Maynooth University
National University of Ireland Maynooth

Nonparametric multivariate survival analysis of activated lymphocyte cell fates.

Harry Tideswell

supervised by
Professor Ken R. Duffy

September 30, 2023

Nonparametric survival analysis of activated lymphocyte cell fates.

Harry Tideswell

Abstract

Upon challenge, lymphocytes multiply and diversify to combat the infection, however, the mechanisms that drive this process are not well understood. A theoretical model has been proposed to explain how a diverse selection of cell fates is achieved, the Cyton model [Hawkins et al, 2007, PNAS]. In that model the censorship caused by competing drives for lymphocytes to undergo certain fates results in complex correlations and impacts the observed distribution of times to cellular events. In [Duffy et al, 2012, Science] the competition hypothesis is tested for consistency with data collected using a novel experimental procedure. Through the implementation and development of a collection of multivariate nonparametric statistical techniques, we create a set of tools that can aid the study of competition hypotheses in biological systems. As a worked example these tools are used to study data collected for the experiments in [Duffy et al, 2012, Science] to challenge some of the underlying assumptions of their parametric analysis. As an additional illustration further unpublished data collected during the experiments is used to study the time at which B cells divide, die and differentiate when they have already undergone class switching, allowing us to address the question of a cell type dependent change.

Contents

Chapter 1

1 Introduction

In this chapter, we introduce important biological and statistical concepts and provide motivation for this thesis.

1.1 Motivation

The immune response is one of the most important processes within our body. Through a collection of tissues, cells and molecules the immune system is able to defend us from a wide variety of threats. Broadly speaking it can be split into two subsystems: the innate and adaptive [21].

Cells of the innate immune system recognise a broad collection of pathogen associated molecules [21]. This allows its many cells, including Natural Killer cells and Macrophages, to rapidly respond to invading pathogen. The innate immune system also provides a key role in alerting and supporting the adaptive immune response via communication through molecules known as cytokines.

In contrast, the adaptive immune response is a highly specialised system, in which just a small number of individual B and T cells are capable of recognizing a specific threat. As an example, it has been estimated in a mouse that between 20 and 200 of its 40 million T helper cells can recognise a specific threat [36]. Specialisation comes from receptors found on cell surfaces and the antibodies that certain cells secrete.

Another key part of the adaptive immune system is its ability to form immunological memory, allowing a much faster response to a pathogen the host has already encountered. This process of immunological memory is the basis of immunisation. A typical pattern of the adaptive immune response is shown in Table 1.

Next we will describe B and T cells in more detail, as well as what happens during an adaptive immune response. B cells develop in the bone marrow; it is during this time that the process responsible for their near unique specificity takes place. V(D)J recombination is unique to B and T cells: different gene segments (known as variable (V), diversity (D) and joining (J)) are rearranged and selected in order to generate a highly diverse set of B cell receptors (BCRs) and T cell receptors (TCRs) [21]. Each cell will have a huge number of copies of just one BCR/TCR type on its surface. To prevent the binding of B cells to harmless antigens naturally found in the body, B cells are screened for auto-reactive BCRs in the bone marrow through a process known as central tolerance [21]. Central tolerance involves the binding of immature B cells to potential self-antigens in the bone marrow. If the binding is successful the B cell will not leave the bone marrow and will die by apoptosis. After cells leave the bone marrow, further checking for autoreactivity is done through the process of peripheral tolerance [21]. B cells then

B Cell Subsets	
Name	**Description**
Plasma cell	Long lived cells that secrete antibodies.
Memory	Long lived cells that will stay in the body after the infection has been cleared. Should the body be infected by the same antigen, these cells will rapidly proliferate to respond to the antigen.
T Cell Subsets	
Name	**Description**
Cytoxic (killer)	Also known as CD8+ T cells as they express CD8 protein on their surface. They destroy cells that have been infected with viruses or other pathogens.
Helper	Also known as CD4+ T cells, as they also express CD4 protein on their surface. By providing expansion or suppression signals, they assist other cells during the immune response.
Memory	Long lasting cells that will stay in the body after the infection has been cleared. Should the body be infected by the same antigen, these cells will rapidly proliferate to respond to the antigen. Memory T cells can either be CD4+ or CD8+.

Figure 1: Brief description of the properties of some important B and T cell subsets [21]. Many of the subsets listed here can be further split into subsets, and have many extra functions.

migrate to different parts of the body awaiting their cognate antigen.

T cells mature in the thymus, where they develop from immature Thymocytes [21]. Like BCRs, the TCRs of a T cell undergo a screening process, including positive selection to determine whether they will bind to 'self' molecules, and negative selection to determine their ability to bind correctly to peptide coupled with major histocompatibility complex (MHC) expressed on the surface of cells. During this time they will become either CD4+CD8- or CD4-CD8+. Both CD4 and CD8 are important co-receptors that play a role in the immune response, and which of the two a given T cell expresses will determine much of its behaviour. Table 1 gives more information on the behaviour of different T cell subsets.

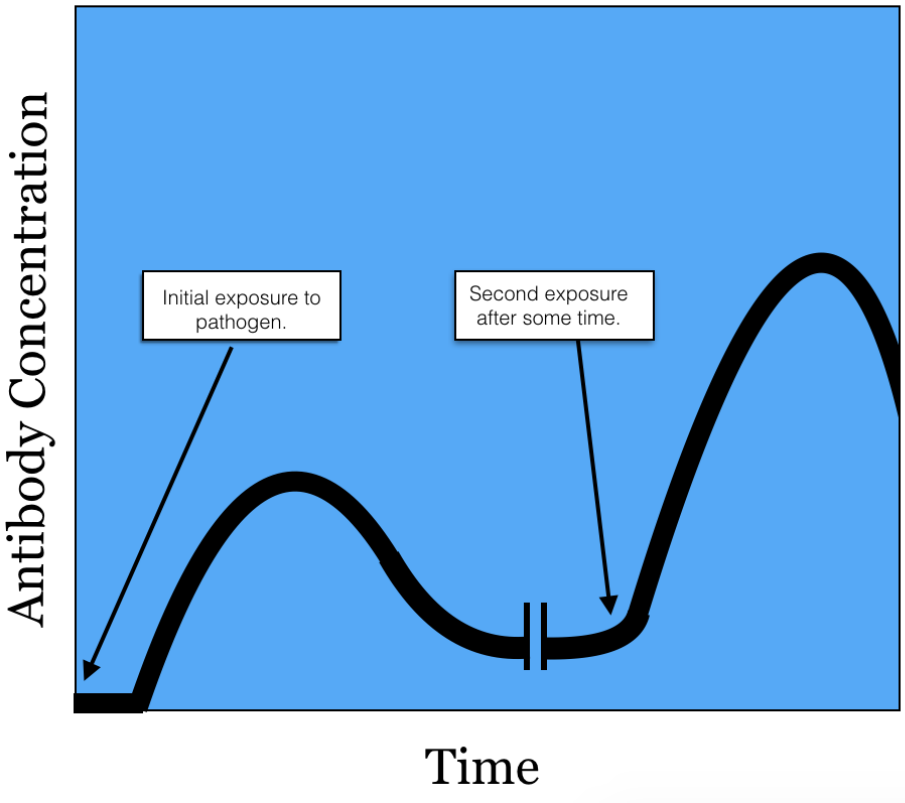


Figure 2: Typical pattern of an humoral immune response (schematic - not data) [21]. Once a pathogen enters the body and encounters a matching B cell, proliferation and differentiation occur leading to an increased number of antibodies specific for that threat. A peak in antibody concentration occurs at around day 7, after which cells begin to die by apoptosis. A small number of memory cells specific for that pathogen will remain in the body in case of reinfection. If this pathogen is encountered again, memory cells will be able to rapidly respond and defeat the pathogen. This is shown by the second peak, which occurs sometime after the initial infection has been cleared. During the secondary response there are more cells that can respond to the antigen initially [21], therefore they will produce a greater concentration of antibodies over the same amount of time.

There are two ways in which B cells can be stimulated to produce an immune response: T cell dependent activation, and T cell independent acti-

Receptor class	Description
IgM	Found on the surface of B cells. Primarily active during the early part of the immune response.
IgG	The most common antibody class found in humans. Can be split into 4 different subsets, IgG1, IgG2, IgG3 and IgG4.

Table 1: Different cell receptor types and their function.

vation. T dependent activation requires B cells to receive two distinct signals. The first occurs when the BCR binds to its cognate antigen; this antigen is then internalised within the cell and digested. Fragments of the antigen are displayed on the cells surface, bound to an MHC class II molecule. Upon binding of a T-Helper cell with the antigen-MHC class II complex, the T-Helper cell will express protein CD40L, which binds to the B cells CD40 receptor causing B cell to become activated. T cell independent activation can occur with certain antigens, for example CpG DNA. These antigens are capable of making the B cell activate without engaging its BCR by binding to a Toll-like receptor (TLR). For example CpG DNA is recognised by TLR 9, which can motivate a T independent response.

After activation, B cells begin a process of proliferation and differentiation, as well as antigen-specific antibody secretion and antibody class switching. During antibody class switching, the constant region of antibodies changes. Different antibody types perform different functions critical for an effective immune response. The process of B cell activation is shown graphically in figure 3. Once activation occurs, a diverse collection of B and T cell subsets develop in order to fight the invading pathogen.

T cell activation occurs when a TCR meets its cognate antigen. It also must receive a second co-stimulatory signal. The costimulatory signal can be provided by the antigen presenting cell (APC) that the T cell is bound to by engagement of its CD27 or CD28 receptor or through the provision of secreted cytokines such as IL-2. From here, activated T cells will divide and release cytokines to motivate more T cells; cytoxic T cells will track down pathogen, destroy them and memory T cells will be formed for future tolerance in case of reinfection.

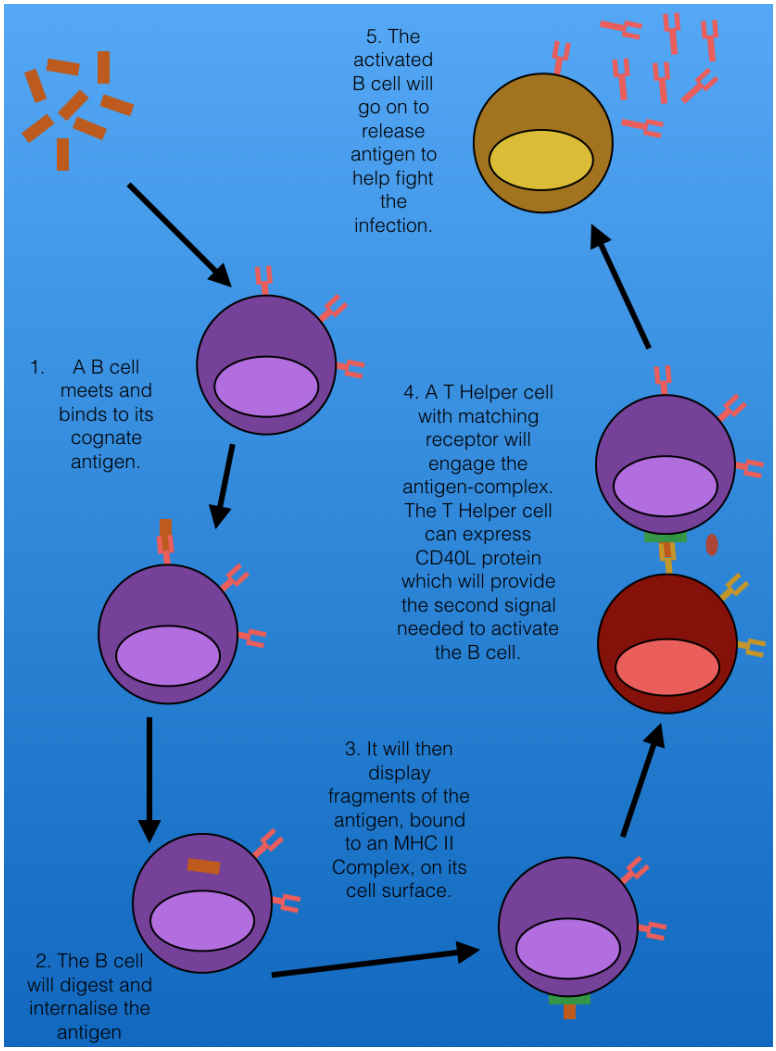


Figure 3: Graphical explanation of B cell activation. This image is simplified for clarity, and is just a small part of the overall adaptive immune response.

With the use of important experimental techniques, including fluorescent dyes [5] and cell sorting technology [21], it is possible to observe B and T cells during a simulated immune response in vitro.

Data collected in these experiments reveals consistent features across both B and T cells [23] [9] [17]. We see a diverse collection of cell types and fates, as well as a broad distribution of times to reach those fates [9] even when regulating agents (including cytokines or interactions between cells)

are removed [9]. From this we understand that internal mechanisms within the cell are contributory factors to the observed heterogeneity [9].

The studies mentioned above, and others, have also noted that despite heterogeneity at the single cell level, the population level response is robust and consistent across multiple experiments [10]. Given the diverse range of cell types observed during an immune response, and the high variability in time to fate, it is important to reconcile these two opposing views. How can heterogeneity at the single cell level, lead to a robust and consistent immune response at the population level?

A model known as the Autonomous Competition Hypothesis (ACH) [9] attempts to answer this question. Not only does it account for the cell level variability, it also models the complex correlation patterns found at both the intra and inter cellular levels. The ACH describes a scenario in which each cell contains independent processes which, through activation or beginning at birth, determine when a cell will undergo division, death, class switching or differentiation. The time at which each process will occur is given by a probability distribution, with cell fates governed by a competition in which the process having the smallest time will win. Competition-based models have been used to study multiple biological systems, including sporulation vs competence in bacillus subtilis [30], and time-to-fate in B cells [10]. In this thesis we aim to study the ACH further by developing multivariate non-parametric statistical tools to aid its investigation.

1.2 Data Source

The data source for this thesis comes from an extended data set of that published in the 2012 paper [10]. Using a collection of experimental techniques the authors of [10] collected the times to death, division, differentiation to plasmablast and IgM to IgG1 class switch of B cells activated in vitro. Here we discuss how the data was collected, and the experimental challenges that needed to be solved.

To collect data that is suitable for testing the ACH, it is necessary to track the time at which each fate occurs in every cell. Through direct observation and use of fluorescent proteins, it was possible to optically track individual B cells and observe a time for the occurrence of several fates. To determine when differentiation occurred, B cells from Blimp1-GFP reporter mice were used. Blimp1 is a transcription factor required for differentiation to plasmablast [25], and so through the use of these specific mice, where a

green fluorescent protein is made every time Blimp1 is made, differentiation can be observed by noting green fluorescence in a cell. To identify IgM to IgG1 class switching, anti-IgG1-APC is added, which fluoresces in the red spectrum when switching from IgM to IgG1 occurs. The beginning of cytokinesis indicates division time, and the start of membrane rupture indicates cell death [10].

Naive murine B cells stimulated in vitro with either LPS or anti-CD40 aggregate and so cannot be optically tracked to create full family trees. Instead B cells were stained with CTV, stimulated with anti-CD40, IL-4 and IL-5, then FACS sorted from generations 0, 2, 4 an 6, and deposited into micro wells at a density such that many wells were singly seeded. This gets around the problem caused by aggregation of cells, however one can only follow the cells through one round of division as further cells cannot be uniquely identified past this point. Because of this limitation the data consists of times-to-fate for siblings cells and no further into the family tree.

Using anti-CD40, as well as cytokines IL-4 and IL-5, a T dependent immune response can be simulated. IL-4 is known to stimulate B cell differentiation into plasmablasts, IL-5 stimulates B cell growth and anti-CD40 will bind to the CD40 receptor of B cells, short-cutting the B-T cell interaction required for T dependent activation. A full account of the experimental procedure can be found in the supplementary material of [10].

1.3 Statistical Methodologies

Motivated by the earlier discoveries including [17], the analysis of B cell data in [10] assumes the existence of four independent random variables per cell, denoted $T_{diff}, T_{die}, T_{switch}, T_{div}$. These random variables describe the time, from cell birth, at which differentiation to plasmablast, death, IgM to IgG1 class switching and division will occur in a given cell. However, as B cells have been observed to not always undergo some processes, they assume that all events excluding death have a positive probability of being infinite. As described by the ACH, some of these variables can be censored by each other, and are in competition to be the first to occur. For example, if a certain cell has $T_{div} < T_{die}$ we would observe division, and the death event is not observed; it is said to be censored. These random variables are independent of each other, as noted by the ACH. However, the observed distribution that has been modified by the competition and censorship processes can be correlated.

They [10] then define pairs of random variables to describe siblings, for example (T^1_{div}, T^2_{div}). While pairs of random variables can be correlated with each other, a cell's individual random variables are independent. Each pair is assumed to be described by a bivariate log-normal distribution with symmetric marginals, giving the model 15 parameters representing the means, variances, correlations and probability of infinity respectively,

$$\theta = \{\mu_{div}, \mu_{die}, \mu_{switch}, \mu_{diff}\, \sigma^2_{div}, \sigma^2_{die}, \sigma^2_{switch}, \sigma^2_{diff},$$
$$\rho_{div}, \rho_{die}, \rho_{switch}, \rho_{diff}, p_{div}, p_{switch}, p_{diff}\}.$$

For a given data set D, containing the fates of all sibling within a given generation, Matlab was used to numerically solve the maximum likelihood problem $\max_\theta L(D|\theta)$. Further information on the modelling procedure can be found in the supplementary material of [10].

1.4 Contribution overview

The primary contribution of this thesis is the implementation, modification and use of non-parametric statistical techniques to provide an alternate analysis of the B cell data set [10]. These techniques do not assume that time-to-event is a member of a particular class of distributions. Using unpublished primary data collected during the [10] experiments, we provide further analysis of B cells born as IgG1+ or ASC+ (Antibody-Secreting Cell).

The statistical techniques come from a branch of statistics known as survival analysis. Survival analysis [24] is concerned with the study of random variables describing the time at which a particular event occurred. For example, it is quite often applied in clinical trials to describe the time of death or cure.

The tools we borrow, implement and adapt from survival analysis include the Kaplan-Meier estimator [26], the Dabrowska estimator [7], and the Log-Rank test [32]. When applied to the B cell data set, and assuming independence between event and censoring, the Kaplan-Meier estimator allows us to calculate the probability of survival beyond a given time for a specific event of interest. Figure 4 shows an example. This gives us an empirical estimate of the uncensored distribution of time-to-event, instead of the observed distribution resulting from competition [9] as described by the ACH.

The Dabrowska estimate is a multivariate generalisation of the Kaplan-Meier estimator which allows non-parametric estimation of the joint distribution between two paired events; from this we can further understand and quantify the intra and inter-cellular correlations between a pair of sibling B cells. Unlike the Kaplan-Meier estimator, the Dabrowska estimator is not available in most standard statistical packages (for example Matlab or R) and so in this case we developed our own implementation, which we adapted to deal with assumed symmetric distributions. An example of the Dabrowska survival function estimate applied to B cell data is shown in Figure 5.

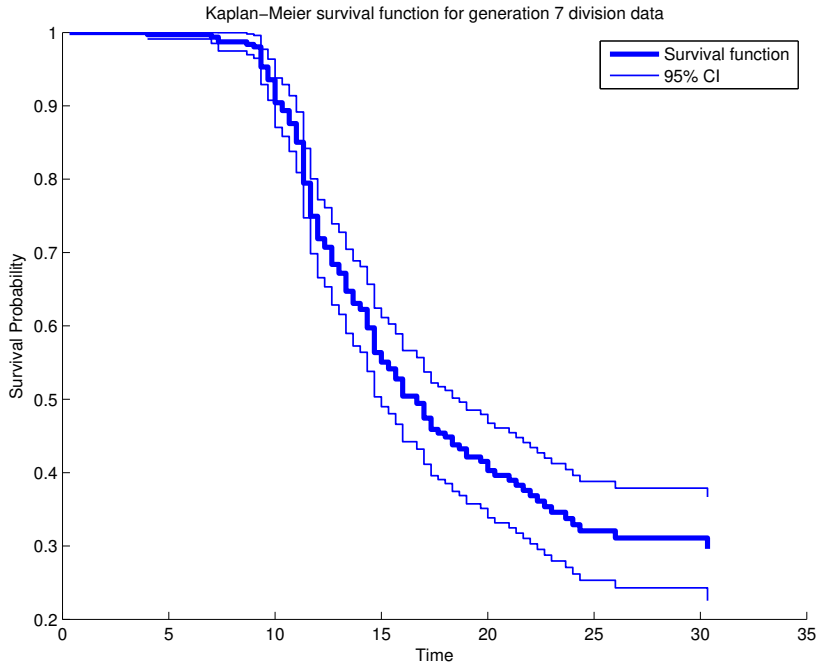


Figure 4: Kaplan-Meier estimate of the survival functions for generation 7 B cell data [10]. This curve shows the probability that differentiation to plasma blast will occur after the specified time. Created using the Matlab ECDF function.

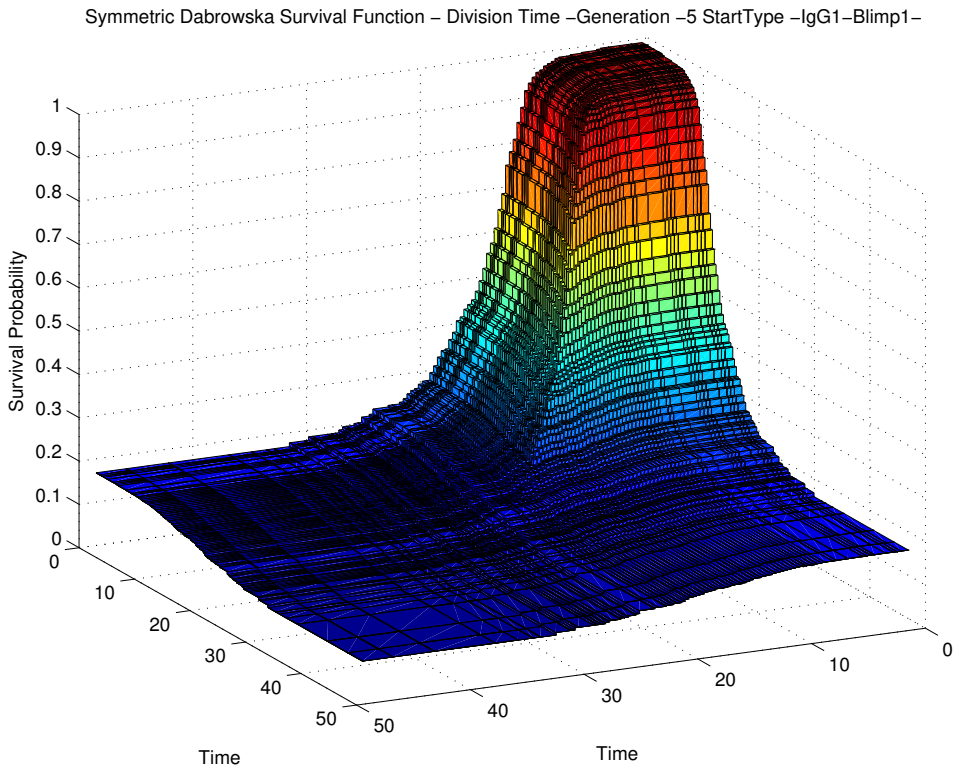


Figure 5: Symmetric modified Dabrowska estimate using B cell data from [10]. This surface shows the joint survival distribution for time to divide for a pair of sibling B cells. Created using a MATLAB implementation for this thesis.

We can use statistical tests, including the Log-Rank test, to provide comparisons between the parametric techniques used in [10] and the survival analysis techniques developed in this thesis. We can then determine how reasonable the parametric techniques used in the model are. For example [8] notes that the time to divide distribution of B cells, should follow a Log-Normal distribution; since our techniques do not follow a specified distribution, we can test if this is a good assumption. Furthermore, the time to class switch and differentiate had not been measured before, and so this work serves as a good test for the type of distribution it should follow.

One interesting feature of the analysis we have performed in this thesis takes advantage of assumed symmetric properties of the underlying model to improve estimation power. Assuming asymmetric cell division is not at

work, the sibling cell data collected in [10] have no inherent rank order (i.e. there is no 'first' cell to undergo an event between the siblings). Leveraging this assumption allows us to create a symmetric Dabrowska estimate, $sym-Dabrowska$. This allows us to create a more accurate survival distribution from a given data set.

We produce a further analysis using some of the B cell data not published in [10], but kindly provided by the Hodgkin Lab (WEHI). This allows us to make interesting estimates of the time-to-fate in B cells that have already undergone certain processes. For instance what does the time to division distribution look like in cells that have already undergone class switching or differentiation? Does it differ from the un-switched or un-differentiated distribution? We know that cells differentiated are more likely to stop dividing [44].

A recent paper [6] questions the use of some survival analysis techniques, for example using the Kaplan-Meier estimate, when studying time to event data for cells. This is because of the possibility of competing risks. Gooley et al [15] define a competing risk as, *'An event whose occurrence either precludes the occurrence of another event under examination or fundamentally alters the probability of occurrence of this other event'*. An assumption of Kaplan-Meier analysis is non-informative censoring [33], this is where the censoring event occurs independently of the outcome of interest. For example if we are performing an analysis of time to isotype switch in B-Cells and there are cells that have not isotype switched at the end of the study then the censoring is non-informative and Kaplan-Meier analysis is suitable. Events such as death or division are also censoring events but are not independent and would be considered a competing risk. As a result we wish to make the reader aware that Kaplan-Meier analysis can provide an overestimated probability of time-to-event. However, in the present thesis, we do not address this matter as, motivated by earlier studies, our aim was to create the tools necessary to perform non-parametric survival analysis for these data.

Finally, with the collection of statistical tools we have developed, we can create a package that, in the future, could be used for similar experiments that apply a competition based hypothesis. For example, the time to sporulation data in [30] or, indeed, for data in other fields.

Chapter 2

2 Statistical Review

In this chapter we review important statistical definitions and techniques used in the rest of the thesis. We introduce the survival function estimators, with numerical examples to show their use. We present the univariate Kaplan-Meier estimate and the multivariate Dabrowska estimate. Lastly we discuss the different hypothesis tests needed to compare survival functions, and modifications that are needed.

The statistical tools used in this thesis come from a branch of statistics known as survival analysis. Broadly speaking survival analysis is a collection of statistical approaches in which the main concern is the duration of time until an event occurs. This event is normally referred to as a failure, and the time of its occurrence a failure time. For example in engineering the tools of survival analysis may be used to determine the lifetime of an electrical component.

Usually an experiment will be performed in which n independent objects are observed from a starting time t_0 to an end time t_m. During this time the objects will be observed until a failure occurs, this could be for example, the breakdown of an electrical component. Sometimes events occur which obscure the observance of the failure event. These are known as censoring events.

