A BAYESIAN DECISION THEORETIC APPROACH TO FIXED SAMPLE SIZE DETERMINATION AND BLINDED SAMPLE SIZE RE-ESTIMATION FOR HYPOTHESIS TESTING

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment
of the Requirements for the Degree
DOCTOR OF PHILOSOPHY

---

by
Dwaine Stephen Banton
May 2016

Examining Committee Members:

Marc Sobel, Advisor, Statistics
Zhigen Zhao, Advisory Chair, Statistics
Xu Han, Statistics
Alexandra Carides, Statistics
Eleni Anni, External Member, Temple University

ABSTRACT

This thesis considers two related problems that has application in the field of experimental design for clinical trials:

- fixed sample size determination for parallel arm, double-blind survival data analysis to test the hypothesis of no difference in survival functions, and
- blinded sample size re-estimation for the same.

For the first problem of fixed sample size determination, a method is developed generally for testing of hypothesis, then applied particularly to survival analysis; for the second problem of blinded sample size re-estimation, a method is developed specifically for survival analysis. In both problems, the exponential survival model is assumed. The approach we propose for sample size determination is Bayesian decision theoretical, using *explicitly* a loss function and a prior distribution. The loss function used is the intrinsic discrepancy loss function introduced by Bernardo and Rueda (2002), and further expounded upon in Bernardo (2011). We use a conjugate prior, and

investigate the sensitivity of the calculated sample sizes to specification of the hyper-parameters. For the second problem of blinded sample size re-estimation, we use prior predictive distributions to facilitate calculation of the interim test statistic in a blinded manner while controlling the Type I error. The determination of the test statistic in a blinded manner continues to be nettling problem for researchers. The first problem is typical of traditional experimental designs, while the second problem extends into the realm of adaptive designs. To the best of our knowledge, the approaches we suggest for both problems have never been done hitherto, and extend the current research on both topics. The advantages of our approach, as far as we see it, are unity and coherence of statistical procedures, systematic and methodical incorporation of prior knowledge, and ease of calculation and interpretation.

DEDICATION

Like Janus, with this work I look simultaneously at the past and to the future. I dedicate this to my mother and father's incredible drive to transcend their conditions and become successful members of society. I also dedicate this to my daughter, Saraya, and my son, Brock; may they eclipse and outstrip me in every positive way.

ACKNOWLEDGEMENTS

This thesis is made possible by the graces of Temple University, and in particular the Statistics department. I would like to thank everyone in the department – faculty, staff members, and fellow students – for their support and encouragement. My thesis advisor, Marc Sobel, has been a constant advocate and champion in my corner, encouraging me and guiding me throughout the entire process. This would not be possible without his support; Marc, thank you. I would also like to thank the other members of my committee: Xu Han, Zhigen Zhao, Alexandra Carides, and Helen "Eleni" Anni. Finally, I would like to thank Louise Jones, without whom I probably would not have made it through the program.

Contents

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

Experimental design is a pivotal part of statistical practice, and ipso facto scientific advancement and discovery. Sample size calculations are an important part of experimental design, and often they are the only consideration. In this thesis, we consider fixed sample size determination for hypothesis testing, and blinded sample size re-estimation, from a Bayesian paradigm. The US Food and Drug Administration (2004) recognizes the need for new methods that look at sample size design from a more adaptive approach:

> There are many important additional opportunities in the area of clinical trial design and analysis ... Enrichment designs have the potential for providing much earlier assurance of drug activity. Bayesian approaches to analysis need to be further explored.

The main purpose of this thesis is to take steps towards filling that void of Bayesian approaches to traditional design problems. The methods developed

herein are applied illustratively to survival analysis for testing the hypothesis of no difference of survival rate between two independent groups. We assume that the event distribution function follows an exponential distribution, that is, we assume constant hazard functions for both groups. Apart from, or probably because of, being the simplest parametric model for survival analysis, the exponential model is the most widely used model for sample size calculation.

The question of how many observations are needed to decide between two competing hypotheses is fundamental in science. The solution can be approached from two perspectives: traditional sample size determination, where a fixed sample size is determined before any experimentation and data collection begin and is never changed, and an adaptive sample size determination that allows the sample size to change as data becomes available. The Bayesian framework seems most suited for adaptive sample size determination, in particular the type of adaptive design called sequential design. Fixed sample size determination, usually the purview of the frequentist framework, seems to be unnecessarily constraining in the Bayesian paradigm. This is due to the fact that the frequentist methods are usually based on optimization of Type II error $(1 - \text{power})$ given the constraint of a Type I error (size), where these probabilities are interpreted as relative frequencies. This framework usually requires fixed sample sizes determined a priori in order to control the Type I error. The frequentist paradigm of controlling Type I error, while important and relevant in some cases, may not be necessary, or natural, in

all cases. In contradistinction, the Bayesian framework considers the particular experiment only, and the probabilities involved are a mathematical way of characterising the degree of uncertainty about the parameters, which are usually unknowable. There is no concern about Type I error. We believe that both approaches are useful, and in some cases can complement each other.

