminating event of interest, is always observed. Equivalently, it can be said that in the length-biased sample, the entire fibre is always observed, which is not the case with subjects afflicted with a certain disease. These individuals may be censored before they are observed to "fail". Censoring is an unavoidable feature in studies where there is follow-up of subjects and length of survival is of interest.

Blumenthal [3] considered slightly different questions in dealing with the study of electron tube life. A primarily parametric analysis was employed, focusing on

the gamma and Weibull distributions, to explore the utility of various sampling schemes in a renewal process setting. Although this setting is quite different from that of a prevalent follow-up study, there are distributional similarities between the two situations. Tubes in operation are identified at a point in time and thus do not form a random sample of all tubes. This is obviously analogous to the natural history of disease setting, but there are a few variations described by Blumenthal that deserve mention. Blumenthal explained three main ways of obtaining a sample of tubes. The first is the standard one in which the total length of operation is observed for the identified tubes. The second is one in which only the backward recurrence times are used from the identified tubes. Backward recurrence time refers to the time from start of operation to identification. This would be similar to using only the time from onset of disease to entry into the study. The third method is one in which every tube identified is replaced by a new tube and the time observed is the backward recurrence time of the identified tube plus the full lifetime of the new tube. An analogous procedure in a natural history of disease study would include in the sample an incident case for every prevalent case identified. The survival time noted would then be the time from onset to entry of the prevalent case plus the complete survival time of the incident case. In practice, this may not be realizable.

Blumenthal, like Cox, examined the efficiency of estimating the mean of the unbiased distribution using the length-biased sample or an unbiased sample. Their results which suggest an advantage of intentionally obtaining a length-biased sample are discussed in the "Closing Remarks" chapter of this thesis. Finally, Blumenthal also did not consider the issue of censoring. However, in the medical field, it is of fundamental importance and must be considered.

1.3 More recent work in the one sample problem:

Although the papers that have been discussed thus far are of historical interest, they have two features which are special cases of a broader area of study. Firstly, there is no censoring involved in the sampling of textile fibres. Blumenthal also did not examine this crucial practical complication. Moreover, the tacit assumption of stationarity was made both by Blumenthal and Cox, although for Blumenthal stationarity was assumed to mean stationarity of the underlying renewal process. In the medical setting, however, stationarity has to do with the incidence of disease over time, and means that disease occurrence is uniform over time before the cases are identified. For example, in an epidemic, this assumption would clearly be violated (see Figure 1.4). In the absence of this assumption, survival times that are said to be length-biased are often termed left-truncated. That is, length-bias is merely a special case of left-truncation with the additional assumption that the onset times are uniform over time. Thus, prevalent follow-up data that are frequently observed in natural history of disease studies are generally subject to left-truncation and right-censoring. The truncation time of a subject is defined to be the time from onset of disease to the start of the study. The term truncation refers to the fact that if an individual's survival time is shorter than the time from onset to potential entry, then it will be impossible for this individual to be part of the sample under study. Without the stationarity assumption, one is forced to condition on the observed truncation times (backward recurrence times). If one does not condition on the observed truncation times the model will be overparameterized and hence non-identifiable. For example, it is impossible to determine whether survival times are "long", say, because the true survival is "long", or because incidence of the disease is increasing close to the sampling point. A multitude of papers examining the area of left-truncation have been published [12] [14] [15] [22]

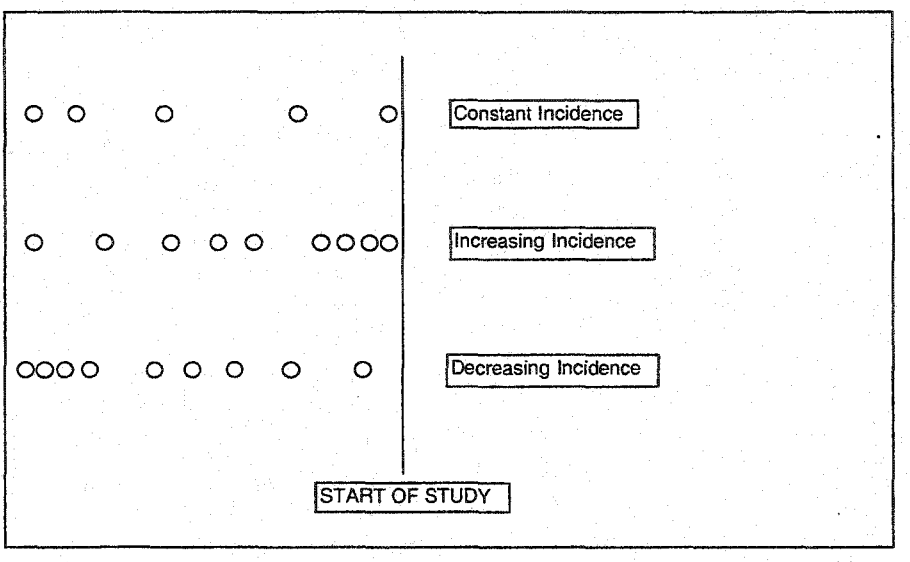


Figure 1.4: Incidence comparison

[23]. Among the results derived in the literature is the nonparametric maximum likelihood estimator (NPMLE) of the failure time distribution.

The paper by Wang [19] is central in the discussion of estimation from left-truncated and right-censored data. Prior to Wang's paper, the estimator which had been proposed for the failure time distribution of data subject to left-truncation and right-censoring had only been heuristically justified. This proposed estimator is analogous to the usual product-limit type estimator under right-censoring, except with modified risk sets. The risk set at an observed failure time includes only individuals who have not failed or been censored, but who are under active follow-up. In this paper, Wang justifies this estimator by showing that it is the NPMLE conditional on the observed truncation times, when all the potential censoring times are known. Thus, the estimator seems intuitively plausible even when the potential censoring times are unknown, as is the case in a majority of studies. Wang also conjectures that when stationarity does indeed hold, Vardi's unconditional estimator [17] has greater efficiency than the product-limit type NPMLE.

As Wang points out, there are many practical instances where the stationarity assumption is reasonable or is known to hold. In this thesis, we assume stationarity

for the simulations in Chapter 4 and hence Vardi's estimate [17] deserves further investigation. In a renewal process context, Vardi [16] derives the unconditional NPMLE of the failure time distribution from a mixture of prevalent and incident cases, assuming stationarity. Importantly, Vardi [16] also derives the asymptotic properties of the NPMLE. His results extend those of Cox [7] who assumed only prevalent cases are observed and who discussed only the pointwise asymptotic behaviour of $\widehat{S(t)}$, the NPMLE of the survivor function $S(t)$. Nevertheless, Vardi [16] does not consider censoring.

Vardi [17] shows how the unconditional NPMLE may be obtained via the EM algorithm, even when there is censoring but no asymptotics are presented. Recently, Asgharian et al [2] have derived the asymptotic properties of the unconditional NPMLE and confirmed, at least through an example, that it produces more efficient estimates than does the conditional NPMLE. Vardi's paper [17] describes a general model arising from data that are said to be multiplicatively censored. Such a model yields a likelihood that is proportional to the likelihood that arises from a prevalent follow-up study. A more detailed examination of the work by Wang and Vardi in this area is presented in Chapter 2.

1.4 The two sample problem:

At this stage, only difficulties arising from the use of length-biased data in the estimation of quantities pertaining to a single group have been examined. Nevertheless, the primary goal of a study will often be to compare the survival experience of two or more groups with a certain condition. The two sample problem will be the main focus of this thesis and an overview of the literature in this area is now given.

An important complication which has not yet been considered is that of unknown backward recurrence times. Although interest will almost always lie in the

time from an initiating event to a failure or terminating event, the time from the initiating event to entry into a cohort may not be known. For example, in the domain of infectious diseases, some studies are concerned with investigating the time from infection to onset of a disease. In this case, the "failure" event is, in fact, onset of the disease. However, the time from infection to entry into the study may not be known. That is, the time at which an individual was infected is sometimes not available in such studies. Alternatively, the instances of onset for diseases with insidious onset, such as Alzheimer's disease, may be difficult to determine. Thus, inference must sometimes be carried out on the times from entry to failure alone, the so called follow-up times. From here on, we refer to these analyses as taking place on the follow-up time scale.

Brookmeyer and Gail [4] discuss several issues which arise from such a complication in the context of infectious diseases. This paper explores the biases that may occur in the estimation of unbiased or incident quantities when the analysis is forced on the follow-up time scale. Although Brookmeyer and Gail examine these biases in the one sample problem, they focus primarily on the biases in the comparison of two groups through the assumption of Cox's proportional hazards model on the incident time-from-infection scale. Prior to commencing any such discussion, the authors raise the issue of onset confounding. Onset confounding refers to the confounding of the effect of a covariate on the relative risk with the effects of this covariate on the duration of infection. When onset confounding is present, no reliable inference can be made about the covariate of interest. The authors state the assumption which is necessary to ensure that onset confounding does not occur. Even when onset confounding is not present, biases may still occur when assessing covariate effects. The nature and extent of bias which may occur in the estimation of relative risk when using only the follow-up times is investigated. This is done for both fixed and binary time-varying covariates. A closer examina-

tion of the one and two sample sections of this paper will be given in Chapter 2 and Chapter 3 respectively.

In natural history of disease studies, the backward recurrence times of subjects are often known. Moreover, interest frequently still lies in the incident time scale in these studies. It is therefore natural to wonder whether it is possible to estimate the effects of covariates when the backward recurrence times are observed. Wang [20] proposes an approach for this problem when prevalent data are observed and the proportional hazards model is assumed on the incident time scale. In this paper, Wang assumes stationarity of the onset times, and thus her discussion is restricted to length-biased data, and not to the more general left-truncated data. Unfortunately, it is not possible to simply perform an analysis using the traditional partial likelihood argument proposed by Cox [8]. The difficulty lies in the prevalent sampling scheme which causes a bias in the risk sets if they are defined in the usual manner. Wang [20] samples from the traditional risk sets in order to remove the inherent bias within them. Wang then uses the newly created unbiased risk sets and forms a pseudo likelihood, which is maximized to obtain estimates of the regression coefficients. Since the unbiased risk sets are random subsets of the biased risk sets, Wang attempts to improve the efficiency of the estimates by repeating the procedure and taking the average of all the estimates. Unfortunately, a major weakness of this method is that it does not allow for censoring which is almost always present in any kind of follow-up study, including natural history studies. Thus, practical applications of this method are limited mainly to data that are size-biased and where censoring does not occur naturally.

In prevalent follow-up studies, the subjects are sometimes given a treatment for the condition at entry into the study. This type of study is called a prevalent treatment study. We may be interested is studying the effect of this treatment or of other covariates in prevalent treatment studies. Clearly, since we are entering

a population "cross-sectionally" in these studies, we are interested in the prevalent time scale. That is, we are interested in examining the effect of covariates on subjects who are prevalent with a condition and not on incident subjects. Hence, it is natural to assume a proportional hazards model on this prevalent time scale. Cnaan and Ryan [5] outline a modified proportional hazards analysis for the estimation of covariate effect on the prevalent time scale, in the presence of prevalent (left-truncated) data. That is, they propose a procedure analogous to the usual partial likelihood argument, only using adjusted risk sets. The risk sets are identical to the ones used in the one sample problem for the conditional NPMLE of the survivor function. Wang et al [21] give an in depth discussion of the underlying assumption required to carry out this analysis along with its severe limitations for practical purposes. The assumption essentially requires that any two subjects have identical hazards after their respective treatment, irrespective of when in the progression toward the terminating event the treatment was administered. It is possible to weaken the assumption needed to carry out the analysis, but in the end this alternative proves to be unsatisfactory as well. A more detailed inspection of this assumption along with the other work performed in the two sample problem will be provided in Chapter 3.

1.5 Organization of thesis:

In this thesis, the central point is that of comparing groups in the presence of length-biased data. Of course, before approaching this problem, the one sample problem must first be addressed. Chapter 2 will focus primarily on the work of Wang [19] and Vardi [17] in the one sample estimation problem. Wang [19] does not assume stationarity and thus, she necessarily conditions on the observed truncation times. On the other hand, Vardi [17] assumes stationarity and this allows for the development of a more efficient estimator. A detailed examination

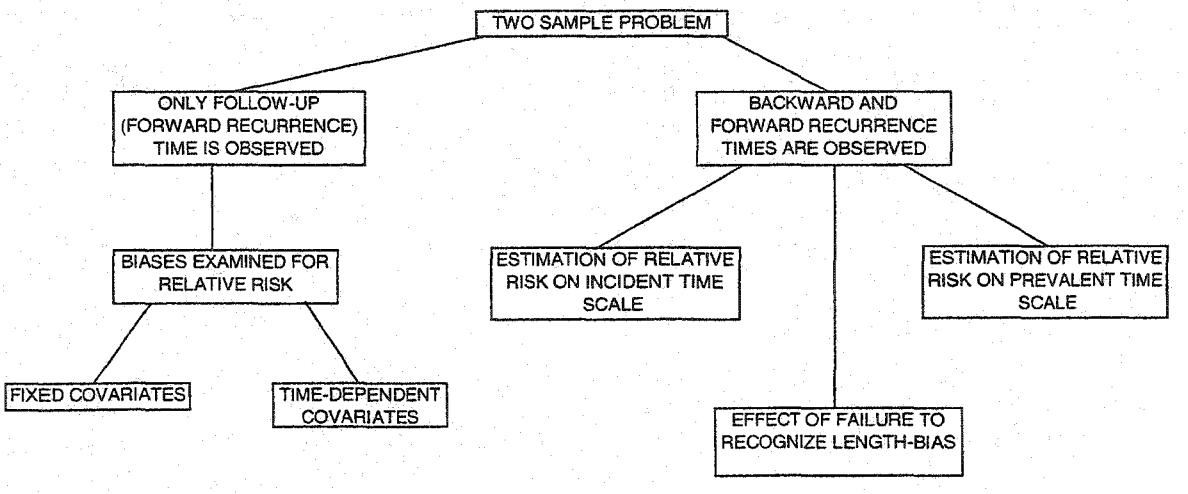


Figure 1.5: Outline of work in Chapter 3 and Chapter 4

of the work in the two sample problem will then be given. Chapter 3 will be based primarily on the papers by Brookmeyer and Gail [4], Wang et al [21] and Wang [20]. It will address the issues raised in the overview given in Section 1.4 more thoroughly. To continue on this theme, Chapter 4 will investigate other possible pitfalls associated with the comparison of two groups in the presence of length-biased data. Specifically, when using the mean or median of the groups as a basis for comparison, one must account for length-bias in order to avoid potentially incorrect inference about the two groups. Finally, Chapter 5 will touch upon some interesting points, including some brought up throughout the thesis.

It should be noted that much of the literature addressing the two sample problem is quite confusing. This is in part due to the inherent difficulty of the material, but also because the authors involved frequently do not refer to earlier work in the field. To aid the reader, we include a tree diagram that outlines the work done in Chapter 3 and Chapter 4 of this thesis (see Figure 1.5). **Moreover, we provide a glossary of terms used in this thesis, in Appendix A, for quick reference, and hopefully, to help the reader.**

Chapter 2

More detailed examination of the one sample problem

2.1 Work of Wang and Vardi assuming known backward recurrence times:

As mentioned, Wang's paper [19] constitutes an important component of any discussion of the one sample estimation problem in the presence of left-truncated and right-censored data. One contribution of this paper lies in its recognition of the product-limit type estimator as the conditional NPMLE in the case where all the potential censoring times are known. We now return to this paper for a closer inspection.

Let X denote the true failure time of an individual, with associated distribution function F and survivor function S. Let C be a subject's potential censoring time and let T be the truncation time of an individual. Denote the distribution of T by G. If Y is the observed event time, then $Y = min(X, C)$. Moreover, let δ be the censoring indicator, where $\delta = 1$ signifies a true failure and $\delta = 0$ means that an individual is censored. That is, let $\delta := I(X < C)$. In a prevalent cohort study, data on an individual is often of the type (T, Y, δ). In addition, these data have the implicit assumption that X and C are greater than T. Otherwise, the subject has not survived until the start of the study and obviously would not have

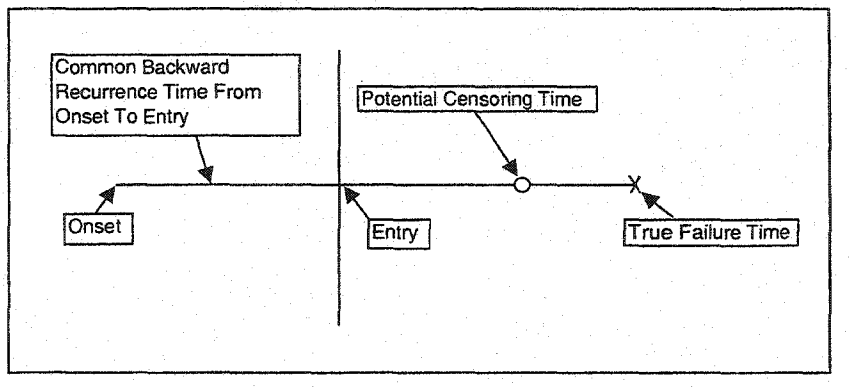


Figure 2.1: Informative censoring

been included in the sample. Wang assumes that X is independent of (T, C), which is important in the nonparametric estimation of F and G. The validity of this assumption is essentially justified as a combination of the assumptions of independence between X and T and between X and C. The former follows from the assumption of independence between the failure times and the calendar times of onset. This assumption may be violated with advances in treatment, say, but often seems reasonable. The assumption of independence between X and C is as in the random censorship model. Even though this assumption is made, the observed length-biased event time $Y = min(X, C)$ is not independent of the censoring time, C, since they share a common backward recurrence time (see Figure 2.1). Thus, informative censoring is, in fact, present and the usual asymptotic results that hold under the assumption of independent censoring do not necessarily hold here.