We denote t_i as the time at which failure would occur if there were no censoring, and c_i the time at which censoring occurs assuming the failure event has not yet occurred. If only one of the two events can happen, then we observe an event at $\min(t_i, c_i)$ and assuming it is possible to distinguish between failure and censoring, we use the indicator function defined as,

$$\mathbb{I}(t_i < c_i) = \begin{cases} 0 & \text{if } c_i \leq t_i \\ 1 & \text{if } t_i < c_i, \end{cases} \tag{1}$$

to record which event occurred. In summary, a pair of observable variables for the ith object is given by,

$$(y_i, \delta_i) = (\min(t_i, c_i), \mathbb{I}(t_i < c_i)), \tag{2}$$

If we were to observe an electrical component and at the end of the study it is still working we would say the event was right censored. We don't know what time the electrical component will break, just that it is sometime after t_m. The different types of censoring are shown graphically in Figure 6.

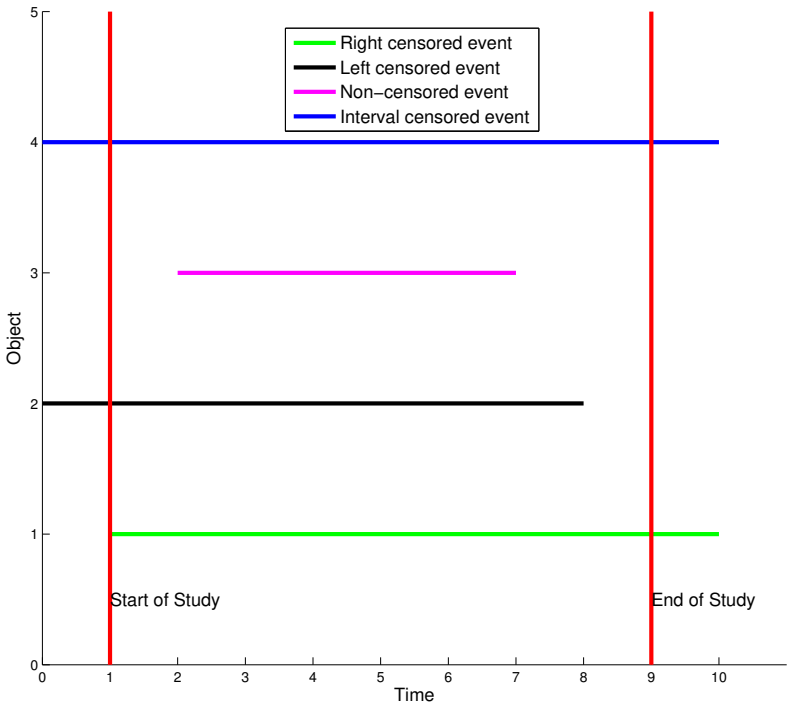


Figure 6: Different types of censoring. This figure could represent the testing of the lifetime of electrical components. At the start of the study left censoring could occur in components that were broken sometime before the study began, and so we only know the approximate time of failure. If an electrical component is still functioning at the end of the study then right censoring has occurred, we know that component will eventually fail but can only say at best that it will happen sometime after the end of the study.

When using survival analysis techniques to study time-to-event data, two functions are commonly estimated to describe the data, the survival function and the hazard function. We define these below.

Let T be a nonnegative random variable representing the time at which failure occurs. This could be the time, from birth, for a cell to divide or the time for a patient to become disease free after administration of medication.

Definition The Survival Function of a random variable T is its complementary cumulative distribution function, denoted $S(t)$. It is a monotonically

decreasing function that describes the probability T takes a value greater than t,

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) \qquad t \in [0, \infty), \tag{3}$$

where $F(t)$ is the cumulative distribution function.

Definition The hazard function, $\lambda(t)$, describes the instantaneous rate failure, conditioned upon survival up to that time [24]. Assuming the random variable T has probability density function (PDF) $f(t)$, it is defined to be,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{S(t)}.$$

A useful related function is the cumulative hazard function, $\Lambda(t)$, given by the integral of $\lambda(t)$ over the length of study $[t_0, t_m]$,

$$\Lambda(t_m) = \int_{t_0}^{t_m} \lambda(t) dt.$$

The cumulative hazard function gives the total hazard over the length of study $[t_0, t_m]$. It is useful when determining how many failures we would expect in a time interval. It also has an important relationship with the survival function and the two functions can be written in terms of each other.

If we define the survival function using product integral notation [14]. For a function $A : (0, t] \to \mathbb{R}$ that is right continuous everywhere with left limits,

$$\prod_{x \in (0,t]} (1 + A(dx)) = \lim_{\max_i |u_i - u_{i-1}| \to 0} \prod_i (1 + (A(u_i, u_{i-1}])), \tag{4}$$

where $0 = u_0 < u_1, ..., < u_n = t$ is a partition of the time interval $(0, t]$. We can express the survival function in terms of the cumulative hazard function [14] as,

$$S(t) = \prod_{s<t} (1 - \Lambda(ds)). \tag{5}$$

2.1 Data sets

In order to perform calculations based on the failure time data (y_i, δ_i) we will briefly explain some notation, and a reformatting of the data set that will be useful for later proofs and calculations. We have a data set where the

superscript (1) indicates univariate observations and the subscript n indicates the number of elements in the data set,

$$D_n^{(1)} = \{(y_i, \delta_i)\}_{i=1}^n, \tag{6}$$

where y_i is the event time and δ_i takes the value 1 if a failure occurred, and 0 if right censoring occurred.

For the purpose of performing calculations we will reformat the data. We define a set $\{x_i\}_{i=1}^m$ to contain the m unique elements of the set $\{y_i\}_{i=1}^n$. At each of the unique time points we define a new set of observations given by,

$$\mathcal{D}_m^{(1)} = \{(x_i, d_i, e_i, r_i)\}_{i=1}^m, \tag{7}$$

where $d_i = \sum_{j=1}^n (1-\delta_j)\mathbb{I}_{y_j=x_i}$ records the number of failures that occurred at x_i, $e_i = \sum_{j=1}^n \delta_j \mathbb{I}_{y_j=x_i}$ is the number of censored events that occurred at x_i and finally $r_i = \sum_{j=i}^m (d_j + e_j)$ is the number of objects that were at risk just before x_i.

2.2 Univariate Kaplan-Meier estimate

The Kaplan-Meier [26] estimator is a nonparametric survival function estimate used in cases of censored data where the time to censoring events are assumed to be i.i.d and independent of the assumed i.i.d times to failure events. When censoring is not observed it reduces to the complement of the empirical distribution function (EDF) of the failure time random variable. Given a data set $\mathcal{D}_m^{(1)}$ as in equation 7 the Kaplan-Meier estimate of the survival function is given by,

$$\hat{S}(x) = \prod_{\{i:x_i<x\}} \left(\frac{r_i - d_i}{r_i}\right). \tag{8}$$

Below is a short proof showing that given the data set $\mathcal{D}_m^1$ of the form shown in equation 7 the maximum likelihood estimate (MLE) of the survival function S for T is given by the Kaplan-Meier formula in equation 8 [24]. What follows is an expanded version of the proof presented in [26].

Proof. The likelihood of the data, $\mathcal{D}_m^{(1)}$ given S is,

$$L(\mathcal{D}_m^{(1)} \mid S) = \prod_{i=1}^m [S(x_{i-1}) - S(x_i)]^{d_i} S(x_i)^{e_i},$$

Now consider the following. For a given x_i,

$$\begin{aligned} S(x_i) &= \mathbb{P}(T > x_i) = \mathbb{P}(T > x_i \mid T > x_{i-1})\mathbb{P}(T > x_{i-1}) \\ &= \mathbb{P}(T > x_i \mid T > x_{i-1})S(x_{i-1}), \end{aligned}$$

if we denote,

$$\pi_i = \mathbb{P}(T > x_i \mid T > x_{i-1}),$$

it allows us to write,

$$S(x_i) = \prod_{j=1}^{i} \pi_j,$$

inserting (6) into (5) we can rewrite L in terms of π_i,

$$\begin{aligned} L(\mathcal{D}_m^{(1)} \mid S) &= \prod_{i=1}^{m} [\prod_{j=1}^{i-1} \pi_j - \prod_{j=1}^{i} \pi_j]^{d_i} (\prod_{j=1}^{i} \pi_j)^{e_i} \\ &= \prod_{i=1}^{m} [\prod_{j=1}^{i-1} \pi_j (1-\pi_i)]^{d_i} (\prod_{j=1}^{i-1} \pi_j)^{e_i} \pi_i^{e_i} \\ &= \prod_{i=1}^{m} [(1-\pi_i)^{d_i} \pi_i^{e_i}] (\prod_{j=1}^{i-1} \pi_j^{d_i+e_i}) \\ &= \prod_{i=1}^{m} (1-\pi_i)^{d_i} \pi_i^{r_i-d_i}, \end{aligned}$$

A useful technique when trying to maximise likelihood functions is to take the logarithm of $L(\mathcal{D}_m^{(1)} \mid S)$ which will give the same result as if we'd tried to maximise L since the log function is a strictly monotonically increasing function. This gives,

$$\mathcal{L}(\mathcal{D}_m^{(1)} \mid S) = \log L(v \mid S) = \sum_{i=1}^{m} [d_i \log[(1-\pi_i)] + (r_i - d_i) \log(\pi_i)].$$

We can then differentiate this with respect to π_i and set it equal to zero in order to maximise the function.

$$\frac{\partial \mathcal{L}(\mathcal{D}_m^{(1)} \mid S)}{\partial \pi_i} = \frac{d_i}{\pi_i} - \frac{r_i - d_i}{1-\pi_i} = 0,$$

rearranging this leaves,

$$\pi_i = \frac{r_i - d_i}{r_i} = 1 - \frac{d_i}{r_i},$$

giving us the Kaplan-Meier estimate,

$$\begin{aligned}\hat{S}(x) &= \prod_{\{i:x_i<x\}} \left(\frac{r_i - d_i}{r_i}\right), \\ &= \prod_{\{i:x_i<x\}} \left(1 - \frac{d_i}{r_i}\right).\end{aligned}$$

.

□

We can use the estimated survival function to calculate the mean survival time, $\mathbb{E}(T)$. If $\hat{S}(x_m) = 0$, where x_m is the time of the last observation, then we can define the expectation of T as,

$$\int_0^\infty \hat{S}(x)dx = \mathbb{E}(T) \approx \sum_{i=1}^{m}(x_i - x_{i-1})\hat{S}(x_i).$$

In a set of censored time-to-event data, if the last time point of observation is censored, then the Kaplan-Meier estimate of the survival distribution will not reach zero. This leads to the above integral being infinite. In order to fix this problem Efron [27] suggests setting $\hat{S}(x) = 0$ beyond the last time point. This is the same as saying that the final survivor would die straight after the censoring occurred. Gill [27] suggests estimating $\hat{S}(x)$ by $\hat{S}(x_m)$ for all times greater than x_m. This corresponds to the idea that some of the objects under study will not undergo the event of interest. This problem is discussed further in [34].

2.2.1 Synthetic data example

Here we use synthetic data to show how the Kaplan-Meier estimate is calculated, and numerically demonstrate its important features. Table 2 below uses a set of observations $(t_i, c_i)_{i=1}^n$ to produce a Kaplan-Meier estimate. Figure 7 shows four different Kaplan-Meier estimates based on different changes

to the data in Table 2. Figure 7(a) shows the usual Kaplan-Meier survival function based on the data. Figure 7(b) shows the survival function estimate when all censoring events are changed to failure events; in this case the Kaplan-Meier estimate returns the empirical distribution function. Figure 7(c) shows a defective distribution caused by changing the data so that the last event is a censored one. Figure 7(d) shows convergence as the amount of data used in the estimate increases.

t_i	c_i	$y_i = \min(t_i, c_i)$	$\delta_i = I(t_i < c_i)$	r_i	d_i	$\frac{n_i - d_i}{n_i}$	$\hat{S}(t)$
17.1507	13.6501	13.6501	0	10	0	1.0000	1.0000
22.3355	18.0349	18.0349	0	9	0	1.0000	1.0000
5.9646	15.7254	5.9646	1	8	0	1.0000	1.0000
18.4487	14.9369	14.9369	0	7	1	0.8571	0.8571
16.2751	15.7147	15.7147	0	6	1	0.8333	0.7143
9.7692	14.7950	9.7692	1	5	1	0.8000	0.5714
13.2656	14.8759	13.2656	1	4	0	1.0000	0.5714
16.3705	16.4897	16.3705	1	3	1	0.6667	0.3810
29.3136	16.4090	16.4090	0	2	1	0.5000	0.1905
26.0777	16.4172	16.4172	0	1	1	0	0

Table 2: Synthetic data generated from a normal distribution. Using this data we calculate the observable variables described above, the risk set r_i, the number of events d_i that occur and the Kaplan-Meier estimate $\hat{S}(t)$ for each t_i.

Kaplan-Meier estimates using synthetic data.

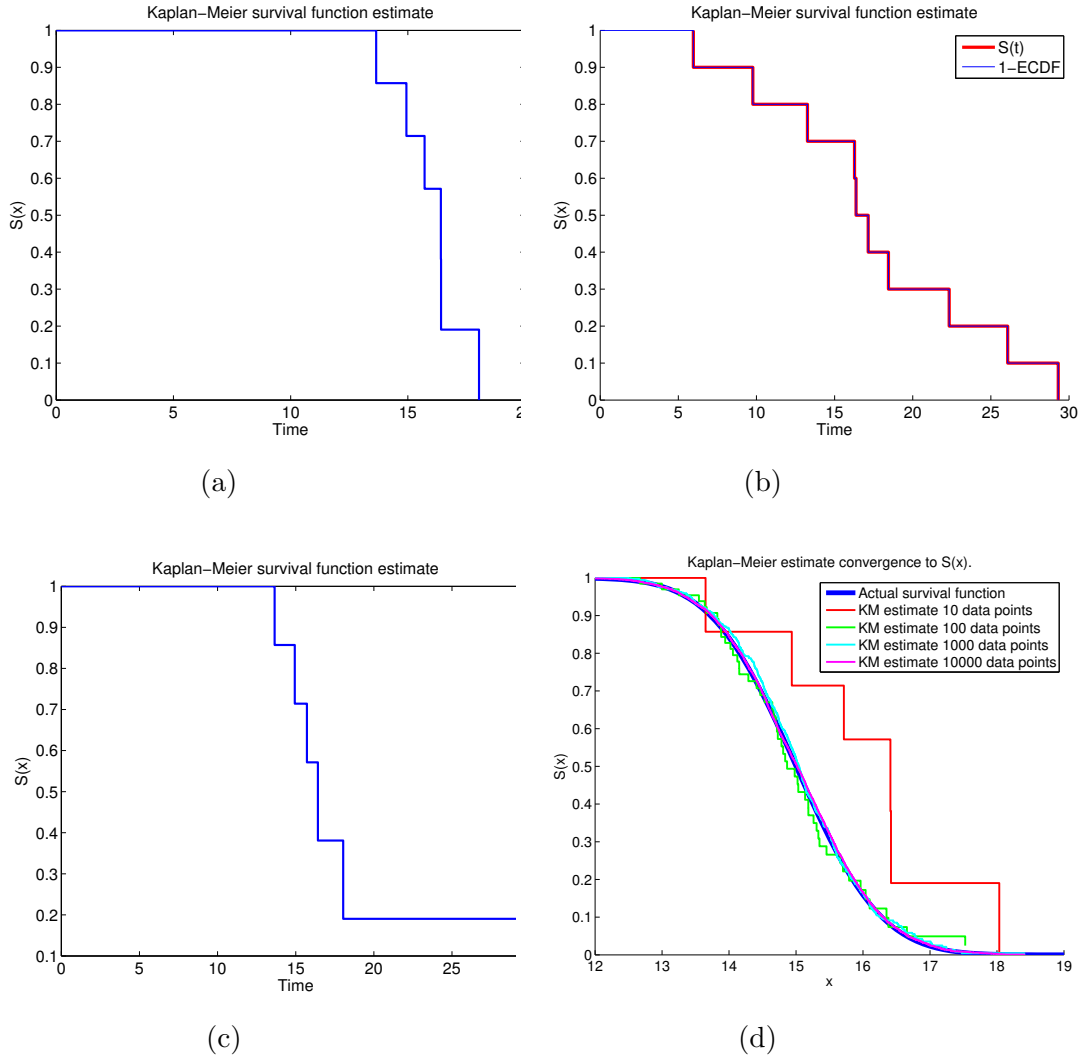


Figure 7: Kaplan-Meier estimates using synthetic data in a variety of cases. (a) Standard plot based on the data in Table 2. (b) Estimate with no censored observations reduces to the complement of the empirical distribution function. (c) Defective distribution estimate when final time point censored. (d) Convergence to true survival function as amount of data increases.

2.3 Bivariate Dabrowska estimate

Suppose that we would like to study the failure times in a pair of objects whose time to failure could be correlated. For example the recovery time of eyes after surgery. Let $T = (T_1, T_2)$ be a pair of nonnegative random variables representing the time to failure of each of the objects in the pair. The time of these failure events are then subject to the possibility of independent censoring, described by random variables $C = (C_1, C_2)$, which may be correlated. Given n observations of censored pairs of random variables in a data set where the superscript (2) indicates bivariate observations,

$$D_n^{(2)} = \{(\min(t_{1i}, c_{1i}), \min(t_{2i}, c_{2i}), \mathbb{I}(t_{1i} < c_{1i}), \mathbb{I}(t_{2i} < c_{2i}))\}_{i=1}^n \tag{9}$$

$$= \{(y_{1i}, y_{2i}, \delta_{1i}, \delta_{2i})\}_{i=1}^n \tag{10}$$

we want to estimate the bivariate survival function,

$$S(t_1, t_2) = \mathbb{P}(T_1 > t_1, T_2 > t_2).$$

If (T_1, T_2) are independent and so are (C_1, C_2) then the Dabrowska estimate is not needed since we would have,

$$S(t_1, t_2) = \mathbb{P}(T_1 > t_1, T_2 > t_2) = \mathbb{P}(T_1 > t_1)\mathbb{P}(T_2 > t_2),$$

which can be estimated with Kaplan-Meier formula. The Dabrowska estimator [7] is a multivariate generalisation of the Kaplan-Meier estimate that allows us to deal with pairs of failure times subject to censoring. Here we will look at the bivariate case, using this we can study the time-to-event relationship in pairs of sibling B cells.

Unfortunately there is no nonparametric maximum likelihood estimator of the bivariate survival function for censored data [7] [40]. The likelihood function cannot be written in a unique way. Instead survival function estimators have been proposed based on their useful properties. Many bivariate survival function estimators have been suggested with varying strengths and weaknesses, including the Yin-Ling [31], Prentice-Cai [39], Pruitt [41] and Van der laan [45] estimators.

Dabrowska's [7] is one of the most cited estimators of a bivariate survival function for censored data. It overcomes many of the problems other proposed estimators have. It is unique, consistent and converges weakly [7]. It can also be extended to higher dimensions. More details on this can be found in [7].

However it does have some drawbacks. Under certain conditions it will produce an estimate with negative probability, a problem that does not go away as the sample size becomes larger [40]. Further information on Dabrowska estimate and comparisons with other estimators can be found in [39] [7].

The starting point for the Dabrowska estimate comes from a definition of the survival function in terms of a product integral, by Gill and Johansen [14]. In [14] it is shown that the survival function can be defined by a product integral of the cumulative hazard function, as presented in equations 4 and 5. Dabrowska shows that a bivariate survival function can be defined in terms of a suitable set of cumulative hazard functions.

Using paired failure time random variables (T_1, T_2) and without yet considering censoring, Dabrowska shows that the survival function can be written using the following three cumulative hazard functions,

$$\Lambda_{11}(dt_1, dt_2) = \frac{P(T_1 \in dt_1, T_2 \in dt_2)}{P(T_1 \geq t_1, T_2 \geq t_2)},$$
$$\Lambda_{10}(dt_1, t_2) = \frac{P(T_1 \in dt_1, T_2 \geq t_2)}{P(T_1 \geq t_1, T_2 \geq t_2)},$$
$$\Lambda_{01}(t_1, dt_2) = \frac{P(T_1 \geq t_1, T_2 \in dt_2)}{P(T_1 \geq t_1, T_2 \geq t_2)},$$

and a term known as the L-measure,

$$L(dt_1, dt_2) = \frac{\Lambda_{10}(dt_1, t_2^-)\Lambda_{01}(t_1^-, dt_2) - \Lambda_{11}(dt_1, dt_2)}{(1 - \Lambda_{10}(dt_1, t_2^-))(1 - \Lambda_{01}(t_1^-, dt_2))},$$

as,

$$S(t_1, t_2) = \prod_{u \leq t_1} (1 - \Lambda_{10}(du, 0)) \prod_{v \leq t_2} (1 - \Lambda_{01}(0, dv)) \prod_{(u,v) \leq (t_1, t_2)} (1 - L(du, dv)), \tag{11}$$

where the first two terms are the respective univariate survival functions of T_1 and T_2.

Dabrowska then uses this representation of the bivariate survival function, showing that each term can be estimated from a set of bivariate right censored data, and goes on to show that it has desirable properties. A formal proof of this can be found in [7]. Figure 8 shows the process of estimating the bivariate survival function from a data set in the presence of right censoring.

Dabrowska estimate algorithm.

Step 1. Take the data set $(t_{1i}, t_{2i})_{i=1}^n$ and $(c_{1i}, c_{2i})_{i=1}^n$, and convert to observables as shown by equation 7.
Step 2. Compute $S(t_1, 0)$ and $S(0, t_2)$ using the Kaplan-Meier estimator using the modified data set.
Step 3. For every point of the lattice given by $[0, t_{1m}] \times [0, t_{2m}]$ compute the following four counting processes,

$$N_{10}^n(ds_1, s_2) = \sum_{i=1}^n \mathbb{I}(y_{1i} \in ds_1, y_{2i} \geq s_2, \delta_1 = 1),$$

$$N_{01}^n(s_1, ds_2) = \sum_{i=1}^n \mathbb{I}(y_{1i} \geq ds_1, y_{2i} \in s_2, \delta_2 = 1),$$

$$N_{11}^n(ds_1, ds_2) = \sum_{i=1}^n \mathbb{I}(y_{1i} \in ds_1, y_{2i} \in s_2, \delta_1 = 1, \delta_2 = 1),$$

$$Y^n(s_1, s_2) = \sum_{i=1}^n \mathbb{I}(y_{1i} \geq ds_1, y_{2i} \geq s_2).$$

Where ds_i represent the interval give by $[s_i, s_{i+1}]$.
Step 4. Calculate $R_n(y_1, y_2)$ with the formula,

$$\prod_{y_{1i} \leq y_1}^n \prod_{y_{2j} \leq y_2}^n \frac{Y_n(y_{1i}, y_{2i})(Y_n(y_{1i}, y_{2i}) - N_{10}^n(dy_{1i}, y_{1i}) - N_{01}^n(y_{1i}, dy_{2i}) + N_{11}^n(dy_{1i}, dy_{2i}))}{(Y_n(y_{1i}, y_{2i}) - N_{10}^n(dy_{1i}, y_{2i}))(Y_n(y_{1i}, y_{2i}) - N_{01}^n(y_{1i}, dy_{2i}))}.$$

Step 5. Combine $S(t_1, 0)$, $S(0, t_2)$ and $R_n(t)$ to calculate the Dabrowska estimate S_n^D,

$$S_n^D(t_1, t_2) = S(t_1, 0)S(0, t_1)R_n(t_1, t_2).$$

Figure 8: Algorithm used to calculate the Dabrowska estimate [46].

In section 2.3.1 we use synthetic data to illustrate features of the estimator and outline the calculation process.

2.3.1 Synthetic data example

Here we use synthetic data to show how the Dabrowska estimate is calculated, and numerically demonstrate some important features. Table 3 shows a data set of paired observations that we will use to produce the estimates. These were chosen so as to be small enough to easily understand the calculations, but also still be able to observe important features of the Dabrowska estimate. Figure 9 shows the calculation of the four counting processes defined in Figure 8 above using the synthetic data set. Figure 10 highlights important features of the Dabrowska estimate using the synthetic data set. Figure 10(a) shows the usual Dabrowska estimate of the data. Figure 10(b) shows the Dabrowska estimate when all events in the data set are censored. Figure 10 (c) shows a defective estimate which does not reach zero because both pairs in the data set have their last observations censored. Figure 10(d) shows the Dabrowska estimate when all of the observations are not

censored which gives the empirical distribution function. Figure 10(e) shows the Dabrowska estimate using the assumption the underlying failure time distribution is symmetric, $F(t_1, t_2) = F(t_2, t_1)$, symmetric estimator are discussed further in chapter 3.

Figure 11 shows convergence of the Dabrowska estimate as the amount of data used to produce the estimate is increased. Given a pair of failure time random variables (T_1, T_2) and censoring time random variables (C_1, C_2) both with bivariate normal distributions with parameters given by,

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1^2\sigma_2^2 \\ \rho\sigma_1^2\sigma_2^2 & \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 2 & 1.5 \\ 1.5 & 3 \end{pmatrix}, \tag{12}$$

and PDF [3],

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} e^{-\frac{z}{2(1-\rho^2)}}, \tag{13}$$

where,

$$z = \frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2},$$

Figure 11(a) shows the bivariate normal survival function given by the parameters in equation 12.

We sequentially generate data sets of the form given by equation 6 with increasing lengths ($n = 50$, $n = 150$, $n = 500$) then calculating the Dabrowska estimate. In Figure 11 we see that as the number of observations used to produce the Dabrowska estimate increases (Figures 11 (b)–(d)) convergence to the distribution the failure time observations were sampled from occurs (Figure 11 (a)).

Another view of convergence is shown in Figure 12. Firstly, we calculate the Dabrowska estimate of failure and censoring time observations generated from bivariate normal distributions with parameters given by equation 12. We then calculate the maximum distance between the Dabrowska estimate and bivariate normal survival distribution the failure time observations were generated from. To do this, we take our data set given by equation 12, taking the two sets given by $\{y_{1i}\}_{i=1}^n$ and $\{y_{2i}\}_{i=1}^n$, rank order them from smallest to largest. Then we define the maximum distance between two bivariate survival functions as,

$$K_n^{(2)} = \max_{(y_1, y_2) \in \{y_{1i}\}_{i=1}^n \times \{y_{2i}\}_{i=1}^n} |\hat{S}(y_1, y_2) - S(y_1, y_2)|. \tag{14}$$

Using the same parameters and distribution as above, estimates were produced as the amount of data sequentially increased from $n = 10$ to $n = 650$. At each n we calculated the statistic 1000 times to produce a box plot showing the minimum, maximum, median, upper and lower quartiles for the distribution of the statistic $K_n^{(2)}$. The mean of $K_n^{(2)}$ was used to produce the black curve shown in Figure 12 and is connected between data lengths.

t_1	c_1	y_1	δ_1	$S_1(t)$	t_2	c_2	y_2	δ_2	$S_2(t)$
0.11	0.23	0.11	1	1	0.82	0.11	0.11	0	1
0.28	0.68	0.28	1	1	0.32	0.22	0.22	0	0.75
0.31	0.62	0.31	1	0.6667	0.71	0.33	0.33	0	0.75
0.34	0.14	0.14	0	0.6667	0.22	0.66	00.22	1	0.375

Table 3: Table of synthetic data used to calculate the Dabrowska estimates in Figure 10.