It is with this in mind, and also due to the popularity and prevalence of sample size determination using size and power methods, that we investigate how we can adapt the Bayesian framework to provide alternative solutions to the fixed sample size problem. We show that the Bayesian solution proposed will give smaller sample sizes than the frequentist solution, depending on the question being asked. A direct comparison between the methods is specious, since, as stated above, their aims and criteria are different. We argue that the criterion used in the Bayesian method may be the one that an investigator actually wants, and if that is the case then the Bayesian method will result in smaller sample sizes than the frequentist method. We also show that, for a given sample size, if one uses the Bayesian testing criterion, derived from the Bayesian Reference Criterion in Bernardo and Rueda (2002), then the resulting test can be more powerful than the frequentist test. This is due to the benefit of using prior information in conducting the test. This is in line with the FDA objective of making clinical trials more economical and efficient. The advantages of the Bayesian method become even more preponderous when we consider in particular precise hypothesis testing, where

the frequentist method does not have a unified approach to the problem of nuisance parameters, and in some cases has no solution at all. Whereas the Bayesian approach can deal with any sampling distribution, since nuisance parameters are not an issue.

In the case of survival study, once an initial fixed sample size is determined by any method, recruitment begins and data can be collected. Before the close of recruitment, it may be of interest to use the data collected thus far, in a blinded manner, in order to update previous uncertainty in quantities used to calculate the initial fixed sample size. By blinded we mean that group identification is unknown to us at this point; all we have are a series of measurements. This is known as blinded sample size re-estimation (SSR), and ICH guideline documents support this type of calculation, as long as it is done in a manner that maintains the blind, and that the effect, if any, on Type I error is noted. We will use a Bayesian approach for blinded sample size re-estimation, using the prior predictive distribution of the measurements to determine group association. For non-Bayesians, the usual method for SSR is recalculating the plug-in estimate of a variance parameter, considered as a nuisance parameter, used in the sample size determination formula. However, for survival data, the variance parameter is usually a function of the variable of interest, and re-estimating this parameter would imply re-estimation of the parameter of interest. Frequentist methods do not accommodate re-calculation of the parameter of interest easily, since it is set by the alternative hypothesized treatment difference. Furthermore, the calculation of these

variance estimates, due to lack of group information, usually proceeds in a rather ad hoc manner. Hence it is not surprising that there is a dearth of methods available to the researcher who wishes to re-estimate sample size for survival data.

Our contributions to the discussion of sample size determination and blinded re-estimation for survival analysis are:

- A simple, comprehensive sample size formula for hypothesis testing that is based on decision theoretic foundations, and requires at most only Monte Carlo integration for even the most complicated sampling distribution

- A new approach to blinded sample size re-estimation developed within the Bayesian framework, acknowledging and accounting for all uncertainty in the re-estimation process.

Our investigation will proceed as follows. Chapter 2 will focus on the problem of fixed sample size determination for hypothesis testing in general, and then the results will be applied to the particular case of exponential survival analysis. In chapter 3, we investigate blinded sample size re-estimation for exponential survival data analysis. Chapter 4 will entail our conclusion of the current work, plus future research plans.

Chapter 2

FIXED SAMPLE SIZE DETERMINATION FOR HYPOTHESIS TESTING

Abstract

This chapter considers the problem of fixed sample size determination for hypothesis testing. There have been multitudinous approaches proposed in the literature for this problem, both from a Bayesian and non-Bayesian perspective. The approach we propose is a Bayesian decision theoretic framework, using a continuous intrinsic loss function. The loss function used is the intrinsic discrepancy loss function introduced by Bernardo and Rueda (2002). The advantage of using a loss function is mathematical and statistical coherency in the decision rule for hypothesis testing. The prior allows for the proba-

bilistic incorporation of previous knowledge into the process, and such knowledge is especially pertinent in clinical studies. The advantages of our approach over other methods are that both the loss function and the prior are generated from the sampling distribution assumption, the cutoff used for the decision function has an intuitive interpretation, and the calculation necessary in the most general cases is Monte Carlo integration. The method developed is compared to the frequentist method. While the method may be generally applied to sample size determination for any hypothesis testing experiment, we focus mainly on the exponential survival data model under the assumptions of no censoring, and administrative censoring only. An exact closed form solution is developed under the first condition, while a numerical solution is used for the optimal sample size under the second assumption. We show that our method will generally require smaller sample sizes than the frequentist method, even though our decision rule is a more stringent test of the null hypothesis.

2.1 Introduction

Sample size determination methods can be viewed on a spectrum, ranging from a one-time fixed calculation, to a per-observation sequential design. Clinical trials experimental design has traditionally favored the former approach, and to a large extent this attitude is still dominant and prevalent. The problem of sample size determination has been studied traditionally from

the frequentist perspective. There are two general approaches: either to determine the sample size that optimizes the power of a given hypothesis test for a fixed size, or compute the sample size for a given confidence interval, the interval usually chosen by inversion of an optimal test. A size α test is defined as:

$$\text{Type I error} \equiv Pr(\text{reject } H_0 | H_0 \text{ is true}) = \alpha, 0 < \alpha < 1.$$