Although a product-limit type estimator had been proposed for left-truncated and right-censored data, it had only been heuristically justified prior to Wang's paper [19]. This estimator is analogous to the usual product-limit estimator under right-censoring, except with modified risk sets. The risk set at an observed failure time includes only individuals who have not failed or been censored, and who are under active follow-up. The estimator is shown below:

Suppose (t_i, y_i, δ_i) for $i = 1, ..., n$ are observed. Let $y_{(1)}, ..., y_{(k)}$ be the distinct

ordered event times from the uncensored y's (i.e. the ordered true failure times which have been observed). Then,

$$\widehat{S}(u) = \begin{cases} 1 & \text{if } u < \min(y_i : \delta_i = 1) \\ \prod_{y_{(i)} \leq u} \left(1 - \frac{card\, \epsilon_i}{card\, A_i}\right) & \text{Otherwise} \end{cases} \tag{2.1}$$

where for $i = 1, ..., k$, $A_i = \{j : t_j \leq y_{(i)} \leq y_j\}$ is the modified risk set, and $\epsilon_i = \{j : y_j = y_{(i)}\}$ is the number of failures at a particular failure time.

Wang proceeds in the following fashion. The full likelihood L is written as a product of two functions, L_1 and L_2. $L = L1 \cdot L2$ where,

$$L_1 = \prod_i \frac{dF(y_i)^{\delta_i} S(y_i)^{1-\delta_i}}{S(t_i)} \tag{2.2}$$

$$L_2 = \prod_i \left[S(t_i) dH(t_i, y_i)^{1-\delta_i} \left(\int_{y_i}^{\infty} h(t_i, u) du \right)^{\delta_i} \left(\frac{1}{\beta} \right) \right]$$

$$\beta = P(X \geq T) = \int S(u) dG(u) \tag{2.3}$$

and H is the joint truncation and censoring distribution with associated density h.

The function L_1 is then maximized nonparametrically to obtain the product-limit type estimator for S shown in (2.1). However, L_1 may not always be a conditional likelihood conditional on the observed truncation times. Hence, this method requires justification since an estimate is being obtained simply by maximizing some function of the quantity of interest which is neither a conditional nor a full likelihood.

Suppose now that the data observed are not of the form (T, Y, δ), but are instead of the form (T, Y, C) for every subject. That is, (t_i, y_i, c_i) for $i = 1, ..., n$ are observed. This data set is similar to the previous one, but contains more information since the potential censoring times are known. In the case where Y does in fact represent a true failure time, the observation of C will allow one to still

know the potential censoring time of that individual. One scenario where this type of data arises is in a delayed entry study which terminates at some fixed time and in which no subjects are lost to follow-up. In this case, although the censoring times are a priori random because of the random entry times, the potential censoring times become fixed constants by conditioning on these entry times (see Fleming and Harrington p.100 [11]). Assuming that such data can be collected, Wang writes the likelihood as the product of the conditional likelihood,

$$L_c = \prod_i \frac{dF(y_i)^{I(y_i<c_i)} S(y_i)^{I(y_i=c_i)}}{S(t_i)} \tag{2.4}$$

conditional on the (t_i, c_i)'s, and the marginal likelihood of the (t_i, c_i)'s,

$$L_m = \prod_i \frac{S(t_i) dH(t_i, c_i)}{\beta}$$

It is clear that L_c in (2.4) is identical to L_1 in (2.2) since $\delta := I(X < C) \equiv I(Y < C)$ and $(1-\delta) := I(X \geq C) \equiv I(Y = C)$. Thus, Wang demonstrates that L_1 is indeed the conditional likelihood when all the potential censoring times are known. Hence, in this special case, Wang [19] shows that the product-limit type estimator is the NPMLE conditional on the observed truncation times. For this reason, the estimator seems intuitively plausible even when the potential censoring times are unknown.

It is interesting to observe that this product-limit type estimator may give poor results near 0, as will be discussed in the "Closing Remarks" chapter of this thesis. Wang also speculates that when stationarity does in fact hold, Vardi's unconditional estimator [17] has greater efficiency than the product-limit type NPMLE. That is, if the truncation time distribution, G, is uniform then one may prefer Vardi's estimator. It is worth noting that since Wang [19] shows a method in which G may be estimated, it is possible to check the assumption of stationarity. Aside from this, estimation of G can be useful in other ways. For example, one may want to estimate the number of individuals who are being truncated in a study.

By estimating G, one can obtain an estimate of β (the proportion of untruncated data) by simply substituting $\hat{S}$ and $\hat{G}$ into (2.3):

$$\hat{\beta} = \hat{P}(X \geq T) = \int \hat{S}(u) d\hat{G}(u)$$

An estimate of the number of truncated individuals for every observed or untruncated individual (i.e. the odds of truncation) can then be obtained by $(1-\hat{\beta})/\hat{\beta}$. Thus, multiplying this odds by the number of subjects in the study gives an estimate of the total number of individuals who were truncated.

As Wang [19] conjectured and Asgharian et al [2] demonstrated via an example, if one is prepared to assume stationarity, Vardi's unconditional NPMLE [17] will be more efficient than the conditional NPMLE. Furthermore, stationarity is a reasonable assumption in many practical circumstances and will be assumed for the simulations in Chapter 4 of this thesis. Therefore, Vardi's paper [17] deserves closer inspection and its details will now be discussed.

Suppose $X_1, ..., X_m$ and $Z_1, ..., Z_n$ are i.i.d. random variables from some distribution F^L. $X_1, ..., X_m$ are fully observed while $Z_1, ..., Z_n$ are censored in the following manner.

Let $U_1, ..., U_n$ be i.i.d Uniform(0,1) random variables independent of $X_1, ..., X_m$, $Z_1, ..., Z_n$. The Z_i's are said to be multiplicatively censored upon multiplication by the U_i's, to yield $Y_1, ..., Y_n$, if $Y_i = Z_i U_i$, for $i = 1, ..., n$.

Of course, if we are assuming stationarity and if F^L itself is a length-biased distribution, then the data $X_1, ..., X_m$, $Y_1, ..., Y_n$ are subject to length-bias and multiplicative censoring.

Vardi was concerned with estimation of the distribution function F^L. Let Y, U, and Z represent any of the above Y_i's, U_i's, and Z_i's respectively, then since $Y = UZ$ we have,

$$
\begin{aligned}
F_Y(y) &= P(UZ \leq y) \\
&= \int_0^\infty P(UZ \leq y | Z = z) F^L(dz) \text{ (By the law of total probability)} \\
&= \int_0^\infty P(U \leq \frac{y}{z} | Z = z) F^L(dz) \\
&= \int_0^\infty P(U \leq \frac{y}{z}) F^L(dz) \text{ (By independence of the } U's \text{ and } Z's) \\
&= \int_0^y P(U \leq \frac{y}{z}) F^L(dz) + \int_y^\infty P(U \leq \frac{y}{z}) F^L(dz) \\
&= \int_0^y 1 \cdot F^L(dz) + \int_y^\infty \frac{y}{z} F^L(dz) \\
&= F^L(y) + y \int_y^\infty \frac{1}{z} F^L(dz) \qquad (2.5) \\
\therefore f_Y(y) &= \frac{d}{dy} \int_0^y F^L(dz) + \frac{d}{dy} \int_y^\infty \frac{y}{z} F^L(dz) \\
&= F^L(dy) + y \left(-\frac{1}{y} F^L(dy) \right) + \int_y^\infty \frac{1}{z} F^L(dz) \\
&= \int_y^\infty \frac{1}{z} F^L(dz) \qquad (2.6)
\end{aligned}
$$

which implies that the full likelihood for $(X_1, ..., X_m, Y_1, ..., Y_n)$ is:

$$
L(F^L) = \prod_{i=1}^m F^L(dx_i) \prod_{i=1}^n \int_{z \geq y_i} \frac{1}{z} F^L(dz) \qquad (2.7)
$$

In a nonparametric setting, $L(F^L)$ must be maximized with respect to F^L in order to obtain the NPMLE of F^L. If a parametric analysis is desired, one need simply replace F^L by the parametric distribution of interest and perform the maximization over the parameters of this distribution.

From a nonparametric viewpoint, it can be shown that only discrete distributions need to be considered as possible maximizers of (2.7). Let $A := \{x_1, ..., x_m, y_1, ..., y_n\}$, and consider a set E to which a potential maximizer of (2.7) assigns mass. If E is comprised of a disjoint union of intervals and singletons then the following argument can be used on each of these individually. Hence, we can assume that E is simply one interval or singleton. If E is to the left of the smallest value in A,

then by shifting the mass to this smallest value, the likelihood is increased. If the smallest value in A is one of the x's then the result is clear since the first product in the likelihood is of the $F^L(dx_i)$'s. Thus, giving more mass to one of the x's will increase the likelihood. If the smallest value in A is one of the y's, then the likelihood is still increased since $F^L(dy)$ will be greater for this y and the corresponding integral increases since the region of integration includes y. Otherwise, if E is somewhere in between two observations in A, or to the right of the largest value in A, then the likelihood can be increased by redistributing the mass to the nearest point from A to the left of E. Again, if the nearest point to the left is an x then the preceding argument still holds. If this nearest point to the left is a y, then the corresponding integral will be increased since y is less than any element in E and hence $1/y$ is bigger than $1/x^*$ for any $x^* \in E$.

Let $0 < t_1 < ... < t_h$ denote the distinct values of $x_1, ..., x_m, y_1, ..., y_n$ where $h \leq n + m$ ($h = n + m$ if the underlying distribution is continuous).

Maximizing $L(F^L)$ over the space of all distribution functions F^L, therefore, reduces to maximizing $L(\underset{\sim}{p})$ over the space of discrete probability functions $\underset{\sim}{p}$ which only assign mass to the observed values $t_1, ..., t_h$. That is, we must maximize:

$$L(\underset{\sim}{p}) = \prod_{j=1}^{h} p_j^{\epsilon_j} \left(\sum_{k=j}^{h} \frac{1}{t_k} p_k \right)^{\eta_j} \tag{2.8}$$

where ϵ_j and η_j are the multiplicities of the x's and y's respectively, $j = 1, ..., h$, and (2.8) must be maximized subject to: (i) $p_j \geq 0$ for $j = 1, ..., h$ and (ii) $\sum_{j=1}^{h} p_j = 1$ where $\underset{\sim}{p} = (p_1, ..., p_h)$ and $p_j \equiv P(t_j) \equiv F^L(dt_j) \equiv$ the "jump" or mass at t_j.

Vardi [17] chooses to use an iterative estimate based on the EM algorithm to maximize (2.8), although other maximization methods can be used as well. Specifically, it is the p_j's which are simultaneously estimated by the EM algorithm. Had the complete data $X_1, ..., X_m$, $Z_1, ..., Z_n$ been observed then the NPMLE of F^L would simply be the empirical distribution function. The EM algorithm is

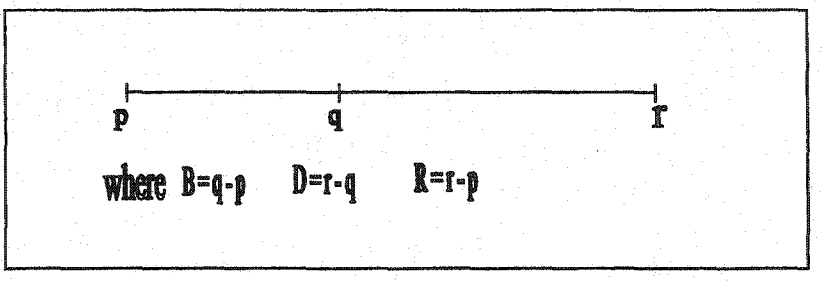


Figure 2.2: Backward and forward recurrence times

classically used for missing or incomplete data problems. Here, the incomplete data can be thought of as being the multiplicatively censored Z_i's, or alternatively, the U_i's could be viewed as the missing data. This interpretation provides an intuitive justification for using the EM algorithm. In this manner, Vardi estimates p_j for $j = 1, ..., h$, and thus obtains $\widehat{F^L}$, the NPMLE of F^L, as:

$$\begin{aligned} \widehat{F^L}(t) &:= \sum \hat{P}(t_i)\, I(t_i \leq t) \\ \text{or } \widehat{F^L}(t) &:= \sum_{\{t_i : t_i \leq t\}} \hat{P}(t_i) \end{aligned} \tag{2.9}$$

Vardi also gives an outline of the proof which illustrates the consistency of this estimator. It will now be shown that the estimator in (2.9) can also be used in the situation of interest in this thesis.

Let B denote the backward recurrence time, let D denote the forward recurrence time and let $R = B + D$ in a renewal process setting (as seen in Figure 2.2). The data described by Vardi [17] would correspond to having observed n values of B (previously labeled as Y_i's) and m values of R. In the medical setting, this would be analogous to identifying n+m prevalent cases at a point in time, censoring n of them immediately and deciding to follow the other m until failure. In practice, this does not seem feasible, nor is it desirable. Hence, it must be acknowledged that this model is not the same as that which is of primary interest in this thesis although multiplicative censoring does induce a form of informative censoring. It is also clear that the number of censored observations is fixed a priori in the

multiplicative censoring model. This is not the case in a natural history of disease study. However, as Vardi shows, the likelihoods which arise from the two different scenarios are proportional to one another, as we now proceed to illustrate.

Vardi [17] compares the multiplicatively censored data likelihood to the one obtained from $n+m$ values of R where m of them are fully observed and n are censored. Conveniently, the distribution of R is identical to that of the failure time distribution in a prevalent cohort study. The likelihood for the latter case is:

$$L(F) = \prod_{i=1}^{m} \frac{f(x_i)}{\mu_F} \prod_{i=1}^{n} \left[\frac{1 - F(y_i)}{\mu_F} \right] \tag{2.10}$$

where f and F are respectively the density and distribution of the failure times in a prevalent cohort.

But in (2.10) the x's are treated as constants, thus multiplying (2.10) by $(\prod_{i=1}^{m} x_i)$ will not change where the maximum is attained. Doing this yields:

$$L^*(F) = \prod_{i=1}^{m} \frac{x_i f(x_i)}{\mu_F} \prod_{i=1}^{n} \left[\frac{1 - F(y_i)}{\mu_F} \right] \tag{2.11}$$

Now, as mentioned, if F^L is itself taken to be length-biased then we have left-truncated data and:

$$F^L(dx) = \frac{x f(x) dx}{\mu_F} \tag{2.12}$$

$$\text{Also,} \quad \frac{1 - F(y)}{\mu_F} = \int_{z \geq y} \frac{1}{z} F^L(dz) \quad \text{where } y > 0 \tag{2.13}$$

Substituting (2.12) and (2.13) into (2.11) gives:

$$\prod_{i=1}^{m} F^L(dx_i) \prod_{i=1}^{n} \int_{z \geq y_i} \frac{1}{z} F^L(dz)$$

which is identical to the likelihood in (2.7).

Since F^L is being considered as the length-biased distribution, and interest in most studies lies in the unbiased distribution, F, it is also interesting to note that (2.12) and (2.13) suggest how to transform $\widehat{F^L}$ to $\widehat{F}$.

$$\text{From (2.12)}, \quad \widehat{F}(dx) = \left[\frac{\widehat{F^L}(dx)}{x}\right]\widehat{\mu_F} \tag{2.14}$$

$$\text{Further, from (2.13)}, \quad \widehat{\mu_F} = \frac{1}{\int_0^\infty \frac{\widehat{F^L}(du)}{u}} \tag{2.15}$$

$$\text{So that ,} \quad \widehat{F}(dx) = \left[\frac{\widehat{F^L}(dx)}{x}\right] \Big/ \int_0^\infty \frac{\widehat{F^L}(du)}{u} \tag{2.16}$$

Now, for the medical setting the values of n and m are not fixed prior to the start of the study. Nevertheless, it is clear that for any specific data set n and m take on fixed values. Hence, Vardi's unconditional NPMLE [17] can still be used with the usual prevalent cohort data of interest here. However, in order to derive the asymptotic properties of this NPMLE in the prevalent cohort setting, the argument provided by Vardi and Zhang [18] for multiplicative censoring does not suffice. Vardi [17] makes the point that the sampling properties depend on the sampling scheme that leads to the likelihood. Indeed, in multiplicative censoring, n and m are fixed a priori. As mentioned earlier, Asgharian et al [2] derived the asymptotic properties of the unconditional NPMLE for the situation of primary interest in this thesis.

2.2 Possible biases when backward recurrence times are unknown:

As we mentioned earlier, in some studies the backward recurrence time of an individual may be unknown due to the unknown calendar time of the initiating event. Although prevalent data are still present, the methods previously used are not applicable because the backward recurrence times are unknown. In such studies, analyses are thus performed on the follow-up times alone. That is, analyses are forced on this follow-up time scale, even though one is still usually interested in making a statement about survival on the incident time scale. We now return

to the paper by Brookmeyer and Gail [4] which was briefly mentioned in Chapter 1. This paper discusses the biases which may arise in the one sample problem when the backward recurrence times are not known in the context of an infectious disease. It is important to note that in this paper the time to infection and the time from infection to onset are assumed independent.

For a study that begins with entry into a population at time Y, let $S(t)$ denote the survivor function on the unbiased or incident time scale and let $S^*(t)$ be the survivor function on the follow-up time scale. That is, $S(t)$ gives the probability of surviving more than t units from infection for an incident case, where $S^*(t)$ gives the probability of surviving more than t units from Y, given that a subject is prevalent at time Y. Moreover, define $I(s)$ as the density of prior infection times or the epidemic density for s $\in (-\infty, Y]$. An expression for $S^*(t)$ in terms of $S(t)$ and $I(s)$ can be obtained. A development of this relationship is given below. It relies on the fact that if an individual is to have onset (the failure event) at calendar time $Y + t$, then infection must have occurred at some time s before Y and that subject must have been prevalent at time Y.

Let T^* = survival time from entry, and T = survival time from infection, S. Then $T = T^* + Y - S$,

$$\begin{aligned} S^*(t) &= P(T^* > t | T^* > 0) = \frac{P(T^* > t \cap T^* > 0)}{P(T^* > 0)} = \frac{P(T^* > t)}{P(T^* > 0)} \\ &= \frac{\int_{-\infty}^{Y} P(T > Y - s + t \cap S = s)ds}{\int_{-\infty}^{Y} P(T > Y - s \cap S = s)ds} = \frac{\int_{-\infty}^{Y} P(T > Y - s + t | S = s)I(s)ds}{\int_{-\infty}^{Y} P(T > Y - s | S = s)I(s)ds} \\ &= \frac{\int_{-\infty}^{Y} P(T > Y - s + t)I(s)ds}{\int_{-\infty}^{Y} P(T > Y - s)I(s)ds} \text{ (By the assumed independence of S and T)} \end{aligned} \tag{2.17}$$

For arbitrary $I(s)$, it is also true that if the hazard of failure is constant then, $S^*(t) \equiv S(t)$. Also, for an increasing hazard $S^*(t) < S(t)$, and if the hazard is decreasing then, $S^*(t) > S(t)$. These results are quite intuitive since for an

increasing hazard, say, prevalent individuals are at an increased risk in comparison to incident individuals. A similar argument holds for a decreasing hazard. A constant hazard corresponds to the exponential failure time distribution, and the memoryless property ensures that prevalent and incident cases are at equal risk of failure. However, if the hazard is not strictly monotone, no general conclusion can be arrived at concerning the direction of this bias. Therefore, one must be careful in the reporting of findings from analyses performed on the follow-up time scale.