1	0	0	0	0
0	0	0	0	0
1	1	0	0	0
1	1	1	0	0

(a)

0	1	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0

(b)

4	3	1	0	0
3	3	1	0	0
2	2	1	0	0
1	1	1	0	0

(c)

0	0	0	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

(d)

Figure 9: Tables showing the values of all 4 counting processes needed to calculate the Dabrowska estimate (a) and (b) singly censored observations N_{10}^n and N_{01}^n (c) number of objects at risk Y^n (d) doubly censored observations N_{11}^n.

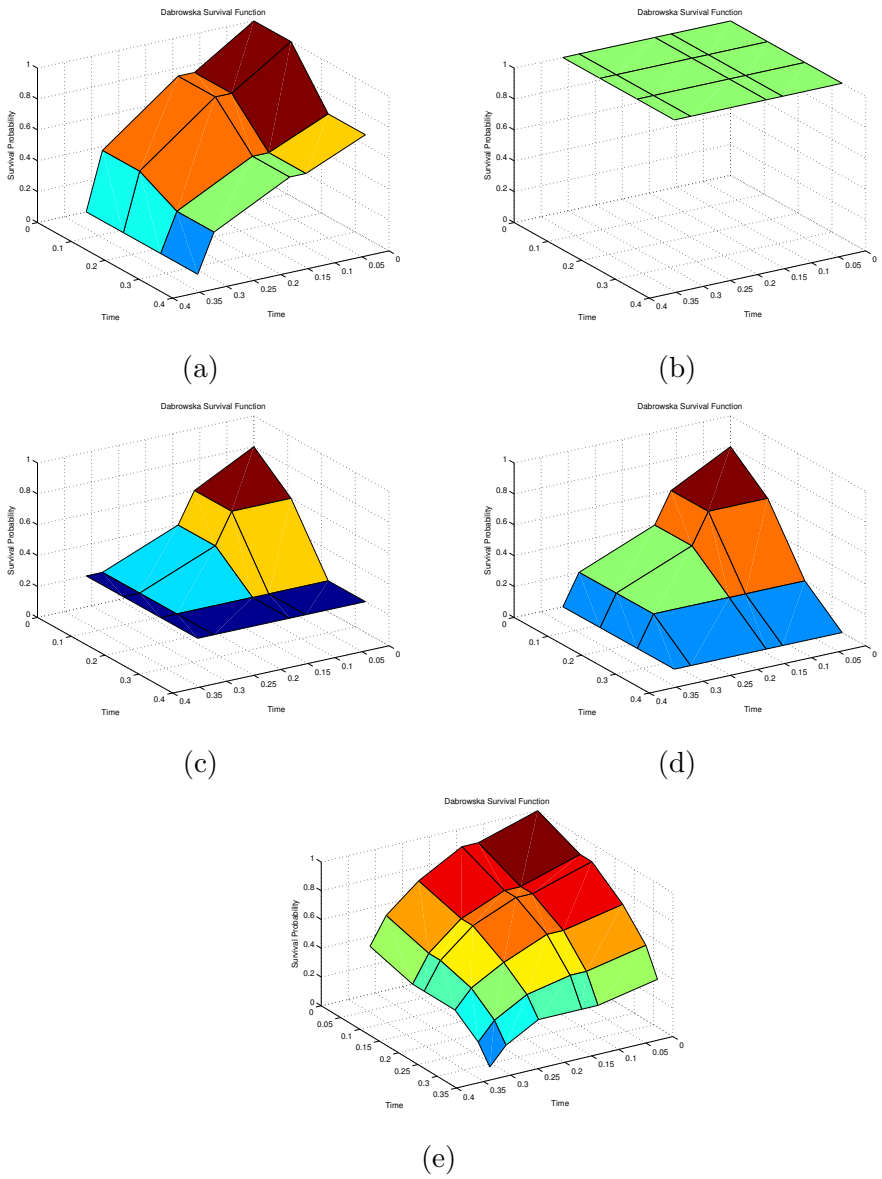


Figure 10: Dabrowska estimate using synthetic data. (a) The usual Dabrowska estimate of the data. (b) The Dabrowska estimate when all events in the data set are censored. (c) A defective distribution which does not reach zero because both pairs in the data set have their last observations censored. (d) The Dabrowska estimate when all of the observations are not censored. (e) Dabrowska estimate modified to make use of the assumption that the underlying failure time distribution is symmetric, $F(t_1, t_2) = F(t_2, t_1)$ for all t_1, t_2. Creation of symmetric estimators is discussed further in Chapter 3.

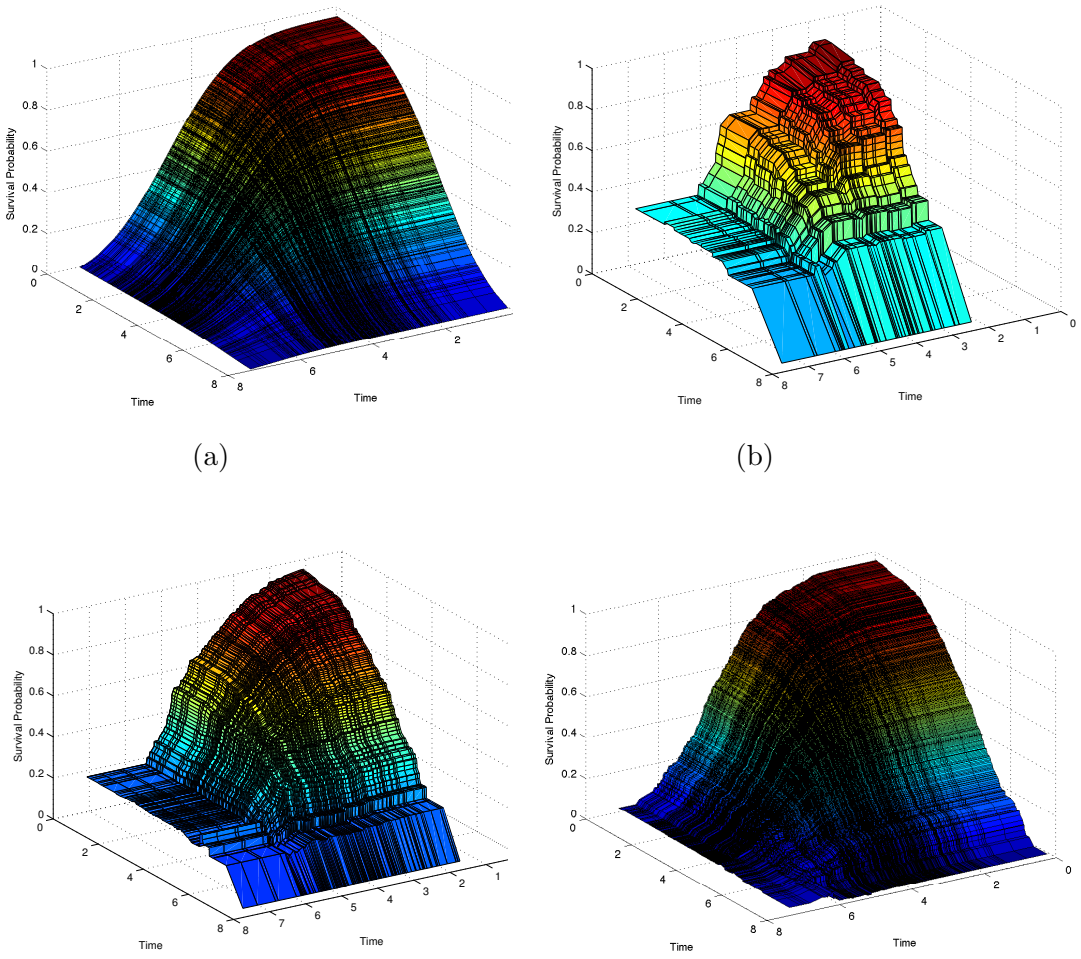


Figure 11: Convergence of the Dabrowska estimate to true failure time distribution using synthetic data. (a) Survival function of a bivariate normal distribution using parameters defined in 12. (b)–(c) Dabrowska estimates using failure time and censoring time data generated from two bivariate normal distributions with parameters described by equation 12. Amount of data increased sequentially ((b) $n = 50$, (c) $n = 150$, (d) $n = 500$). Distance from Dabrowska estimate to the true distribution as defined in 14 given by 0.2401, 0.1667 and 0.0904 respectively.

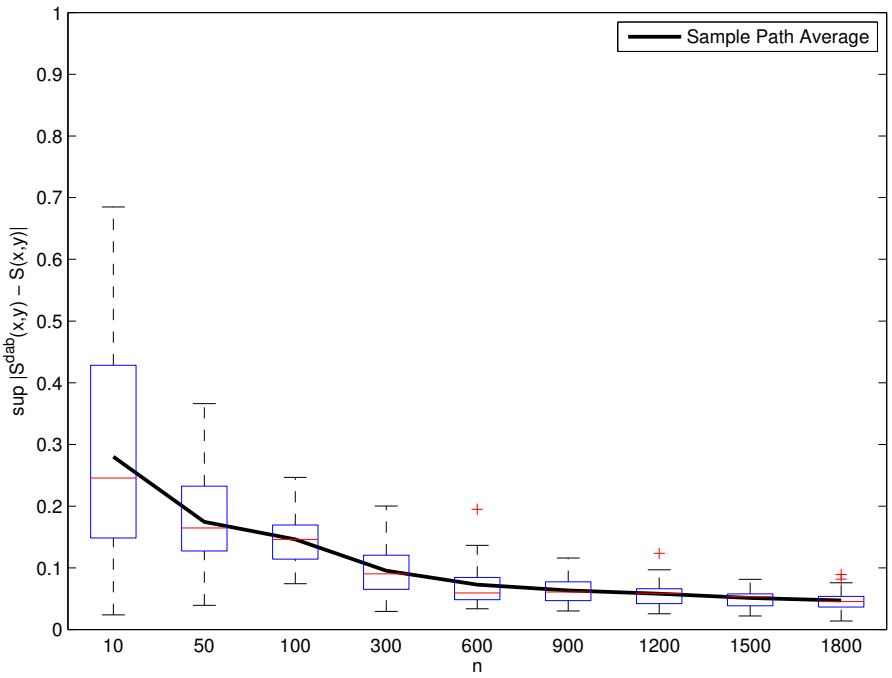


Figure 12: Boxplots of the distribution of the maximum distance between the Dabrowska estimate of the failure time distribution and the true distribution as the amount of data used to produce each estimate increases. Failure time and censoring time data is generated from two bivariate normal distributions with parameters described by equation 12. Box plots describe the distribution of the quantity in equation 14 for that n. The black line shows the mean of the distribution for that n.

2.4 Hypothesis Tests

In this section we introduce hypothesis tests that will be used to compare survival function estimates [26] [7].

Section 2.4.1 introduces the Kolmogorov-Smirnov statistical test for the hypothesis that a set of observations comes from a specific distribution in the case of non-censored univariate data. It can be extended to test bivariate data.

Sections 2.4.3 and 2.4.4 look at two different statistical tests: the Log-Rank (LR) and Weighted Kaplan-Meier. The LR test compares hazard estimates at each event time between empirical survival curves as a way to determine if the two estimates have the same underlying distribution. The WKM test produces a statistic by integrating the difference between the two Kaplan-Meier survival functions estimates over the time period of study. While the former is most commonly used, the latter test can perform as well or better in many situations [38] such as when the hazard estimates of the two survival functions cross each other over the time interval of study, and so can be a good alternative or complementary test in these cases.

Section 2.4.6 discusses the family-wise error rate (FWER) and the Holm-Bonferroni method (HBM). The FWER is the error given by the increasing number of rejections of a true null hypothesis when multiple hypothesis tests are performed. The HBM is a corrective tool used to reduce false rejections through the modification of the rejection threshold α based on number of test performed.

Lastly section 2.4.7 shows a small set of numerical examples used to indicate that the implementations used in this thesis are working correctly and are suitable for use as analysis tools of the B cell data set [10].

2.4.1 Univariate Kolmogorov-Smirnov test

Given observations drawn independently from a random variable X and a univariate continuous cumulative distribution function $F(x)$, the Kolmgorov-Smirnov test [49] is a goodness-of-fit test for the hypothesis,

$$H_0 : P(X \leq t) = F(t) \quad \text{for all } t \in \mathbb{R}.$$

Given a collection of observations $\{t_i\}_{i=1}^n$ we produce the EDF,

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(t_i \leq t) \tag{15}$$

with which the KS statistic is defined as,

$$K_n^{(1)} = \sup_{t \in \mathbb{R}} |F_n(t) - F(t)|$$

Under H_0, as n goes to infinity,

$$\sqrt{n} K_n^{(1)},$$

follows a Kolmogorov distribution [49]. This fact can be used to perform the statistical test. The CDF of a Kolmgorov distribution is given by,

$$G(x) = 1 - 2\sum_{i=1}^{\infty}(-1)^i e^{-2i^2x^2}$$

allowing the calculation of a p value as,

$$p = 1 - G(\sqrt{n}K_n^{(1)}).$$

Under the null hypothesis the p value tells us the probability of observing a statistic as large or larger than $\sqrt{n}K_n^{(1)}$. That is $p = \mathbb{P}(\sqrt{n}K_n^{(1)} \geq x)$. A small p value, where 'small' is determined by the α level, commonly set to be 0.05, tells us that either H_0 is false or that the data set we observed was uncommon.

2.4.2 Kolmogorov-Smirnov test for right censored data

In [12] the authors present a modified version of the Kolmogorov-Smirnov test procedure that allows goodness-of-fit testing when dealing with censored events. Given a failure time random variable T and a censoring time random variable C, with estimated survival function $\hat{S}(t)$ from a data of set as described by equation 7, which we suspect comes from a reference survival function $S^0(t)$, censoring random variable C with survival distribution $\hat{G}(t) = \mathbb{P}(C > t)$, and function $N(t)$ which counts the number of items at risk at time t, we define a function,

$$Y_N(t) = \frac{1}{2}\{\hat{S}(t) - S^0(t)\}\int_0^t \{N\hat{G}(s^-)\}^{\frac{1}{2}}\mathbb{I}_{(N(s)>0)}d\{\Lambda^0(s) - \hat{\Lambda}(s)\}$$

from which we calculate the statistic over the interval of study give by $[0, t_m]$,

$$\hat{\mathcal{Y}} = \sup_{0\leq t\leq t_m} Y_N(t).$$

Let $\Theta(x)$ be the CDF of the standard normal distribution, we define a function q,

$$q(x,y) = 1 - \Theta\left(\frac{y}{(x-x^2)^{\frac{1}{2}}}\right) + \Theta\left(\frac{y(2x-1)}{(x-x^2)^{\frac{1}{2}}}\right)e^{-2y^2},$$

with which for large N and under H_0 we compute the p value as follows,

$$p = \mathbb{P}(\hat{\mathcal{Y}} > y) = q(y, 1 - S(t_m)).$$

2.4.3 The Log-Rank Test

The Log-Rank [29] test is a univariate statistical test that, given two data sets with possibly censored failure times like those shown in equation 6, allows the testing of the null hypothesis that there is no difference between the distributions of the two underlying failure times,

$$H_0 : S_1(t) = P(T_1 > t) = S_2(t) = P(T_2 > t) \quad \text{for all } t \geq 0. \tag{16}$$

Given two univariate data sets based on observations of failure and censoring times of the form,

$$D_{n_1}^{(1)} = \{y_{i1}, \delta_{i1}\}_{i=1}^{n_1}, \quad \text{and} \quad D_{n_2}^{(1)} = \{y_{i2}, \delta_{i2}\}_{i=1}^{n_2},$$

we begin by formatting the data for the test. We need both data sets to be over the same time periods, so we define $\{x_{i1}\}_{i=1}^{l} = \{x_{i2}\}_{i=1}^{l}$ as two equal sets made up of the unique elements of $\{y_{i1}\}_{i=1}^{n_1} \cup \{y_{i2}\}_{i=1}^{n_2}$. Then, for $j \in \{1, 2\}$ as usual we can calculate $d_{ij} = \sum_{k=1}^{n_j} \delta_{kj} \mathbb{I}(x_{ij} = y_{kj})$, $e_{ij} = \sum_{k=1}^{n_j} (1 - \delta_{kj}) \mathbb{I}(x_{ij} = y_{kj})$ and $r_{ij} = \sum_{k=1}^{n_j} (d_k + e_k)$ as the number of failures, censoring events and elements at risk respectively at time x_{ij} for data set j.

The Log-Rank test works by comparing the hazard estimate at every time point where an event occurs. At each unique time point x_{ij} we calculate the observed number of events for each survival curve, denoted $o_{ij} = (d_{ij} + e_{ij})$. Using the number of objects at risk at x_{ij}, denoted r_{ij} we calculate the expected number of events for $S_1(t)$ as,

$$E_{i1} = \frac{o_{i1} + o_{i2}}{r_{i1} + r_{i2}} o_{i1},$$

We then calculate Log-Rank statistic as,

$$\hat{\theta}_{LR} = \frac{\sum_{i=1}^{l} (o_{i1} - E_{i1})}{\sum_{i=1}^{l} v_{i1}}$$

where,

$$v_{i1} = \sum_{i=1}^{l} \frac{r_{i1} r_{2i} (o_{i1} + o_{i2})(r_{i1} + r_{i2} - o_{j2} - o_{j2})}{(r_{i1} + r_{i2})^2 (r_{i1} + r_{i2} - 1)}.$$

Under H_0 and for sufficiently large n_1 and n_2 the Log-Rank statistic follows a χ^2 distribution with 1 degree of freedom [3]. If we define $F_1(x)$ as

a chi-squared CDF with one degree of freedom we can calculate the p value as follows,

$$p = 1 - F_1(\hat{\theta}_{LR}).$$

2.4.4 Weighted Kaplan-Meier test

Here we present an overview of the Weighted Kaplan-Meier (WKM) test, a univariate statistical test for comparing two Kaplan-Meier survival function estimates [38]. In some cases, described below, the WKM is a more appropriate statistic to use to compare two survival function estimates than the Log-Rank test. The WKM tests the null hypothesis,

$$H_0 : S_1(t) = S_2(t) \quad \text{for all } t > 0.$$

Given two survival functions estimates $\hat{S}_1(t)$ and $\hat{S}_2(t)$ based on two univariate data sets $\{t_{1i}, c_{1i}\}_{i=1}^{n_1}$ and $\{t_{2i}, c_{2i}\}_{i=1}^{n_2}$, we define $G_i(t) = \mathbb{P}(C_i > t)$ as the survival function of the censoring time random variable for survival curve $i = 1, 2$. We define $\hat{S}(t)$ and $\hat{G}(t)$ as the failure time and censoring time Kaplan-Meier estimates of the combined data set used to make estimates $S_1(t)$ and $S_2(t)$.

When comparing two survival functions $S_1(t)$ and $S_2(t)$ in a situation where,

$$\lambda_1(t) \geq \lambda_2(t) \quad \text{for all } t \geq 0, \tag{17}$$

or

$$S_1(t) \geq S_2(r) \quad \text{for all } t \geq 0, \tag{18}$$

we know that population 1 always has a better survival time than population 2. Statistical tests used to compare survival curve estimates should be sensitive to these cases. The Log-Rank test explained above, which compares hazard differences, is sensitive to the first case, but will not necessarily be able to distinguish the second case [38]. As such the authors of [38] suggest the use of a statistic based on the integrated difference between two survival functions that has the ability to distinguish between both of the above cases,

$$\int_{t_0}^{t_n} (\hat{S}_1(t) - \hat{S}_2(t))dt,$$

however they note that this statistic can have a large variance in situations where t is near t_n and there is a large amount of censoring. Consequently

they suggest a weighting function which will discount the impact of heavily censored time points. They define the Weighted Kaplan-Meier (WKM) statistic for a weight function $w(t)$ as,

$$WKM = \sqrt{\Big(\frac{n_1 n_2}{n}\Big)} \int_0^{T_c} \hat{w}(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt, \tag{19}$$

where $T_c = \sup\{t : \min(\hat{C}_1(t), \hat{C}_2(t)) > 0\}$. The weight function is chosen such that for constants α and Γ,

$$|w(t)| \leq \Gamma[C_i^-(t)]^{1/2+\alpha} \quad \text{and} \quad |\hat{w}(t)| \leq \Gamma[\hat{C}_i^-(t)]^{1/2+\alpha},$$

for all $t < \tau = \sup\{\min(S(t), C_1(t), C_2(t))\}$ and $i = 1, 2$. Given the WKM statistic in equation 19, we can use the fact that under $H_0 : S_1(t) = S_2(t)$,

$$WKM = \sqrt{\Big(\frac{n_1 n_2}{n}\Big)} \int_0^{T_c} \hat{w}(t)[\hat{S}_1(t) - \hat{S}_2(t)]dt \rightarrow^d \mathcal{N}(0, \sigma^2),$$

where,

$$\sigma^2 = -\int_0^{\tau} \frac{[\int_t^{\tau} w(u)S(u)du]^2}{S^2(t)} \frac{p_1 G_1^-(t) + p_2 G_2^-(t)}{G_1^-(t) G_2^-(t)} dS(t),$$

which can be estimated,

$$\hat{\sigma}_p^2 = -\int_0^{T_c} \frac{[\int_t^{T_c} \hat{w}(u)\hat{S}(u)du]^2}{\hat{S}(t)\hat{S}^-(t)} \frac{\hat{p}_1 \hat{G}_1^-(t) + \hat{p}_2 \hat{G}_2^-(t)}{\hat{G}_1^-(t)\hat{G}_2^-(t)} d\hat{S}(t).$$

A proof of the above can be found in [38]. We determine the p value for a WKM statistics given by $\hat{\theta}_{WKM}$ using the CDF, $F(x)$, of the normal distribution described above as,

$$p = 1 - F(\hat{\theta}_{WKM}).$$

2.4.5 Bootstrap Monte Carlo hypothesis test

In this section we present a hypothesis test that allows us check if two bivariate Dabrowska estimates share the same underlying distribution. Given two Dabrowska estimates $\hat{S}_1(x, y)$ and $\hat{S}_2(x, y)$ defined on a common grid of time points given by $[0, t_1] \times [0, t_2]$ we defined H_0,

$$H_0 : S_1(x, y) = S_2(x, y) \text{ for all } (x, y) \in [0, t_1] \times [0, t_2] \tag{20}$$

Since no standard test exists for comparing two Dabrowska estimates we do not have a statistic and reference distribution from which to calculate a p value, as we did with the univariate tests described above. To perform this test we will produce a distribution numerically for a sensible choice of test statistic by sampling with replacement [11] from the data sets used to produce the Dabrowka estimates a large number of times, and then for each resampled data set, we calculate the chosen statistic, from this set of statistics we will produce a distribution from which a p value can be calculated. We will also present a set of numerical tests to provide evidence for the viability of this method of hypothesis testing in Section 2.4.7 once the data sets are sufficiently large.

Given two data sets of the form described in equation 9 of length n and m denoted by $D_n^{(2)}$ and $D_m^{(2)}$ we calculate the Dabrowska estimates $\hat{S}_1(x, y)$ and $\hat{S}_2(x, y)$ respectively as well as the Dabrowska estimate of combined distribution $D_n^{(2)} \cup D_m^{(2)}$ given by $\hat{S}_{1+2}(x, y)$. From these distributions we calculate the following statistic,

$$\hat{\theta} = d(\hat{S}_1(x, y), \hat{S}_{1+2}(x, y)) + d(\hat{S}_2(x, y), \hat{S}_{1+2}(x, y)), \tag{21}$$

where $d(S_1(x, y), S_2(x, y))$ is the bivariate maximum absolute difference defined in equation 14. We then sample with replacement from the combined distribution $D_n^{(2)} \cup D_m^{(2)}$ to produce a data set of event times and censoring times of length $n + m$.

This data set is then used to produce a new set of Dabrowska estimates where the first n elements are used to produce $\hat{S}_1^i(x, y)$ and elements $n+1$ to $n + m$ are used to produce $\hat{S}_2^i(x, y)$, finally the entire resampled data set is used to produce $\hat{S}_{1+2}^i(x, y)$. From these three estimates we produce a statistic $\hat{\theta}^i$. Repeating this procedure a large number of times, we can calculate the empirical CDF of this sets of statistics and from this distribution compute the p value for $\hat{\theta}$.

2.4.6 Family-Wise Error Rate

Given a collection of n statistical tests with true null hypothesis H_0, p values $\{p_i\}_{i=1}^n$, and a significance level α, we define the number of 'false positives',

that is, situations in which $p_i < \alpha$, resulting in a rejection of the true hypothesis H_0 as,

$$E_n = \sum_{i=1}^{n} \mathbb{I}(p_i < \alpha). \tag{22}$$

The family wise error rate (FWER) is defined as the probability at least one false positive occurs when performing a collection of statistical tests. That is,

$$FWER = \mathbb{P}(E_n > 1). \tag{23}$$

The FWER can be calculated as follows,

$$FWER = \mathbb{P}(E_n > 1) = 1 - \mathbb{P}(E_n = 0) = 1 - (1 - \alpha)^n. \tag{24}$$

Figure 13 shows how the FWER changes as we perform and increasing number of tests at a rejection threshold $\alpha = 0.05$.

To compensate for the fact that increasing the number of tests performed increases the FWER we use the Holm-Bonferroni method (HBM) [20]. The HBM works by modifying the individual rejection threshold based on the number of tests performed. An example calculation for a family of three tests is shown in Figure 14.

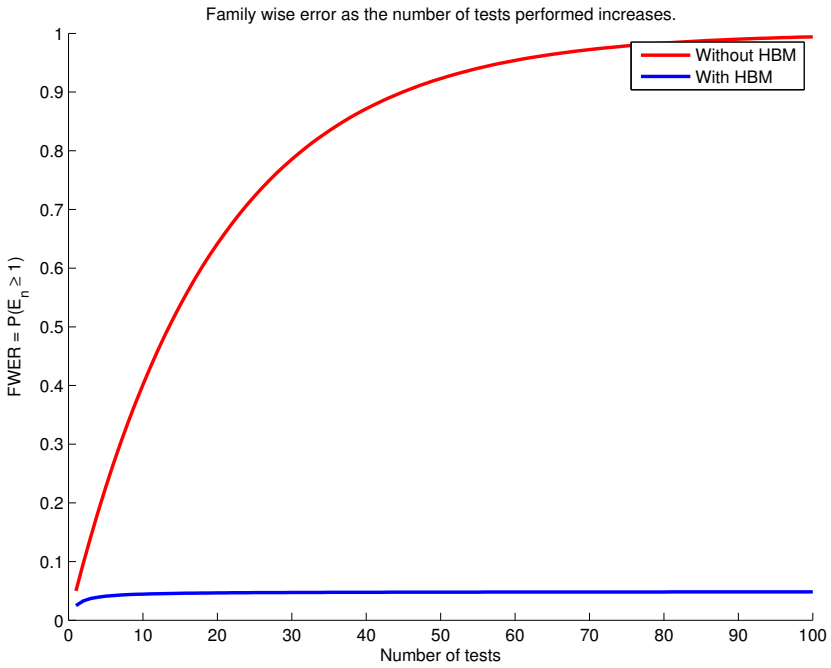


Figure 13: Family wise error rate (FWER) as the number of tests, n, increases at a rejection threshold of $\alpha = 0.05$ with and without the Holm-Bonferroni method (HBM) procedure. As we can see FWER is always less than or equal to α when HBM has been used.

Holm-Bonferroni method.