A test is defined to have a power of $1 - \beta$ if

$$Pr(\text{ reject } H_0 | H_1 \text{ is true}) = 1 - \beta, 0 < \beta < 1.$$

The main idea, see for example Lachin (2009), is to assume that the test statistic is normally distributed, at least asymptotically. This will be the case for survival data analysis. As such, sample sized determination for survival data, under certain assumptions to be made clear below, is a special case of sample size determination for normally distributed data. The hypothesis testing situation is then generally set up as follows:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu = \mu_1$$

$$T|H_0 \sim \mathrm{N}(\mu_0, {\sigma_0}^2)$$

$$T|H_1 \sim \mathrm{N}(\mu_1, {\sigma_1}^2)$$

where $T(x)$ is the test statistic. Assume ${\sigma_0}^2 = \frac{{\psi_0}^2}{n}$, and ${\sigma_1}^2 = \frac{{\psi_1}^2}{n}$. Define $\Delta = \mu_1 - \mu_0$, and Z_k as the k-quantile of the standard normal distribution. Then the sample size that is required to control Type I error at level α while give a test with power $1 - \beta$ is

$$n = \left(\frac{Z_{1-\frac{\alpha}{2}}\psi_0 + Z_{1-\beta}\psi_1}{\Delta} \right)^2. \tag{2.1}$$

This equation is the root for almost all frequentist sample size calculation.

The sample size for a given interval length, when one considers confidence interval estimation of the parameter of interest rather than hypothesis testing, is basically the same equation with $Z_{1-\beta} = 0$. This implies a power of 0.5. To wit, assume that the estimator for the parameter is normally distributed as $T \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. We seek an interval such that:

$$Pr\left(T - Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < T + Z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \mid \mu \right) = 1 - \alpha.$$

Define the length of the interval as l. Then the required sample size is:

$$n = 4\left(\frac{Z_{1-\frac{\alpha}{2}}\sigma}{l} \right)^2 = \left(\frac{Z_{1-\frac{\alpha}{2}}\sigma}{e} \right)^2$$

where e is known as the margin of error, and is equal to $\frac{l}{2}$.

A few concerns have precipitated the pursuit of non-frequentist methods for fixed sample size determination. As pointed out by Wang and Gelfand (2002), it is not clear that one wishes to use power as an optimality criterion in all situations, since we may be concerned about judging the test based on its

performance on a single experiment, and not based on hypothetical infinite repetition of the same experiment. Another concern is that, in equation (2.1), ψ_0, ψ_1, and Δ are treated as fixed parameters that are known. Of course, in practice these values are not known, and estimates are used, plug-in values, for the purpose of sample size determination. The problem is that there is no accounting for the uncertainty in these estimates, and as such the resulting calculation can be rather misleading. After all, statistical inference deals with taking into account and quantifying uncertainty. Lastly, there is no room for utilizing prior knowledge about the parameters, which may be quite significant in practice. These concerns have spurred an interest in approaches that utilize prior distributions.

Techniques that can incorporate prior information and account for uncertainty in the parameters will in general be Bayesian. Bayesian methods have thus entered the discourse, but mainly by focusing on the confidence interval approach for sample size determination. However, there have been some Bayesian methods that determine sample size for hypothesis testing. These methods try to keep within the confines of the frequentist approach, but use a prior distribution to overcome some of the perceived short-comings of the frequentist approach. They have been termed "semi-Bayesian" by Pezeshk (2003). Methods based on a Bayesian decision theoretic approach, which utilizes both a prior and a loss function, have been proposed less frequently by various authors for sample size determination. These methods have usually focused on sample size determination for the purpose of interval estimation.

There have been even fewer attempts to utilize the Bayesian decision theoretic approach for sample size determination for hypothesis testing.

Our contributions to sample size determination for hypothesis testing are as follows:

- a simple and elegant sample size procedure, based on decision theoretic foundations, that depends mainly on the sampling distribution assumption, and requires very little computational effort for even the most complicated likelihoods;

- a procedure that outperforms the frequentist approach on some levels for survival data analysis, while being just as, or even more, facile and interpretable; and

- application of a continuous, parametric invariant loss function for sample size determination that allows the use of the same continuous prior distribution for determining the sample size for testing both precise and composite null hypotheses.

The remainder of this chapter will be organized as follows. In section 2, we will do a literature review on fixed sample size determination, paying particular attention to problems apropos to survival analysis. In section 3, we introduce our method and present our main result. In section 4 we apply our result to two related survival analysis sample size problems, and compare our results to those of the frequentist procedure.

2.2 Literature Review

Sample size determination has been investigated from a frequentist perspective, a Bayesian perspective, and perspectives that lie somewhere in between. By frequentist perspective, we do not mean anything pejorative; we mean a certain branch of statistical practice that is concerned with the sampling distributions of test statistics, the concept of a fixed non-random parameter, and the interpretation of probability as long run frequency. In particular, the frequentist method determines sample size by controlling Type I error while maximizing power. Both these probabilities, Type I error and power, are given a frequency interpretation based on an infinite repetition of the experiment.

In contradistinction, the Bayesian perspective, as argued by Berger (1985), Bernardo and Smith (1994), and Robert (2007), for example, is a decision theoretical approach. The unknown parameter is initially quantified via a prior probability distribution; this distribution is ideally subjectively chosen to give a measure of the initial uncertainty about the parameter. All subsequent actions, like estimation, hypothesis testing, or even stating the posterior distribution, can be seen as a decision process that is being made with various consequences depending on the true but unknown parameter value. Hence to make a "best" decision, some quantification of the various utility/loss of the various decisions must be taken into account. The decision that maximizes/minimizes the expected utility/loss is considered the best

decision. This is the approach we advocate in general, and will be taking in this paper.