Chapter 3

The two sample problem

In Chapter 2, we investigated the phenomenon of left-truncation where all the individuals in a cohort come from a single group. The main objective of a study in the medical setting will often be to compare the survival experience of two or more groups. Here, and in the next chapter, the focus will be on the situation where there are only two groups under study. For practical reasons which were discussed in Chapter 1, the comparison of these two groups is often carried out through the observation of prevalent cases. The discussion on group comparison with these prevalent (left-truncated) data will be divided into two main branches. The first is when the calendar time of the initiating event is not known. That is, the backward recurrence times of the prevalent subjects remain unknown. The second is simply when both the backward recurrence time and follow-up (forward recurrence time) are observed (see Figure 3.1).

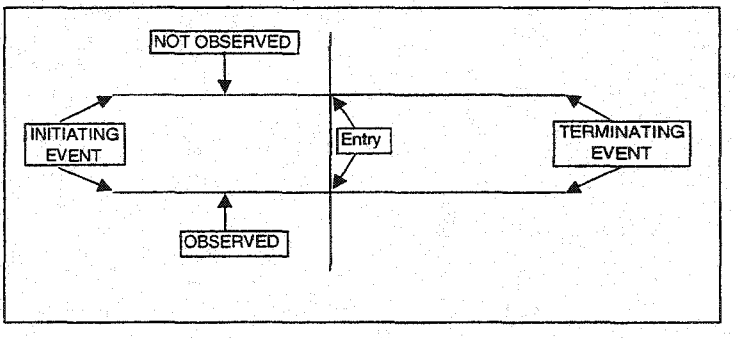


Figure 3.1: Unobserved and observed backward recurrence times

3.1 Biases associated with inference based on forward recurrence times:

In this section, we shall restrict our attention to the situation where only the forward recurrence times are observed, as a continuation of the one sample problem we discussed at the end of the Chapter 2, where the backward recurrence times of the subjects are not known. In fact, Brookmeyer and Gail's main objective [4] was to examine the biases associated with the added complexity introduced by the two sample problem. We now turn to this issue.

This section will be based on Brookmeyer and Gail's paper [4] which examines the possible biases arising from the use of forward recurrence times alone when interest lies in the underlying incident time scale. Much of the early work in survival analysis in a medical setting that addressed the issue of inference with unknown backward recurrence times was motivated by data collected early in the AIDS epidemic. More specifically, researchers were frequently concerned with the latent period between HIV-infection and onset of AIDS. In this setting, the initiating event was HIV-infection, whose calendar time was, generally, unknown, and the terminating event was onset of AIDS. For the remainder of this section, the terms infection and onset will thus be used instead of initiating and terminating event, respectively. Of course, the results of this section can be generalized to other situations as one can easily imagine various initiating and terminating events. For the most part, only heuristic arguments will be presented; the mathematical details are developed in the appendix of Brookmeyer and Gail [4].

3.1.1 Fixed covariates:

For simplicity of exposition, a binary fixed covariate, $Z = 0, 1$, say, is considered first. Let f and h represent the density and hazard function, respectively, on the incident time scale and let f^* and h^* represent the density and hazard on the

follow-up time scale. Also, let θ and θ^* denote the relative risks on the incident and follow-up time scale respectively. We note here, that without further assumption both θ and θ^* will, in general, be time dependent. Furthermore, for all quantities introduced here and at the end of Chapter 2, let the subscripts 0 and 1 represent the two levels of the covariate Z. In this section, we shall further assume Cox's proportional hazards model for the incident time scale. That is, we assume that $h_1(u) = \theta h_0(u)$, where u is the time from infection and now θ is, by assumption, independent of u. Since the backward recurence times are unknown, however, biases may be anticipated for the parameter, θ, of Cox's model if only the follow-up times are used and the analysis is carried out on this time scale. We proceed to examine these biases.

The hazard on the follow-up time scale, h^*, is expressed below, and can be developed following a similar argument to the one given for $S^*(t)$ at the end of Chapter 2.

$$\begin{aligned} h_Z^*(t) &= \frac{f_Z^*(t)}{S_Z^*(t)} = \frac{\int_{-\infty}^{Y} f_Z(t+Y-s|S=s)I_Z(s)ds}{\int_{-\infty}^{Y} S_Z(t+Y-s|S=s)I_Z(s)ds} \\ &= \frac{\int_{-\infty}^{Y} f_Z(t+Y-s)I_Z(s)ds}{\int_{-\infty}^{Y} S_Z(t+Y-s)I_Z(s)ds}, \; Z=0,1 \qquad (3.1) \\ &\quad \text{(By the assumed independence of S and T)} \end{aligned}$$

Since $I_Z(s)$ is usually unknown for $Z = 0, 1$, unless it is assumed that $I_0(S) \equiv I_1(S)$, "onset confounding" may occur and no dependable inference can be made about θ. The following example illustrates this point.

Example 3.1: Suppose that we wish to compare the relative risk, θ, of developing a disease for two groups on the incident time scale. We assume that the baseline hazard, h_0, is increasing. Further, suppose that one group is systematically infected before the other. An expression for $\theta^*(t)$ can be written as:

$$\theta^*(t) = \frac{h_1^*(t)}{h_0^*(t)} = \frac{h_1(Y - s_1 + t)}{h_0(Y - s_0 + t)} \quad \forall\, t \geq 0 \tag{3.2}$$

where s_0 and s_1 are the calendar times of infection of the two groups and $s_1 < s_0$, say. This, of course, assumes that everyone in each group is infected at the same time. Here, onset confounding would occur since the backward recurrence times are unknown and the two groups appear to have different risks simply because they were infected at different times. That is, since h_0 is increasing, we have:

$$\theta^*(t) = \theta \left[\frac{h_0(Y - s_1 + t)}{h_0(Y - s_0 + t)}\right] > \theta \quad \forall\, t \geq 0 \tag{3.3}$$

The magnitude of θ and $\left[\frac{h_0(Y-s_1+t)}{h_0(Y-s_0+t)}\right]$ are confounded when assessing the magnitude of $\theta^*(t)$.

When the effect of Z on θ is confounded with the effects of Z on the duration of infection, we call this onset confounding. More generally, onset confounding will occur whenever $I_0(S) \not\equiv I_1(S)$. We therefore avoid this difficulty by assuming $I_0(S) \equiv I_1(S)$. We recognize, however, that estimates of $\theta^*(t)$ may still be biased for θ even when onset confounding is not present.

Result 1: *For a "true" risk factor ($\theta > 1$), if h_0 is strictly increasing, then $\theta^*(t) \leq \theta \; \forall\, t \geq 0$.*

We may argue heuristically as follows: Suppose that individuals who are unexposed to the covariate Z (i.e. $Z = 0$) are at reduced risk in comparison to those who are exposed to the covariate (i.e. $Z = 1$). Since the exposed group is at a higher risk, it is more likely that they will have experienced onset of AIDS (the failure event) before the time of entry into the study, Y, if their truncation times are long (i.e. if they were infected well before Y). These subjects would therefore be ineligible to enter the study at time Y, since only HIV+, non-AIDS subjects

would be of interest here. Therefore, at entry (Y), those with $Z = 0$ and with early infection will be similar in stage of progression to those with $Z = 1$, who must have been infected later; both will be at high risk of failure. Since we are unable to observe the time of infection, the similarity of the forward recurrence times alone will attenuate any between-group effect towards the null value $\theta = 1$. That is, $\theta^*(t) \leq \theta \; \forall \; t \geq 0$.

Result 2: *For a "true" risk factor ($\theta > 1$), if h_0 is strictly decreasing, then* $\theta^*(t) \geq \theta \; \forall \; t \geq 0$.

In the case of a strictly decreasing hazard, a similar argument gives that $\theta^*(t) \geq \theta \; \forall \; t \geq 0$. That is, there is a bias away from the null value $\theta = 1$.

3.1.2 Is it possible that for some t, $\theta^*(t) < 1$ while $\theta > 1$?:

These biases can never make a "true" risk factor appear protective (or make a protective factor appear to be a "true" risk factor). That is, if $\theta > 1$ ($\theta < 1$), then $\theta^*(t) \geq 1 \; \forall \; t \geq 0$ ($\theta^*(t) \leq 1 \; \forall \; t \geq 0$). These results hold irrespective of the epidemic density $I(s)$ [4].

The source of these biases is termed differential length-biased sampling. At Y, the forward recurrence times are sampled differentially from the two different backward recurrence time distributions ($Z = 0, Z = 1$). Cnaan and Ryan [5] obtain results which are almost identical to those of Brookmeyer and Gail [4] with regards to the biases which may occur when using estimates of $\theta^*(t)$ for θ. We now illustrate these points through an example on the natural history of dementia.

Example 3.2: Consider two groups of subjects with dementia, those with vascular dementia and those with probable Alzheimer's disease. Suppose the aim of the

study is to compare survival, from onset with dementia to death, between these two groups. For this example we take onset with dementia and death to be the initiating and terminating events respectively. We assume a proportional hazards model on the incident time scale and suppose that the hazard of dying is increasing. Further, suppose that those with vascular dementia are at a higher risk than those with probable Alzheimer's. Since these conditions have insidious onset, it can easily be imagined that the calendar times of onset could not be ascertained. Many of those with onset long before the start of the study will have died. However, this will occur more frequently in the vascular dementia group since they are at an increased risk. Since the hazard is increasing, those from the probable Alzheimer's group who make it into the study will have high risk at entry having lived for a long time, at the time of entry. Thus, the Alzheimer's group will seem to be at similar risk levels in comparison to the high risk vascular dementia group, causing a bias toward the null risk value of 1.

3.1.3 Can it ever happen that $\theta^*(t) = \theta \; \forall \; t \geq 0$?:

When the baseline hazard is constant, $\theta^*(t) = \theta \; \forall \; t \geq 0$. This follows from the forgetfulness property of the exponential distribution, as mentioned at the end of Chapter 2. There are two additional circumstances in which $\theta^*(t) \approx \theta \; \forall \; t \geq 0$, even though $\theta^*(t)$ is estimated from data collected on the follow-up time scale whereas the proportional hazards model is assumed on the incident time scale. The first arises when $I(s)$ is concentrated on a small interval, say, at the beginning of an epidemic. The backward recurrence times will be forced to be similar for the two groups, thus avoiding differential length-biased sampling. That is, if initiation takes place on a small interval then all the subjects, high and low risk, will start at essentially the same place. In the extreme case, where everyone starts at the exact same point, there is no bias at all and $\theta^*(t) = \theta \; \forall \; t \geq 0$. This can

be seen, formally, by letting $s_1 = s_0$ in (3.3). This is a special case of $I_0(S) \equiv I_1(S)$.

Example 3.3: Consider vCJD (new variant Creutzveldt Jacob disease) and suppose that exposure to BSE (mad cow disease) occurred over some small unknown time period. Interest lies in the time from exposure to BSE to the development of vCJD, the initiating and terminating events, respectively, in this example. Further, suppose that the two groups being considered are those who ate organ meat (e.g. brain) and those who did not. If eating organ meat is truly a risk factor this will be detected since individuals from both groups who are in the study will have essentially the same backward recurrence times. Hence, differences in their follow-up times will indicate a difference between the two groups.

Of course, it is difficult to imagine a situation where we can assert that $I(s)$ has mass only on a small interval, yet we do not know when in time this took place. Hence, in the situation of unknown backward recurrence times, this scenario hardly seems useful.

A second circumstance which would lead to $\theta^*(t) \approx \theta \; \forall \; t \geq 0$ is when the disease is very rare. Recall that in this section the terminating event is the occurrence of disease. Hence, even if one is infected before time Y, there is a very high probability that this individual will still be at risk of getting the disease at time $Y + t$. Thus, we would observe essentially all incident cases. Any disparity which exists between the two groups will thus be discernible. However, again, this second situation seems to have limited applicability as a rare disease will produce very few observed occurrences of disease onset (the failure event). Thus, inference for such a disease will be "low-powered".

For $\theta \geq 1$ (analogous results hold for $\theta < 1$)			
Covariate Type	Increasing Hazard	Decreasing Hazard	Constant Hazard
Fixed	$1 \leq \theta^*(t) \leq \theta$	$\theta^*(t) \geq \theta$	$\theta^*(t) = \theta$
Time-Dependent	$1 \leq \theta^*(t) \leq \theta$	$1 \leq \theta^*(t) \leq \theta$	$\theta^*(t) = \theta$

Table 3.1: Summary of biases when backward recurrence times are unknown

3.1.4 Time-dependent covariates:

Let $Z(t) = 0(1)$ if the covariate is absent (present) at time t, be a time-dependent covariate. While the results for such covariates are similar to those for fixed covariates, there are differences that must be discussed. This covariate could be, for example, a treatment given to subjects only after entry into the study. Assuming $I_0(S) \equiv I_1(S)$, we can see from Table 3.1 that for time-dependent covariates, estimates of $\theta^*(t)$ will always be biased toward unity, in contrast to the fixed covariate case. That is, in the case $\theta < 1$, $\theta \leq \theta^*(t) \leq 1 \ \ \forall\, t \geq 0$ for both increasing and decreasing hazards. The following example illustrates why time-dependent covariates induce slightly different biases from fixed covariates.

Example 3.4 (Refer to Figure 3.2): Suppose that for some infectious disease we are interested in studying the time from infection to disease onset. We enter a population of infected individuals at calendar time Y. We select a sample from those individuals who have not yet received a certain treatment. The treatment is then randomly assigned at time Y to some of the infected subjects. Now, suppose that this treatment is truly protective against development of the disease ($\theta < 1$), and that the baseline hazard of developing the disease is monotonically increasing after infection. We follow all of these individuals forward and note when they develop the disease under study. Since the hazard is increasing, those individuals who were infected long before Y will be at high risk, and, in fact, many of them will have already developed the disease, thus making them ineligible for the study.

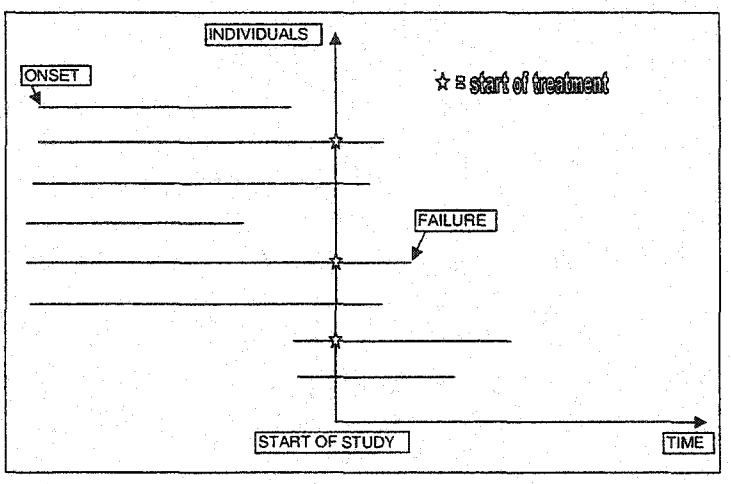


Figure 3.2: Frailty selection

These "deleted" individuals are said to be "frail" and a phenomenon known as "frailty selection" occurs. Thus, of those with infection times in the distant past, only the more robust ones will survive to the entry point Y. Some of these will eventually be treated, but even if the treatment is protective the survival experience of the treated and untreated subjects will appear to be similar, as the treatment is given to a robust group of subjects. Of course, for individuals who are infected close to Y, the treatment effect will be detectable. However, overall, there will be an attenuation of the treatment effect toward the null value of 1.

It is very important to note that if the hazard were decreasing, then the "frail" subjects would still tend to be depleted, the only difference being that these would be the ones with short truncation times. Hence, whether the hazard is strictly increasing or decreasing, the cohort will always have been depleted of "frail" individuals. The bias is thus always toward unity, unlike for fixed covariates where the direction of the bias depends on whether the hazard is strictly increasing or decreasing. An analogous result can be given for a true risk factor ($\theta > 1$). An example will be provided to demonstrate the situation in this case.

Example 3.5 (Refer to Figure 3.3): Suppose we are interested in a population

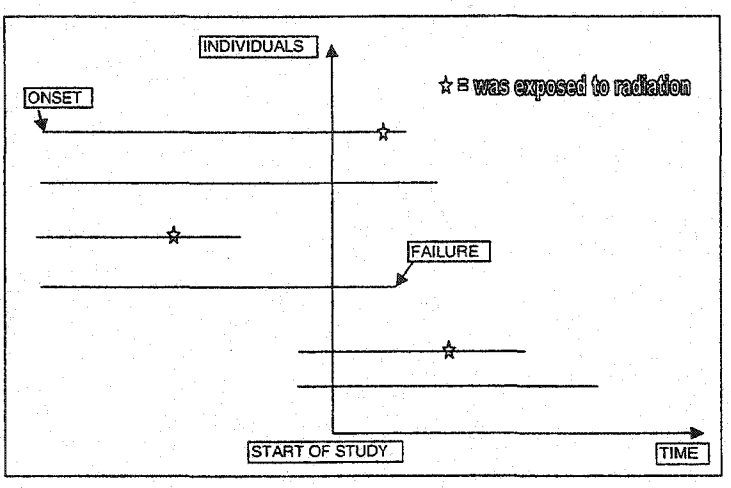


Figure 3.3: Attenuation of effect toward the null.

of adults who are prone to getting some form of leukemia. We assume that their hazard of getting leukemia is increasing from the time they enter this predisposed population, although we do not know the moment at which they enter it. Further, we speculate that exposure to radiation is a risk factor for getting leukemia. The two groups are those who are, and are not, exposed to radiation, respectively, in this population of prone individuals. We obtain a sample of such adults "cross-sectionally", at time Y, and follow them forward until some develop leukemia. Along the way, we note when, if ever, these subjects become exposed to radiation. If radiation is truly a risk factor, then this cohort will be depleted of subjects who were predisposed to getting leukemia long before Y, including, some who may have been exposed to radiation. This will make survival in the exposed and unexposed groups seem similar because many of the survivors will be resistant to leukemia. For subjects entering the predisposed population near Y, however, the detrimental effect of radiation will be more apparent. Nevertheless, overall, there will be an attenuation of the effect toward the null.