Step 1. Order p values smallest to largest. $p = \{p_1, p_2, p_3\}$

Step 2. Perform test with rejection threshold α using modified p value $p_i^{HB} = (C - i + 1)p_i$, where C is the size of the test family and i is the number of the current test.

Step 3. If the test result is not significant, we are finished. If the first test is significant, we continue until we reach a none significant result or we run out of tests.

Note: If p_i^{HB} becomes greater than 1 set its value to 1.

Figure 14: Holm-Bonferroni method for a family of 3 tests.

2.4.7 Hypothesis Test Examples with Synthetic Data

Figure 15a shows the EDF for 150 elements of univariate normal data with parameters $\mu = 5 \quad \text{and} \quad \sigma = 1$, the second curve labelled 'reference distribution' is a normal CDF with parameters $\mu = 5 \quad \text{and} \quad \sigma = 1$. Here we used the Kolmgorov-Smirnov test to check the hypothesis that the underlying distribution of the data was sampled from is the same as the reference distribution. The test results gave a p value of 0.5928 indicating we do not reject the null hypothesis at a threshold of $\alpha = 0.05$. This was a test to give evidence that the code used is working correctly. The Kolmogorov-Smirnov tests in this section were produced using the Matlab function KSTEST.

Figure 15b shows a similar situation to that of the above except here we would like to show that under correct conditions the KSTEST function will correctly reject a null hypothesis. The EDF is generated from $n = 150$ elements of univariate normal data with parameters $\mu = 5$ and $\sigma = 1$. The built in Matlab function KSTEST was used to check if the data had a normal distribution with the parameters $\mu = 6$ and $\sigma = 1$. The result was a rejection at $\alpha = 0.05$ with a p value of 1.3590×10^{-26} showing that the test is working correctly.

Figure 16a shows two Kaplan-Meier estimates using the same configuration as above but here the parameters used for the failure time distribution are given in the first estimate by, $\mu = 16 \quad \text{and} \quad \sigma = 4$, and in the second estimate by, $\mu = 15 \quad \text{and} \quad \sigma = 4$. As the survival curves here are different we expect that the statistical tests will reject H_0. The Log-Rank and Weighted Kaplan-Meier tests return p values of 0.0025 and 0.0073 respectively which is in line with what we would expect at a confidence level of $\alpha = 0.05$.

Figure 16b shows two Kaplan-Meier estimates produced from two data sets in which the underlying distributions are identical and as such we expect that both tests should fail to reject H_0. The failure time data was generated from a normal distribution with parameters, $\mu = 16 \quad \text{and} \quad \sigma = 4$. The censoring time data was generated from a normal distribution with parameters given by, $\mu = 14 \quad \text{and} \quad \sigma = 1$. In both cases 50 data points were generated. The Log-Rank and Weighted Kaplan-Meier tests returned p values of 0.1113 and 0.6952 respectively which is in line with what we would expect at a confidence level of $\alpha = 0.05$.

The above tests represent two different instances of the Kolmogorov-Smirnov producing the correct results given the underlying data that the two tests are based on. However a more extensive examination can be ap-

plied that allows us to check the code is producing correct results across $m = 1000$ different hypothesis tests. To do that we must outline the following mathematics below. Given a true hypothesis H_0, and given a generalized statistic $\hat{\theta} \overset{d}{\sim} X$, with p value given by,

$$p = \mathbb{P}(X \geq \hat{\theta}), \tag{25}$$

for a rejection threshold α all p must values have the property that,

$$\mathbb{P}(p \leq \alpha) = \alpha \quad \text{for all } \alpha \in [0, 1]. \tag{26}$$

That is, the p value is uniformly distributed. If we use synthetic data to produce many p values based on true hypothesis tests, we can check the EDF is uniform and thus confirm our test is working over a much larger sample size than just one test.

Figure 17 is a simulated illustration of uniform p values for the Kolomgorov-Smirnov test. Here 10000 hypothesis tests were performed under a scenario in which H_0 is true. The p values produced were then used to create an empirical distribution function,

$$F_n^p(\alpha) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(p_i \leq \alpha). \tag{27}$$

Figures 17 also shows a simulation of the p value distribution for the Weighted Kaplan-Meier and Log-Rank tests to provide evidence supporting their correct functioning beyond just the testing of one false and one true hypothesis as with other figures. The WKM and LR tests were implemented specifically for this thesis. Figure 18 plots a set of p values under true H_0 for the modified Kolmogorov-Smirnov test for right censored data [12] described in section 2.4.2. Here we see a p value distribution that is close to uniform.

Figure 19 shows what happens to the distribution of p when H_0 is not true. Here multiple KS tests were performed to produce p value distributions using survival curves that are estimated from data with increasingly different 19(a) means and 19(b) variances of the underlying distribution. As we can see the distributions diverge from uniformity and the probability of a rejection of H_0 becomes much higher.

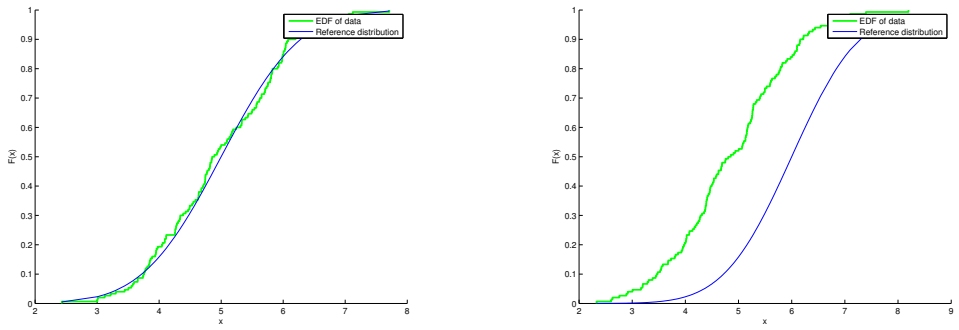
In Figure 20 we produced two identical survival distributions. Both failure and censoring time data was generated from a univariate normal distribution with parameters $\mu = 10$ and $\sigma = 1$. Hypothesis tests were performed,

$H_0 : S_1(t) = S_2(t)$ at a significance level $\alpha = 0.05$, using both the Log-Rank and Weighted-Kaplan Meier tests. We repeated this experiment 1000 times to produce Figure 20(a) in which we see that there are a few instances of rejection of H_0 in line with the chosen α value of 0.05. As we increase the number of tests we perform the chance that we produce false positive results increases. Figure 20(b) shows the same test again but here HBM was used to adjust the rejection threshold. As we can see there are almost no rejections.

Figure 21 presents a test to provide evidence that the numerical hypothesis test presented in section 2.4.5 works correctly. Figure 21 shows the uniform hypothesis test described above. Here the hypothesis test has been performed 2000 times using data generated from a bivariate normal distribution as described by equation 13 using parameters given by,

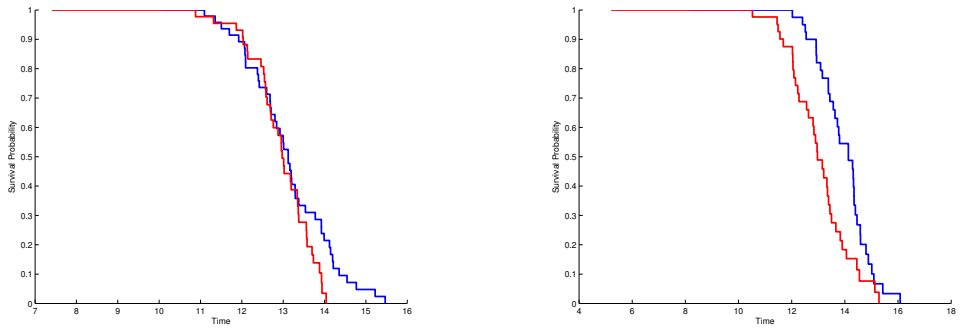
$$\mu = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix}, \tag{28}$$

the distribution produced is uniform as we would expect. However we note that there is some discretization due to the fact we are resampling with replacement from a data set to produce estimates.



(a) (b)

Figure 15: KS Test, Section 2.4.1. Here we use the KS Test to check if a set of data has a specific distribution. Data was generated ($n = 150$) from a normal distribution with parameters $\mu = 5$, $\sigma = 1$. (a) We then used the built in Matlab function KSTest to check if the data generated had a normal distribution with the same parameters. The result was a failure to reject that the data has the specified distribution at $\alpha = 0.05$ with a p value of 0.5928. (b) We then used the built in Matlab function KSTest to check if the data generated had a normal distribution with the parameters $\mu = 6$, $\sigma = 1$. The result was a rejection at $\alpha = 0.05$ with a p value of 1.3590×10^{-26}.



(a) (b)

Figure 16: Log-Rank Test, Section 2.4.3 and Weighted Kaplan-Meier test Section 2.4.4. Here we compare two Kaplan-Meier survival curve estimates using synthetic failure time and censoring data with both the Log-Rank and Weighted Kaplan-Meier tests. In both cases $n = 50$. (a) Both observations use the data generated from the same underlying distributions. The Log-Rank test returns a statistic of 2.5361, which gives a p value of 0.1113. The Weighted Kaplan-Meier test returns a statistic of 0.2667 which gives a p value of 0.6952. Thus in both cases we would correctly fail to reject H_0 at a confidence level of $\alpha = 0.05$. (b) Here both observations have the same censoring distribution, but one has a failure time distribution which differs by a mean of 1. The Log-Rank test returns a statistic of 9.1401, which gives a p value of 0.0025. The Weighted Kaplan-Meier test returns a statistic of 1.2357 which gives a p value of 0.0073. Thus in both cases we would correctly reject H_0 at a confidence level of $\alpha = 0.05$

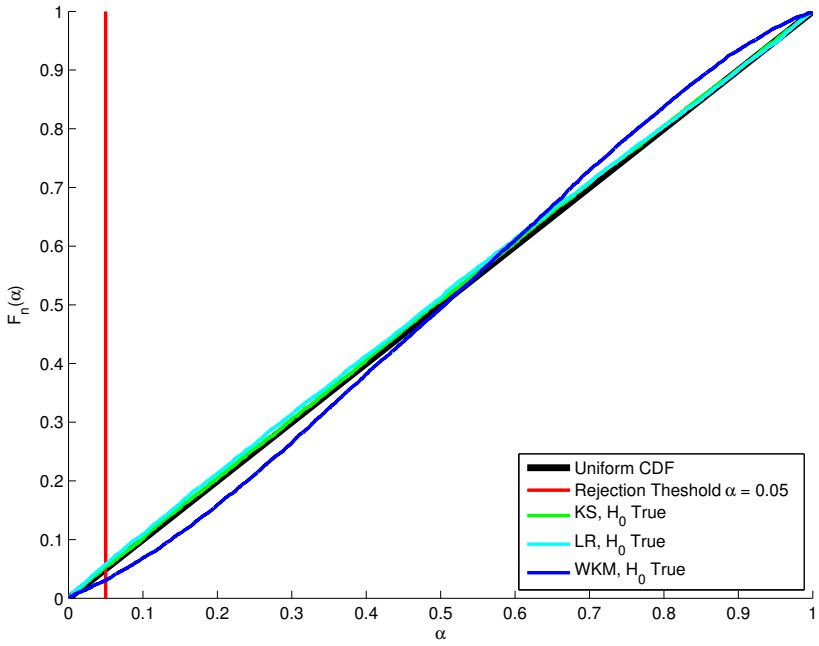


Figure 17: KS Test, Section 2.4.1. LR Test, Section 2.4.3. WKM Test, Section 2.4.4. Under the null hypothesis, the p-value distribution is uniform, shown here as the black line. Comparison with it is provided as evidence that the implementation of hypothesis tests are working correctly. (Green) Kolmogorov-Smirnov test H_0 true, (Black) Kolmogorov-Smirnov test H_0 false (Red) Log-Rank test H_0 true (Blue) Weighted Kaplan-Meier test H_0 true. Tests under false H_0 have a distribution centred around a much smaller value as, assuming the correct working of the test, the p values should be much smaller.

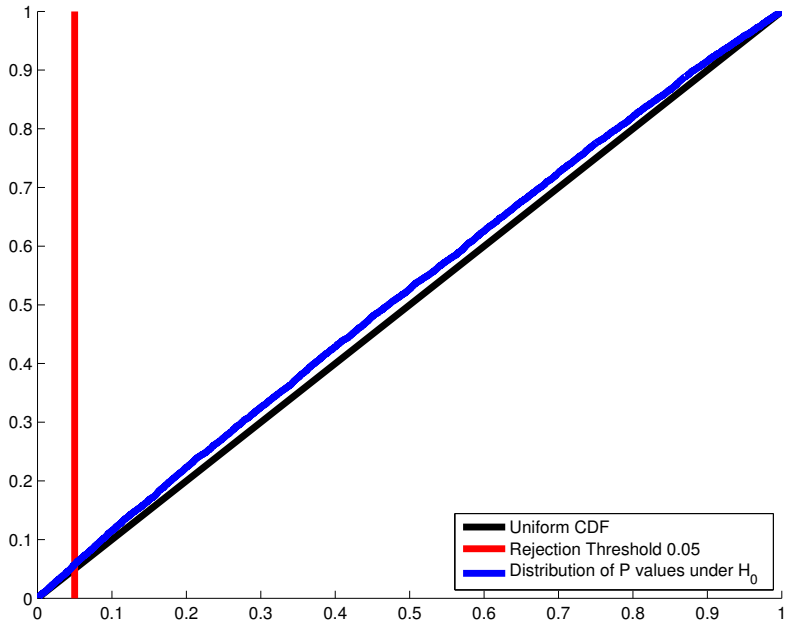
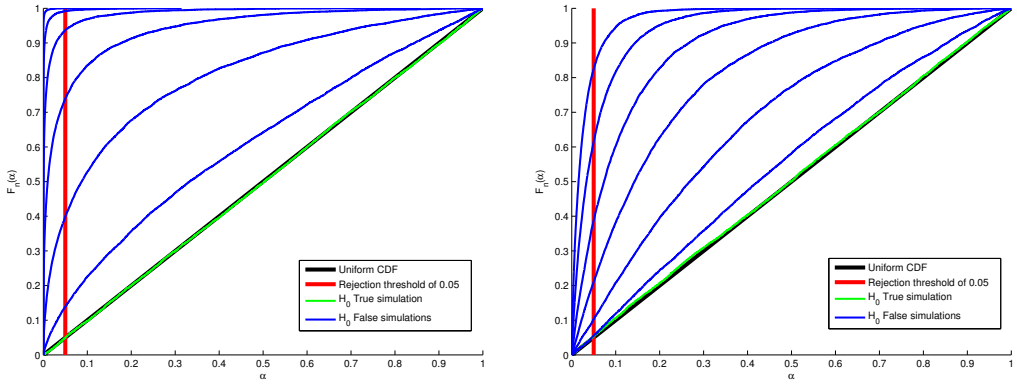
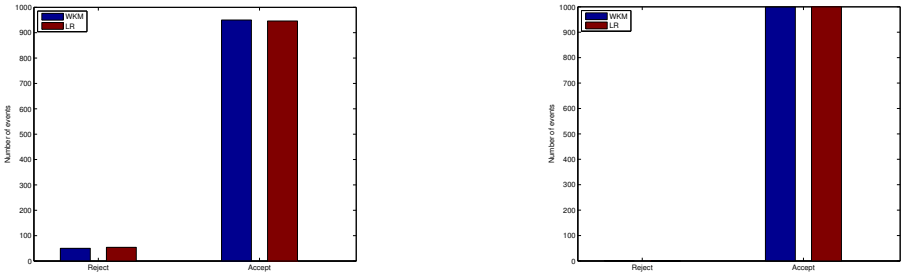


Figure 18: Modified KS Test for right censored data, section 2.4.2. Under the null hypothesis, the p-value distribution is uniform, shown here as the black line. Comparison with it is provided as evidence that the implementation of hypothesis tests are working correctly for the modified Kolmogorov-Smirnov test for right censored data described in section 2.4.2.



(a) (b)

Figure 19: KS Test, Section 2.4.1. Here we show further evidence for a working Kolmgorov-Smirnov test via uniform p values. Using data to produce the estimates from an underlying distribution such that H_0 begins true, and so produces a uniform distribution. (a) Here, the green curve shows the test when both estimates were produced from underlying normal data was identical, with $\mu = 5$ and $\sigma = 1$, the curves then have a progressively larger rejection rate as μ goes from 5 to 4.4 while $\sigma = 1$ stays the same. (b). Here $\sigma = 1$ increases to 1.6 producing uniform p initially when H_0 is true, with a rapidly increasing threshold of rejection as the distributions begins to differ further.



(a) (b)

Figure 20: FWER and HBM, Section 2.4.6. Two survival distributions were compared based on synthetic data with the same underlying distribution. We repeated this experiment 1000 times. This plot shows the number of times the WKM test accepted and rejected $H_0 : S_1(t) = S_2(t)$ both without the HBM (a) and with the HBM (b). As we can see the number of false positives has been reduced when the HBM method is applied for both the WKM and LR tests. This is as we would expect from looking at Figure 13, we see that when performing 100 tests the FWER is close to 1. When the HBM is applied the FWER is close to 0.05 and so we except fewer false positives.

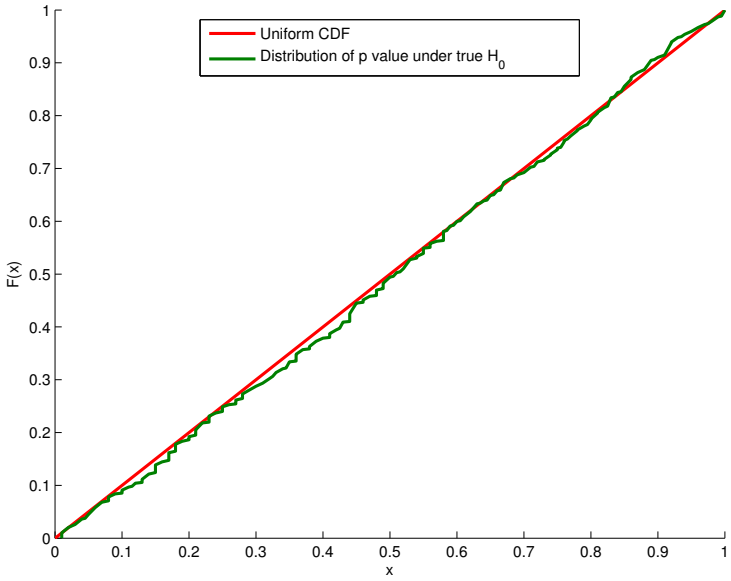


Figure 21: Under the null hypothesis, the p-value distribution is uniform as evidence that the implementation of bootstrap hypothesis test described in section 2.4.5 is working correctly. As expected the distribution of p values under true H_0 has a uniform distribution but with some discretization due to the resampling process.

Chapter 3

3 Statistical Extensions

In this chapter we cover defective distributions and symmetry. These two topics are important modifications to the survival analysis tool set. As mentioned in both the Kaplan-Meier and Dabrowska sections, when the final event is censored, the survival curve does not reach 0. When this occurs we need a way to deal with it that allows us to still perform statistical tests, and calculate expectations and variances. The second section, which covers symmetry, allows us to leverage the fact that siblings have no inherent rank-order. We use this fact to generate improved estimates when calculating survival functions.

This chapter introduces two extensions to the statistical tool kit we have described in chapter 2 that will be used to inform and improve the analysis of the B cell data set [10] in chapter 4.

Firstly, section 3.1 introduces an estimator for the probability that cells are not motivated to undergo certain events. The model developed in [10] discussed the fact that some B cells will never for example, divide, regardless of the length of time given. Under this assumption we propose a statistical estimator for the probability that the failure time distribution, T, can take on the value infinity. This is denoted by,

$$P(T = \infty) = p_\infty. \tag{29}$$

We then apply numerical simulations to give evidence for estimators having properties required to perform our analysis.

Secondly, in Section 3.2 we propose a modified Dabrowska estimator, *sym*-Dabrowska that leverages the assumption of symmetry in the underlying bivariate failure time distribution. The use of this estimator will improve both the quality of the survival distributions we estimate, and the reliability of the statistical tests we perform on the B cell data set [10].

3.1 Defective Distributions

In [10] three parameters are defined to describe the probability cells never undergo division, differentiation to PB, and IgM to IgG1 class switching: $p_{div}, p_{diff}, p_{switch}$. This is because not all cells will necessarily activate the machinery to undergo these processes [10]. Ideally we would like to define a statistic that can be used to calculate this parameter in terms of the data. Below we develop a maximum likelihood estimator of p_{div}. To account for the possibility that division does not occur, we will modify non-negative real valued random variable representing the division time, T, such that with probability p it can take on value ∞, otherwise it will take on the usual division time, represented here by the non negative real valued random variable A,

$$T = (1 - B)A + B\infty, \tag{30}$$

where B is a Bernoulli random variable taking values in $\{0, 1\}$ with $P(B = 1) = p$. The censoring event is given by non-negative real valued random variable C. We define the observable random variables as $Y = \min(T, C)$

and $\Delta = \mathbb{I}(T < C)$. A data set made up of observations from these random variables is given by $D_n^{(1)} = \{(y_i, \delta_i)\}_{i=1}^n$ in a similar way to equation 6. We express the likelihood over the n data points in terms of the PDFs and survival functions of T, C and A as follows,

$$\begin{aligned}\mathcal{L}(D_n^{(1)}|\theta) &= \prod_{i=1}^{n}[f_T(y_i)S_C(y_i)]^{\delta_i}[f_C(y_i)S_T(y_i)]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[(1-p)f_A(y_i)S_C(y_i)]^{\delta_i}[f_C(y_i)(p+(1-p)S_A(y_i))]^{1-\delta_i},\end{aligned}$$

we can then take the logarithm,

$$\begin{aligned}\log \mathcal{L}(D_n^{(1)}|\theta) &= \sum_{i=1}^{n}\log([f_T(y_i)S_C(y_i)]^{\delta_i}[f_C(y_i)S_G(y_i)]^{1-\delta_i}) \\ &= \sum_{i=1}^{n}(\delta_i)\log[(1-p)f_A(y_i)S_C(y_i)] + (1-\delta_i)\log[f_C(y_i)(p+(1-p)S_A(y_i))] \\ &= \sum_{i=1}^{n}(\delta_i)\log(1-p) + (\delta_i)\log f_A(y_i) + (\delta_i)\log S_C(y_i) + (1-\delta_i)\log f_C(y_i) \\ &+ (1-\delta_i)\log(p+(1-p)S_A(y_i)),\end{aligned}$$

then we differentiate with respect to p,

$$\frac{\partial \log \mathcal{L}(D_n^{(1)}|\theta)}{\partial p} = \sum_{i=1}^{n}\frac{\delta_i}{p-1} + \frac{(\delta_i - 1)(S_A(y_i)-1)}{p + S_A(y_i) - pS_A(y_i)}. \tag{31}$$

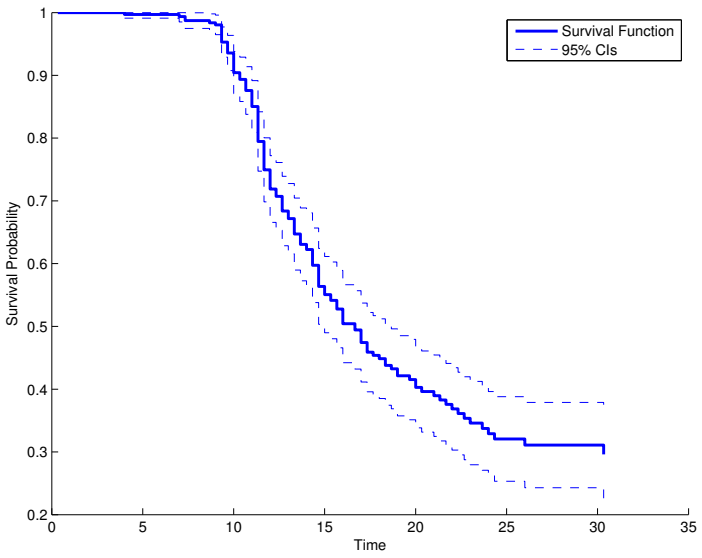


Figure 22: Kaplan-Meier survival function estimate for generation 7 division data. Here we see that the survival curve does not reach zero and $S(t_m) > 0$. Dashed lines show 95% confidence intervals generated using Matlabs ECDF function.

This equation can not be solved for p explicitly and so we cannot get an explicit expression for the probability of infinity in terms of the data via maximum likelihood method. However the above equation could be evaluated numerically. If we take a look at an example of a defective distributions for B cell data like the Kaplan-Meier estimate for time to divide in Figure 22 we see that this curve does not reach zero. At the end point there is some probability mass left over. We use this to propose an estimator of the probability that the failure time event does not occur as,

$$\hat{p}_n = \min_{t \in \{y_i\}_{i=1}^n} \hat{S}_n(t). \tag{32}$$

In order to check this estimate is reasonable we will provide simulations to suggest it is both consistent [43],

$$\lim_{n \to \infty} P(|\hat{p}_n - p| < \epsilon) = 1, \tag{33}$$

and that it is unbiased,

$$(\mathbb{E}[\hat{p}_n] - p) = 0 \quad \text{for all } n. \tag{34}$$

In Figure 23 we provide evidence to support the hypothesis that equation 32 has the properties discussed above as an estimator of p, via simulation. Here we generated data using the above random variables such that the censoring distribution C and the failure time distribution T both had log normal distributions with parameters $\mu = 0$ and $\sigma = 1$. We then introduced the Bernoulli distribution B with parameter given by $p = 0.15$ to determine if the failure time would take on the value ∞. Data sets were then generated of different lengths with multiple data sets for each length, allowing the production of distributions of estimates for fixed n. We then used this data to show box plots in Figure 23 as well as give evidence for consistency.

Looking at the upper plot of Figure 23, we see that as the amount of data used to produce the estimate increases, the estimate tends towards the true underlying value given by the distributions from which the data was sampled with $p = 0.15$. This provides evidence for consistency of the estimator.

The bottom plot shows many box plots, each of which represents the distribution of $\hat{p}$ for a specific value of n given by the x coordinate. From this we see that the mean value of all of the estimates of $\hat{p}$ from data of a specific length produce an expected value close to the underlying value of $p = 0.15$. This is to give evidence for the estimator being unbiased as the expected value for all $\hat{p}_n$ for a given n average around the true value of p.

The last property from the plot 23 shown here shows that as n increases the distributions centre more closely around the true value of $p = 0.15$ suggesting that the estimator is asymptotically unbiased.