A compromising approach, which can be termed semi-Bayesian, utilizes a prior distribution explicitly while a loss function is only implicitly used. Sometimes it operates mostly in the frequentist paradigm, using the control of Type I and Type II errors to make decisions, but using the prior distribution to average the unknown parameter of the Type I and Type II error probabilities. Other times it operates in a more Bayesian paradigm, utilizing credible regions and posterior distributions to make decisions, and is not concerned with Type I and Type II errors. The literature review will therefore be organized along these three broad categories.

2.2.1 Sample Size Determination – Frequentist Perspective

Traditional compendiums, such as Desu and Rhagavaro (1990), treat the problem of sample size determination in a very general context. Julious (2004) focuses more on sample size determination as it relates to clinical trials for crossover and parallel groups, assuming the data is normally distributed. Lachin (1981) reviews sample size determination for clinical trials in general, and survival analysis in particular.

For survival analysis, the data is in the form of the pair (t_i, δ_i), where for patient i, t_i represents the observation at that time, and $\delta_i = 1$ if an

event occurred at that time, $\delta_i = 0$ if the patient was lost to follow-up. The latter is generally referred to as right censoring. If censoring is assumed to occur at random, as is usually the case, then the censoring stochastic process is statistically independent of the event stochastic process. This implies that the event process can be applied to all observations, whether there is censoring or not. The survival function is the complement of the event time distribution function: $S(t) = Pr(T > t)$. The hazard function describes the instantaneous rate of the event, given that the subject has survived up to that point: $\lambda(t) = -\frac{d}{dt}\left(ln\left[S(t)\right]\right)$. Under the assumption of a constant hazard function, $\lambda(t) = \lambda$, the survival function is: $S(t) = e^{-\lambda t}$. This is known as the exponential model. Other assumptions may give rise to different parametric models for the survival function.

The first case of interest is comparing the survival functions of two independent groups, under the assumption of constant hazard functions, and assuming no censoring. The null hypothesis can be expressed in terms of the hazard difference:

$$H_0 : \lambda_1 - \lambda_2 = 0$$

Define $L = \frac{1}{M}$, where M is the mean survival time calculated from the sample. Let Q_i = proportion of subjects assigned to group i, n_i = number of subjects assigned to group i, $\overline{L} = Q_1 L_1 + Q_2 L_2$, and $S^2 = \overline{L}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)$. The test statistic is defined as $Z = \frac{(L_1 - L_2)}{S}$. Define $\overline{\lambda} = Q_1\lambda_1 + Q_2\lambda_2$. Then sample size is determined by

$$\sqrt{n}\,|\lambda_1 - \lambda_2| = Z_{1-\frac{\alpha}{2}}\sqrt{\overline{\lambda}^2\left(Q_1^{-1} + Q_2^{-1}\right)} + Z_{1-\beta}\sqrt{\lambda_1^2 Q_1^{-1} + \lambda_2^2 Q_2^{-1}}$$

Note that the null hypothesis of no difference between the survival function of the two groups can be expressed alternatively in terms of the log hazard ratio:

$$H_0 : \theta \equiv ln\left(\frac{\lambda_1}{\lambda_2}\right) = 0$$

Lachin and Foulkes (1986) noted that calculations from the hazard difference produces larger sample sizes, and lower power, when compared to those calculated from the log hazard ratio. In addition, Lachin (2009) suggests that the log hazard ratio formulation be used because of its generalization to the Cox proportional hazards model, which allow for modeling survival rates with covariates. Thus the log hazard ratio formulation is the one most used in practice, and that is the formulation we will assume throughout the paper.

For the log hazard ratio parametrization of the null hypothesis, under the assumption of constant hazard rates, the sample size formula, as derived in Lachin (2009) is:

$$\sqrt{n}\left|ln\left(\frac{\lambda_1}{\lambda_2}\right)\right| = Z_{1-\frac{\alpha}{2}}\sqrt{(Q_1 Q_2 E(\delta|\lambda))^{-1}} + Z_{1-\beta}\sqrt{(Q_1 E\,(\delta|\lambda_1))^{-1} + (Q_2 E\,(\delta|\lambda_2))^{-1}} \quad (2.2)$$

where $E(\delta|\lambda)$ is the probability that the event will be observed as a function of the hazard rate λ, and the total exposure of the group. Equation (2.2)

forms the basis for almost all sample size determination for survival analysis. Under the assumption of no censoring, which implies $E(\delta|\lambda) = 1$, equation (2.2) simplifies to:

$$n_f = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{Q_1 Q_2 \theta^2} \tag{2.3}$$

Lachin (1981) showed that under the additional assumptions of no random losses to follow up, uniform patient entry over an accrual period of time T_R, and total maximum follow up time of T_S, then the probability of observing an event is given by:

$$E(\delta|\lambda) = \left[1 - \frac{e^{-\lambda(T_S - T_R)} - e^{-\lambda T_S}}{\lambda T_R}\right]. \tag{2.4}$$

This can be plugged into equation (2.2) for determining the sample size due to administrative censoring only; that is, censoring due to a fixed follow up period only, and no random losses of subjects before the end of the study.