3.1.5 Testing:

In spite of the biases present when using estimates of $\theta^*(t)$ for θ, the two-sample nonparametric survival tests of $H_0 : \theta = 1$ are valid even when carried out on the follow-up time scale. This follows from the fact that $f_0^* = f_1^* \Leftrightarrow f_0 = f_1$, which can be observed from the numerator of (3.1). Thus, $\theta^*(t) \equiv 1 \Leftrightarrow \theta = 1$.

We have seen in the one sample problem that we can estimate the survivor function, either conditionally or unconditionally, by observing possibly right-censored, prevalent (left-truncated) data. In natural history studies, one often has access solely to such data and one is interested in examining not only survival from an initiating event, but also the effect of covariates on this time scale as well. Thus, it is of interest to investigate whether it is possible to estimate covariate effects by observing such prevalent data having observed both backward and (possibly censored) forward recurrence times.

3.2 Estimation when both backward and forward recurrence times are observed:

Since we are often interested in the effect of covariates on the incident time scale, it is therefore natural to assume a proportional hazards model on this time scale. However, often the only data that are available, have been obtained from the follow-up of prevalent cases. Moreover, the dates of initiation (and thus the backward recurrence times) are often obtained, at least approximately, when the prevalent cases are first identified.

This is conceivable even with the infectious disease scenario for AIDS considered in Section 3.1. For example, suppose we are studying hemophiliacs who were infected during a blood transfusion, then their infection times could be ascertained. That is, the dates of their transfusions could be obtained and their infection time could be deduced in this fashion. Another example is in the natural history study

of Alzheimer's disease to death. Although Alzheimer's has an insidious onset, many researchers approximate onset dates by close questioning of the caregivers of the patients. The patients are then followed until failure or censoring in the usual manner. Hence, an important question is whether covariate effect on the incident time scale can be estimated using left-truncated, right-censored data with known backward recurrence times.

At first glance, it seems as though Wang's paper [20] addresses this problem for the simplified case of length-biased data. We recall that length-biased data is merely left-truncated data under the stationarity assumption of the initiation times, which is the assumption of a uniform truncation distribution. This assumption is reasonable in many circumstances. Nevertheless, the method developed in Wang's paper has a major shortcoming for follow-up studies, which will be revealed shortly. For the moment we discuss, uncritically, Wang's approach. The possible biases in the estimation of the relative risk, θ, on the incident time scale when using follow-up data, excluding the backward recurrence times, were already considered in Section 3.1. Now, however, with the assumed availability of these augmented follow-up data, we proceed with an investigation of actual estimation in Cox's proportional hazards model defined, naturally, on the incident time scale. In summary, we are concerned with estimation for an unbiased model based on length-biased data.

Unfortunately, under the assumption of the incident proportional hazards model, the partial likelihood approach introduced by Cox [8] is not directly applicable for left-truncated data. For incident cases, the traditional risk sets would contain subjects who, after adjustment for their covariate effects, would have equal chances of failing at a particular failure time. This follows, since their hazard of failing would simply be the baseline hazard of failure. For left-truncated data, the risk sets defined in the usual manner no longer exhibit this property since the subjects are

sampled in a prevalent fashion. The consequence of this is that, even after removing covariate effects, two subjects who have survived until "t" will have different hazards of failing if they have different backward recurrence times; those with long backward recurrence times will, generally, be at greater risk to fail. However, in the special case of length-biased data, the explicit relationship (see (1.1)) between the unbiased and length-biased distributions allows us to sample cleverly from the observed "biased" risk sets thereby creating risk sets that mimic conventional unbiased risk sets. The sampling procedure which we will make explicit shortly, tends to exclude subjects in the risk set with longer survival times, as they are the source of the length-bias. Using these new unbiased risk sets, a pseudo-likelihood, $L^*(\beta)$, is formed and an estimate of β can be obtained by maximizing $L^*(\beta)$. We now provide a more formal mathematical development of the procedure.

Let T denote the unbiased failure time distribution for which the proportional hazards model is assumed. The hazard function is then

$$h(t;Z) = h_0(t)\exp(Z(t)\beta) \tag{3.4}$$

where h_0 is the unspecified baseline hazard function, and $Z(t)$ is a vector of possibly time-varying covariates. The term $\exp(Z(t)\beta)$ represents the relative risk (previously denoted by θ) and the unknown parameter of interest is now β.

Suppose $t_1, ..., t_n$ is observed, and let $R_i = \{j : t_j \geq t_i\}$ be the risk set at t_i for $i = 1, ..., n$. Let $t_{(1)}, ..., t_{(n)}$ be the order statistics, assuming no ties, with corresponding covariate vectors $Z_{(1)}, ..., Z_{(n)}$. Denote p_i as the density for the history $H_i = \{(Z_{(1)}, t_{(1)}), ..., (Z_{(i)}, t_{(i)})\}$ with $p_0 = p$. The full likelihood may be written as:

$$L = \left\{\prod_{i=1}^{n} p_{i-1}(Z_{(i)}|t_{(i)})\right\}\left\{\prod_{i=1}^{n} p_{i-1}(t_{(i)})\right\} \tag{3.5}$$

where the first term, L_p, is the partial likelihood.

For unbiased data, Cox [8] proposed that L_p be maximized in order to obtain an estimate of β, since in this case, L_p depends on the unknown parameters (β, h_0)

only through β. However, for length-biased data this method cannot be used since this partial likelihood depends on h_0 as well as β. As discussed, the bias is induced by the presence of survivors with length-biased survival times in the risk set R_i. That is, individuals in R_i no longer have an equal chance of failure at t_i after adjustment for $\exp(Z(t)\beta)$ since the hazard on the prevalent time scale no longer satisfies the proportional hazards assumption in (3.4). Hence, the unspecified baseline hazard, h_0, does not cancel when assessing the relative hazard of two individuals. Thus, h_0 remains in L_p as an unknown and this partial likelihood can no longer be used to estimate β without knowledge of h_0. To account for these biased risk sets we proceed as follows:

Define a random variable, $\delta_j(u)$, for $0 \leq u \leq t_j$ as follows:

$$\delta_j(u) = \begin{cases} 1 & \text{with probability } (u/t_j) \\ 0 & \text{with probability } (1 - u/t_j) \end{cases} \tag{3.6}$$

letting $\delta_j(t_i)$ for $i = 1, ..., n$ be independent.

Now, form a new risk set, $R_i^* = \{j : t_i \leq t_j \text{ and } \delta_j(t_i) = 1\}$ for $i = 1, ..., n$. It should be noted that since $\delta_j(t_i)$ is random, the size of R_i^* will also be random for $i = 1, ..., n$. R_i^* always includes an individual at his/her failure time. R_i^* is also more likely to include individuals with shorter failure times, since (t_i/t_j) is closer to 1 for t_j's which are closer to t_i, thereby correcting for the bias. Wang [20] demonstrates that by using these newly created risk sets, the risk set structure in the unbiased population is being artificially duplicated. That is, individuals in R_i^* have equal chance of failing at a particular failure time after adjustment for their relative hazard, even though these subjects were originally identified as prevalent cases. A familiar partial likelihood $L^*(\beta)$ can thus be defined as:

$$L^*(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\{Z_i(t)\beta\}}{\sum_{j \in R_i^*} \exp\{Z_j(t_i)\beta\}} \right] \tag{3.7}$$

This likelihood can only be considered as a pseudo-likelihood as it has been artificially created by a contrived random mechanism, following observation of the data. Nevertheless, proceeding formally, $\hat{\beta}$, can be obtained through maximization of $L^*(\beta)$. Importantly, this procedure is justified by Wang [20], who establishes asymptotic pseudo-properties of partial likelihood estimators, that mimic those of ordinary partial likelihood estimators.

Since the method just described requires a sampling of the biased risk sets R_i, one way of improving the estimator of β is by repeating the sampling procedure many times and using the average of all the $\hat{\beta}$'s. That is, if the procedure is repeated K times, then $\hat{\beta} := \sum_{i=1}^{k} \hat{\beta}_i/K$. It can be shown, however, that although the repetition process may reduce variation when the sample size is reasonably small, it does not aid the asymptotic efficiency for estimation of β.

In spite of the ingenuity of the method developed in Wang's paper [20], it has a major weakness for the analysis of follow-up data in that it does not allow for censoring. Whenever there is follow-up involved in a study, there will almost always be censoring. However, the problem is that the censoring accompanying follow-up studies is informative, as was discussed in Chapter 2, and the partial likelihood methods used in Cox's proportional hazards model breakdown when there is informative censoring. Thus, the estimation of covariate effects in natural history studies, when the data are left-truncated (even length-biased) and right-censored, remains an open problem for semi-parametric models. Of course, for fully parametric models there is no problem as the partial likelihood approach is not necessary; one can write down the full likelihood.

Wang recognizes the deficiency in the method as she gives an example and a simulation study where no follow-up time is required. Nevertheless, this method is quite limited as far as practical applications are concerned. A situation where this risk set sampling procedure can be used is described by Patil and Rao [13] and is

explained in Example 3.6.

Example 3.6: Suppose we wish to estimate the number of individuals in a group, say, for example, the number of albino children in a family which is prone to having albino children. One way of sampling such groups is to record the size of a group of individuals, only when at least one member of the group is sighted. That is, when an albino child is observed, the number of albino children in his/her family is recorded. Assuming that each child has an equal chance of selection, it is clear that families with a large number of albino children will be more likely to be observed than families with smaller numbers of albino children. Hence, a sample of group sizes representing the number albino children from families prone to having albino children obtained in this manner will be sized-biased (or length-biased). Now, if we assume that once an albino child is sighted we will be able to observe the entire group of albino children from that family, then there is no censoring involved. That is, under the assumption that the observed group sizes will be known exactly, Wang's method [20] will be applicable here since there is no censoring present. A related paper which may be of interest is by Davidov and Zelen [10].

Thus far, we have mainly been interested in natural history of disease studies for which the incident time scale is natural. That is, in spite of observing prevalent data, which is done out of practical necessity, our concern is with covariate effects for incident individuals. We now turn to another type of study which gathers prevalent data but for which inference on the prevalent time scale is the focus. This is a time scale for individuals who are identified in a cross-sectional study and who clearly do not correspond to a group of incident cases. In the next section, we discuss the difficulties involved in carrying out inference in such studies.

3.3 Prevalent treatment studies and estimation procedure on the prevalent time scale:

Since a treatment for a condition is often administered to individuals who already have the disease (prevalent cases), studies seeking to investigate the effectiveness of such treatments are often based on prevalent cases. Thus, prevalent cases are first identified, and if possible, their backward recurrence times ascertained, as was discussed in the previous section. The treatment is then administered to these prevalent individuals at entry into the study and they are followed forward until failure or censoring. This type of study is called a prevalent treatment study. A prevalent treatment study forces the extra requirement that a subject must not have received the treatment prior to entry, along with the usual condition that a subject has experienced initiation but has not yet experienced failure at the start of the study.

Suppose we are interested in examining the effect of covariates, in combination with the treatment, on survival, for prevalent subjects. This investigation can be carried out through a prevalent treatment study. It is thus clear that the incident time scale is not of interest here. In this type of study, a proportional hazards model on the follow-up time scale is the standard model since the treatment only starts at entry. If one is interested in the effect of covariates on this follow-up time scale, then, of course, a standard partial likelihood analysis is appropriate [8] with the risk sets defined in the usual manner. That is, at a particular failure time all subjects with larger failure times are included in the risk set.

Now, we may be interested in treatment-covariate effects on survival from initiation, rather than from the time of case ascertainment. For example, the time of the initiating event may be biologically defined, such as the date of infection with HIV. If this were the case it would be reasonable to impose a proportional hazards

model on the prevalent time scale. That is, a proportional hazards model would be assumed on the backward plus forward recurrence times of prevalent individuals, and not only on the follow-up times (forward recurrence times).

With this in mind, Cnaan and Ryan [5] suggest an analysis similar to the standard partial likelihood method described by Cox [8], only using modified risk sets. The modified risk sets that are suggested are exactly the same as those used in the one sample estimation of the survival distribution with left-truncated and right-censored data. That is, we include in the risk set at a particular failure time, only those subjects with larger failure times who are under active follow-up. However, Cnaan and Ryan provide no formal justification for performing such an analysis, nor is there any discussion of the assumption required that ensures its validity.

Wang et al [21] formally justify Cnaan and Ryan's ad hoc procedure. Making a crucial assumption, they carefully construct the partial likelihood. We now discuss this construction, paying particular attention to the main assumption needed for its validity.

Let u denote the calendar time of initiation for an individual and let v' be the calendar time of failure. Define the point of entry for a subject to be the calendar time τ. The prevalent proportional hazards model can be written as:

$$h'(t; Z'(\cdot)) = h'_0(t) \exp(Z'(t)\beta) \text{ for } t \geq \tau - u, \tag{3.8}$$

where h', h'_o are, respectively, the hazard and baseline hazard on the prevalent time scale. $Z'(\cdot)$ represents a time-varying covariate and β is the regression coefficient or the log of the relative risk on the prevalent time scale. We are interested in the estimation of the β.

Let $h'(t; u, \tau)$ be the hazard function for failure t units after u for a prevalent individual who was enrolled at time τ, where $t \geq \tau - u$. Wang et al [21] make the

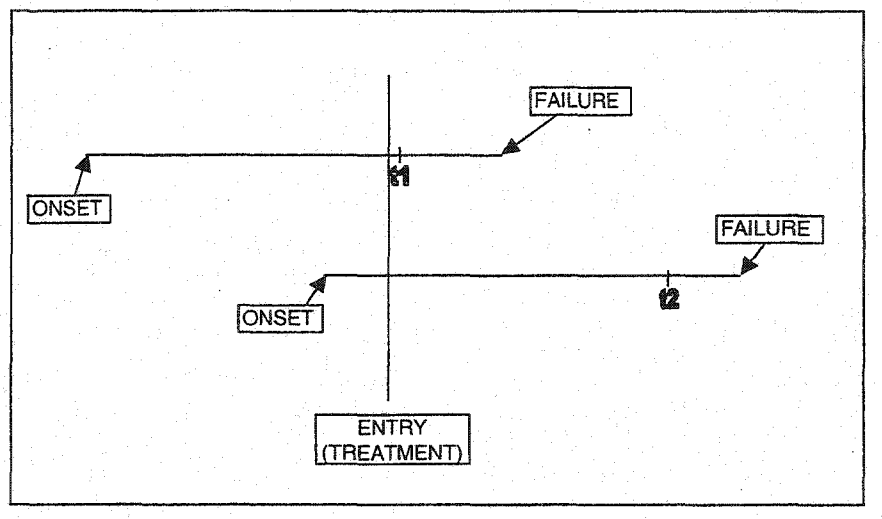


Figure 3.4: Quasi-stationarity

Under quasi-stationarity, these two subjects are assumed to have the same hazard at t1 and t2 since both subjects have been treated and since t1 and t2 are equidistant from their respective onset times. Note, however, that in the second case treatment was started sooner after onset than in the first case.

following important assumption:

Quasi-stationarity: *There exists a baseline hazard function* h'_0 *such that* h' *satisfies* $h'(t;u,\tau) = h'_o(t)$ *for* $t \geq \tau - u$. *That is,* $h'(t;u,\tau)$ *is independent of* (u,τ) *when* $t \geq \tau - u$.

Quasi-stationarity is a very strong, unrealistic assumption in a prevalent treatment study. It states that a subject's hazard function after entry is not affected by their calendar date of initiation or by the amount of time from initiation to entry.

In fact, by the authors' own admission, this assumption will rarely hold in a prevalent treatment study. The following example demonstrates one of many situations where assuming quasi-stationarity would be inappropriate.

Example 3.7: Consider a treatment which is developed for some form of cancer. We enter a population and identify prevalent cases with this form of cancer. At entry, we treat the subjects, who are obviously at different stages in the progression of the cancer. That is, some have had the cancer for long periods of time before being identified, while some may have just recently had onset of the cancer. Assuming quasi-stationarity would mean that any two subjects are thought to have identical hazards at some point in the progression of the cancer if, at this point of comparison, both subjects have entered the study and thus been treated (see Figure 3.4). Clearly, the hazard of failure for those subjects who have had the cancer for a longer period of time at entry (treatment) will be much greater than the hazard for those who are "caught" in the early stages of the cancer. That is, at entry the cancer may have developed to a stage where the treatment is not as effective as if it were administered immediately after onset of the cancer.

Clearly, quasi-stationarity does not hold in the situation described in Example 3.7. In this example, we assumed that all individuals were identified at the same point in time. If subjects were allowed to enter the study at different points in time, it may be reasonable to assume that their hazard after entry was independent of their time of onset. However, it almost always seems wrong to assume that the length of the backward recurrence time does not affect a subject's hazard after entry (treatment).

Wang et al [21] state that in the one sample problem the stronger assumption is often made that the entire survival time, $v' - u$, is independent of (u, τ) for $(v' - u) \geq 0$. However, this is not a fair comparison since this assumption is frequently made in natural history of disease studies, where there is no treatment involved. While it is true that often survival is assessed from prevalent cases, (leading to inference about incident cases) at intervention there is assumed to be no effective treatment. We can think of "treatment" as simply the act of case

ascertainment, in which case quasi-stationarity and even its stronger counterpart, is frequently reasonable; how long after onset a subject is identified should not affect their survival on the incident time scale. An alternative weaker assumption is thus needed when there is a treatment involved and quasi-stationarity does not seem plausible in any such study.

We shall shortly describe an attempt to weaken the assumption of quasi-stationarity. However, for the moment we assume quasi-stationarity and proceed with the analysis.

Wang et al [21] show that the full likelihood is proportional to a product of two functions, one of which is analogous to Cox's partial likelihood except for the modified risk sets.