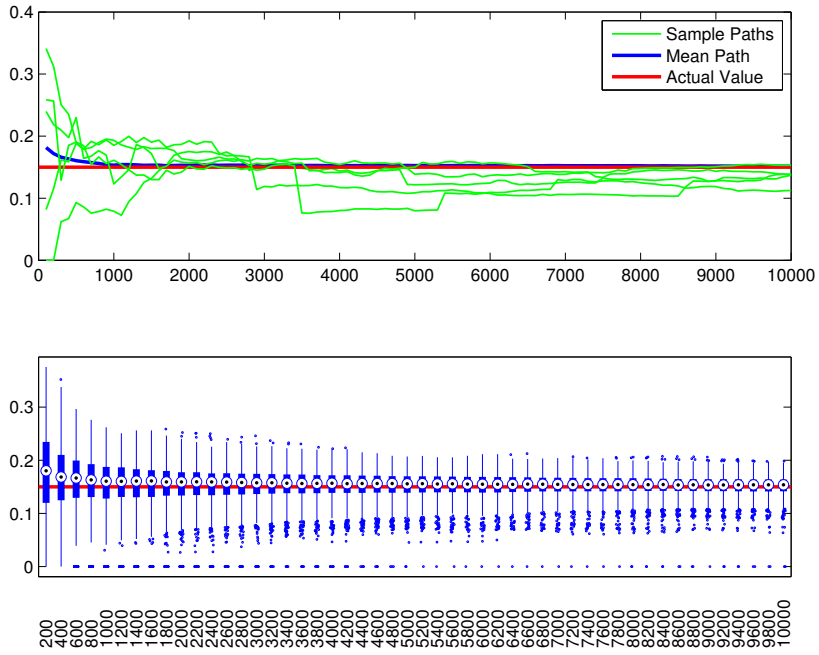


Figure 23: Consistency and unbiasedness of $\min S_n(t)$ as an estimator of the probability of a defective distribution. The upper sub plot shows five sample paths of 1000 that were generated. The average path is also shown which can be seen to rapidly convergence to the true underlying value of 0.15. The lower sub plot in this figure shows a collection of box plots each representing the empirical distribution of 10000 $\hat{p} = \min_t \hat{S}(t)$ estimates. The length of the data used to produce the collection $\hat{p}$ estimates then increases sequentially for each box plot. As we can see convergence is taking place towards the true value of 0.15 through a reduction in the overall variance of the estimates and the mean value as n increases.

3.2 Symmetric Dabrowska estimate

As discussed in the introduction, under the hypothesis that the data set collected in [10] has no inherent rank order we expect the underlying marginal distributions to be symmetric.

While the symmetry is motivated here by the data source [10], it is reasonable to assume that this idea could apply in any situation where there is

symmetry in distribution of the the pairs. As an example, the time, after surgery for a pair of eyes to become fully healed, or the lifetime of identical light bulbs produced in pairs. In both of these cases we see that see there is no item within the pair that is inherently the first item and so we propose that the underlying marginal distributions will be the same.

Firstly we will cover the uncensored a MLE in Section 3.2.1 and then adapt this to the censored case in 3.2.2.

3.2.1 Bivariate EDF

In this section we elaborate on an existing proof describing the MLE of a bivariate symmetric distribution that does not consider censoring [35]. The core idea here can then carried over to solve the case involving censored data.

Proof. Given n IID observations $\{(x_i, y_i)\}_{i=1}^n$ from a bivariate distribution $F(x, y)$ where the underlying random variables (X, Y) are exchangeable and so,

$$F(x, y) = F(y, x),$$

the MLE is given by,

$$\hat{F}_n^{sym}(x, y) = \frac{1}{2}(\hat{F}_n(x, y) + \hat{F}_n(y, x)), \tag{35}$$

where,

$$\hat{F}_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(x_i \leq x, y_i \leq y),$$

is the MLE without the assumption of symmetry.

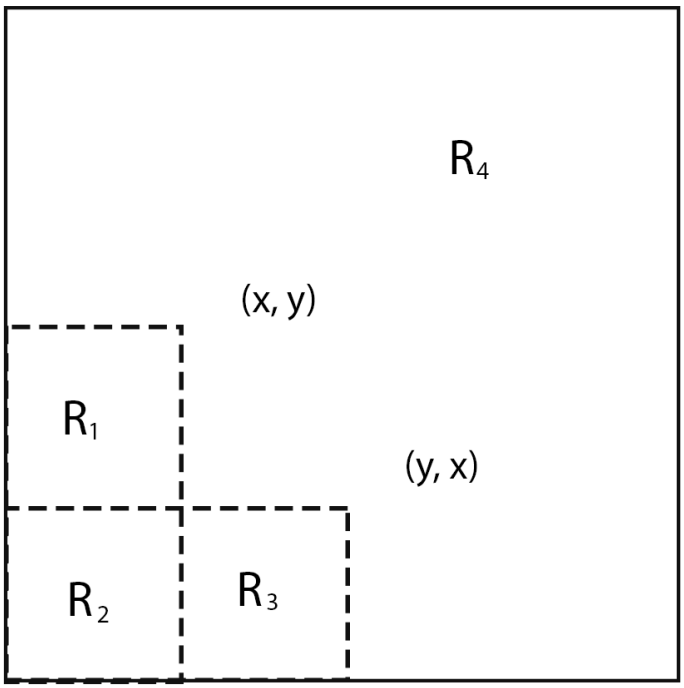


Figure 24: Probability regions R_i with $n = \sum_{i=1}^{4} n_i$ and $n_i = \sum_{i=1}^{n} \mathbb{I}((x_i, y_i) \in R_i)$ represents the observed frequency of random variable N_i probability $P_i = \mathbb{P}((x_i, y_i) \in R_i)$ in region .

We split the range of (X, Y) into regions such that $n_1 = \sum_{i=1}^{n} \mathbb{I}((x_i, y_i) \in R_1)$ is the observed frequency of random variable N_1 in region R_1 with probability $P_1 = P((X, Y) \in R_1)$ as shown in Figure 24. We wish to estimate $F(x, y) = P_1 + P_2$ using maximum likelihood where the MLE is given over all the regions as,

$$L(\{n_i\}|F) = \prod_{i=1}^{4} P_i^{n_i} = P_1^{n_1} P_2^{n_2} P_3^{n_3} P_4^{n_4},$$

using symmetry, $P_1 = P_3$ and so,

$$L(\{n_i\}|F) = P_1^{n_1+n_3} P_2^{n_2} (1 - 2P_1 - P_2)^{n_4}$$

taking logarithms we have,

$$\log\left(L(\{n_i\}|F)\right) = (n_1 + n_3)\log(P_1) + n_2\log\left(P_2\right) + n_4\log(1 - 2P_1 - P_2)$$

and then differentiating w.r.t P_1 and P_2,

$$\frac{d\log(L)}{dP_1} = \frac{n_1 + n_3}{P_1} - \frac{2n_4}{1 - 2P_1 - P_2}$$

$$\frac{d\log(L)}{dP_2} = \frac{n_2}{P_2} - \frac{n_4}{1 - 2P_1 - P_2}$$

$\frac{d\log(L)}{dP_1} = 0$ gives,

$$P_1 = \frac{1}{2n}(n_1 + n_3)$$

$\frac{d\log(L)}{dP_2} = 0$ gives,

$$P_2 = \frac{n_2}{n}$$

to get the combined symmetric estimate,

$$F_n^s(x, y) = P_1 + P_2 = \frac{1}{2n}(n_1 + 2n_2 + n_3) = \frac{1}{2}(F_n(x, y) + F_n(y, x))$$

□

This estimator now places probability mass $1/2n$ at point of observations (x_i, y_i) unlike the EDF which places probability mass $1/n$.

3.2.2 Dabrowska estimate

Given bivariate failure times (T_1, T_2) with corresponding independent bivariate censoring times (C_1, C_2), observable random variables are defined as $(Y_1, Y_2) = (\min(T_1, C_1).\min(T_2, C_2))$ and $(\Delta_1, \Delta_2) = (\mathbb{I}(T_1 < C_1), \mathbb{I}(T_2 < C_2))$, Dabrowska provides an estimate for $S(t_1, t_2)$ by defining bivariate hazard in terms of the joint distribution of (Y_1, Y_2) and (Δ_1, Δ_2). Given functions,

$$K_1(t_1, t_2) = \mathbb{P}(Y_1 > t_1, Y_2 > t_2, \Delta_1 = 1, \Delta_2 = 1), \tag{36}$$
$$K_2(t_1, t_2) = \mathbb{P}(Y_1 > t_1, Y_2 > t_2, \Delta_1 = 1), \tag{37}$$
$$K_3(t_1, t_2) = \mathbb{P}(Y_1 > t_1, Y_2 > t_2, \Delta_2 = 1), \tag{38}$$
$$H(t_1, t_2) = \mathbb{P}(Y_1 > t_1, Y_2 > t_2). \tag{39}$$

Which are used to define univariate cumulative hazard functions and a bivariate cumulative hazard function as,

$$\Lambda_{11}(t_1, t_2) = \int_0^{t_1} \int_0^{t_2} \frac{K_1(du, dv)}{H(u-, v-)},$$
$$\Lambda_{10}(t_1, t_2) = -\int_0^{t_1} \frac{K_2(du, t_2)}{H(u-, t_1)},$$
$$\Lambda_{01}(t_1, t_2) = -\int_0^{t_2} \frac{K_3(t_2, dv)}{H(t_2, v-)}.$$

Given iid right censored observations, as in equation 9. Dabrowska [7] uses the following empirical functions as estimates of the above,

$$\hat{K_1}(t_1, t_2) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_{1i} > t_1, y_{2i} > t_2, \delta_{1i} = 1, \delta_{2i} = 1), \tag{40}$$
$$\hat{K_2}(t_1, t_2) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_{1i} > t_1, y_{2i} > t_2, \delta_{1i} = 1), \tag{41}$$
$$\hat{K_3}(t_1, t_2) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_{1i} > t_1, y_{2i} > t_2, \delta_{2i} = 1), \tag{42}$$
$$\hat{H}(t_1, t_2) = n^{-1} \sum_{i=1}^{n} \mathbb{I}(y_{1i} > t_1, y_{2i} > t_2). \tag{43}$$

We will now apply the logic from section 3.2.1 to produce symmetric empirical estimates of equations (39)–(42), which will result in a symmetric Dabrowska estimate through the relationship to the cumulative hazard functions described above. As with the proof shown above, we begin by splitting the range of failure times (T_1, T_2) into regions such that $n_1 = \sum_{i=1}^{n} \mathbb{I}(y_{1i}, y_{2i} \in R_1, \delta_{1i} = 1)$ is the observed frequency of random variable N_1 in region R_1, the probability of which is given by $P_1 = P(N_1 \in R)$.

Here however, we split the regions by components such that only the data with the first element censored is used. This will allow the estimation of $K_2(t_1, t_2)$.

As before we can express the maximum likelihood in terms of the probability observed frequencies and their respective probabilities,

$$L(\{n_i\}|S) = P_1^{n_1} P_2^{n_2} P_3^{n_3} P_4^{n_4},$$

which due to symmetry we have $P_1 = P_3$, as before we can also take logarithms and maximise over parameter P_i. We wish to estimate the region given by,

$$K_2(t_1, t_2) = 1 - (P_1 + P_2),$$

which, using the results from before, leads to,

$$\hat{K}_2^{sym}(t_1, t_2) = 1 - \left(\frac{1}{2n}(n_1 + 2n_2 + n_3) \right),$$

where n_1 and n_2 are the observed number of elements in R_1 and R_2 in which the first element is censored. We can rewrite this equation as,

$$\begin{aligned} \hat{K}_2^{sym}(t_1, t_2) &= \frac{1}{2}(\hat{K}_2(t_1, t_2) + \hat{K}_2(t_2, t_1)) \\ &= \frac{1}{2}(\hat{K}_2(t_1, t_2) + \hat{K}_3(t_1, t_2)) \end{aligned}$$

.

we can also define the second marginally censored estimator as follows,

$$\begin{aligned} \hat{K}_3^{sym}(t_1, t_2) &= \frac{1}{2}(\hat{K}_3(t_1, t_2) + \hat{K}_3(t_2, t_1)) \\ &= \frac{1}{2}(\hat{K}_3(t_1, t_2) + \hat{K}_2(t_1, t_2)) \end{aligned}$$

.

and by similar logic, we obtain the following, leaving us with the follows symmetric estimators,

$$\hat{K}_1^{sym}(t_1, t_2) = \frac{1}{2}(\hat{K}_1(t_1, t_2) + \hat{K}_1(t_2, t_1)), \tag{44}$$

$$\hat{K}_2^{sym}(t_1, t_2) = \frac{1}{2}(\hat{K}_2(t_1, t_2) + \hat{K}_3(t_1, t_2)), \tag{45}$$

$$\hat{K}_3^{sym}(t_1, t_2) = \frac{1}{2}(\hat{K}_3(t_1, t_2) + \hat{K}_2(t_1, t_2)), \tag{46}$$

$$\hat{H}^{sym}(t_1, t_2) = \frac{1}{2}(\hat{H}(t_1, t_2) + \hat{H}(t_1, t_2)). \tag{47}$$

As with the symmetric empirical distribution these estimates place probability mass $1/2n$ at each point (x_i, y_i) where their counterparts would place probability mass $1/n$.

3.2.3 Synthetic Example

Here we present synthetic data simulations using the symmetric Dabrowska estimate.

Figure 25 shows a simple comparison between the Dabrowska and symmetric Dabrowska estimates. Both were generated using the exact same underlying failure time and censoring time data from a bivariate normal distribution as given by equation 13 with 50 points of data and parameters show in equation 48. Distances between true distribution, as defined by equation 14 were 0.2725 for (a) and 0.2502 for (b).

Figure 26 shows convergence of the symmetric Dabrowska estimate using parameters from a bivariate normal distribution as described by the PDF in equation 13. The parameters for this distribution are,

$$\mu = \begin{pmatrix} 5 \\ 5 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 3 & 1.5 \\ 1.5 & 3 \end{pmatrix}, \tag{48}$$

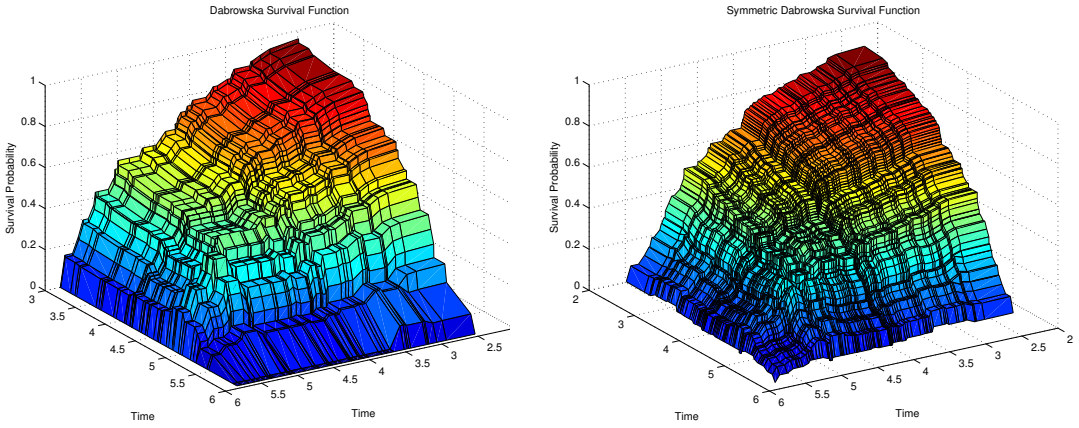
The data increased sequentially with each plot from $n = 50, 150$ and $n = 500$. (b)–(c). Here we see that as the amount of data increases the distance to the true distribution is decreases.

Figure 27 shows what happens when the symmetric Dabrowska estimate is used when the underlying data is not symmetric. Here failure and censoring time data were generated from a normal distribution, however here the parameters were not symmetric.

$$\mu = \begin{pmatrix} 4 \\ 6 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & 1.5 \\ 1.5 & 2 \end{pmatrix}, \tag{49}$$

As we can see from the true distribituon in Figure 27 (a), as the amount of data increases in Figure 27 (b)—(c) convergence is not taking place to the true distribution.

Finally Figure 28 compares the convergence of the Dabrowska and Symmetric Dabrowska estimates when the underlying distribution is symmetric (upper panel) and not symmetric (lower panel) for sequentially increasing amounts of data. In the upper panel both estimates converge to the true failure time distribution the data was sampled from where as in the lower panel only the Dabrowska estimate converges since the underlying failure time distribution is not symmetric.



(a) (b)

Figure 25: Comparison between the Dabrowska estimate which makes no assumptions about the marginal distribution, (a) and the new symmetric version (b). Here 50 points of data from a symmetric bivariate normal distribution with equally distributed censoring data were used. Distribution given by equation 48. Distances between true distribution, as defined by equation 14 given by 0.2725 and 0.2502 respectively.

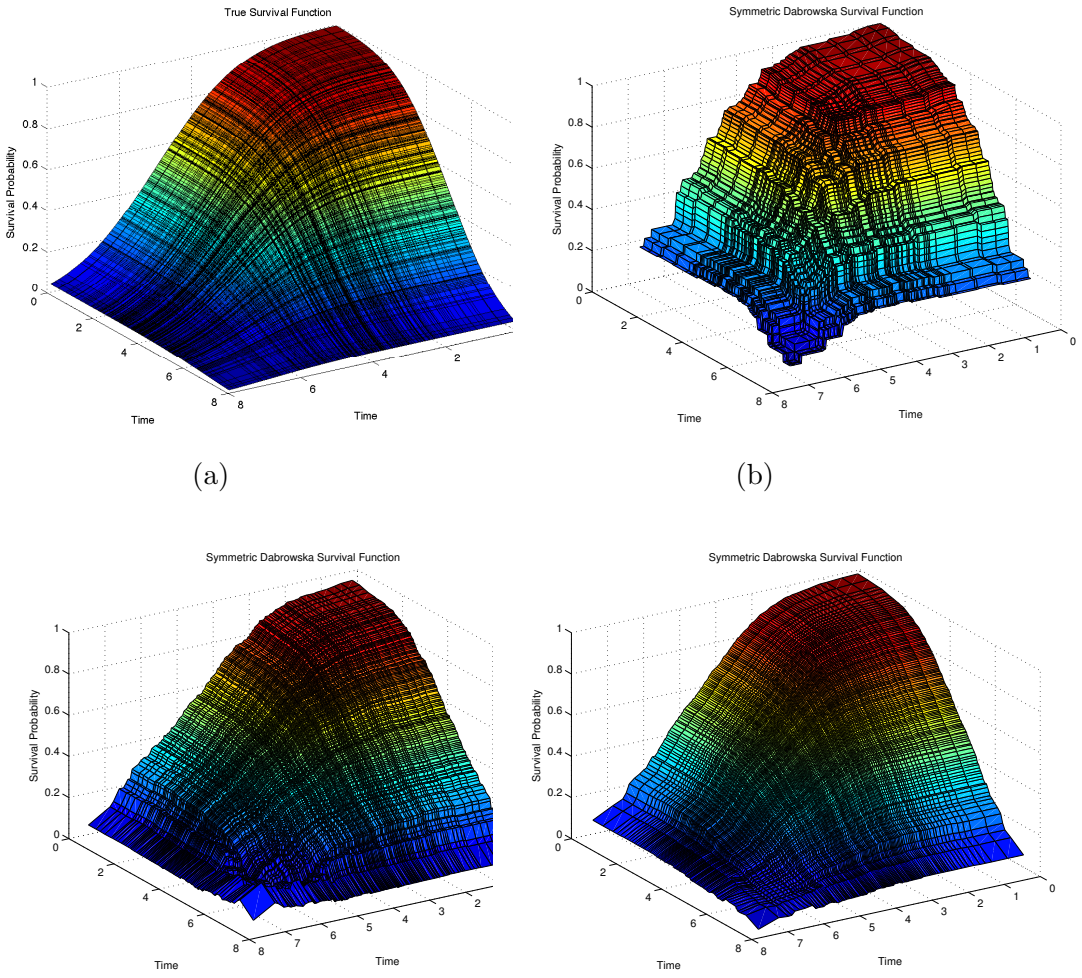
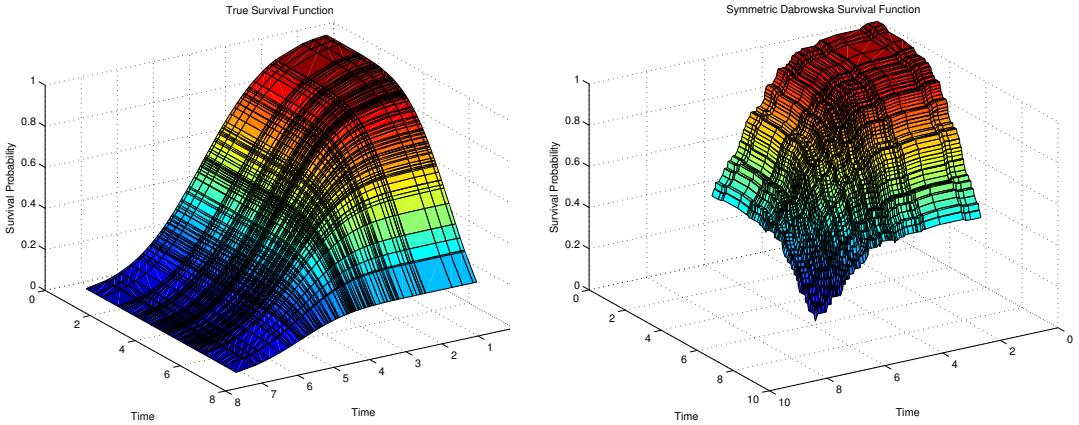
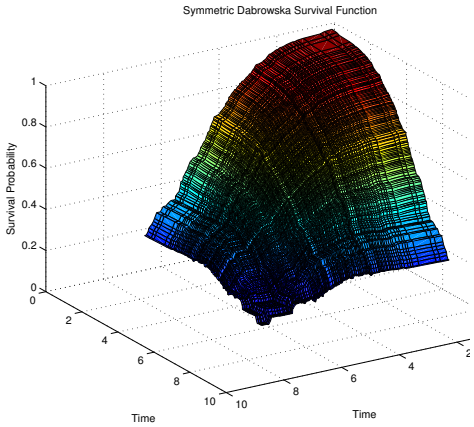


Figure 26: Convergence of the symmetric Dabrowska estimate to true failure time distribution using synthetic data. (a) Survival function of the actual normal distribution as defined with the parameters in equation 48. (b)–(d) Symmetric Dabrowska estimates using failure time and censoring time data generated from two bivariate normal distributions with parameters described by equation 48. Amount of data increased sequentially ((b) $n = 50$, (c) $n = 150$, (d) $n = 500$). Supremum norm between Dabrowska estimate to true survival function give by 0.1528, 0.1323 and 0.0556 respectively.

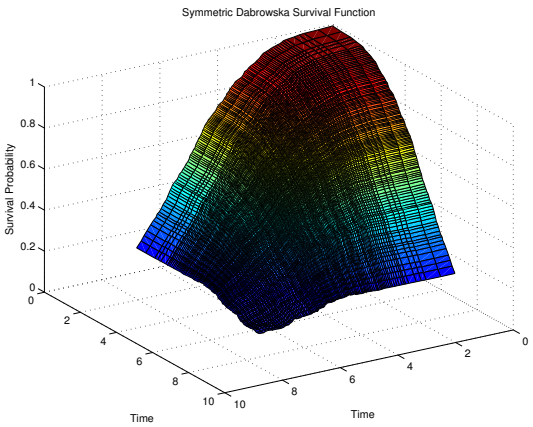


(a)

(b)



(c)



(d)

Figure 27: Convergence of the symmetric Dabrowska estimate to the wrong distribution when the underlying failure time distribution is not symmetric. Failure and censoring data generated from a bivariate normal distribution with parameters given by 49. (a) True failure time distribution (b)–(d) Symmetric Dabrowska estimate as amount of data used increases sequentially ((b) $n = 50$, (c) $n = 150$, (d) $n = 500$. Supremum norm between Dabrowska estimate and true survival function given by 0.3154, 0.2976 and 0.2777 respectively.

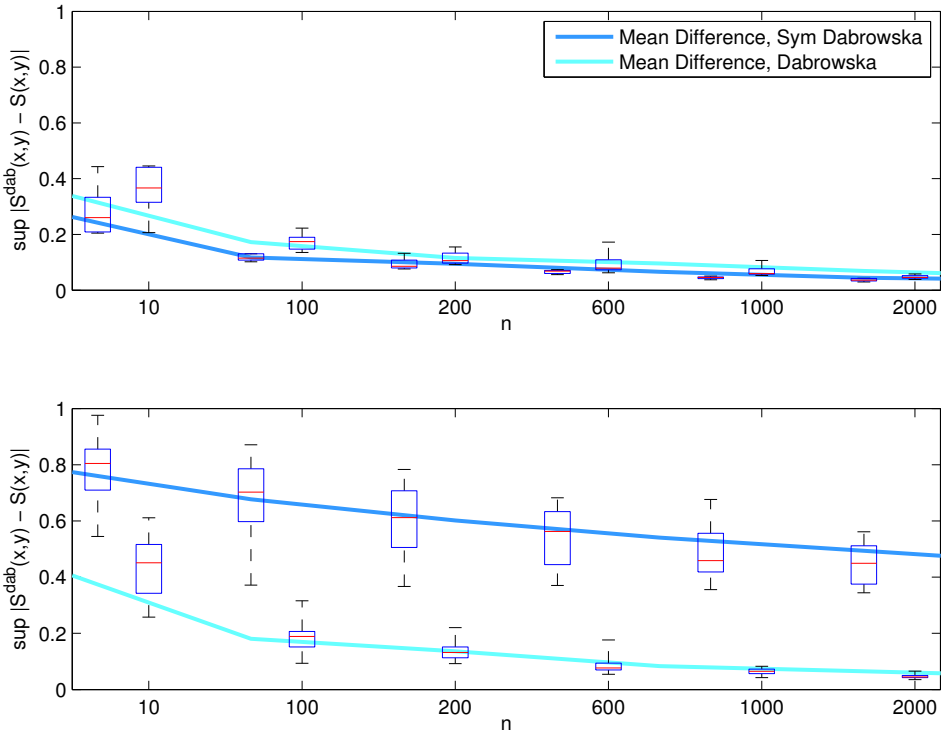


Figure 28: Comparison of convergence for Dabrowska and Symmetric Dabrowska estimates when the underlying distribution is symmetric (upper panel) not symmetric (lower panel). The maximum difference between estimate and reference distribution is calculated as in equation 14 for sequentially increasing amounts of data as in the Chapter 2 Figure 12 to produce a distribution for each n. In the upper panel, both estimates converge to the failure time distribution the data was sampled from. In the lower panel the Dabrowska estimate converges as n becomes large while the symmetric estimator does not. The symmetric Dabrowska estimate will only converge to the correct distribution when the underlying failure distribution is symmetric.

Chapter 4

4 Non-parametric Survival Analysis of Published Data

Here we use the statistical techniques discussed and developed in chapters 2 and 3 to provide analysis of the already published data in [10]. Given this data set, we will try to answer two questions. 1) Is the assumption of log normality reasonable as a description of the marginal distribution of time to IgM to IgG1 class switch and for the marginal distribution of time to differentiate to plasmablast in B cells in this data set? If so, is it reasonable assumption for the joint distribution of time-to-event in pairs of sibling B cells? 2) How does the time to event vary with generation?

4.1 Testing the Assumption of Log-Normal Marginal Distributions

In the introduction we discussed testing the assumption of log-normality for the distribution of time-to-event. Existing work [8] [18] shows that the marginal distribution of time to divide and time to die in B cells are well described by a log normal distribution.

Using the data collected during [10] we can further test similar assumptions by looking at the marginal distribution of the time taken for B cells to IgM to IgG1 class switch, and the marginal distribution of time to differentiate into a plasmablast. We can further look as this distribution in different generations of B cells and provide analysis for each of them.