Lachin and Foulkes (1986) considered the case of randomly censored observations due to loss to follow-up, as well as administratively censored observations due to fixed follow-up period. If, in addition to the aforementioned assumptions, we further assume that random losses due to follow-up times are exponentially distributed with constant hazard rate η, which is equal for both groups, then:

$$E(\delta|\lambda, \eta) = \frac{\lambda}{\lambda + \eta}\left[1 - \frac{e^{-(\lambda+\eta)(T_S - T_R)} - e^{-(\lambda+\eta)T_S}}{(\lambda + \eta)T_R}\right].$$

This can be substituted into equation (2.2) for sample size determination.

Lakatos (1988) extended these results further, taking into consideration such things as compliance, lag times, staggered entry, and stratification using a Markov model. Rather than restricting attention to the exponential model, as is usually done for ease of computation, he models the survival curve via a stochastic process. The asymptotic expectation and variance of the usual log rank test statistic is then used for sample size calculations. While much is gained in terms of modelling freedom, there are correspondingly high costs in terms of complexity of concepts and calculations. Snappin and Iglewicz (2005) extended Lakatos's results to take into account the effect of informative noncompliance on sample size determination. Cantor (1992) also eschewed the simple exponential model and determined sample size under the Gompertz model. This has not been very popular, perhaps due to the fact that the sample size solution is not explicit, and it requires the specification of four additional parameters when compared to the exponential survival model.

Our main concerns with the frequentist methods for sample size determination are:

- The use of a plug-in values for all the parameters, without consideration of the uncertainty inherent in using these values. For example, it is always somewhat unsettling to discuss how we determine the value of a nuisance parameter, like the variance, in the sample size formula, since it is generally unknown.

- The use of power as an optimality criterion, since it must be calculated at a particular value under the alternative hypothesis, and it is not obvious what this value should be in most applications.

- The interpretability of size and power. From our experience, there are many practitioners who find these concepts hard to grasp, since they depend on an infinite repetition of the experiment under the same conditions, and the idea of long run frequencies. They tend to misapprehend these frequentist concepts and give them a Bayesian interpretation, that is, the degree of current certainty based on this particular experiment.

It should be noted that, for a certain choice of loss function and prior distribution, the frequentist methods for sample size determination can be made equivalent to the Bayesian decision theoretic method.

2.2.2 Sample Size Determination – Semi-Bayesian Perspective

In direct response to the aforementioned concerns of using the frequentist method, the motivation for the semi-Bayesian approach is that it mitigates the use of a plug-in estimate of the alternative hypothesized value of the parameter of interest in equation (2.2), acknowledging the uncertainty in this value. Another benefit of using a prior distribution on the parameter is that genuine prior knowledge about the parameter, whether historical, expert,

or some such combination, can be incorporated in the current design of the experiment. This can have a significant impact on the power of an test, or the number of events necessary for rejecting the null hypothesis. What has been long acknowledged in the field of physics, for example see Jeffreys (1998) and Jaynes (2003), is becoming more recognized in the pharmaceutical industry, as highlighted by Berry (2006): prior knowledge described via probability distributions is a real thing that can be used to improve experimental design and analysis.

Semi-Bayesian methods, can be bifurcated into two categories: those based on interval estimation, by far the more popular, and those based on hypothesis testing. There is not much research on the latter, most likely due to the view that in most cases what is desired is an interval estimate of the parameter. Parenthetically, hypothesis testing tends to be abused and misapplied. There is one interval estimation method that can be extended for hypothesis testing, so we review this procedure next.

The Average Coverage Criterion: Semi-Bayesian Method Based on Interval Estimation

Following Adcock (1997), we define the following: Let the data consist of n independent observations, $\mathbf{x} = (x_1, x_2, ..., x_n)$. The parameter vector is $\omega = (\theta, \lambda)$ where θ is the parameter of interest, and λ is the so-called nuisance parameter. The likelihood function is $L(\omega|\mathbf{x}) = \prod p(x_i|\omega)$, with some associated prior distribution, $\pi(\omega)$. The prior predictive distribu-

tion is $p(\mathbf{x}) = \int p(x_i|\omega)\pi(\omega)d\omega$, and the posterior distribution is given by $\pi(\omega|\mathbf{x}) = \dfrac{L(\omega|\mathbf{x}) * \pi(\omega)}{p(\mathbf{x})}$. The quintessential idea of the semi-Bayesian interval based methods is to calculate a test statistic based on the observed data, $T(\mathbf{x})$, then require it to meet some given condition, on average, over all possible samples. Namely,

$$E_{p(\mathbf{x})}\left[T(\mathbf{X})\right] = \int T(\mathbf{x})p(\mathbf{x})d\mathbf{x} \leq \varphi$$

where φ is given. Sample size is determined by:

$$\min_{n}\left[E_{p(\mathbf{x})}\left[T(\mathbf{X})\right] - \varphi \leq 0\right]. \tag{2.5}$$

It should be anticipated that the choice of test statistic, $T(\mathbf{x})$, will give rise to different methods. Since there is nothing within this theoretical framework that can determine any type of best choice of statistic, it is unsurprising that there are divers types of sample size determination methods based on equation (2.5).