Let y_i and δ_i represent the observed event time and the censoring indicator, respectively, for the i^{th} individual in the study, for $i = 1, ..., n$. Moreover, define the modified risks sets as $R(y) := \{j : \tau_j - u_j \leq y \leq y_j\}$, as is done in the one sample conditional approach (see Section 2.1). Writing the prevalent proportional hazard model as:

$$h'(t; Z(s), 0 \leq s < \infty) = h'_0(t) \exp(Z(t)\beta) \text{ for } t \geq \tau - u, \tag{3.9}$$

and conditioning on the $(u, \tau, Z(\cdot))$'s, the full likelihood based on the (y, δ)'s is:

$$L \propto \prod_{i=1}^{n} \left[\frac{f(y_i; Z_i(\cdot))^{\delta_i} S(y_i; Z_i(\cdot))^{1-\delta_i}}{S(\tau_i - u_i; Z_i(\cdot))} \right] \tag{3.10}$$

where f and S are, respectively, the density and survivor function of the failure time distribution on the prevalent time scale.

Under the model (3.9), L is proportional to $L_p * L_R$ where:

$$L_p(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(Z_i(y_i)\beta)}{\sum_{j \in R(y_i)} \exp(Z_j(y_i)\beta)} \right]^{\delta_i} \text{ and,} \tag{3.11}$$

$$L_R(\beta, h_0') = \left[\prod_{i=1}^{n}\left\{h_0'(y_i)\sum_{j\in R(y_i)}\exp(Z_j(y_i)\beta)\right\}^{\delta_i}\right]\cdot\exp\left[-\int h_0'(w)\sum_{j\in R(w)}\exp(Z_j(w)\beta)dw\right] \tag{3.12}$$

L_p is analogous to Cox's partial likelihood and L_R is termed the residual likelihood. The authors motivate the consideration of $L_p(\beta)$ alone for estimation of β by showing that $L_R(\beta, h_0')$ is ancillary for β. This result suggests that $L_R(\beta, h_0')$ does not provide any information in the estimation of β without knowledge of h_0'. Further justification for preceding in this manner is provided. That is, Wang et al [21] establish the usual properties of the partial likelihood. Namely, they show that the score function has zero expectation and, importantly, that $\widehat{\beta}$ converges in distribution to a multivariate normal distribution with a diagonal covariance matrix.

We now return to the quasi-stationarity assumption to determine if it can be relaxed in such a way so as to make the model useful in practice. Wang et al [21] suggest weakening quasi-stationarity by introducing (u, τ) through a function of (u, τ) that is then regarded as a time-dependent covariate. Ignoring the other covariates $Z(t)$, they propose that the prevalent proportional hazards model in (3.9) be replaced by,

$$h'(t; u, \tau) = h_0'(t)\exp(\phi(u, \tau)\alpha) \text{ for } t \geq \tau - u \tag{3.13}$$

where $\phi(\cdot, \cdot)$ is a some specified function and α is an unknown constant to be estimated from the data. Clearly (3.13) shows the (baseline * function-of-covariates) form is retained and the dependence on (u, τ) is incorporated through the function ϕ. We note, however, that the domain of $h'(t; u, \tau)$ still depends on (u, τ) through $(\tau - u)$. This model clearly includes quasi-stationarity as a special case ($\alpha = 0$). However, ultimately this proposal does not help since in prevalent treatment studies the interpretation of α will almost always be difficult owing to

confounding as is discussed by Wang et al in their example on ZVD treatment. Hence, the weakening of quasi-stationarity in this fashion is also not useful in practice. Therefore, if in a prevalent treatment study we wish to assess the effect of covariates on the prevalent time scale we are constrained by the assumption of quasi-stationarity. Unfortunately, this assumption will almost never be realistic in a prevalent treatment study. In fact, Wang et al suggest that prevalent treatment studies are probably not suitable unless an appropriate control group is feasible or there are appropriate historical data for "control" or baseline comparison. We discuss this point further in the "Closing Remarks" chapter of this thesis.

We have thus far focused on group comparisons through the examination of covariates. However, there are major difficulties in the procedures for both the incident and prevalent proportional hazards model. The incident time scale is of greater appeal in a natural history of disease study since in this type of study one often wishes to make a statement about survival from initiation for an incident case. We saw in Section 3.2 how Wang proposed a method for the estimation of covariate effects in this case, but her model does not allow for censoring. Although the prevalent time scale is more relevant in a prevalent treatment study we have seen that attempts to use a semi-parametric model to assess the effects of treatment-covariate combinations have been largely unsuccessful as they rely on the unrealistic assumption of quasi-stationarity. Another paper which may be of interest is by Alioum and Commenges [1].

We briefly discuss the alternative of fitting purely parametric models in the next chapter of this thesis. We assess the effect of erroneously using length-biased data in an unbiased model when comparing two groups. This seems not to have been addressed in the literature although the failure to recognize length-bias in data is common in the applied literature.

Chapter 4

Alternative comparisons and the importance of recognizing length-bias

In Chapter 3, we examined the comparison of two groups in the presence of length-biased data when Cox's proportional hazards model is assumed on a time scale relevant to the type of study being performed. Procedures were described for the estimation of covariate effects in the presence of length-bias when one is interested in the incident time scale and also when one is interested in the prevalent time scale. Unfortunately, the procedures in both these circumstances have critical shortcomings which cannot be overlooked.

In this chapter, in the same spirit as in Chapter 3, we investigate the consequences of making group comparisons in the presence of randomly left-truncated data. Here, however, we restrict our discussion to length-biased data, and, more importantly, to the effect of failing to recognize length-bias in these data. Simply stated, we address the question, not addressed in the literature, "Suppose we wish to compare survival, from initiation, between two groups. Then, what are the consequences of failing to recognize length-bias in the data?" We demonstrate, using parametric models, that at least when attention is restricted to a comparison of mean and median survival, the wrong conclusions may be drawn. These results

are important, particularly in natural history studies, where it is common for researchers to ignore length-bias. We propose an obvious solution for parametric models.

We first examine how the relationship between two biased survivor functions can change when their unbiased counterparts are compared. That is, we begin by making purely theoretical comparisons of the true survivor functions. Later in this chapter, we will illustrate through two simulation studies how incorrect inference can actually occur when data analyses are performed.

4.1 Comparison of the biased and unbiased survivor functions of two groups:

In this chapter, we assume that a natural history of disease study is being performed and that there is no effective treatment available, that can effect length of survival. Hence, in our previous terminology, we are interested in a comparison of survival on the incident time scale. One question of interest is whether the presence of length-bias could cause a complete "reversal" of the two true survivor functions under consideration. We shall explain, shortly, what is meant by a "reversal" of these survivor functions. This investigation is analogous to Brookmeyer and Gail's examination [4] of the biases that can occur for the survivor function in the one sample problem with prevalent data when only the follow-up times are available (see Section 2.2).

Consider some length-biased density $g_i(x)$, as in (1.1), where $f_i(x)$ is the underlying unbiased distribution, and $i = 1, 2$ is used to represent the two groups of interest. Let μ_i^U and μ_i^B represent the means of the unbiased and length-biased distributions respectively, so that μ_i^B is related to μ_i^U by the relationship (1.3). Also, let M_i^U and M_i^B be the medians of the unbiased and length-biased distributions respectively. Furthermore, let S_i^U and S_i^B denote the unbiased and length-biased

survivor functions respectively. That is, S^U represents the survivor function corresponding to $f(x)$ and S^B is the survivor function corresponding to $g(x)$.

Definition 4.1: *We say that a reversal of the survivor functions of the two groups has occurred if* $S_1^U(t) \geq S_2^U(t)\ \forall\ t \geq 0$, *but* $S_1^B(t) \leq S_2^B(t)\ \forall\ t \geq 0$, *or equivalently with the inequalities reversed (see Figure 4.1).*

We may first ask whether the reversal described in Definition 4.1 is possible. This would be an interesting finding since it would mean that length-bias could lead to a reversal of the inferred relationship between the survivor functions, and thus the survival experiences, of the two groups. It transpires that this "reversal" cannot occur:

Lemma 4.1: *Suppose* $S_1^U(t) \geq S_2^U(t)\ \forall\ t \geq 0$. *Then* $\exists$ *some interval* $[0, \delta], \delta > 0$, *such that* $S_1^B(t) > S_2^B(t)\ \forall\ t \in [0, \delta]$. *(An analogous result holds with the inequalities reversed since the assigning of group labels is arbitrary.)*

Note: (i) We suppose that $S_1^U(t) \not\equiv S_2^U(t)$. If the unbiased survivor functions are identical then it follows from (1.1) that the length-biased survivor functions are identical as well.

Note: (ii) In the proof, f_i and g_i, for $i = 1, 2$, can be either density functions or probability functions, with the consequence that Lemma 4.1 is valid in both the continuous and discrete cases.

Proof: We have that $g_i(x) = \frac{x f_i(x)}{\mu_i^U}$ for $i = 1, 2$. Since $S_1^U(t) \geq S_2^U(t)\ \ \forall\ t \geq 0$, $\Rightarrow \exists$ an interval $[0, \delta]$ where $f_2(t) \geq f_1(t)\ \forall\ t \in [0, \delta] \Leftrightarrow t f_2(t) \geq t f_1(t)\ \forall\ t \in [0, \delta]$ $\Leftrightarrow \frac{t f_2(t)}{\mu_2^U} \geq \frac{t f_1(t)}{\mu_2^U}\ \forall\ t \in [0, \delta]$ since $\mu_2^U \geq 0$. But $\frac{t f_1(t)}{\mu_2^U} \geq \frac{t f_1(t)}{\mu_1^U}\ \forall\ t \geq 0$ since $\mu_1^U > \mu_2^U$.

$$\therefore \frac{t f_2(t)}{\mu_2^U} > \frac{t f_1(t)}{\mu_1^U}\ \forall\ t \in [0, \delta] \tag{4.1}$$

Figure 4.1: Unbiased and biased survivor functions

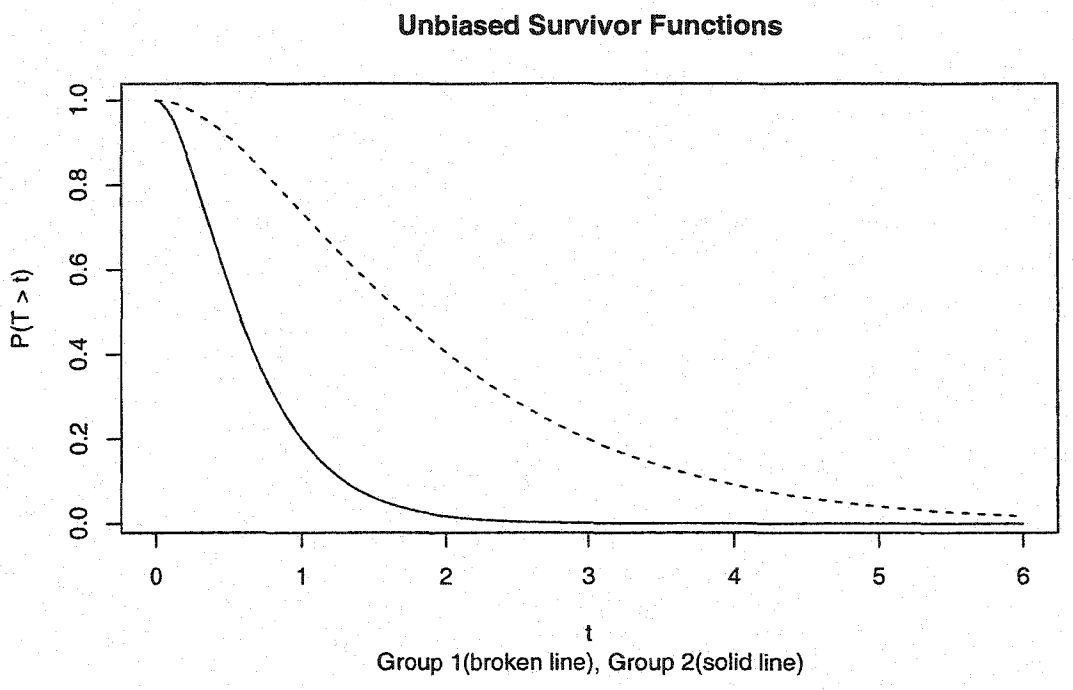


Figure 4.1a Unbiased survivor functions

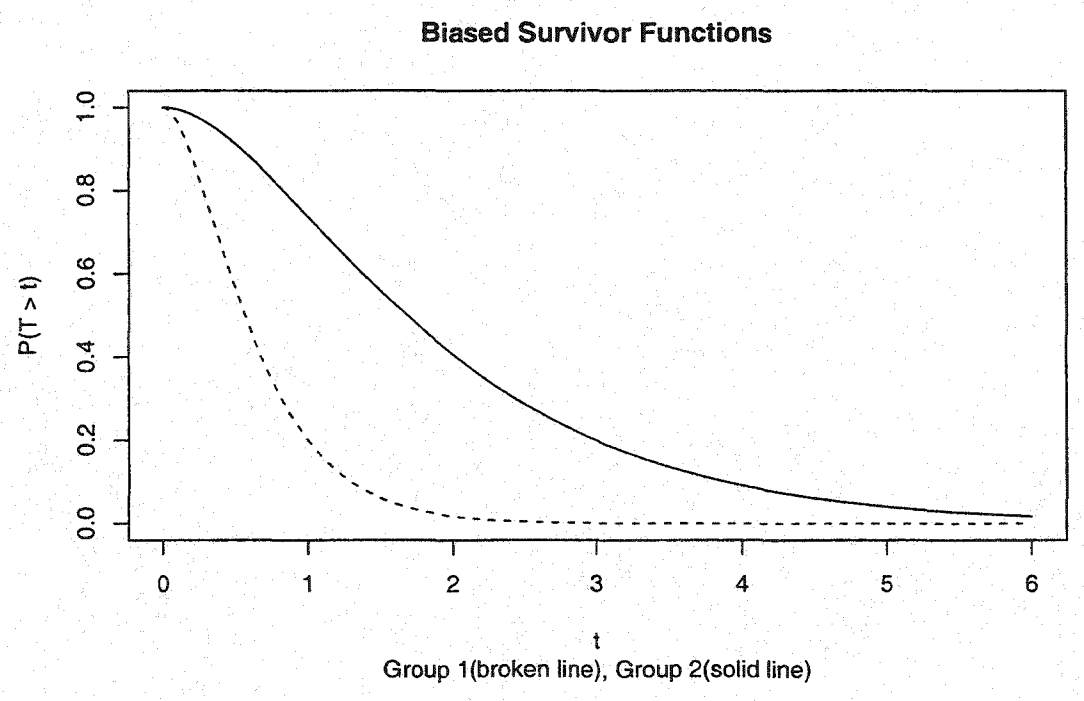


Figure 4.1b Biased survivor functions

Also, since the L.H.S of (4.1) $= g_2(t)$ and the R.H.S of (4.1) $= g_1(t)$, it follows that $g_2(t) > g_1(t)\ \forall\ t \in [0, \delta] \Leftrightarrow S_1^B(t) > S_2^B(t)\ \forall\ t \in [0, \delta]$, which proves Lemma 4.1.

Although Lemma 4.1 shows that the "reversal" described for the survivor functions cannot occur, the example below shows that this phenomenon can "approximately" hold, which, as we shall see, means that vigilance must be maintained in the presence of length-bias.

Example 4.1 (Refer to Figure 4.2): Let the underlying unbiased distributions be Weibull$(0.5, 0.5)$ and Weibull$(0.75, 0.75)$ for group 1 and group 2 respectively, where the Weibull(α, β) distribution is parameterized as:

$$f(x|\alpha, \beta) = \frac{\alpha}{\beta} x^{\alpha-1} e^{-x/\beta} \text{ where } (0 \leq x \leq \infty), (\alpha, \beta > 0) \tag{4.2}$$

We see from Figure 4.2(a) that $S_1^B(t) > S_2^B(t)$ for practically all values of t, while in Figure 4.2(b), $S_2^U(t) > S_1^U(t)$ for nearly all t.

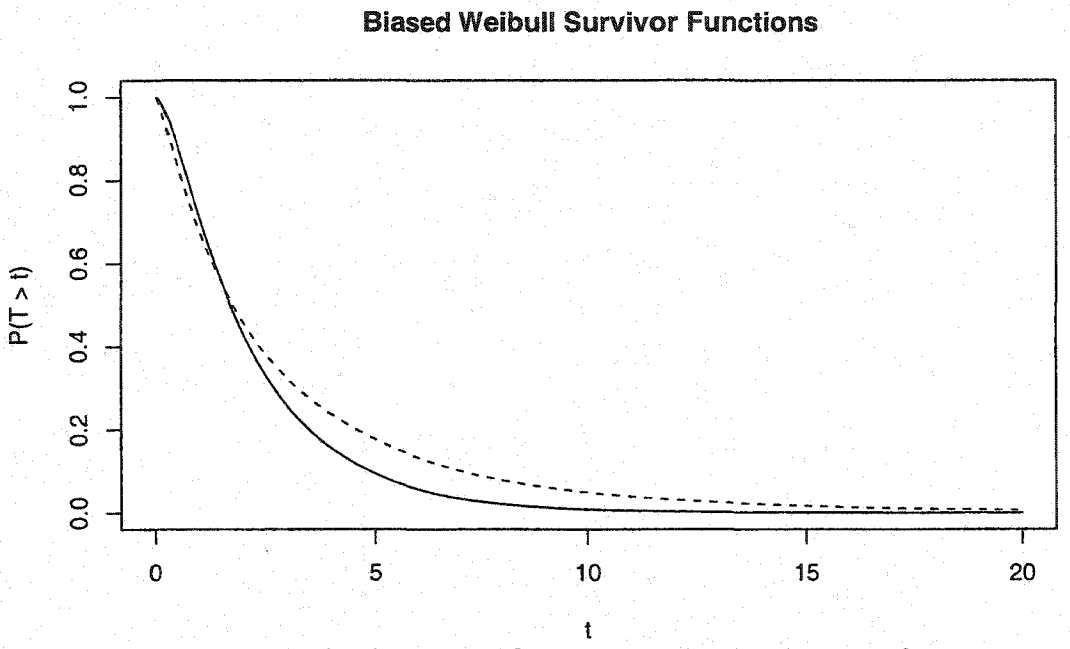
The next example shows that even when there is not "approximate" reversal, incorrect inference is still possible if inference is based on the length-biased survivor functions instead of the unbiased equivalents. Furthermore, the example demonstrates that these difficulties are not restricted to the class of Weibull distributions.