If we are able to determine that a log normal distribution is appropriate in the marginal case, then it makes sense to use the bivariate statistical tools we have developed in chapters 2 and 3 to then consider the assumption that time-to-event in pairs of sibling cells should follow a bivariate log normal distribution. If evidence suggests marginal distributions are not appropriate or given the available data we can not determine with a high enough degree of certainty if it is possible, then it does not make sense to consider this type of distribution for the time to event in pairs of B cells.

Figure 29 shows the survival distributions for the time to differentiate to plasmablast for all generations of available data. Each subplot shows a set of survival curves for one of the generations for which data was collected in [10] comparing the Kaplan-Meier estimate of the time to differentiate to plasmablast and a Log Normal survival function computed using the parameters in [10].

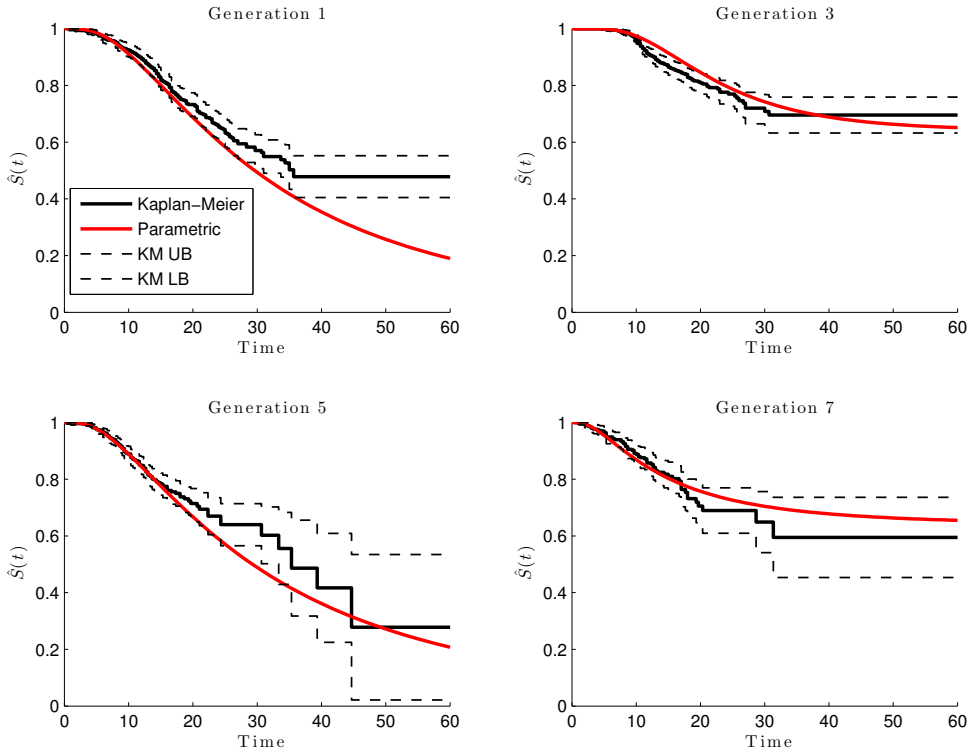


Figure 29: The above plots show the log normal survival function estimates [10], and Kaplan-Meier survival function estimates for the time to differentiate to a plasmablast for all four generations of available B cell data. KM UB and KM LB show the 95% confidence intervals.

As well as performing hypothesis tests we can make a quantitative comparison between the two survival curves through statistics including the expected time to event and the probability of finite event time p. Since the survival distributions are often expected to be defective and so do not go to zero, we will compare the conditional expectation up to a given time point which we define in terms of our dataset in equation 6 as

$$t^* = \max(\{y_i\}_{i=1}^n : \delta_i = 1). \tag{50}$$

We then calculate the conditional expectation from the Log-Normal param-

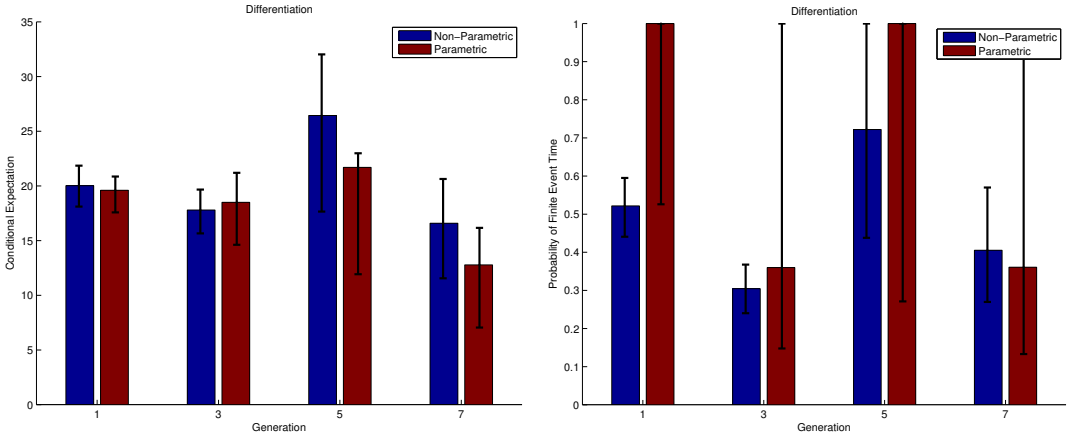
eters, given PDF $f_{LN}(x)$ and CDF $F_{LN}(x)$ as

$$\mathbb{E}(T|T < t^*) = \int_0^{t^*} \frac{f_{LN}(x)}{F_{LN}(t^*)} dx. \tag{51}$$

and calculate the conditional expectation from the Kaplan-Meier survival function given by $\hat{S}(x)$ as

$$\mathbb{E}(T|T < t^*) = \int_0^{t^*} \frac{\hat{S}(x) - \hat{S}(t^*)}{1 - \hat{S}(t^*)} dx. \tag{52}$$

Figure 30 shows similar results across different generations for the conditional expectation of time-to-event with larger differences occurring in generations 5 and 7. The probability that cells don't undergo differentiation varies with generation but shows no real trend in either models, with both models giving estimates that do not compare well in generations 1 and 5.



(a) (b)

Figure 30: Conditional expectation of time to differentiate to plasmablast and probability p that cells do not undergo the event for all generations using both Kaplan-Meier (Non-Parametric) estimate and estimates produced in [10] (Parametric). Bootstrapping was used to generate 95% confidence intervals for these estimates.

We can now take a visual look at the survival distribution estimates for time to IgM to IgG1+ class switch. These are shown for all four generations

in Figure 31. In most cases there is some agreement between the two methods, especially at early time points, with generation 3 showing the worst correspondence between the two methods of estimation. In Generation 1 there is almost no estimate available, due to the small amount of switching events observed in generation 1 B cells which can be seen from the event data plot in Figure 33.

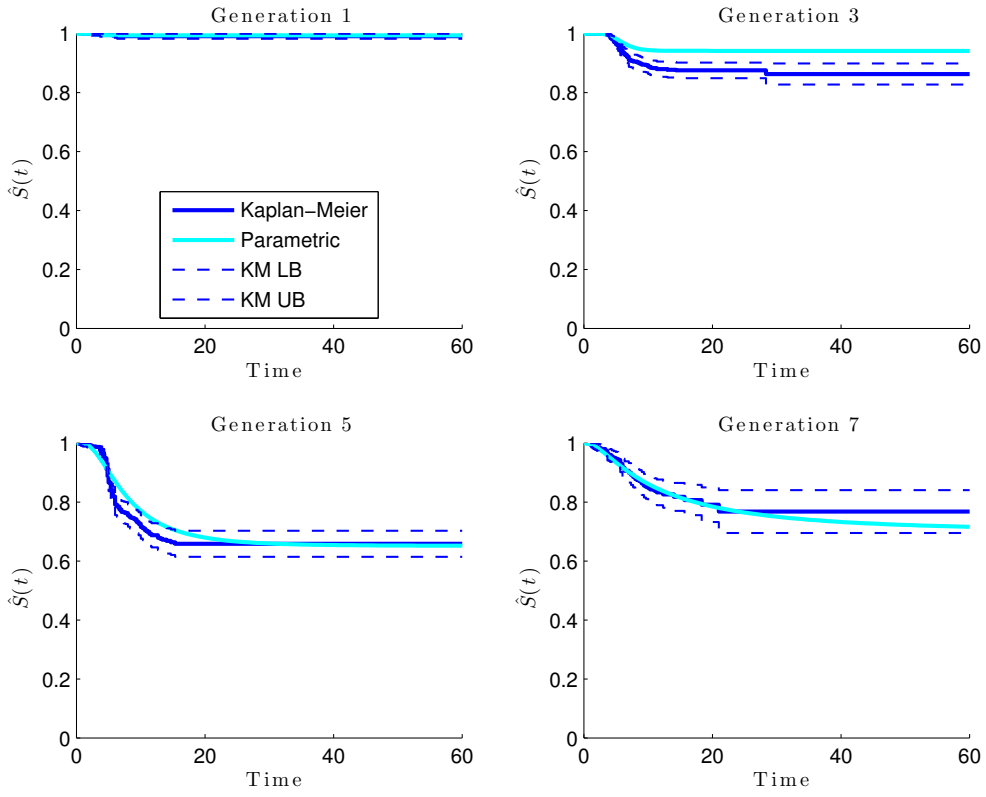
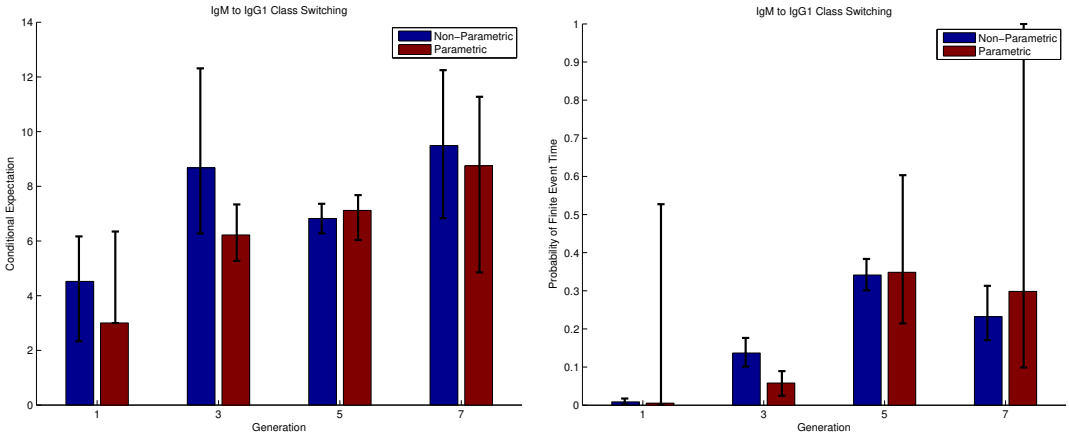


Figure 31: The above plots show the log normal survival function estimates [10], and Kaplan-Meier survival function estimates for the time to IgM to IgG1+ class switch for all four generations of available B cell data.

Figure 32 shows the conditional expectation as defined above and the probability of finite event time for both estimation methods. Here we see some correspondence between the two methods. We see a high degree of

similarity between the two methods of estimation for the probability that switching does not occur, p.



(a) (b)

Figure 32: Conditional expectation of time to IgM to IgG1 class switch and probability p that cells do not undergo the event for all generations using both Kaplan-Meier (Non-Parametric) estimate and estimates produced in [10] (Parametric). Bootstrapping was used to generate 95% confidence intervals for these estimates.

As mentioned above there was not enough data available to produce a good estimate for generation 1 switching. Figure 33 shows the number of observations of each type available in each generation. This shows us the number of plasmablast differentiation events and number of IgM to IgG1 class switching events available to produce survival function estimates across all generations. As a side note, we see a characteristic features observed in [19]. They [19] note that IgG1+ class switching does not begin until after the first division has occurred, with a plateau occurring after six divisions. This helps explains the low number of observations in generation 1.

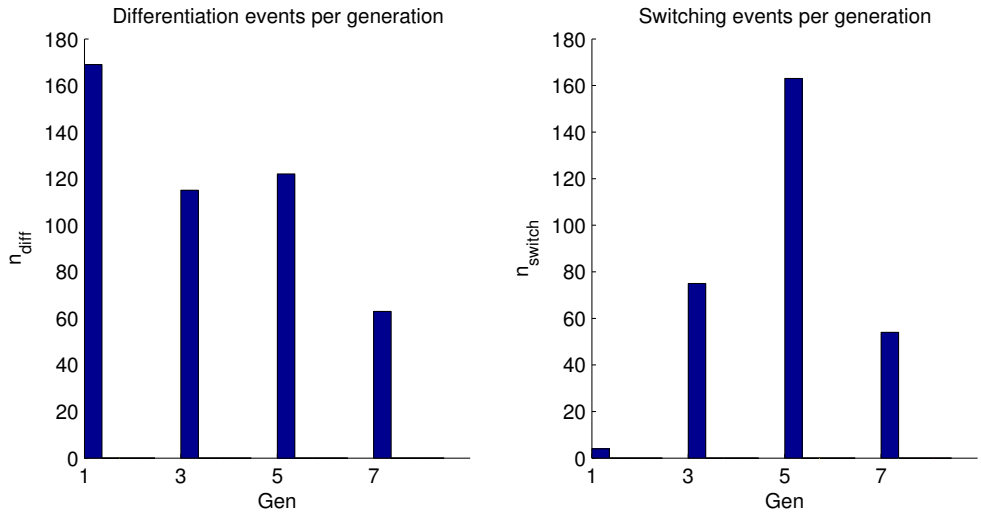


Figure 33: The number of plasmablast differentiation events and number of IgM to IgG1 class switching events available to produce survival function estimates across all generations.

This behaviour of cell switching helps support the idea that in most cases before generation two we expect to see very little isotype switching, which explains the low amount of event count for generation 1 switching and the corresponding low probability of finite event time.

Next, we performed a goodness of fit test to check the hypothesis that time to switch and time to differentiate data is well described by a log normal distribution using the parameters estimated for a log normal distribution in [10]. The test chosen was a modified version of the Kolmogorov-Smirnov test for right censored data [12] described in 2.4.2. Here we take a collection of right censored observations, shown in uncensored form by the Kaplan-Meier estimate, and perform a two sided statistical test to check if these observations follow a log-normal distribution with parameters given by the estimates in [10].

As we cannot compare estimates based upon the same data set when performing hypothesis tests, we had to split the available data using one half to produce the log normal distribution parameters (instead of using the parameters in [10] which used the full data set), and the other half as the data whose distribution we suspect to be log normal. The results of the test are shown in the table below.

We can see in most cases at the level $\alpha = 0.05$, the tests would reject

Generation	P Value - Differentiation	P Value - Switching
1	0.0070	0.0059
3	1.6606e-7	0.0004
5	0.0050	0.4635e-9
7	0.0953	0.0043

Table 4: Results of statistical test to check the possible log normality of distribution of time to IgM to IgG1 class switch and time to differentiate to plasmablast. This test was performed using a modified Kolmogorov-Smirnov test for right censored data [12].

the hypothesis that a log normal distribution is an appropriate choice for distribution of time-to-event for differentiation to plasmablast and IgM to IgG1 class switch. However we must be careful in reading too much into these tests, as not only did we half the data set, in some instances there would not have been appropriate data to begin with (for example generation 1 switching).

Given the results of the hypothesis tests performed and the statistics calculated, we can conclude that while the log normal distribution may not be the perfect choice for the distribution of time to class switch and differentiate in B cells, it may be an appropriate choice is coarse models where a parametric model is necessary or preferable. Ideally it would be best to collect a larger data set in order to produce more accurate statistics, and perform hypothesis tests with which we can have greater confidence before making definitive conclusions about the class of distributions B cells should follow.

4.2 Testing the Change in Time-To-Event as a Function of Generation

In this section we investigate how time-to-event varies with cell generation. Firstly we will take a qualitative look at how the survival curves vary with generation, we compare the expected time-to-event across generations, and the probabilities of finite event time. Lastly we performed Log-Rank hypothesis tests to see if it is possible to reject the hypothesis that any two survival distributions share the same underlying time-to-event distribution.

Figure 34 shows the Kaplan-Meier survival function estimates of the time-to-event for all possible generations and cell fates for which data was col-

lected. Figure 35 shows the number of events available for all generations and event types.

In the previous section, when comparing the conditional expectation between the parametric and non-parametric models, we conditioned the expectations from both models such that $T < t^*$ where t^* is defined in equation 50. In order to make the comparison fair across generations such that each conditional is the same we define a new t^*. If t_i^* gives the value of t^* for generation i then we conditioned all expectations of a given fate on t^{**} which is defined as,

$$t^{**} = \min\{t_1^*, t_3^*, t_5^*, t_7^*\}. \tag{53}$$

An import consideration when using this definition is that t^* may reflect the censoring distribution more than the event distribution if mean censoring time is less than mean event time. Tables 5, 6, 7 and 8 show these conditional expectations for all event types and generations as well as the probabilities of finite event time.

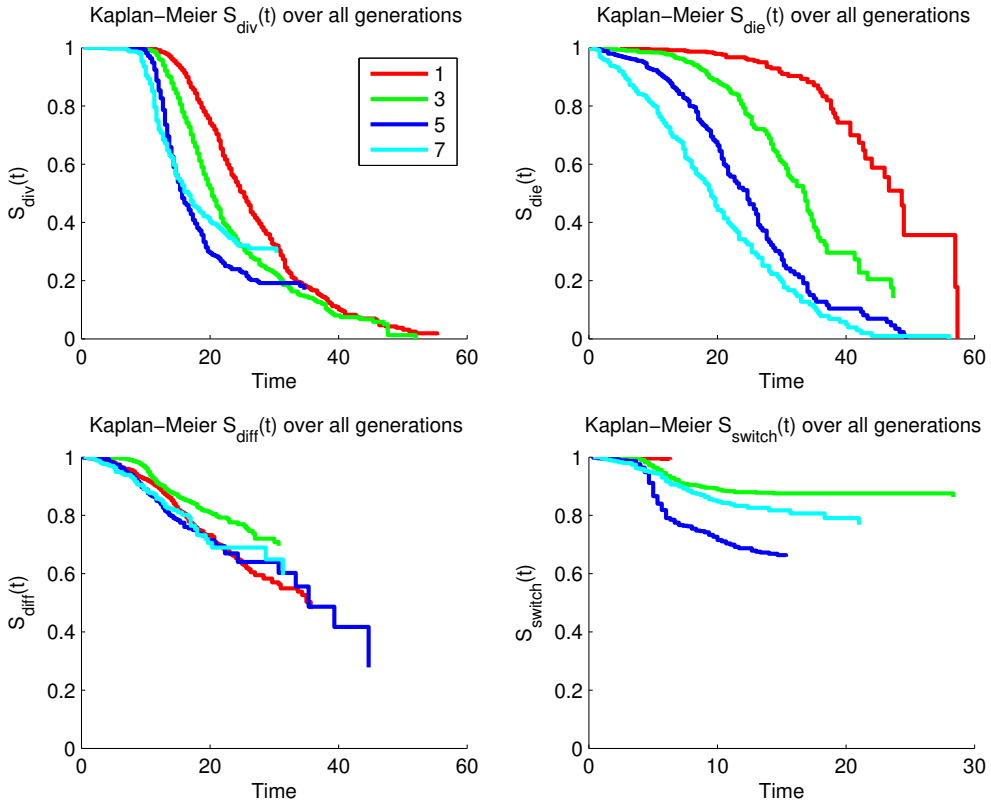


Figure 34: Kaplan-Meier survival function estimates showing the time-to-event for the B cell data set [10]. Each plot shows all generations within a specific cell fate.

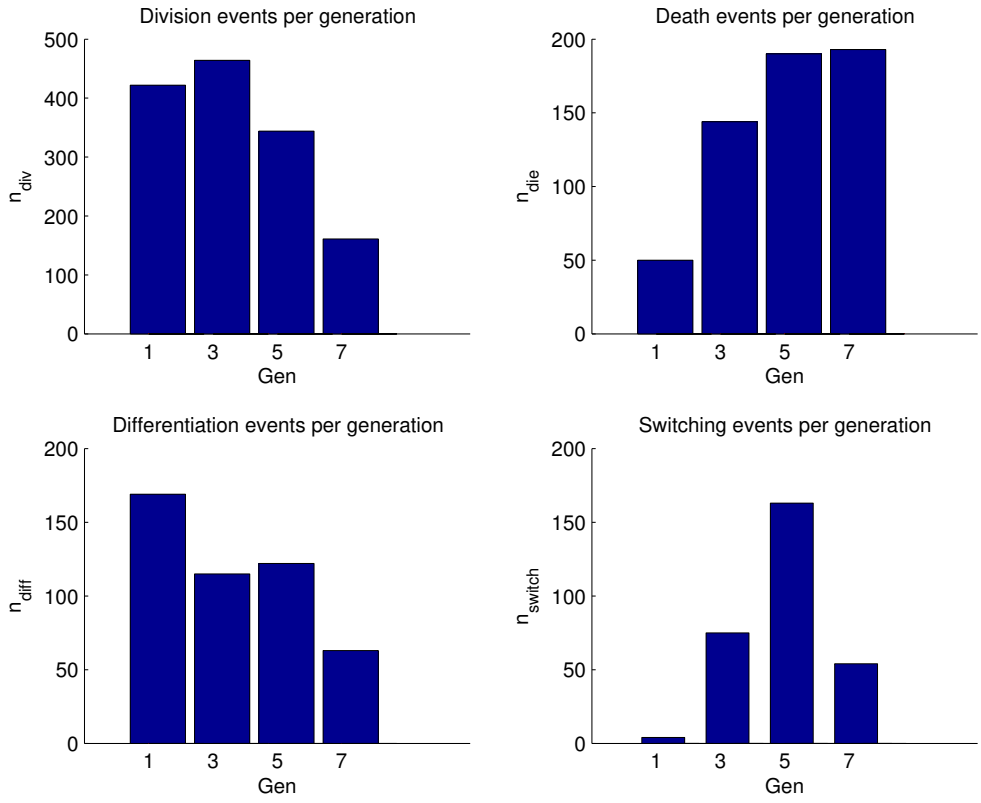


Figure 35: Number of observations for every event type and generation for the data published in [10].

We begin by considering the time to cell division. The Kaplan-Meier estimates for time to division are shown in Figure 34, here we see that all curves begin with a period of guaranteed survival, this is because there is a minimum time it takes for a cell to begin the division process of approximately 8 hours [21]. We then see a steady decline in survival over the next 15—60 hours with generations 1 and 3 eventually reaching a value close to zero, at which point most cells have undergone division. On the other hand survival curves for generation 5 and 7 end at a value much greater than zero, indicating that cells in later generations are less likely to divide. This is in line with expectations as towards the end of an in vivo immune response cells stop proliferating and start dying by apoptosis [21].

Fate	**Division**			
Gen	1	3	5	7
p	0.987	1.000	0.831	0.704
$\mathbb{E}(T\|T<30.33)$	21.676	19.025	15.610	14.799

Table 5: Table of statistics for time to division for all generations. The column labelled p shows the probability that the even time is finite. The bottom column shows the conditional expectation conditioned on time t^{**} which is defined in equation 53.

In Table 5 we see that the conditional expectation for time to division, conditioned on $T < 30.33 = t^{**}$ where t^{**} is defined in equation 53 and conditional expectation for Kaplan-Meier is defined in 52. Having conditioned the expectation for the purpose of comparison, we do not capture all of the information as generations have a lot of division events that occur after this time point. From these values we see a slow decline in the conditionally expected time to divide as the generation increases. This is something we would expect as it is observed that earlier generations of division take longer, with later division times being smaller. In this table we also see a deceasing probability of finite event time which as explained above is something we would expect.

Fate	**Death**			
Gen	1	3	5	7
$\mathbb{E}(T\|T<47.33)$	37.543	30.605	23.646	19.683

Table 6: Table of statistics for time to die for all generations. Conditional expectation conditioned on time t^{**} which is defined in equation 53. This table does not show a probability of finite event time as death will occur in every cell and so event time is always assumed to be finite even if the estimate does not currently have enough data to estimate it.

Next we consider how time to die changes with generation. Figure 34 shows a plot of the Kaplan-Meier estimates for time to cell death for every generation. We can see here that there is a clear trend, as the generation number increases, the probability of survival decreases. This is clear because the survival functions become smaller with each generation. This is in line with our expectations during an immune response, initially cells are dividing

and diversifying in order to fight an infection, and in later generations these processes stop as cells rapidly die by apoptosis [21]. Table 6 shows the conditional expectation for the time to die with the condition as defined in equation 53. We see the value decreases as the generation increases, again this confirms what we would expect given a normal in vivo immune response.

Fate	**Differentiation**			
Gen	1	3	5	7
p	0.522	0.305	0.722	0.405
$\mathbb{E}(T\|T<30.67)$	17.245	17.798	15.466	14.318

Table 7: Table of statistics for time to differentiate for all generations. The column labelled p shows the probability that the even time is finite. The bottom column shows the conditional expectation conditioned on time t^{**} which is defined in equation 53.

Next we compare the time for cells to differentiate into plasmablasts. Again Figure 34 shows the Kaplan-Meier estimate of time to differentiate for all generations. Here we see that there is much less variability between generations, with most of survival curves intersecting at different time points. Conditional expectations in Table 7 show relatively similar values across generations, with generations 5 and 7 being 2—3 hours smaller than generations 1 and 3. The probabilities of finite event time show no real trend across generations. Thus we conclude that unlike cell division and death it seems that the uncensored times to differentiation do not change in a generation dependent manner for this data set.

Fate	**IgM to IgG1 Class Switching**			
Gen	1	3	5	7
p	0.008	0.137	0.341	0.232
$\mathbb{E}(T\|T<15.333)$	N/A	6.697	6.822	6.897

Table 8: Table of statistics for time to IgM to IgG1 class switch for all generations. The column labelled p shows the probability that the even time is finite. The bottom column shows the conditional expectation conditioned on time t^{**} which is defined in equation 53.

Figure 35 shows the number of switching events that occurred across each generation. Here we see characteristic features observed in [19]. They note

that IgG1+ class switching does not begin until after the first division has occurred, with a plateau occurring after six divisions.

The Kaplan-Meier survival curves for class switching in Figure 34 show similar values at early time points, with differences becoming more pronounced over time. Generation 1 cells had only a small number of events, with more data is available to produce estimates in later generations and the most being in generation 5. All survival curves have a period during the first few hours with a high probability of survival, eventually leading to a plateau. Since the amount of data available to produce a survival curve for generation 1 switching is so low, it has been excluded when calculating a value for t^{**}. We conclude that as the generation number increases, more switching events occur, the expectation conditioned on t^{**} is consistent across generations, but the survival curves themselves vary with generation showing survival probability decreasing faster in later generations.

In order to further quantify the difference in survival as the generation increases, we present the results of Log-Rank statistical tests. Figure 36 shows the results of Log-Rank hypothesis tests comparing the unconditional survival function estimates for all generations of a given fate, with all other generations of that fate. The null hypothesis is that the two unconditional survival functions are the same, as given by equation 16. To account for the fact we are performing multiple tests, we have used the Holm-Bonferroni method discussed in section 2.4.6 to adjust an α value of 0.05.