The most common interval based method is the Average Coverage Criterion (ACC). For the ACC,

$$T(\mathbf{x}) = \begin{cases} 1, & \text{if } \theta \in R(\mathbf{x}) \mid \mathbf{x} \\ 0, & \text{otherwise} \end{cases}$$

where $R(\mathbf{x})$ is a specified region for containing θ. This implies that

$$E_{p(\mathbf{x})}\left[T(\mathbf{X})\right] = \int \left(\int_{R(\mathbf{x})} \pi(\theta \mid \mathbf{x})d\theta\right) p(\mathbf{x})d\mathbf{x} = 1 - \alpha.$$

The simplest case for this type is when $R(\mathbf{x}) = E[\theta \mid \mathbf{x}] \pm \frac{l}{2}$, where l is given.

As shown in Adcock (1997) and Pezeshk (2003), the ACC can be adopted for hypothesis testing in the particular situation of testing a normal mean with known variance. Assume that the sampling distribution is $N(\theta, \sigma^2)$, and that the prior for θ is $N\left(\mu, \frac{\sigma^2}{n_0}\right)$, where σ^2 is assumed known. The two hypotheses are:

$$H_0 : \mu^{'} = \mu_0^{'}$$
$$H_1 : \mu^{'} = \mu_0^{'} + e$$

where $\mu^{'}$ is the posterior mean, and e is the half length of the highest posterior density (HPD) interval, and is assumed given. Under the following two constraints:

$$\int \left\{ Pr\left[|\theta - \mu^{'}| \leq e \mid \overline{x}_n \text{ and } H_0 \right] \right\} f(\overline{x}_n) d\overline{x}_n = 1 - \alpha$$
$$\int \left\{ Pr\left[|\theta - \mu^{'}| \geq e \mid \overline{x}_n \text{ and } H_1 \right] \right\} f(\overline{x}_n) d\overline{x}_n = 1 - \beta$$

the resulting sample size formula is $n = \frac{\sigma^2 (Z_{\frac{\alpha}{2}} + Z_\beta)^2}{e^2} - n_0$. We are usually interested in testing hypotheses about the mean of the sampling distribution, θ, not about the posterior mean $\mu^{'}$. Therefore, this result will be of little use for sample size determination. What is note worthy is the fact that Type I and Type II error constraints were used to determine the sample size. This will be a common technique for the semi-Bayesian approaches.

Semi-Bayesian Methods Based on Hypothesis Testing

Semi-Bayesian methods based on hypothesis testing usually involving averaging the Type I and Type II errors over the prior distribution of the parameter. Spiegelhalter and Freedman (1986) suggested that we calculate the sample size needed for testing based on the averaged power, rather than the power at a single point value of the parameter, as is usually done in the frequentist paradigm. Their idea is very particular to clinical trials, where one wishes to test for inferiority, equivalence, or superiority of a new treatment over some standard. Rather than using $H_0 : \delta = 0$, they suggest using $H_0 : \delta_w = 0$, where δ_w is the worthwhile clinical difference, which is usually well established. Then, rather than calculating power at some δ_A, which is usually ill-defined and quite arbitrary, they suggest using a prior for δ, based on consensus expert opinion on the matter. Then the average power is determined as follows:

$$P_S = \int_{\delta_w}^{\infty} p(S \mid \delta)\pi(\delta)d\delta$$

where $p(S \mid \delta)$ is the probability that the new treatment is superior, for example. The sample size may be determined using the average power. This approach undoubtedly informed Weiss (1997), and Reyes and Ghosh (2013), among others.

The main semi-Bayesian hypothesis testing method, introduced by Weiss (1997), combines Bayes factor with control of Type I and Type II errors for sample size determination. Define $B_{01} = \dfrac{p(\mathbf{x} \mid H_0)}{p(\mathbf{x} \mid H_1)}$ and $ln\,(B_{01}) = b_{01}$. Large

values of b_{01} support the null hypothesis, H_0. If the hypotheses are defined as follows:

$$H_0 : \theta = \theta_0$$
$$H_1 : \theta \sim N\left(\theta_0 + e, \frac{\sigma^2}{n_0}\right)$$

then for given α and β, find the values of n and b_{cut} that simultaneously satisfy the following:

$$\begin{cases} Pr\left(b_{01} > b_{cut} \mid H_0\right) & = 1 - \alpha \\ Pr\left(b_{01} \leq b_{cut} \mid H_1\right) & = 1 - \beta. \end{cases}$$

The solution usually requires iterative computations. If the variance σ^2 is unknown, which is the usual case, then the computations are further complicated due to the form of the marginal distributions under the null and the alternative. This is also a feature of all the Bayesian methods that we will review: they tend to require complicated and complex numerical solutions for all but the simplest cases.