Example 4.2 (Refer to Figure 4.3): Let the unbiased distributions be gamma$(1, 1)$ and gamma$(5, 0.32)$ for group 1 and group 2 respectively, where the gamma(α, β) distribution is parameterized as:

$$f(x|\alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \text{ where } (0 \leq x \leq \infty), (\alpha, \beta > 0) \tag{4.3}$$

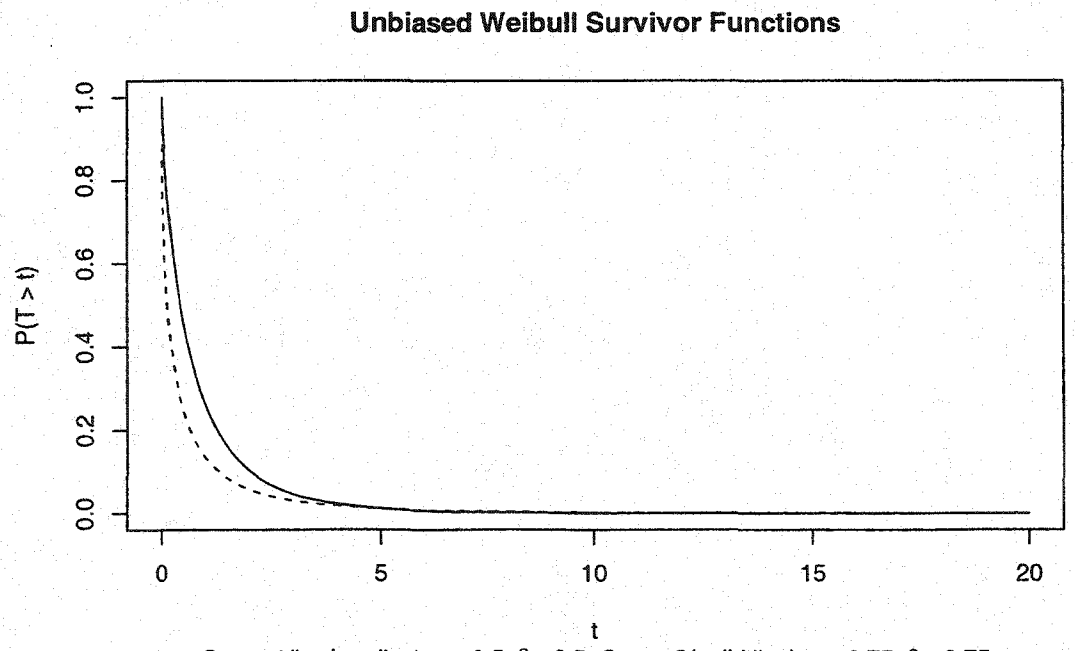
In Figure 4.3, the biased survivor functions seem to indicate that the two groups experience similar survival, with group 2 enjoying slightly better survival early on and group 1 having better survival later. However, the unbiased equivalents

Figure 4.2: Biased and unbiased Weibull survivor functions



Group 1(broken line): α_1=0.5, β_1=0.5, Group 2(solid line): α_2=0.75, β_2=0.75

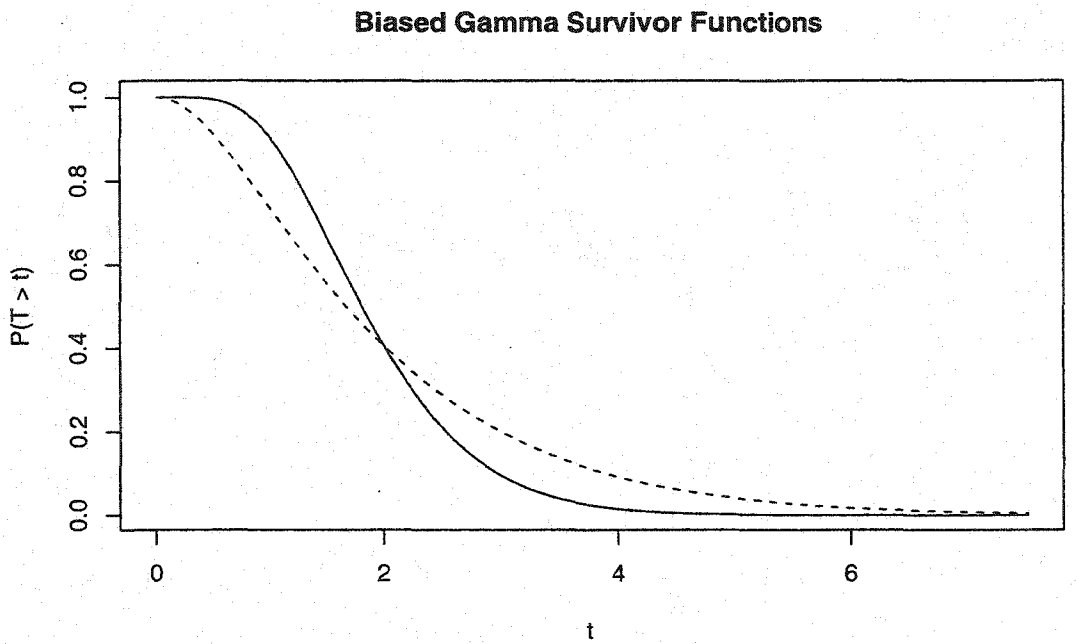
Figure 4.2a Biased Weibull survivor functions



Group 1(broken line): α_1=0.5, β_1=0.5, Group 2(solid line): α_2=0.75, β_2=0.75

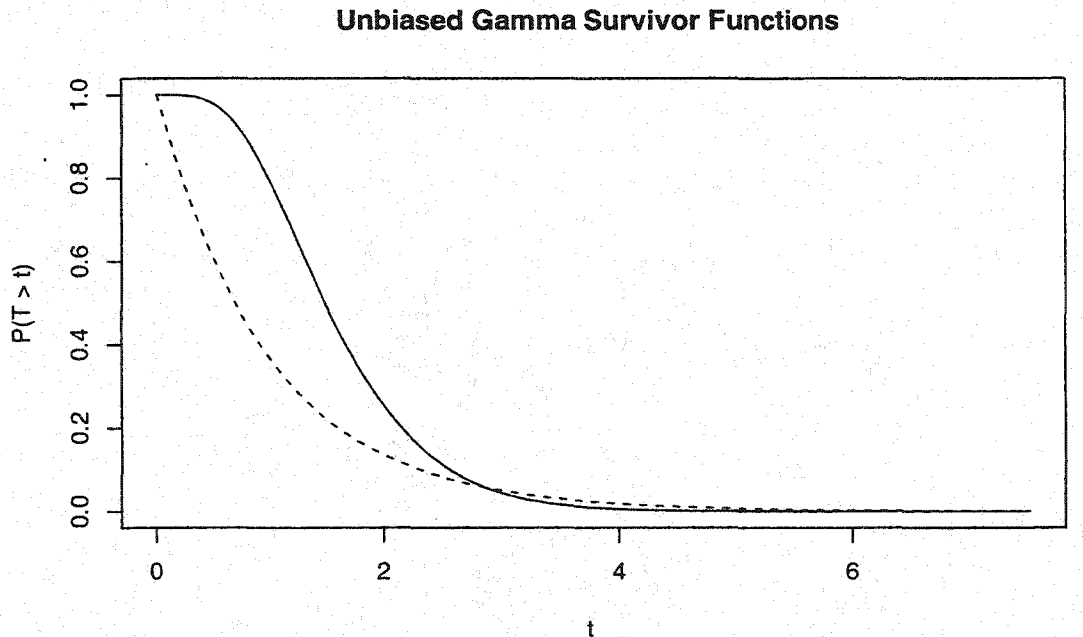
Figure 4.2b Unbiased Weibull survivor functions

Figure 4.3: Biased and unbiased gamma survivor functions



Group 1(broken line): α_1=1, β_1=1, Group 2(solid line): α_2=5, β_2=0.32

Figure 4.3a Biased gamma survivor functions



Group 1(broken line): α_1=1, β_1=1, Group 2(solid line): α_2=5, β_2=0.32

Figure 4.3b Unbiased gamma survivor functions

show that group 2 has much better survival in the early stages, while survival is essentially identical afterwards.

Thus, Example 4.1 and Example 4.2 demonstrate that the potential seriousness of using length-biased survivor functions instead of the corresponding unbiased survivor functions is not dispelled by Lemma 4.1.

Finally, we give an example where length-bias does not cause much change in the relationship between the survivor functions of the two groups.

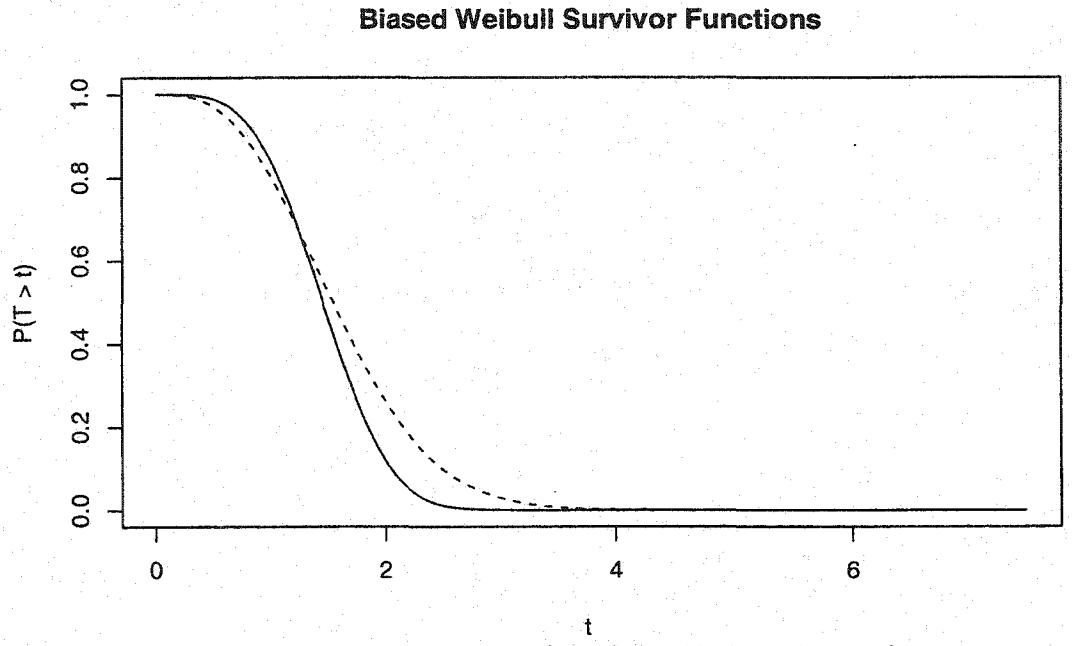
Example 4.3 (Refer to Figure 4.4): Let the unbiased distributions be Weibull$(2, 2)$ and Weibull$(3, 3)$ for group 1 and group 2 respectively. From Figure 4.4, we can see that the relative survival of the two groups is similar whether the biased or unbiased survivor functions are used for inference.

Example 4.3 is given to illustrate that length-bias need not affect the inference in every problem. Nevertheless, one must account for its consequences carefully since the effect (or lack of effect) of length-bias will not be known a priori.

4.2 Comparison of the means (medians) of two groups:

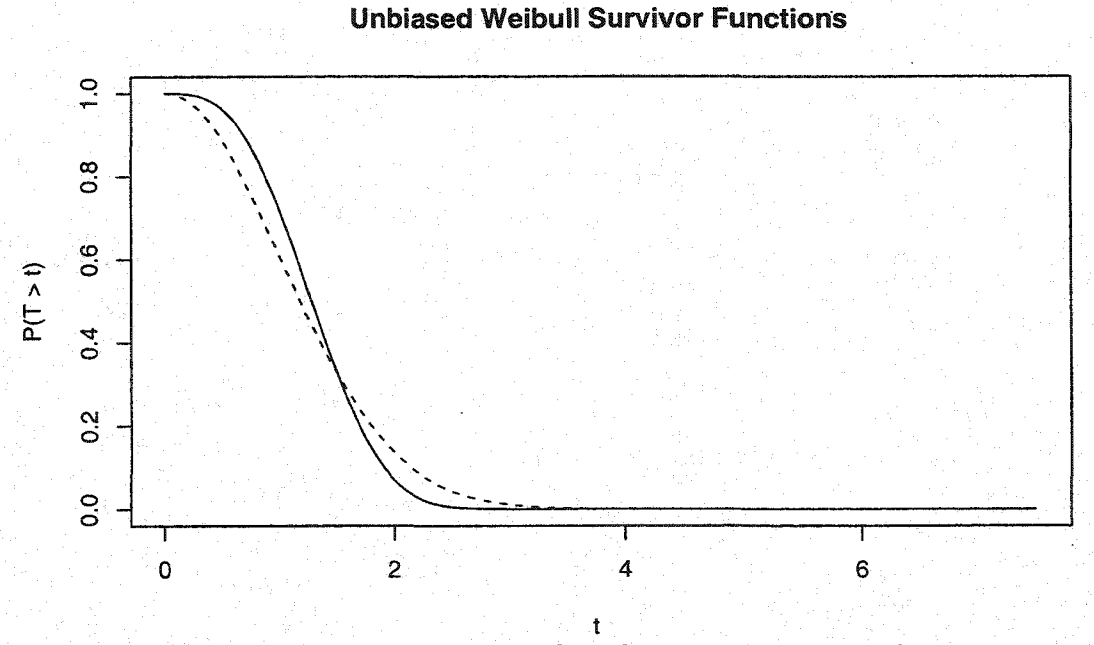
One reason for having considered possible "reversal" of the survivor functions is that it would have caused a "reversal" of the medians of the two groups. Even though we have shown that "reversal" cannot occur for the survivor functions of the two groups, it may still be possible that it occurs for their medians. Therefore, although the "reversal" of the survivor functions is clearly sufficient for the "reversal" of the medians, we need to determine whether it is necessary as well. In fact, it is not difficult to provide an example where reversal of the medians has occurred even though the reversal of the survivor functions is impossible. Figure 4.5 demonstrates this "reversal" of the medians where the unbiased distributions are assumed to be gamma$(1, 1)$ and gamma$(5, 0.21)$ for group 1 and group 2 respectively.

Figure 4.4: Biased and unbiased Weibull survivor functions



Group 1(broken line): α_1=2, β_1=2, Group 2(solid line): α_2=3, β_2=3

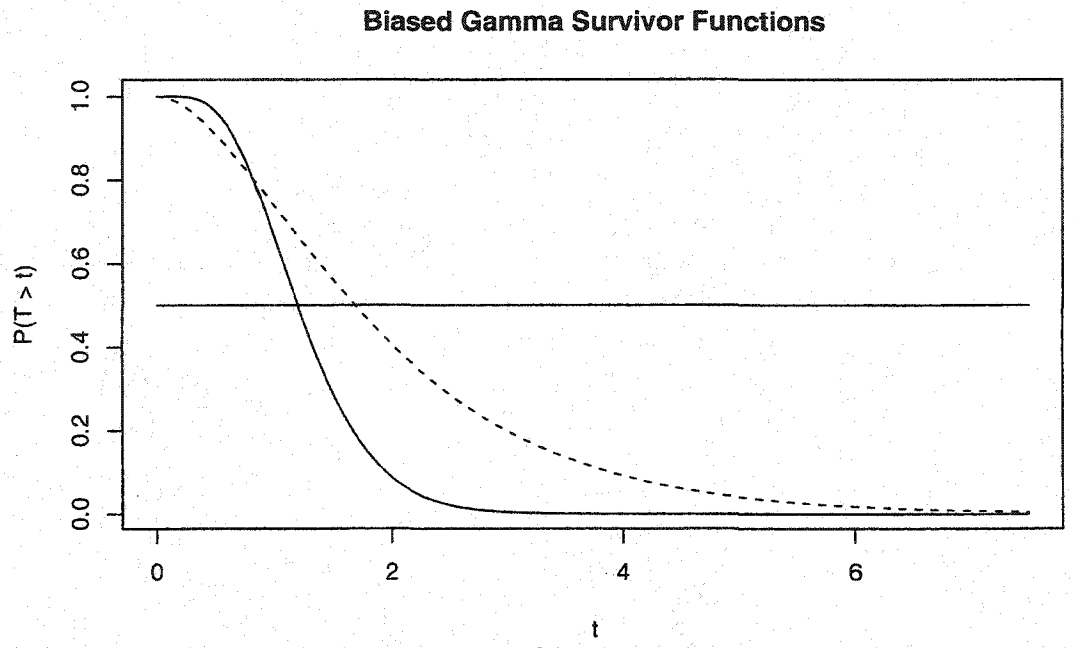
Figure 4.4a Biased Weibull survivor functions



Group 1(broken line): α_1=2, β_1=2, Group 2(solid line): α_2=3, β_2=3

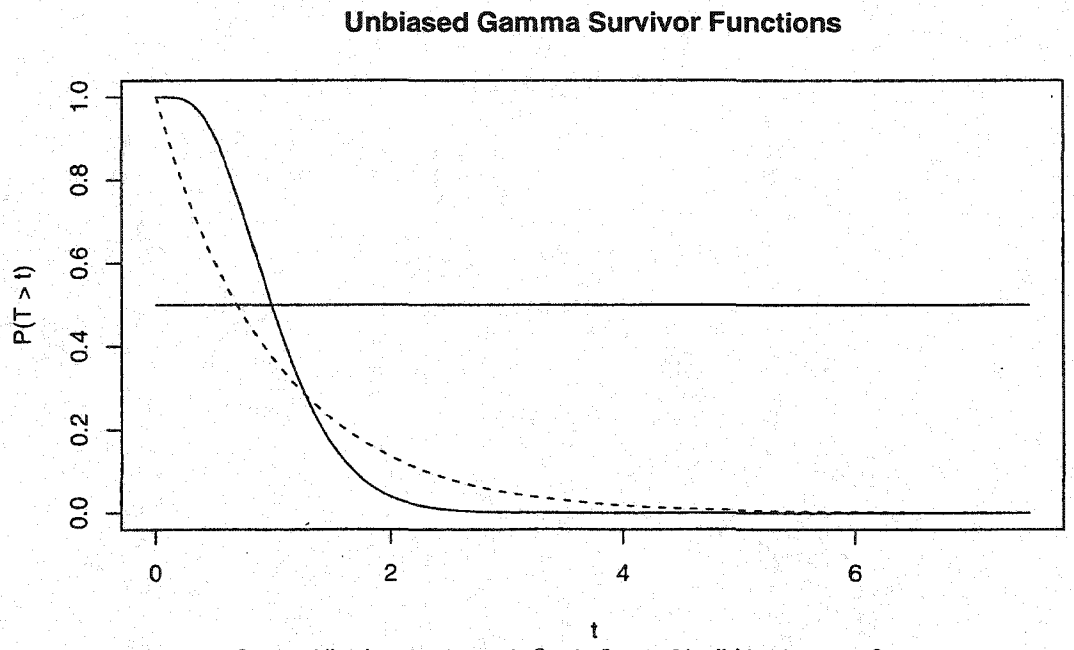
Figure 4.4b Unbiased Weibull survivor functions

Figure 4.5: Biased and unbiased gamma survivor functions



Group 1(broken line): α_1=1, β_1=1, Group 2(solid line): α_2=5, β_2=0.21

Figure 4.5a Biased gamma survivor functions



Group 1(broken line): α_1=1, β_1=1, Group 2(solid line): α_2=5, β_2=0.21

Figure 4.5b Unbiased gamma survivor functions

Moreover, everything that holds for the median is also true for the mean of the two groups. Recall that for a positive-valued random variable, the mean of the distribution is equal to the area underneath the survivor function. Hence, if we had that $S_1^U(t) \leq S_2^U(t)\ \forall\ t \geq 0$, and that $S_1^B(t) \geq S_2^B(t)\ \forall\ t \geq 0$ then we would have that $\mu_1^U < \mu_2^U$, but that $\mu_1^B > \mu_2^B$ (or equivalently with the inequalities reversed). Thus, interest now turns to whether it is possible that $\mu_1^U < \mu_2^U$ but that $\mu_1^B > \mu_2^B$ (or with the inequalities reversed) even though the "reversal" of the survivor functions is impossible. The answer to this question is in the affirmative.