Figure 36 shows the results of all hypothesis tests performed and the resulting p values. We see in all cases, the tests reject H_0, saying that no two generations follow the same distribution. In the cases of division and death we would expect the distributions to change with generation, B cells are more likely to die and less likely to divide as the generation increases [21], this is also demonstrated through the estimates and figures presented above.

Gen	1	3	5	7
1	1.00	0.00	0.00	0.00
3	0.00	1.00	0.00	0.00
5	0.00	0.00	1.00	4.97e-6
7	0.00	0.00	4.97e-6	1.00

(a) Division

Gen	1	3	5	7
1	1.00	6.07e-8	0.00	1.77e-8
3	6.07e-8	1.00	5.22e-9	1.25e-2
5	0.00	5.22e-9	1.00	2.35e-1
7	1.77e-8	1.25e-2	2.35e-1	1.00

(b) Death

Gen	1	3	5	7
1	1.00	0.00	0.00	0.00
3	0.00	1.00	0.00	0.00
5	0.00	0.00	1.00	4.29e-4
7	0.00	0.00	4.29e-4	1.00

(c) Differentiation

Gen	1	3	5	7
1	1.00	9.60e-6	0.00	0.00
3	9.60e-6	1.00	0.00	0.00
5	0.00	0.00	1.00	1.68e-4
7	0.00	0.00	1.68e-4	1.00

(d) Class Switching

Figure 36: Log-Rank Hypothesis tests comparing all generations within a given fate. The null hypothesis is shown in equation 16. A green cell indicates the test accepts H_0 and a red cell indicates its rejection. Each cell contains the p values produced by the Log-Rank test. The rejection level of $\alpha = 0.05$ was modified for multiple testing using the Holm-Bonferroni procedure described in section 2.4.6.

Differentiation and class switching show that despite similar conditional expectation estimates across generations, the hypothesis tests reject H_0 in all cases. We must be careful when comparing statistics, as the expectation is conditioned on $T < t^{**}$, where the hypothesis test uses the unconditional distribution and the underlying hypothesis in the parametric model is that all distributions bar T_{die} can be defective with a positive probability mass at $+\infty$. Despite the similarities between the survival functions over the conditional range, when tested over the full range of time points, the survival curves are sufficiently different over $T \geq t^{**}$ that H_0 is rejected.

To investigate how the Log-Rank tests performs when the two survival functions are estimated from data with identical underlying distributions, but with one being possibly defective, we produced the numerical experiment below. Unsurprisingly we see that the Log-Rank test has a higher probability of rejection as the defective distribution places more probability mass at $+\infty$. We first generated two sets of failure time and censoring time data from identical log normal distributions, similar to the distribution for generation 3 division, with parameters given by,

$$\mu = 3 \text{ and } \sigma = 0.25, \tag{54}$$

we then modify one of the data sets, such that a certain amount of the data takes on the value infinity, corresponding to a chosen value for the probability of finite event time. We then calculate the survival functions and perform a Log-Rank test. Repeating this 10000 times, we produced an empirical distribution of the p values generated by the Log-Rank test. This was then repeated while keeping the underlying distributions identical but increasing the probability of finite event time from 0.0 to 0.4 in steps of 0.1 for one of the two distributions.

Figure 37 shows the result of this simulation. We see that when there is no difference between the probabilities of finite event time, the distribution of p values produced by the Log-Rank test is uniform as expected. When we modify the probability of finite event time for one of the data sets, the Log-Rank test produces p values that would result in a rejection of H_0 at the $\alpha = 0.05$ level with greater probability as shown in the plot. This highlights how the probability of finite event time can impact the result of hypothesis tests, and shows that, while the conditional statistics can lead us to believe two distributions are similar, if they are conditioned such that they miss import aspects of the distribution then they can not be reliably used to fully quantify the difference between two distributions.

In this section we have seen that in most cases the distribution of time-to-event in B cells does vary with the generation number, something we would expect given knowledge of the in vivo immune response. However we have also seen that in some cases, while the overall distribution of time-to-event are not the same when comparing two generations of a given fate, this may be due to the proportion of B cells that have the mechanism for that event "switched on". In these cases we see that while a hypothesis test like the Log-Rank test will reject the hypothesis that the underlying distributions are similar, the time-to-event in cells that have the event "switched on" may be the same.

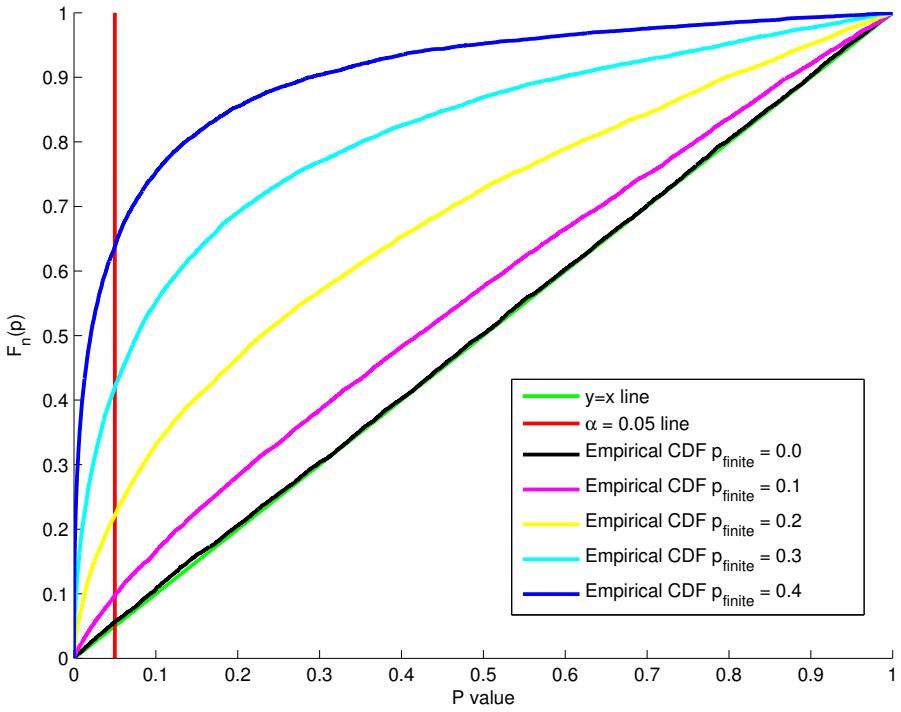


Figure 37: Empirical CDFs of Log-Rank p values showing how the distribution of p values varies as the difference in probability of finite event time between the two survival functions whose distributions are the same when conditioned on being finite. Here 10000 pairs of Kaplan-Meier survival functions are Log-Rank tested to produce an empirical CDF of the p values. This is repeated as the difference in the probability of finite event time is modified from 0.0 to 0.4 in steps of 0.1 by changing the proportion of the data with an infinite value, keeping the underlying distribution of the data the same. Underlying data is normally distributed with parameters shown in equation 54. Here we can see that as the difference becomes larger, the proportion of p values resulting in a rejection at the level $\alpha = 0.05$ increases. We conclude that while the survival functions are identical on a finite interval in all cases, the difference in probability of finite event time causes an increased rejection rate when the unconditional survival function is tested. Thus we would expect that, while B cells can have similar distributions when conditioned on being finite, if they have a significant difference in the probability of finite event time the Log-Rank test will conclude that they do not share the same underlying distributions.

Chapter 5

5 Analysis of Unpublished Data

Using previously unpublished data collected during the experiments reported on in [10] we ask: What is the distribution of the time-to-event data of cells that have already undergone differentiation to plasmablast and/or IgM to IgG1 class switching? Are these distribution different to the published IgG1-Blimp1- data studied in chapter 4?

Here we use unpublished data collected during the experiments for [10] to study the time-to-event in cells that have already undergone class switching (from IgM, now IgG1+), differentiation to plasmablasts (are now Blimp1+), and cells that did both (IgG1+ and Blimp1+). Including IgG1-Blimp1- cells, there are a total of four different initial 'start types' that time-to-event data was collected for.

Figure 38 shows the number of observations for each generation and start type, excluding the already studied IgG1-Blimp1- data which is shown in Figure 35. There were much fewer observations, and in some cases no observations, when compared to the already published data. In some cases however, we would not expect to see any observations. For example with differentiation we would not observe any Blimp1+ events because we are looking at cells that started Blimp1+.

In the best case scenario there are approximately 80 observations for generation 5 IgG1+Blimp1- time to death and time to differentiation. In most cases there are between 0–20. We must also be careful when drawing conclusions based on estimates with small amounts of data.

Figure 39 shows a box plot generated for the time to event for every event type/generation/start type combination showing the minimum and maximum event time that was observed, and the lower, middle and upper quartiles calculated from the uncensored Kaplan-Meier survival function estimate of the respective data set. Across all data sets we see that the extra start type data has a much narrower range of times than their IgG1-Blimp- counterparts, and in most cases the values of the quartiles are smaller, but this is something we would expect this due to the fact we have less data. We can see here that most of the analysis in this section will focus on division and death as we only have 3 data sets for differentiation, and as explained above, no data sets for class switching. In the case of division we have both IgG1+Blimp1- data and IgG1-Blimp1+ data available but no IgG1+Blimp1+ data available. Cell death is the event for which most extra start type data is available, and from this data we were able to produce an estimate for every possible start type.

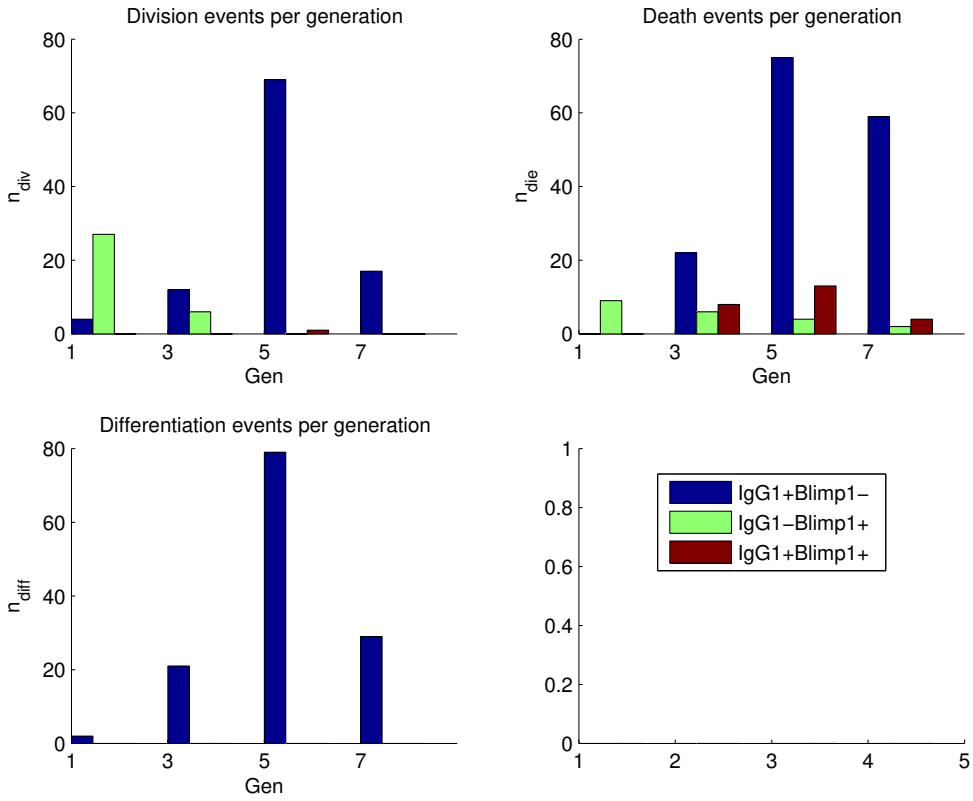


Figure 38: Number of events observed in each generation for every event type for the unpublished data.

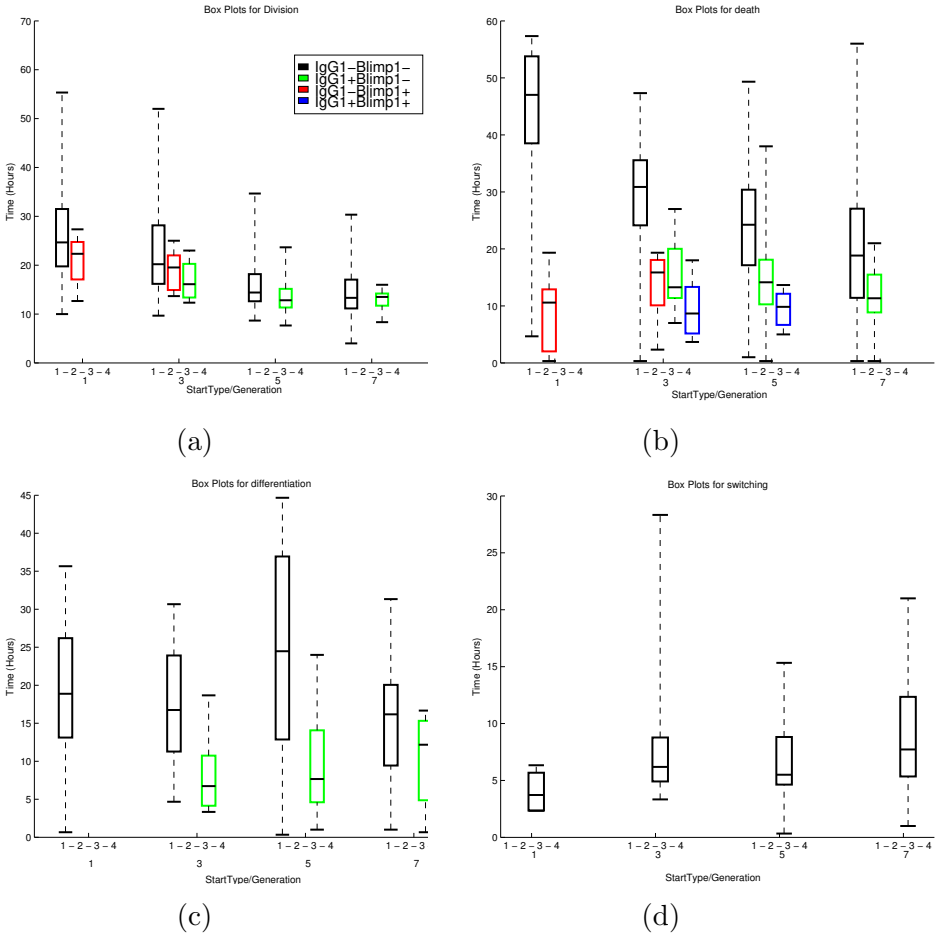


Figure 39: Collection of box plots for times from birth to each fate for all generations and start types. These statistics are calculated from the uncensored Kaplan-Meier survival distributions. Each box plot contains the minimum and maximum event time, upper, lower and middle quartiles.

Figure 40 shows the Kaplan-Meier survival function estimates for time to divide, with each individual plot showing all possible initial start types for available data of that generation. Tables 9 and 10 show the results of Log-Rank tests for the hypothesis that the time-to-division distribution of B cells whose initial start type is IgG1-Blimp1- is the same as the time-to-division distribution of cells whose initial start type is IgG1+Blimp1- and IgG1-Blimp+ respectively.

From the survival curves in Figure 40 we see a period of guaranteed survival, as with IgG1-Blimp1- data, as we would expect. For generation 1 and 3 we see the survival functions drop off rapidly, when compared to the IgG1-Blimp1- estimates, eventually falling to zero or close to zero over the next 10 hours. Survival curves for IgG1+Blimp- data in generation 5 and 7 seem to match their IgG1-Blimp1- counterparts more closely than in other generations, but end earlier, likely due to the lack of available data.

The hypothesis tests performed show no clear pattern as to whether the different start type impacts the time-to-division distribution with cases of both rejection of H_0 and failure to reject in both the IgG1+Blimp1- and IgG1-Blimp1+ tests. However, of note here is the fact that the tests which failed to reject H_0 were based on estimates with significantly less data than the tests that were able to reject H_0, suggesting that lack of a trend may be due to the amount of available data.

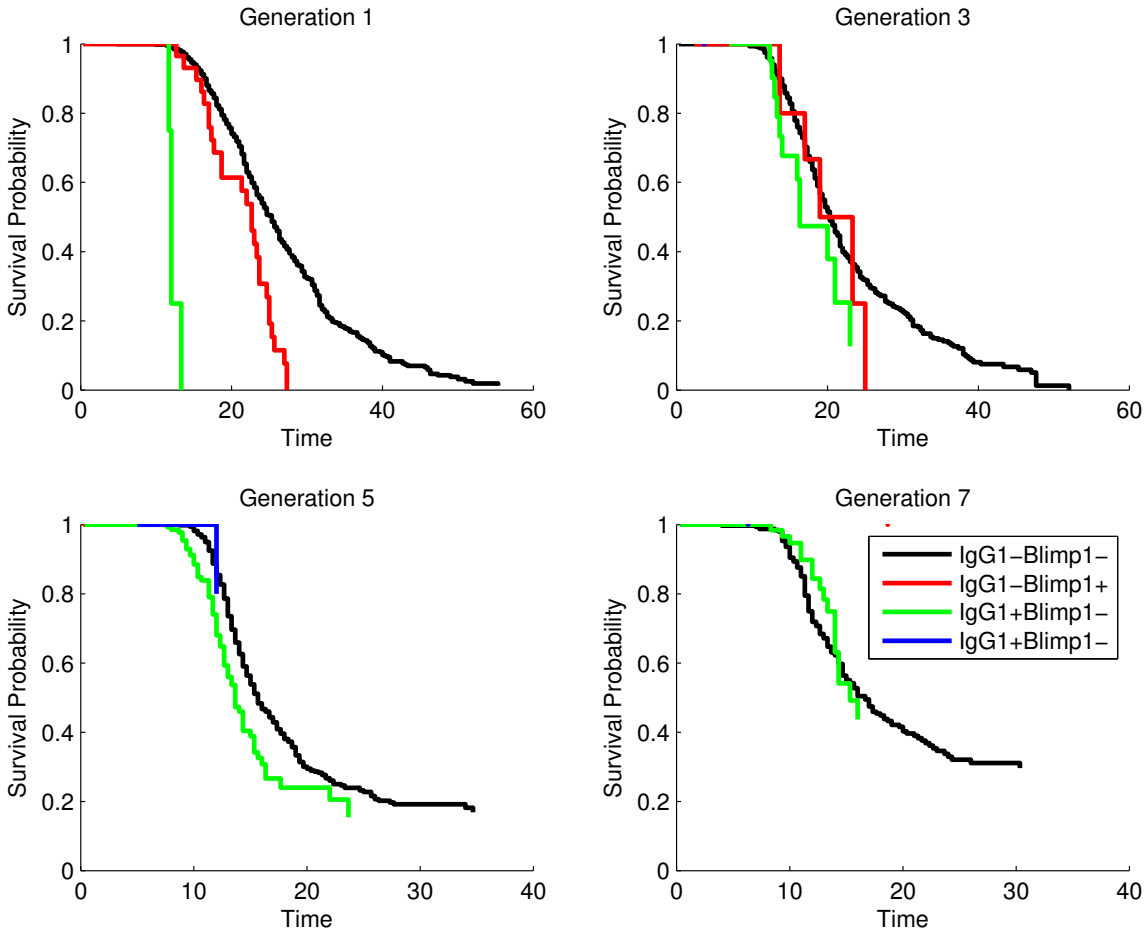


Figure 40: Kaplan-Meier survival function estimates for time to division for all generations and initial cell states for which data was available [10]. Each plot shows an all initial states for a given generation.

Fate	**Division**			
Gen	1	3	5	7
P Value	N/A	0.06965	0.00036	0.48689
HBM Adjusted Result	N/A	Fail to Reject	Reject	Fail to Reject

Table 9: Table of Log-Rank statistical tests. The null hypothesis is given by equation 16. We are testing if the survival distribution for time to division for cells with initial state IgG1-Blimp1- is the same as the time to division distribution for cells with initial state IgG1+Blimp1-. A rejection threshold of 0.05 was used and the Holm-Bonferroni method described in section 2.4.6 was used to account for multiple testing. For some generations there is insufficient data for hypothesis tests to be performed.

Fate	**Division**			
Gen	1	3	5	7
P Value	0.000027	0.551218	N/A	N/A
HBM Adjusted Result	Reject	Fail to Reject	N/A	N/A

Table 10: Table of Log-Rank statistical tests. The null hypothesis is given by equation 16. We are testing if the survival distribution for time to division for cells with initial state IgG1-Blimp1- is the same as the time to division distribution for cells with initial state IgG1-Blimp1+. A rejection threshold of 0.05 was used and the Holm-Bonferroni method described in section 2.4.6 was used to account for multiple testing. For some generations there is insufficient data for hypothesis tests to be performed.

We have also produced the same plots and tests as above, allowing us to study differences in initial start type and its impact on the time to death distribution. Figure 41 shows the Kaplan-Meier survival function estimates for time to die, with each individual plot showing all possible initial start types within one generation.

Tables 11, 12 and 13 show the results of Log-Rank tests again under the hypothesis that the time-to-die distribution of B cells whose initial start type is IgG1-Blimp1- is the same as the time-to-die distribution of cells whose initial start type is IgG1+Blimp1-, IgG1-Blimp1+ and IgG1+Blimp1+ respectively.

In Figure 41 we see that generally the time to die survival function estimates differ significantly from the IgG1-Blimp- survival functions. For some

start types there is not enough data to produce a survival function that can be used for testing, for example the IgG1-Blimp+ in generation 5 and IgG1-Blimp+ and IgG1+Blimp1+ in generation 7. From the Log Rank tests we see in all cases H_0 is rejected. The results of this test, and the survival function estimates, suggest that cells that have undergone IgM to IgG1 class switching and cells that have differentiated to plasmablast have a life span shorter than that of IgG1-Blimp1- B cells.

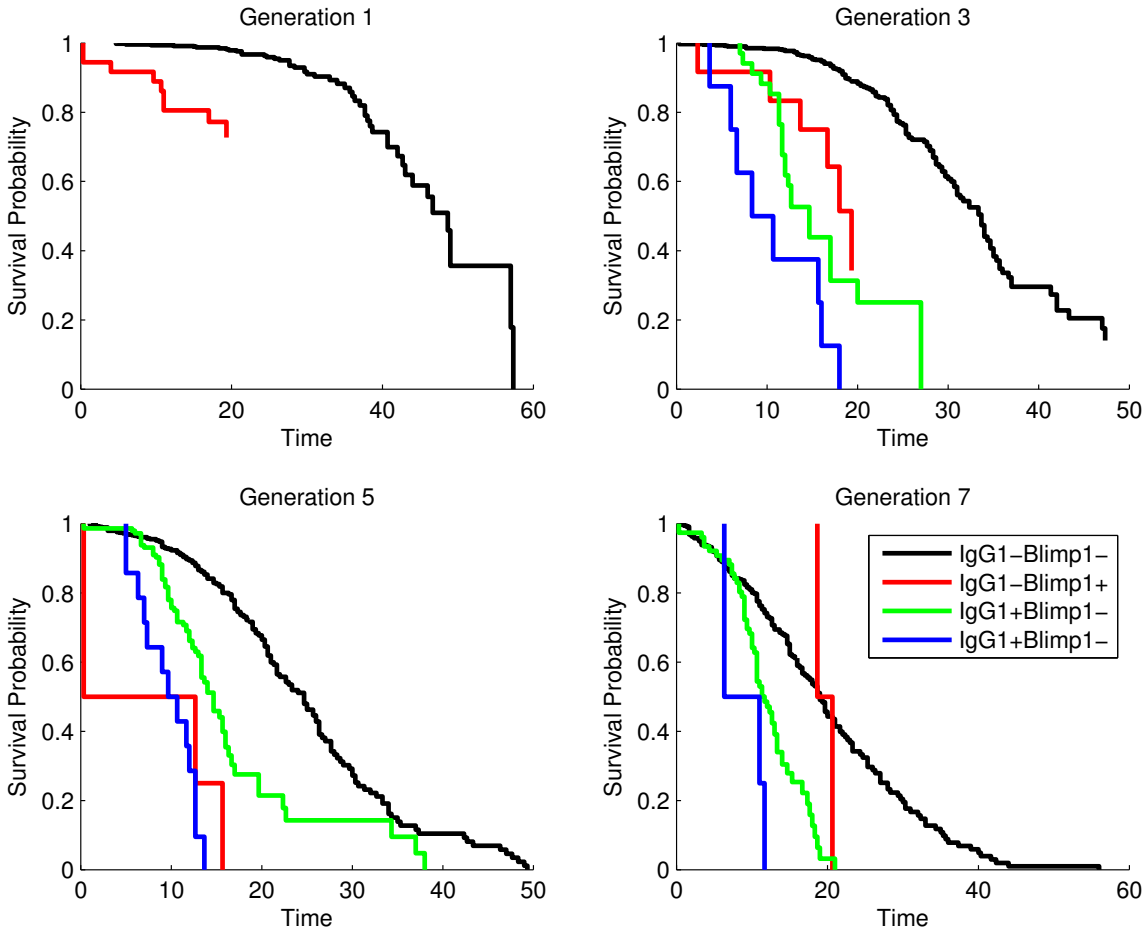


Figure 41: Kaplan-Meier survival function estimates for time to die for all generations and initial cell states for which data was available [10]. Each plot shows an all initial states for a given generation.

Fate	**Death**			
Gen	1	3	5	7
P Value (*1e-10)	N/A	0.00	0.000007	0.401628
HBM Adjusted Result	N/A	Reject	Reject	Reject

Table 11: Table of Log-Rank statistical tests. The null hypothesis is given by equation 16. We are testing if the survival distribution for time to die for cells with initial state IgG1-Blimp1- is the same as the time to die distribution for cells with initial state IgG1+Blimp1-. A rejection threshold of 0.05 was used and the Holm-Bonferroni method described in section 2.4.6 was used to account for multiple testing. For some generations there is insufficient data for hypothesis tests to be performed.

Fate	**Death**			
Gen	1	3	5	7
P Value	0.00000	0.00000	N/A	N/A
HBM Adjusted Result	Reject	Reject	N/A	N/A

Table 12: Table of Log-Rank statistical tests. The null hypothesis is given by equation 16. We are testing if the survival distribution for time to die for cells with initial state IgG1-Blimp1- is the same as the time to die distribution for cells with initial state IgG1-Blimp1+. A rejection threshold of 0.05 was used and the Holm-Bonferroni method described in section 2.4.6 was used to account for multiple testing. For some generations there is insufficient data for hypothesis tests to be performed.

Fate	**Death**			
Gen	1	3	5	7
P Value	N/A	0.00000	0.00000	N/A
HBM Adjusted Result	N/A	Reject	Reject	N/A

Table 13: Table of Log-Rank statistical tests. The null hypothesis is given by equation 16. We are testing if the survival distribution for time to die for cells with initial state IgG1-Blimp1- is the same as the time to die distribution for cells with initial state IgG1+Blimp1+. A rejection threshold of 0.05 was used and the Holm-Bonferroni method described in section 2.4.6 was used to account for multiple testing. For some generations there is insufficient data for hypothesis tests to be performed.