Reyes and Ghosh (2013) proposed a slight modification to Weiss (1997), where they combine the averaged Type I and Type II errors into one total weighted error function. By averaged errors it is meant that they have integrated out the parameters from the sampling distribution based on each hypotheses. That is, they define average Bayes Type I error, $AE_1(t) = \int I_{[T(\mathbf{X})>t]} m_0(\mathbf{x}) d\mathbf{x}$ and similarly $AE_2(t) = \int I_{[T(\mathbf{X})\leq t]} m_1(\mathbf{x}) d\mathbf{x}$, where $m_i(\mathbf{x})$

is the prior predictive or marginal distribution based on the null and alternative parameter space. This is the usual frequentist Type I and Type II errors averaged over their respective prior distributions. A total weighted error can then be defined as $TWE(t, w) = wAE_1(t) + (1 - w)AE_2(t), 0 < w < 1$. The choice of test statistic used is the log Bayes factor, as used in Weiss (1997). The decision rule is to reject H_0 if $T(\mathbf{X}) > t_0(w)$, where $t_0(w)$ is the argument that minimizes $TWE(t, w)$. The optimal sample size is found by finding the minimum n such that $TE\left(t_0(w)\right) \leq \alpha$, where $TE(t) = AE_1(t) + AE_2(t)$. Some issues with this approach are: (i) the use of Bayes factor as a test statistic requires the use of non-continuous priors for precise hypothesis testing; (ii) the cutoff, α, cannot be chosen in an intuitive way, and it is hard to determine what level a combined average error measure should be; and (iii) calculations for the simplest case, normal mean with known variance, requires iterative numerical solutions; the case of unknown variance was not even attempted.

De Santis (2004) also considers sample size for hypothesis testing using Bayes factor. The criterion used, however, is maximizing the probability of obtaining evidence in favor of the true hypothesis. The posterior probabilities of both hypotheses can be expressed in terms of the Bayes factor. The idea then is to make decisive and correct decisions based on some appropriately chosen cutoffs for the Bayes factor under each hypothesis:

$$p^{DC}(k_0, k_1, n) = \Pi_0 p_0^{DC}(k_0, n) + \Pi_1 p_1^{DC}(k_1, n)$$

where $p_i^{DC}(k_i, n)$ is the probability of making correct decision for hypothesis i, and Π_i are prior probabilities of the hypotheses. Sample size is then chosen such that $min\left(n \in \mathbb{N} : p^{DC}(k_0, k_1, n) \geq \varphi\right)$. It should be noted that methods based on Bayes factor and the related posterior probability of the hypotheses are special cases of the Bayesian decision theoretic approach with the use of an appropriately chosen discrete loss function.

Another approach, developed by Rubin and Stern (1998), is to determine the sample size needed to differentiate between a simpler model, labelled the null model, and a more complex alternative model using the posterior predictive distribution. Here it is assume the alternative model is true, and the idea is to determine if the data would support the simpler model labelled by the null hypothesis. Notice that in this approach the models are not necessarily nested, and as such is more general than the methods reviewed thus far. The idea is that, for a given choice of sample size n, data are simulated from the alternative model. This data is compared to data drawn from the posterior predictive distribution of the null model. The fit is determined by the tail area probability, or p-value, calculated with respect to the posterior predictive distribution:

$$p\left(x^{rep} \mid M_0, x_{obs}\right) = \int p\left(x^{rep} \mid M_0, \theta_0\right) \pi\left(\theta_0 \mid M_0, x_{obs}\right) d\theta_0.$$

This probability is calculated for a given sample size, and the process is repeated until the minimum sample size is found that would lead to a rejection of the null model based on suitably chosen p-value. The algorithm is:

1. choose n

2. for $l = 1, \ldots, L$:

   (a) simulate θ from the alternative model

   (b) simulate $\mathbf{x}_l^n$ from marginal distribution of the alternative model

   (c) obtain a single draw from the posterior of the null model of the value of θ_0

   (d) Using that value of θ_0, obtain a draw from the posterior predictive of the null model

3. compare the test statistic calculated from the posterior predictive with that calculated from sampling distribution of data under the null, via p-values

This would be repeated until a minimum sample size is found.

Lee and Zelen (2000) provide us with yet another example of using certain aspects of the Bayesian paradigm to solve problems within the frequentist framework. They suggest using posterior probabilities that are conditioned on the outcome of an experiment, rather than using the Type I and Type II errors probabilities that are conditioned on the unknown parameter, for sample size calculations. Their main idea is to still use the traditional α and β for sample size determination; but rather than simply using standard values, they suggest determining the values of α and β in terms of posterior probabilities. Namely, define $C = +$ or $-$ if we decide to reject or accept

the null respectively, and $T = +$ or $-$ if the true state of affairs is the null is false or the null is true respectively. Then, traditionally, $\alpha = Pr(C = + \mid T = -)$ and $\beta = Pr(C = - \mid T = +)$. Define $\alpha^* = Pr(T = +|C = -)$ and $\beta^* = Pr(T = -|C = +)$, where these are posterior probabilities. This leads to the following definitions: $P_1 = 1 - \alpha^* = Pr(T = -|C = -) = \dfrac{(1-\alpha)(1-\theta)}{[(1-\alpha)(1-\theta) + \beta\theta]}$ and $P_2 = 1 - \beta^* = Pr(T = +|C = +) = \dfrac{(1-\beta)\theta}{[(1-\beta)\theta + \alpha(1-\theta)]}$, where θ is the prior probability of the null hypothesis. From these results, traditional α and β can be rewritten in terms of P_1, P_2, and θ. These values, which can facilitate more subjective information in the formulation of the traditional errors, are used for sample size calculations in the usual manner.