With the gamma(α, β) parameterized as in (4.3), the length-biased density is:

$$g_i(x|\alpha,\beta) = \frac{x}{\alpha\beta} \cdot \left(\frac{x^{\alpha-1}e^{-x/\beta}}{\Gamma(\alpha)\beta^{\alpha}} \right) = \frac{x^{\alpha}e^{-x/\beta}}{\Gamma(\alpha+1)\beta^{\alpha+1}} \sim \text{gamma}(\alpha+1,\beta) \quad (4.4)$$

where $i = 1, 2$

Thus, if the true underlying distribution is gamma(α, β) then the length-biased distribution is gamma$(\alpha+1, \beta)$. A relationship such as this one, where the length-biased distribution takes on the same parametric form as the unbiased distribution, does not hold, in general, for other distributions, although this family invariance also holds for the family of Pareto distributions. For the gamma parameterized as in (4.3), the mean is $\alpha\beta$. Hence, the mean of the length-biased density is $(\alpha+1)\beta$. Now, this reversal of means will occur if:

$$(1)\quad \mu_1^U = \alpha_1\beta_1 < \alpha_2\beta_2 = \mu_2^U \text{ and if}$$

$$(2)\quad \mu_1^B = (\alpha_1+1)\beta_1 > (\alpha_2+1)\beta_2 = \mu_2^B$$

(1) and (2) are simultaneously satisfied if:

$$\frac{\alpha_2+1}{\alpha_1+1} < \frac{\beta_1}{\beta_2} < \frac{\alpha_2}{\alpha_1} \quad (4.5)$$

Hence, whenever the parameters for the gamma distributions of the two groups satisfy (4.5), the mean of group 2, say, will be larger for the unbiased distribution comparison, but the opposite will be true for the length-biased distribution comparison.

A commonly used distribution in survival analysis is the Weibull. If the underlying unbiased distribution of the data is Weibull then the length-biased distribution does not remain Weibull. Unlike the situation with the gamma distribution, we cannot obtain a simple expression for the Weibull parameters which indicates whether "reversal" of means will occur. However, it is not difficult to find examples where reversal does occur. Using expression (1.2), it follows that the mean of the length-biased Weibull is simply the second moment of the unbiased Weibull divided by the mean of the unbiased Weibull. Furthermore, using the parameterization of the Weibull(α, β) given in (4.2), we have that the k^{th} moment of a Weibull is given by:

$$E(X^k) = \beta^{k/\alpha}\Gamma(1 + \frac{k}{\alpha}) \tag{4.6}$$

Example 4.5: Assume the Weibull parameters to be (1, 1) for group 1 and (2, 2) for group 2. Hence, from (4.6),

$$\begin{aligned} \mu_1^U &= 1^{1/1}\Gamma(1+\frac{1}{1}) = 1,\ \mu_2^U = 2^{1/2}\Gamma(1+\frac{1}{2}) \approx 1.25, \text{ and} \\ \mu_1^B &= \frac{1^{2/1}\Gamma(1+\frac{2}{1})}{1} = 2,\ \mu_2^B \approx \frac{2^{2/2}\Gamma(1+\frac{2}{2})}{1.25} = 1.60 \end{aligned}$$

It is clear that what we refer to as a reversal of means has taken place in Example 4.5.

We now give a real-life illustration of the incorrect inference this could cause.

Example 4.6: Suppose, as in Example 3.2, that we are interested in comparing survival, from onset of dementia until death in two groups, namely, individuals with probable Alzheimer's disease and those with vascular dementia. In the literature, it has often been reported that individuals with probable Alzheimer's disease have longer median (and mean) survival from onset than those with vascular dementia. However, for practical reasons, these analyses were performed using prevalent cases. Thus, the data used in these studies were subject to length-bias

which was not recognized. It is therefore conceivable that, in fact, individuals with vascular dementia have longer median or mean survival than those with probable Alzheimer's disease.

We will shortly turn to an examination of what can happen in practice when data are actually analyzed. That is, we will demonstrate how incorrect inference can occur in a comparison of the mean and median of two groups when length-bias is not recognized. Before doing that, we describe the types of analyses that a researcher might perform on an observed data set.

4.3 Comparison of data analyses:

Below we present a comparison of three possible procedures that might be used to analyze data. The first is a naïve approach where length-bias is not recognized, while the second is the correct manner of proceeding. The third approach is restricted only to certain parametric analyses. We concentrate on the first two procedures in our parametric simulations.

4.3.1 "Naïve" approach:

A naïve approach to maximum likelihood estimation which does not recognize length-bias would proceed as follows:

Let $L^B(\underset{\sim}{x};\theta)$ and $L^U(\underset{\sim}{x};\theta)$ be, respectively, the biased and unbiased likelihoods, for the parameter vector θ, given the observed biased data $\underset{\sim}{x}^B$, generated by the biased model L^B.

1. Maximize $L^U(\underset{\sim}{x};\theta)$ with respect to θ.

2. Carry out all inference using the incorrect estimator $\hat{\theta}_U^B$, where the subscript designates the model that is assumed and the superscript the "true" model of the data.

3. All sampling distributions are derived under the assumption that the data were derived under L^U.

4. If there are two independent comparison groups, repeat this procedure for each group and base the inference on $h(\hat{\theta}^B_{1;U}, \hat{\theta}^B_{2;U})$ for some function h, of the two estimators $\hat{\theta}^B_{1;U}$ and $\hat{\theta}^B_{2;U}$.

The method just described uses an unbiased model for inference although the data are length-biased.

For example, suppose that the observed survival times for the two groups, $\underset{\sim}{x}^B_i, i = 1, 2$, arise from length-biased Weibull distributions with length-biased Weibull likelihoods given by,

$$L^B_i(\underset{\sim}{x}^B_i; (\alpha_i, \beta_i)) = \prod_{j=1}^{n} \left[\frac{\alpha_i (x^B_{ij})^{\alpha_i} e^{-x^B_{ij}/\beta_i}}{\beta_i^{1+1/\alpha_i} \Gamma(1 + \frac{1}{\alpha_i})} \right] \tag{4.7}$$

where $(0 \leq x^B_{ij} < \infty)$, $(\alpha_i, \beta_i > 0)$, $(i = 1, 2)$

and n is the number of survival times in $\underset{\sim}{x}^B_i$.

The "Naïve" Approach would specify the maximization of,

$$L^U_i(\underset{\sim}{x}^B_i; (\alpha_i, \beta_i)) = \prod_{j=1}^{n} \left[\frac{\alpha_i (x^B_{ij})^{\alpha_i - 1} e^{-x^B_{ij}/\beta_i}}{\beta_i} \right] \tag{4.8}$$

where $(0 \leq x^B_{ij} < \infty)$, $(\alpha_i, \beta_i > 0)$, $(i = 1, 2)$

and n is the number of survival times in $\underset{\sim}{x}^B_i$,

with respect to (α_i, β_i), $i = 1, 2$, respectively for the two groups, to obtain $(\hat{\alpha}_i, \hat{\beta}_i)^B_U$.

If we are interested in making inference about the difference in the unbiased medians, $(M^U_1 - M^U_2)$, say, then a parametric bootstrap would be (naïvely) performed as follows:

1. Generate k sets of survival times, each of size n from $f^U_i(\underset{\sim}{x}; (\hat{\alpha}_i, \hat{\beta}_i)^B_U)$. Denote these by $\underset{\sim}{x}^U_{i;p}$ for $i = 1, 2$ and $p = 1, ..., k$.

2. For each of $i = 1, 2$, and $p = 1, ..., k$, maximize (4.8) with respect to (α_i, β_i) to obtain $(\hat{\alpha}_{i;p}, \hat{\beta}_{i;p})_U^U$ for $i = 1, 2$, and $p = 1, ..., k$.

3. After substituting $(\hat{\alpha}_{i;p}, \hat{\beta}_{i;p})_U^U$ in $f_i^U(\underset{\sim}{x}; (\alpha_i, \beta_i))$, evaluate $(\widehat{M}_{1;U}^U - \widehat{M}_{2;U}^U)_p$ for $p = 1, ..., k$ to obtain k "naïve" estimates of $(M_1^U - M_2^U)$.

4. Obtain a 95% confidence interval for $(M_1^U - M_2^U)$ by sorting these k estimates and eliminating the 2.5% smallest and biggest estimates.

A correct nonparametric approach would proceed along similar lines except that the estimators used would be nonparametric and bootstrap procedures would be nonparametric as well.

Moreover, using the "Naïve" Approach, if one were to assume Cox's proportional hazards model on the incident time scale, one would unwittingly fit Cox's proportional hazards model to the length-biased data. Unfortunately, proportionality of hazards on the incident (unbiased) scale does not imply proportionality of the hazards for the length-biased distributions.

4.3.2 Correct approach:

A correct approach, which would recognize length-bias and account for it accordingly, would proceed as follows:

1. Using the same notation, maximize $L^B(\underset{\sim}{x}^B; \theta)$ with respect to θ.

2. Carry out all inference using the correct estimator $\hat{\theta}_B^B$.

3. All sampling distributions are derived under the assumption that the data were derived under L^B.

4. If there are two independent comparison groups, repeat this procedure for each group and base the inference on $h(\hat{\theta}_{1;B}^B, \hat{\theta}_{2;B}^B)$ for some function h, of the two estimators $\hat{\theta}_{1;B}^B$ and $\hat{\theta}_{2;B}^B$.

Note, importantly, though, that the function h is a functional of the unbiased distribution, which is of main interest. For example, h might represent $(M_1^U - M_2^U)$, using the parameter estimates obtained by recognizing the presence of length-bias in the data. In addition this method uses the length-biased model, that generated the data, to infer the sampling distribution of any parameter estimator.

For example, suppose that the observed survival times for the two groups, $\underset{\sim}{x}_i^B, i = 1, 2$, arise from length-biased Weibull distributions with length-biased Weibull likelihoods given by (4.7). The Correct Approach would specify the correct maximization of (4.7) with respect to (α_i, β_i), $i = 1, 2$, respectively for the two groups, to obtain $(\hat{\alpha}_i, \hat{\beta}_i)_B^B$.

If we are interested in making inference about the difference in the unbiased medians, $(M_1^U - M_2^U)$, say, then a parametric bootstrap would be performed as follows:

1. Generate k sets of survival times, each of size n from $f_i^B(\underset{\sim}{x}; (\hat{\alpha}_i, \hat{\beta}_i)_B^B)$. Denote these by $\underset{\sim}{x}_{i;p}^B$ for $i = 1, 2$ and $p = 1, ..., k$.

2. For each of $i = 1, 2$, and $p = 1, ..., k$, maximize (4.7) with respect to (α_i, β_i) to obtain $(\hat{\alpha}_{i;p}, \hat{\beta}_{i;p})_B^B$ for $i = 1, 2$, and $p = 1, ..., k$.

3. After substituting $(\hat{\alpha}_{i;p}, \hat{\beta}_{i;p})_B^B$ in $f_i^U(\underset{\sim}{x}; (\alpha_i, \beta_i))$, obtain $(\widehat{M}_{1;B}^B - \widehat{M}_{2;B}^B)_p$ for $p = 1, ..., k$ to yield k estimates of $(M_1^U - M_2^U)$.

4. Using the quantile method obtain a 95% confidence interval for $(M_1^U - M_2^U)$.

A correct nonparametric approach would proceed along similar lines except that the estimators and bootstrap procedures used would be nonparametric.

4.3.3 "Partially naïve" approach:

A third approach is plausible for certain parametric scenarios. Suppose the length-biased aspect of the data is recognized, but it is not accounted for correctly. We

refer to such an approach as a "partially naïve" method. This method proceeds as follows:

1. As in the correct approach, maximize $L^B(\underset{\sim}{x}^B;\theta)$ with respect to θ.

2. Obtain the correct estimator, $\hat{\theta}_B^B$. That is, this method uses the correct biased model for estimation of θ, recognizing that the data are length-biased.

3. If there are two independent comparison groups, repeat this procedure for each group and base the inference, however, on $h'(\hat{\theta}_{1;B}^B, \hat{\theta}_{2;B}^B)$ where h' is the length-biased function corresponding to h, the function of interest.

For example, suppose we are interested in $(M_1^U - M_2^U)$, then in the partially naïve method, we would base inference on estimates of $(M_1^B - M_2^B)$. A parametric bootstrap would be performed exactly as in the correct approach except that $(\hat{\alpha}_{i;p}, \hat{\beta}_{i;p})_B^B$, for $i = 1, 2$, and $p = 1, ..., k$, would be substituted into $f_i^B(\underset{\sim}{x};(\alpha_i, \beta_i))$ in order to obtain the k estimates of $(M_1^B - M_2^B)$.

Of course, although inference is based on estimates of $(M_1^B - M_2^B)$, the quantity of interest is $(M_1^U - M_2^U)$. The partially naïve reasoning is as follows, "Although unbiased and length-biased distributions differ, *comparisons* between pairs of unbiased and pairs of their corresponding length-biased distributions should remain invariant".

This scenario corresponds to the hypothetical scenario described in Section 4.1 whereby one might be tempted to compare only the pair of biased survivor functions to infer the same relationship for the unbiased versions. We have seen this reasoning to be flawed.

This third approach is largely a curiosity, though, as it would only arise in very few situations. It would not be possible in a nonparametric analysis since, here, if length-bias is recognized, then the correct estimates are immediately obtained for the survivor functions of the two groups. Even in a parametric analysis, this

situation does not always occur. For example, suppose we are performing a parametric analysis assuming that the underlying unbiased distributions are gamma. In this case, recognizing length-bias is sufficient in order to carry out a correct analysis since a length-biased gamma distribution retains the parametric form of a gamma distribution. Hence, once the correct estimates of the parameters are obtained, substituting these estimates into the length-biased or unbiased gamma forms yields identical results. However, this "partially naïve" approach may yield erroneous conclusions if the Weibull distribution is assumed to be the unbiased distribution in a parametric analysis since the length-biased Weibull does not retain the Weibull parametric form. As this approach can only occur in very special circumstances, we will focus our discussion on the "naïve" and correct methods.

4.4 Performance evaluations through two simulation studies:

Now, suppose that we are interested in the difference in mean survivals of two groups (or, possibly the difference in median survivals), that is, the quantity $(\mu_1^U - \mu_2^U)$. Any instance of a "reversal" of means will cause the quantities $(\mu_1^U - \mu_2^U)$ and $(\mu_1^B - \mu_2^B)$ to be of opposite sign. Thus, these theoretical results suggest that inference about the mean survival experience of the two groups will be incorrect if length-bias is not accounted for in the analysis. We investigate these considerations by carrying out simulations that enable us to examine the coverage percentages of "naïve" and correct bootstrapped 95% confidence intervals for $(\mu_1^U - \mu_2^U)$ and for $(M_1^U - M_2^U)$. These may be termed "performance analyses".

A comparison of the correct approach with the "partially naïve" method could just as easily have been carried out and similar results can be obtained but these will not be included in this thesis.

Although our simulations generate uncensored data, similar results would be

obtained with censored survival times since, here, censoring only diminishes the amount of information contained in the data and does not change anything conceptually.

4.4.1 Simulation study #1:

The underlying unbiased distribution was assumed to be gamma$(1, 3)$ for group 1 and gamma$(4, 1)$ for group 2. Five-hundred (500) length-biased uncensored survival times were generated for each of the two groups from gamma$(1 + 1 = 2, 3)$ and gamma$(4 + 1 = 5, 1)$ distributions, respectively, which are the length-biased distributions corresponding to the gamma$(1, 3)$ and gamma$(4, 1)$ distributions, respectively. These data were taken as our "observed" data set.

Since the length-biased gamma distribution is itself gamma, the "naïve" and correct approach both maximize the same likelihood. That is, if one believes that the data are unbiased, a gamma likelihood will be maximized, while if one knows that the data are length-biased, a length-biased gamma likelihood, which is again a gamma likelihood, will be maximized. Hence, naïve and correct methods will yield the same parameter estimates for both groups. We obtained parameter estimates, $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$, for the parameters $(\alpha_1 = 2, \beta_1 = 3)$ and $(\alpha_2 = 5, \beta_1 = 1)$.