Lastly we have the alternative start type data for time-to-differentiate. Figure 42 shows the IgG1-Blimp1- survival function as well as the survival function for IgG1-Blimp1+ data. For generation 1 we only have 2 observations and so we cannot perform analysis for this generation. Looking at generations 3, 5 and 7 we see a much smaller probability of survival, with cells undergoing differentiation much earlier than the IgG1-Blimp1- case. Table 14 shows Log-Rank hypothesis tests to test the hypothesis that IgG1-Blimp+ cells have the same survival distribution as IgG1-Blimp1- cells. Here we have excluded generation 1 due to lack of data. As we would expect from looking at the survival functions, in all cases we reject the hypothesis, confirming that, for this data set, the two start types have different survival distributions.

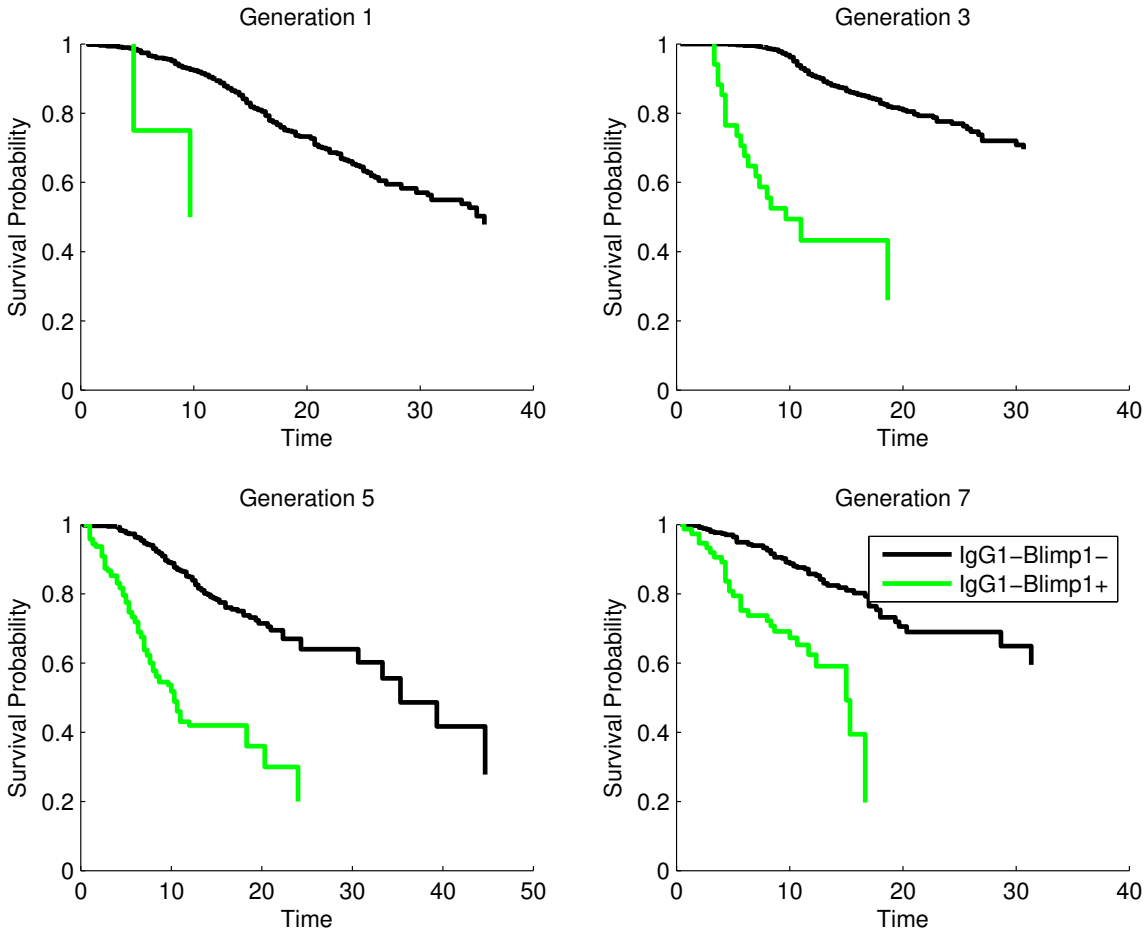


Figure 42: Kaplan-Meier survival function estimates for time to IgM to IgG1 class switch for all generations and initial cell states for which data was available [10]. Each plot shows the survival distributions for cells whose initial start type is either IgG1-Blimp1- or IgG1-Blimp1+ for a specific generation.

Fate	**Differentiation**			
Gen	1	3	5	7
P Value (*1e-3)	N/A	0.41659	0.14101	0.95763
HBM Adjusted Result	N/A	Reject	Reject	Reject

Table 14: Table of Log-Rank statistical tests. The null hypothesis is given by equation 16. We are testing if the survival distribution for time to differentiate for cells with initial state IgG1-Blimp1- is the same as the time to differentiate distribution for cells with initial state IgG1-Blimp1+. A rejection threshold of 0.05 was used and the Holm-Bonferroni method described in section 2.4.6 was used to account for multiple testing. For some generations there is insufficient data for hypothesis tests to be performed.

Having looked at the univariate time to event distributions, we can use the Dabrowska estimate described in Section 2.3 and modified for symmetric data in Section 3.2 to estimate the bivariate survival distributions of pairs of sibling cells. Figure 43 shows the Dabrowska estimates for generation 3, 5 and 7 time to division, for start types IgG1-Blimp1- and IgG1+Blimp1- cells. Figure 44 shows the same generations and start types for time to death, and Figure 45 shows the same generations and start types for time to differentiation.

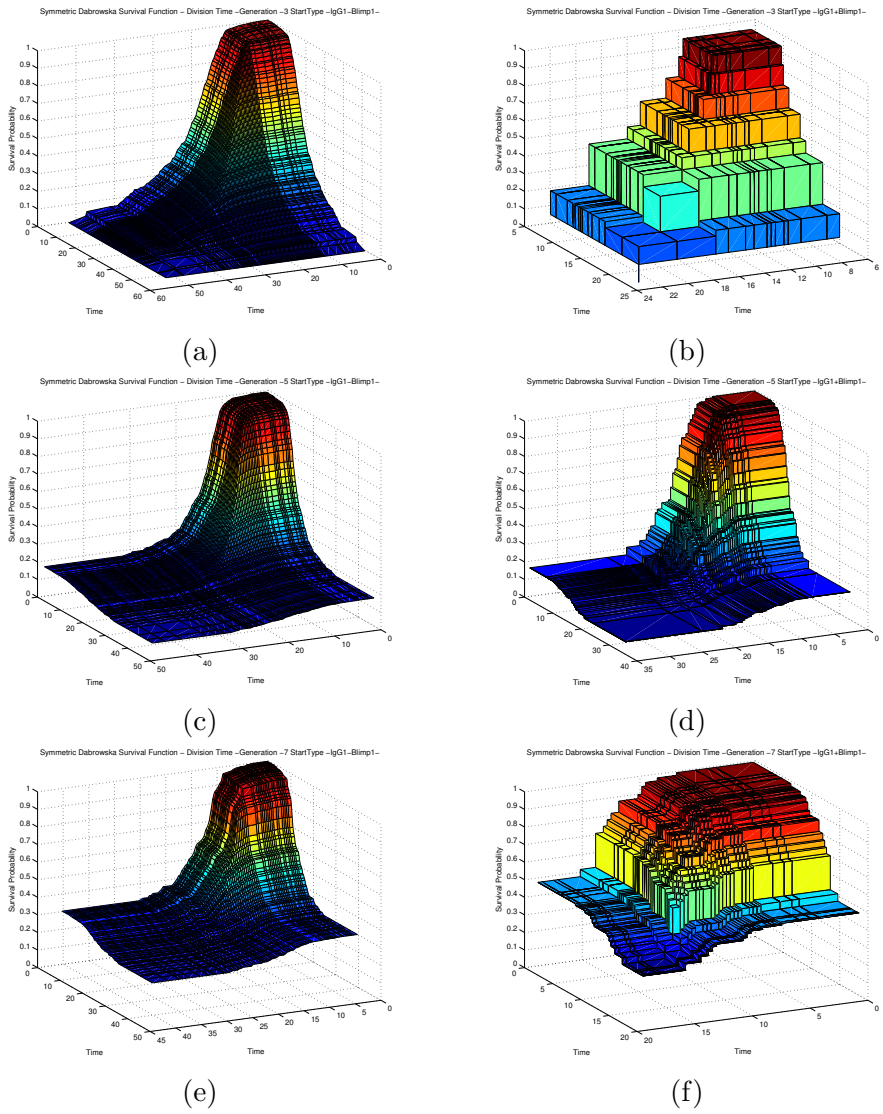


Figure 43: Bivariate survival function estimates for the time to division of pairs of sibling B cells using the symmetric Dabrowska estimate. (a) Generation 3, IgG1-Blimp1-, (b) Generation 3, IgG1+Blimp1-, (c) Generation 5 IgG1-Blimp1-, (d) Generation 5 IgG1+Blimp1-, (e) Generation 7 IgG1-Blimp1-, (f) Generation 7 IgG1+Blimp1-.

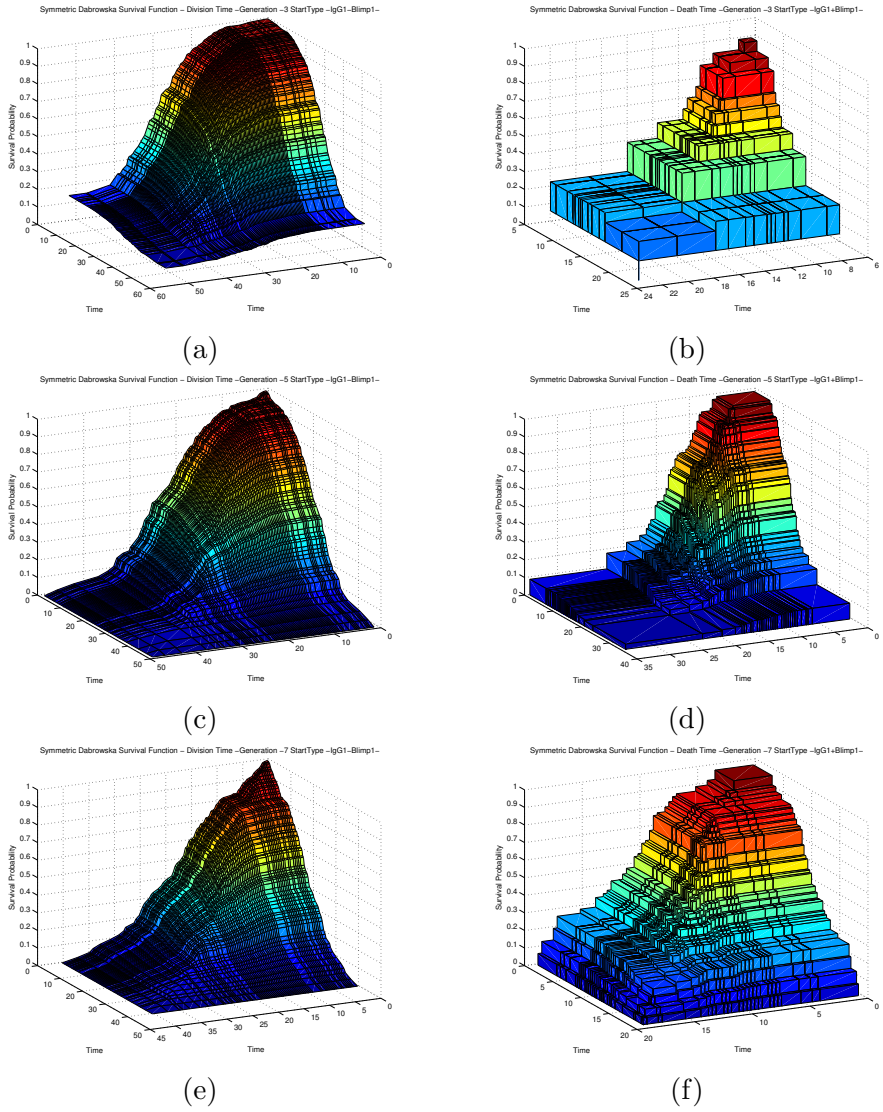
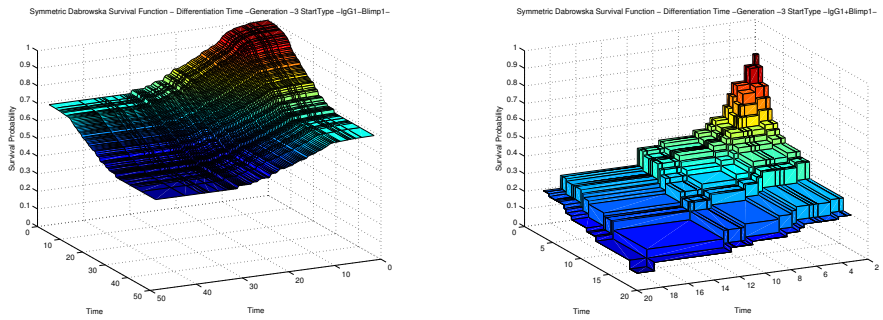
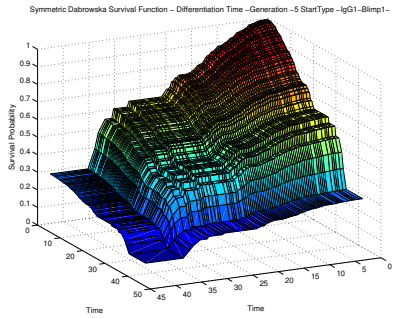


Figure 44: Bivariate survival function estimates for the time to death of pairs of sibling B cells using the symmetric Dabrowska estimate. (a) Generation 3, IgG1-Blimp1-, (b) Generation 3, IgG1+Blimp1-, (c) Generation 5 IgG1-Blimp1-, (d) Generation 5 IgG1+Blimp1-, (e) Generation 7 IgG1-Blimp1-, (f) Generation 7 IgG1+Blimp1-.

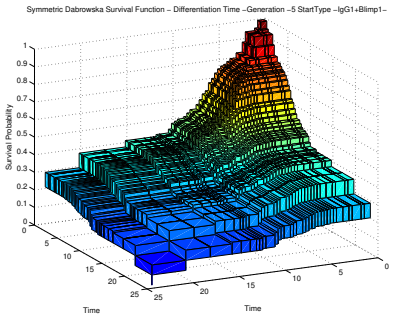


(a)

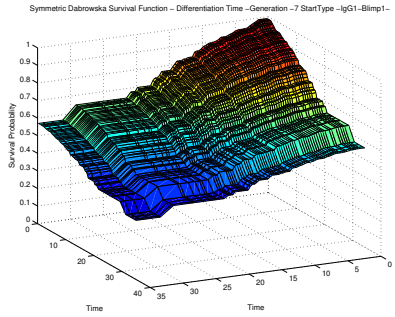
(b)



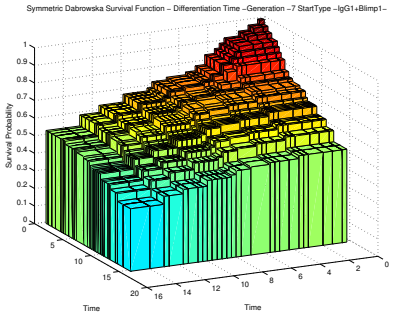
(c)



(d)



(e)



(f)

Figure 45: Bivariate survival function estimates for the time to differentiate of pairs of sibling B cells using the symmetric Dabrowska estimate. (a) Generation 3, IgG1-Blimp1-, (b) Generation 3, IgG1+Blimp1-, (c) Generation 5 IgG1-Blimp1-, (d) Generation 5 IgG1+Blimp1-, (e) Generation 7 IgG1-Blimp1-, (f) Generation 7 IgG1+Blimp1-.

We see here that the results are largely the same as the univariate case with different start types showing a different distribution to the IgG1-Blimp1-

Fate	**PValue/ Generation**			
	1	**3**	**5**	**7**
Division	-	0.0010	0.0004	0.0158
Death	-	0.3806	0.1510	0.7802
Differentiation	-	0.0000	0.0000	0.0072

Table 15: Hypothesis test results comparing Dabrowska survival functions.

cells. This is backed up in the cases of division and death by the hypothesis tests we have performed below.

In order to compare the Dabrowska estimates for cells with start type IgG1-Blimp1- with cells that start IgG1+Blimp1- we use the hypothesis test presented in Section 2.4.5. Here we present the p values for the hypothesis that the distribution of time to event is the same in cells regardless of whether they have undergone IgM to IgG1 class switching.

For division and differentiation, all tests are rejected at the $\alpha = 0.05$ level, providing evidence that different start types have different distributions. In the case of death we see that none of the tests are rejected at the $\alpha = 0.05$. Suggesting that cells of different start types may still have the same death times regardless of whether they have become IgG1+.

6 Conclusion

In this section we will give a few concluding remarks.

Motivated by the data set of [10], where sibling B cells have correlated fate times, under a competition based model, a set of nonparametric statistical tools were developed for the purpose of further investigation and comparison with the parametric techniques that were used in the paper.

The primary tools that were employed included the Kaplan-Meier and Dabrowska survival function estimators, and a set of hypothesis tests to allow comparison of distributions. For some of these, existing packages could be used, while others new implementations were created. Family wise error was considering when performing multiple tests through the use of the Holm-Bonferroni procedure.

As an extension to this work, we addressed the case when distributions are defective, either due to lack of data when the final event time was censored, or due to the underlying features of the system from which the data had been measured. Further development includes the implementation of a symmetric Dabrowska estimate based on the underlying assumptions of [10] allowing the production of more accurate survival distribution estimates when the underlying distributions are symmetric.

When applied to the data from [10] this tool set allowed the comparison of survival distributions estimated using Kaplan-Meier, to the Log-Normal distributions from [10]. We compared the distributions of time from birth to differentiation to plasmablast, and the times to IgM to IgG1 class switch. Here we concluded that while these distributions may not be the best choice, we must account for the small amount of data available due to the method of comparison. The choice of distribution depends on the granularity of the model, meaning it may be an appropriate choice in some cases, but not in others.

The Kaplan-Meier estimate and the Log-Rank statistical test were used to determine how the distribution of time to event varies with the generation of cells. Here we concluded with results showing that, while time to divison and death seem to vary with respect to generation, it seems that time to differentiate to plasmablast does not vary with respect to generation. Here we saw that while the statistical tests confirm that the differentiation distributions are different, it seems the quantity that varied was the number of B cells for which differentiation was "on" in that cell, and not the time at which cells undergo the event.

We looked at extra unpublished data which was collected during the experiment performed for [10] and was kindly provided by Phil Hodgkin's Lab at WEHI. We performed hypothesis tests to compare the distributions of time to event in B cells that had already undergone either class switching to IgG1, differentiation to plasmablast, or both. In most cases we see that the underlying distributions are different to the cells that had not undergo class switching or differentiation. We must be careful here however, as lack of data could be a large contributing factor in some cases.

In the future it would be interesting to apply this statistical methodology to other systems, for example T cells. Further tools from survival analysis could be used and compared, for example the Prentice-Cai [39] or Yin-Ling [31] estimators of the bivariate survival function.

Finally it would be interesting to look at non-parametric statistical techniques more suitable when competing risks are present such as the cumulative incidence function[33] to avoid the possibility of overestimation. For example the Kaplan-Meier analysis of time to IgM to IgG1 class switch given in Figure 42 are likely an overestimate due to the informative censoring events like cell death and division.

Code for this thesis is available through the Matlab Fileworks library.

References

[1] Odd Aalen. Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726, 1978.

[2] Dick M Bakker. Two nonparametric estimators of the survival function of bivariate right censored observations. *Department of Operations Research, Statistics, and System Theory [BS]*, (R 9035):1–44, 1990.

[3] Dimitri P Bertsekas and John N Tsitsiklis. *Introduction to probability*, volume 1. Athena Scientific Belmont, MA, 2002.

[4] Veit R Buchholz, Ton NM Schumacher, and Dirk H Busch. T cell fate at the single-cell level. *Annual Review of Immunology*, 34:65–92, 2016.

[5] Martin Chalfie and Steven R Kain. *Green fluorescent protein: properties, applications and protocols*, volume 47. John Wiley & Sons, 2005.

[6] James A Cornwell, Robin M Hallett, Stefanie Auf der Mauer, Ali Motazedian, Tim Schroeder, Jonathan S Draper, Richard P Harvey, and Robert E Nordon. Quantifying intrinsic and extrinsic control of single-cell fates in cancer and stem/progenitor cell pedigrees with competing risks analysis. *Scientific Reports*, 6(27100), 2016.

[7] Dorota M Dabrowska. Kaplan-Meier estimate on the plane. *The Annals of Statistics*, 16(4):1475–1489, 1988.

[8] Mark R Dowling, Andrey Kan, Susanne Heinzel, Jie HS Zhou, Julia M Marchingo, Cameron J Wellard, John F Markham, and Philip D Hodgkin. Stretched cell cycle model for proliferating lymphocytes. *Proceedings of the National Academy of Sciences*, 111(17):6377–6382, 2014.

[9] Ken R Duffy and Philip D Hodgkin. Intracellular competition for fates in the immune system. *Trends in Cell Biology*, 22(9):457–464, 2012.

[10] Ken R Duffy, Cameron J Wellard, John F Markham, Jie HS Zhou, Ross Holmberg, Edwin D Hawkins, Jhagvaral Hasbold, Mark R Dowling, and Philip D Hodgkin. Activation-induced B cell fates are selected by intracellular stochastic competition. *Science*, 335(6066):338–341, 2012.

[11] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[12] Thomas R Fleming, Judith R O'Fallon, Peter C O'Brien, and David P Harrington. Modified Kolmogorov-Smirnov test procedures with application to arbitrarily right-censored data. *Biometrics*, 36(4):607–625, 1980.

[13] Richard D Gill. Multivariate survival analysis. *Theory of Probability & Its Applications*, 37(2):284–301, 1993.

[14] Richard D Gill and Soren Johansen. A survey of product-integration with a view toward application in survival analysis. *The Annals of Statistics*, 18(4):1501–1555, 1990.

[15] Ted A Gooley, Wendy Leisenring, John Crowley, and Barry E Storer. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in Medicine*, 18(6):695–706, 1999.

[16] Major Greenwood. A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects.*, (33):1–26, 1926.

[17] Edwin D Hawkins, John F Markham, Liam P McGuinness, and Philip D Hodgkin. A single-cell pedigree analysis of alternative stochastic lymphocyte fates. *Proceedings of the National Academy of Sciences*, 106(32):13457–13462, 2009.

[18] Edwin D Hawkins, Marian L Turner, Mark R Dowling, C Van Gend, and Philip D Hodgkin. A model of immune regulation as a consequence of randomized lymphocyte division and death times. *Proceedings of the National Academy of Sciences*, 104(12):5032–5037, 2007.

[19] Philip D Hodgkin, Jae-Ho Lee, and Alan B Lyons. B cell differentiation and isotype switching is related to division cycle number. *The Journal of Experimental Medicine*, 184(1):277–281, 1996.

[20] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[21] Charles Janeway, Kenneth P Murphy, Paul Travers, Mark Walport, and et al. *Janeway's immunobiology.* Garland Science, 2008.

[22] Marc K Jenkins and James J Moon. The role of naive T cell precursor frequency and recruitment in dictating immune response magnitude. *The Journal of Immunology*, 188(9):4135–4140, 2012.

[23] Susan M Kaech and E John Wherry. Heterogeneity and cell-fate decisions in effector and memory CD8+ T cell differentiation during viral infection. *Immunity*, 27(3):393–405, 2007.

[24] John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

[25] Axel Kallies, Jhagvaral Hasbold, David M Tarlinton, Wendy Dietrich, Lynn M Corcoran, Philip D Hodgkin, and Stephen L Nutt. Plasma cell ontogeny defined by quantitative changes in blimp-1 expression. *Journal of Experimental Medicine*, 200(8):967–977, 2004.

[26] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

[27] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[28] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

[29] David G Kleinbaum and Mitchel Klein. *Survival analysis: a self-learning text*. Springer Science & Business Media, 2006.

[30] Anna Kuchina, Lorena Espinar, Tolga Çağatay, Alejandro O Balbin, Fang Zhang, Alma Alvarado, Jordi Garcia-Ojalvo, and Gürol M Süel. Temporal competition between differentiation programs determines cell fate choice. *Molecular Systems Biology*, 7(1):557, 2011.

[31] DY Lin and Zhiliang Ying. A simple nonparametric estimator of the bivariate survival function under univariate censoring. *Biometrika*, 80(3):573–581, 1993.

[32] Nathan Mantel. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports. Part 1*, 50(3):163–170, 1966.

[33] BS Manzoor, S Adlmadhyam, and SM Walton. An introduction to competing risks. *Value Outcomes Spotlight*, 3(2):21–22, 2017.

[34] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[35] Reza Modarres. Estimation of a bivariate symmetric distribution function. *Statistics & Probability Letters*, 63(1):25–34, 2003.

[36] James J Moon, H Hamlet Chu, Marion Pepper, Stephen J McSorley, Stephen C Jameson, Ross M Kedl, and Marc K Jenkins. Naive CD4+ T cell frequency varies for different epitopes and predicts repertoire diversity and response magnitude. *Immunity*, 27(2):203–213, 2007.

[37] Wayne Nelson. Theory and applications of hazard plotting for censored failure data. *Technometrics*, 14(4):945–966, 1972.

[38] Margaret S Pepe and Thomas R Fleming. Weighted Kaplan-Meier statistics: a class of distance tests for censored survival data. *Biometrics*, 45(2):497–507, 1989.

[39] Ross L Prentice and Jianwen Cai. Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79(3):497–507, 1989.

[40] Ronald C Pruitt. On negative mass assigned by the bivariate Kaplan-Meier estimator. *The Annals of Statistics*, 19(1):443–453, 1991.

[41] Ronald C Pruitt. Strong consistency of self-consistent estimators: general theory and an application to bivariate survival analysis. Technical Report 543, University of Minnesota, 1991.

[42] Rahul Satija and Alex K Shalek. Heterogeneity in immune responses: from populations to single cells. *Trends in Immunology*, 35(5):219–229, 2014.

[43] Yuri Suhov and Mark Kelbert. *Probability and Statistics by Example*, volume 1. Cambridge University Press, 2005.

[44] Marian L Turner, Edwin D Hawkins, and Philip D Hodgkin. Quantitative regulation of B cell division destiny by signal strength. *The Journal of Immunology*, 181(1):374–382, 2008.

[45] Mark J Van Der Laan. Efficient estimation in the bivariate censoring model and repairing NPMLE. *The Annals of Statistics*, 24(2):596–627, 1996.

[46] Mark J Van Der Laan. Nonparametric estimators of the bivariate survival function under random censoring. *Statistica Neerlandica*, 51(2):178–200, 1997.

[47] Jiantian Wang and Pablo Zafra. Estimating bivariate survival function by Volterra estimator using dynamic programming techniques. *Journal of Data Science*, 7(3):365–380, 2009.

[48] Lee-Jen Wei and John M Lachin. Two-sample asymptotically distribution-free tests for incomplete multivariate observations. *Journal of the American Statistical Association*, 79(387):653–661, 1984.

[49] R Wilcox. Kolmogorov-Smirnov test. *Encyclopedia of Biostatistics*, 3:2174–6, 1998.