However, someone applying the naïve method would believe that $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$ estimate the parameters of the unbiased gamma distribution since he/she will not have recognized the length-bias. On the other hand, someone using the correct method would realize that, in fact, these parameter estimates correspond to the length-biased gamma. Using the correct approach, the parameter estimates for the unbiased gamma are $(\hat{\alpha}_1 - 1, \hat{\beta}_1)$ and $(\hat{\alpha}_2 - 1, \hat{\beta}_2)$.

In order to find a confidence interval for $(\mu_1^U - \mu_2^U)$, a parametric bootstrap was carried out, using both the "naïve" and correct method. For the "naïve" parametric bootstrap, the gamma$(\hat{\alpha}_1, \hat{\beta}_1)$ and gamma$(\hat{\alpha}_2, \hat{\beta}_2)$ distributions were

used to generate 1000 data sets each consisting of 500 survival times for group 1 and group 2 respectively. The incorrect philosophy behind this bootstrap is to generate data from the unbiased distribution. In a correct parametric bootstrap, one would want to generate data from the length-biased distribution. Coincidentally, the two methods are again identical, since the length-biased gamma distributions with parameters $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$ for group 1 and group 2, respectively, are again gamma with these parameters.

For each of the 1000 data sets, the same procedure was used to obtain the estimates of the parameters as with the "observed" data. Having paired off the data sets for the two groups arbitrarily, 1000 estimates of the difference in means and medians of the two groups were produced. For the naïve approach, these estimates were obtained by using the parameter estimates without modification. The estimates of the difference in means and medians using the correct method were obtained by realizing that the parameter estimates were obtained from a length-biased gamma distribution and need adjusting before they can be used to estimate $(\mu_1^U - \mu_2^U)$ and $(M_1^U - M_2^U)$. That is, in the correct approach, $\hat{\alpha}_{ij}$ became $(\hat{\alpha}_{ij} - 1)$ for $i = 1, 2$ and $j = 1, ..., 1000$.

To obtain a confidence interval for these differences, we simply ordered the 1000 estimates in increasing fashion and selected the 26^{th} and the 975^{th} largest difference. In this way, we obtained two confidence intervals (for the difference in means and medians respectively). This entire procedure was repeated 100 times and each time we verified whether the true difference in means and medians was captured by the corresponding confidence interval.

The results for the two procedures are presented in Table 4.1. We can see from these results that if length-bias were not recognized and the naïve method applied, we never captured the true values of interest. The correct method gave approximately 95% coverage, as expected. Table 4.2 displays the average lengths

Simulation Study #1: Performance of the Two Methods		
Frequency of true value captures out of 100		
Method	Difference in Medians	Difference in Means
Correct Approach	94	93
Naïve Approach	0	0

Table 4.1: Simulation study #1: performance of the two methods

Simulation Study #1: Variability of the Two Methods		
Average length of confidence interval		
Method	Difference in Medians	Difference in Means
Correct Approach	0.98812	0.87485
Naïve Approach	0.75655	0.83661

Table 4.2: Simulation study #1: variability of the two methods

of the confidence intervals produced. One may have expected the intervals to be wider for the naïve approach, but this is not the case in this example. We suspect that this result is particular to the gamma distribution, owing to the property that a length-biased gamma remains gamma.

4.4.2 Simulation study #2:

In this simulation study, the underlying unbiased distributions were assumed to be Weibull$(0.5, 0.5)$ and Weibull$(0.75, 0.75)$ for group 1 and group 2 respectively. We generated 500 uncensored survival times from the length-biased Weibull$(0.5, 0.5)$ and the length-biased Weibull$(0.75, 0.75)$, for group 1 and group 2 respectively. It is interesting to note that length-biased Weibull data can be simulated conveniently by initially generating from a gamma distribution (Correa and Wolfson [6]). These survival times are assumed to be the "observed" data.

Essentially the same method was used as for Simulation study #1, except with different likelihoods to accommodate the different parametric forms. In the correct analysis, a length-biased Weibull likelihood was maximized whereas, for the naïve method, the likelihood that was maximized was simply a Weibull. This led to

different parameter estimates for the two approaches, which is slightly different from the situation in Simulation study #1. This is because the length-biased Weibull does not remain Weibull. Again, in the correct parametric bootstrap, 1000 sets each consisting of 500 length-biased Weibull survival times were generated for each of the two groups using the correct parameter estimates. In the "naïve" method, the incorrect parameter estimates were used to generate 1000 sets each consisting of 500 Weibull survival times for each group. For each of these data sets, parameter estimates were obtained. Estimates of the means and medians were then derived by substituting the parameter estimates into the appropriate functionals of the unbiased Weibull distribution. Of course, the parameter estimates used in the "naïve" method were incorrect. We then proceeded as in Simulation study #1 to obtain confidence intervals for the difference in means and medians, respectively, and to assess the performance of the two methods.

Simulation Study #2: Performance of the Two Methods		
Frequency of true value captures out of 100		
Method	Difference in Medians	Difference in Means
Correct Approach	96	97
Naïve Approach	90	0

Table 4.3: Simulation study #2: performance of the two methods

Simulation Study #2: Variability of the Two Methods		
Average length of confidence interval		
Method	Difference in Medians	Difference in Means
Correct Approach	0.19721	0.32802
Naïve Approach	0.86288	0.66321

Table 4.4: Simulation study #2: variability of the two methods

The results for the two procedures are presented in Table 4.3. The correct method again gave approximately 95% coverage. The naïve method for the difference in means again performed extremely poorly. For the difference in medians, the naïve method performed fairly well in terms of coverage proportion. However,

when examining the average lengths of the confidence intervals, we see that the naïve method produced much wider confidence intervals (see Table 4.4). This explains the apparently adequate coverage of its confidence intervals for the difference in medians.

We can see from the results of these two simulation studies that the failure to recognize length-bias can affect the validity and/or the efficiency of an analysis. That is, one may obtain very poor coverage, and even if one does, in fact, capture the true value frequently, the confidence intervals may be very wide.

Chapter 5

Closing remarks

In this chapter, we discuss some interesting points that have been raised in this thesis. We also mention some topics for further research and a possible procedure for the estimation of covariate effects in the two sample problem, even when there is censoring.

In Section 1.2, we gave a brief overview of the historical literature in the area of length-bias. We discussed how Cox had developed an unbiased estimator of the mean of the underlying distribution from a length-biased sample. In fact, Cox [7] demonstrates that, in some instances, it may be more efficient to use a length-biased sample for estimation of the mean than the unbiased sample. This is, perhaps, a justification for the use of the sampling method of "grabbing" used in the textile industry at that time, and which gives rise to length-biased fibre lengths.

Blumenthal [3] also examines whether it is more efficient to estimate the mean of the unbiased distribution using a length-biased sample or an unbiased sample. For the gamma and Weibull distributions it is, in fact, more efficient to use a length-biased sample. This result is very interesting since the gamma and Weibull are widely used in survival analysis. For the log-normal distribution, the efficiency is always the same, regardless of the parameters of the distribution. Thus, if one believes the data to come from a log-normal distribution, it seems that a length-biased sample should be considered, since there is no loss in efficiency and

it may be easier to obtain such a sample. However, as Cox mentions in his paper, Blumenthal obtains most of his results by assuming a known coefficient of variation. For the gamma and Weibull, the coefficient of variation is only known if the shape parameter of the distribution is known. However, a method for estimating the coefficient of variation is provided by Blumethal when it is unknown.

In the medical setting, the use of prevalent cohorts have often been perceived as being necessary for practical reasons such as time limitations. However, the preceding results show that using a cross-sectional sampling scheme may not only be of practical convenience but may improve the efficiency of mean estimates. Although these papers did not consider censoring, it may be speculated that the same result will hold even when censoring occurs. This question deserves further consideration.

In Section 2.1, we briefly mention that Wang's one sample conditional product-limit type estimator may give poor results near 0. Firstly, Wang admits that the estimator in (2.1) may be non-identifiable before the smallest observed event time, $y_{(1)}$. Hence, this estimator should really be seen as a conditional estimator of survival, given that one's survival is greater than $y_{(1)}$. In practice, this is not usually a major difficulty if $y_{(1)}$ is small. The risk sets for this estimator include only individuals who have not failed or been censored and who are under active follow-up. Hence, even after $y_{(1)}$, there may be difficulties caused by small risk sets. If at any failure time, everyone in the risk fails then the estimator drops to 0 and obviously remains at 0 for all subsequent times since it is comprised of a product of terms. Cnaan and Ryan [5] point out that this may happen in a small study or in the early stages of a study. In one wishes to avoid such difficulties, they suggest an estimator based on the cumulative hazard function. This estimator is also recommended for such circumstances by Cox and Oakes [9].

Cnaan and Ryan [5] also make an interesting non-technical observation regard-

ing the interpretation of fixed and time-varying covariates. Measured covariates are often considered as fixed in a proportional hazards model when performing an analysis on survival from entry. Some of these covariates, such as symptoms or extent of disease, may be used to indicate a patient's status at entry. However, one must be careful since these same covariates measured at entry should not necessarily be viewed as fixed in an analysis of survival measured from onset. For example, a patient with weight loss at enrollment may not have had that symptom at onset. Hence, this covariate must be viewed as time-varying in an analysis from onset.

Wang et al [21] make an important point regarding the evaluation of treatments in a prevalent treatment study. In Example 3.4, a type of prevalent treatment study is described where randomization has been carried out to determine which subjects receive a certain treatment. A prevalent treatment study is one in which the time of entry corresponds with the beginning of a treatment. The effectiveness of treatment cannot be determined in a prevalent treatment study unless there is a control group or some other external information is utilized. That is, if all subjects receive treatment at entry then the effect of treatment cannot be identified, although, of course, information about covariate effects within the treated population can be acquired. Hence, although these studies may be useful in some ways, they should not be used to determine the main effect of a treatment unless randomization takes place. This may have ethical ramifications which must be addressed.

In Chapter 3 we saw that using a proportional hazards model for the estimation of covariate effects is, in many ways, a futile endeavor. When interest lies in the incident time scale, Wang's risk set sampling method [20] does not allow for censoring. Moreover, when one is interested in the prevalent time scale, the adjusted risk sets procedure relies heavily on an unrealistic assumption for prevalent treatment studies. Therefore, an alternative approach should be investigated which circumvents these difficulties. We propose the use of a model which

assumes a piecewise constant baseline hazard function. This eliminates the shortcomings of a semi-parametric model since estimation reduces to the estimation of finitely many parameters. Namely, these parameters are the values of the hazard on the finite number of intervals for which the hazard is assumed to be constant. The introduction of this parametric structure to the hazard should alleviate the problems encountered thus far. Furthermore, the conclusions drawn from any reasonable amount of data should be similar whether a proportional hazards model or a piecewise constant hazard model is adopted. The piecewise constant hazard model therefore deserves future consideration.

Brookmeyer and Gail [4] examined the biases which may occur in the estimation of relative risk when one is interested in the incident time scale, but only the follow-up times are available. In Chapter 4, we briefly look into the incorrect inference which is possible when length-bias is not recognized through two simulation studies using a parametric analysis. It seems worthwhile to examine whether analogous results to those of Brookmeyer and Gail hold regarding the relative risks on the incident and prevalent time scales. For instance, can we state that the bias induced by not recognizing length-bias is never enough to make a true risk factor appear protective (or vice versa)?

In this thesis, we examined the phenomenon of length-bias in the one and two sample problems. We provided a review of the literature for both these circumstances, paying particular attention to group comparisons and the difficulties associated with such ventures when only prevalent data are available. Length-bias continues to be a significant area of research since, in practice, prevalent follow-up data is often the most convenient to observe. We have seen that there are still many open questions in this area which need to be pursued. The importance of this pursuit becomes apparent when one realizes the consequences of making erroneous inference and acting upon these conclusions.

Bibliography

[1] A. Alioum and D. Commenges. A proportional hazards model for arbitrarily censored and truncated data. *Biometrics*, 52:512–524, 1996.

[2] M. Asgharian, C.E. M'Lan, and D.B. Wolfson. Length-biased sampling with right censoring. *manuscript submitted to the Journal of the American Statistical Association*, 2001.

[3] S. Blumenthal. Proportional sampling in life length studies. *Technometrics*, 9(2):205–218, 1967.

[4] R. Brookmeyer and M.H. Gail. Biases in prevalent cohorts. *Biometrics*, 43:739–749, 1987.

[5] A. Cnaan and L. Ryan. Survival analysis in natural history studies of disease. *Statistics in Medicine*, 8:1255–1268, 1989.

[6] J.A. Correa and D.B. Wolfson. Length-bias: some characterizations and applications. *Journal of Statistical Computation and Simulation*, 64:209–219, 1999.

[7] D.R. Cox. *Some sampling problems in technology. In: New Developments in Survey Sampling.* Wiley-Interscience, New York, 1969.

[8] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[9] D.R. Cox and D. Oakes. *Analysis of Survival Data.* Chapman and Hall, London, 1984.

[10] O. Davidov and M. Zelen. Referent sampling, family history and relative risk: the role of length-biased sampling. *Biostatistics*, 2(2):173–181, 2001.

[11] T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analysis.* Wiley-Interscience, New York, 1991.

[12] T.L. Lai and Z. Ying. Estimating a distribution function with truncated and censored data. *The Annals of Statistics*, 19(1):417–442, 1991.

[13] G.P. Patil and C.R. Rao. Weighted distributions and sized-biased sampling with applications to wildlife populations and human families. *Biometrics*, 34:179–189, 1978.

[14] W.-Y. Tsai, N.P. Jewell, and M.-C. Wang. A note on the product-limit estimator under right censoring and left truncation. *Biometrika*, 74(4):883–886, 1987.

[15] B.W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295, 1976.

[16] Y. Vardi. Nonparametric estimation in the presence of length-bias. *The Annals of Statistics*, 10(2):616–620, 1982.

[17] Y. Vardi. Multiplicative censoring, renewal processes, deconvolution and decreasing density: Nonparametric estimation. *Biometrika*, 76(4):751–761, 1989.

[18] Y. Vardi and C.-H. Zhang. Large sample study of empirical distributions in a random-multiplicative censoring model. *The Annals of Statistics*, 20(2):1022–1039, 1992.

[19] M.-C. Wang. Nonparametric estimation from cross-sectional survival data. *Journal of the American Statistical Association*, 86:130–143, 1991.

[20] M.-C. Wang. Hazards regression analysis for length-biased data. *Biometrika*, 83(2):343–354, 1996.

[21] M.-C. Wang, R. Brookmeyer, and N.P. Jewel. Statistical models for prevalent cohort data. *Biometrics*, 49:1–11, 1993.

[22] M.-C. Wang, N.P. Jewell, and W.-Y. Tsai. Asymptotic properties of the product limit estimate under random truncation. *The Annals of Statistics*, 14(4):1597–1605, 1986.

[23] M. Woodroofe. Estimating a distribution function with truncated data. *The Annals of Statistics*, 13(1):163–177, 1985.

Appendix A

Glossary

Backward recurrence time: The time from initiation to entry for a subject identified in a prevalent manner. This differs from the truncation time of an individual which refers to the time from initiation to the start of the study *even for a subject who does not survive long enough to enter the study.*

Differential length-biased sampling: A type of bias induced by sampling the forward recurrence times differentially from two different backward recurrence time distributions.

Follow-up time scale: The time scale from entry to failure. An analysis on the follow-up time scale may be obligatory if the backward recurrence times are unobserved. It is also of interest in many prevalent treatment studies since in these studies the treatment is administered at entry.

Forward recurrence time (Follow-up time): The time from entry to failure for a subject identified in a prevalent manner.

Frailty selection: The unwanted deletion of "frail" subjects from a prevalent cohort.

Incident follow-up study: A study that identifies new cases from initiation as they occur, and follows them until failure or censoring.

Incident/Prevalent/Follow-up proportional hazards model: The proportional hazards model on the incident/prevalent/follow-up time scale.

Incident time scale: The time scale from an initiating event to a failure event. The incident time scale is of interest in natural history of disease studies, since in these studies one wishes to make a statement about the survival experience of an incident individual.

Multiplicative censoring: A type of informative censoring.

Naïve/Correct/Partially naïve approach: Three procedures which are possible when performing data analysis. The naïve approach does not recognize length-bias. The correct approach recognizes the length-bias in the data and accounts for it accordingly in the analysis. The partially naïve method also recognizes the length-bias, but does not account for it adequately in the analysis.

Natural history of disease study: A study concerned with the natural progression of a disease, usually under the assumption that subjects have not been administered a treatment that changes the disease course.

Onset confounding: Onset confounding refers to the confounding of the effect of a covariate on the relative risk with the effect of this covariate on the duration of infection. That is, two groups may appear to have different risks simply because they were infected at different times. This can only occur when the backward recurrence times are unknown.

Prevalent follow-up study: A study that identifies prevalent cases, that is, cases that experienced the initiating event before they were identified.

Prevalent time scale: The time scale from an initiating event to a failure event for those subjects identified in a prevalent manner. Aside from the follow-up

time scale, the prevalent time scale may also be of interest in a prevalent treatment study, if the backward recurrence times are observed, since, for example, the initiating event may be biologically defined.

Prevalent treatment study: A type of study where a treatment for the condition of interest is administered to prevalent cases at entry. This differs from a natural history of disease study since in prevalent treatment studies one is not interested in the natural progression of the condition.

Quasi-Stationarity: An assumption made on the conditional hazard of failure in prevalent treatment studies that is unrelated to the assumption of stationarity of onset times.

"Reversal" of the survivor functions: A *reversal* of the survivor functions of two groups has occurred if $S_1^U(t) \geq S_2^U(t) \; \forall \; t \geq 0$, but $S_1^B(t) \leq S_2^B(t) \; \forall \; t \geq 0$, or equivalently with the inequalities reversed.

Stationarity: In a medical setting stationarity means that the incidence of disease is uniform over time before the cases are identified. It can also refer to the stationarity of the underlying renewal process, when such a process is the focus of attention; this perspective is ignored in this thesis.

Truncation time: The time from initiation to the start of the study for an incident individual. This differs from the backward recurrence time in that all subjects have a truncation time, which may be smaller or larger than their failure time, even if they fail before the start of the study, and are thus not observed as prevalent cases.