The introduction of a prior distribution addresses the problem of taking uncertainty of the parameters into account. However, there are some aspects of the semi-Bayesian methods that may be disconcerting:

- The fact that there are several possible choices of test statistics. In general all these methods will give a different answer for the optimal sample size given the same information; there is no way to choose among them. We believe, similar to Berger (1985), Lindley (1997), and others, that sample size calculation is a decision problem. Any solution that does not take a decision theoretic approach will contain too many arbitrary components, as is evinced by the many criteria used in the semi-Bayesian approaches.

- Semi-Bayesian methods for sample size determination based on hypothesis testing, like Weiss (1997), and Reyes and Ghosh (2013), might be an attractive alternative for frequentist who wish to deal in some way with parameter uncertainty. However, they use the Bayes factor as their test statistic. This requires the use of prior distributions for testing precise null hypotheses that are fundamentally different from those used for estimation. This is due to the fact that hypothesis testing is usually done with a discrete loss function, while estimation is usually done with continuous loss functions. This schism of priors has been a huge embarrassment for Bayesian statisticians, see Bernardo (2011). Basically, the statistician is forced to admit that his/her prior opinions about a parameter is dependent on the type of analysis to be done; this seems hardly defensible.

- Semi-Bayesian methods for hypothesis testing that do not use the Bayes factor, but rather use a chimera of Bayesian methods combined with frequentist procedures, have even less theoretical foundation. These methods are not based on any decision theoretical criteria; it is simply a variation on the frequentist criteria of controlling size and maximizing power. It is our opinion that the philosophy of "do whatever works for the moment" has done statistics a disservice, and has discredited us in the eyes of other disciplines.

- The methods tend to have very complex structures and Byzantine algo-

rithms; this makes them difficult to explain to non-statisticians. This may seem captious on our part, but this last point is very important, since most statistical analyses are collaborative works.

2.2.3 Sample Size Determination – Fully Bayesian Approach

Sample size determination from the Bayesian perspective — using a utility/loss function along with a prior distribution — has been written about in Schlaifer and Raiffa (1961), Berger (1985), Lindley (1997), and Bernardo (1997), among others. Lindley (1997) succinctly summarizes the method as follows:

$$\min_{n} \left(\int \min_{d} \left[\int l(d, \omega) \pi(\omega \mid \mathbf{x}, n) d\omega \right] p(\mathbf{x} \mid n) d\mathbf{x} \right). \tag{2.6}$$

Intuitively, since the data and parameter are unknown stochastic variables, we take expectation over these variables. Since sample size and decision action are non-stochastic variables, we optimize over these. This approach, in its generality, is quite self-contained. Once the loss function and prior distribution are chosen then everything is automatic. These are the primary user-based parameters; other methods only shift the focus downwards to secondary and tertiary parameters, such as test statistics, cutoffs, and control of errors.

The differences in application of this atomistic theoretical result will be due to the choice of loss function and prior distribution. These choices, ide-

ally, should be problem specific. However, for the sake of further academic development and pragmatism, some common loss functions and priors have been suggested. Berger, Lindley, and Bernardo use loss functions that also consider the cost of sampling. The cost of sampling function is case dependent, but it is often assumed to be linear. For example, Lindley (1997) suggests the use of $ln[\pi(\theta)] - cn$ for the utility function, where $\pi(\theta)$ is your prior distribution, and c is the cost of sampling one unit. This utility, known as a proper utility function, is considered in the case where the final action is simply to state a posterior distribution. It does not make sense in any other context. Parenthetically, this extends the use of the decision theoretic framework to all aspects of statistical inference. For cases where the final decision is a credible posterior interval — an interval (a, b) within which $\theta \mid \mathbf{x}$ must lie with some given probability – then the decision takes the form of all such intervals, $d = (a, b)$. The utility function in this case may be written as $u(a, b)+\delta(\theta)$, where $\delta(\theta) = 1$ if $\theta \in (a, b)$, and $\delta(\theta) = 0$ otherwise. The choice of $u(a, b)$ is still open, and the suggestion is to make it a linear function of the length of the interval. The point here is that it is somewhat problematic to find a suitably representative utility/loss function to discuss in any meaningful terms. Appropriate omnibus loss functions for hypothesis testing has traditionally been discrete in nature. We have already stated what problems arise from using such loss functions for precise hypothesis testing. The loss function suggested by Bernardo and Rueda (2002), which we propose to use, overcomes these problems. The choice of prior distribution is usually the

conjugate prior, a very flexible and non-controversial choice.

Sahu and Smith (2006) consider Bayesian sample size determination specifically for hypothesis testing situations. The method developed can only handle tests of the sort: $H_0 : \theta \leq \theta_0$ verses $H_1 : \theta > \theta_0$. This is due to the use of a discrete loss function, and the desire to use a continuous prior distribution. The idea is to find the sample size that minimizes $r\left(\pi, \delta_n^\pi\right) + cn$, where $r\left(\pi, \delta_n^\pi\right) = \int_{\theta_1} L(\theta, a_0) Pr(g(X) < k^\pi(n) \mid \theta)\pi(\theta)d\theta + \int_{\theta_0} L(\theta, a_1) Pr(g(X) \geq k^\pi(n) \mid \theta)\pi(\theta)d\theta$, a_0 is the action that decides that the null hypothesis is accepted, a_1 denotes that the alternative hypothesis is accepted, and $k^\pi(n)$ is a cutoff for deciding whether or not to accept the null, based on the test statistic $g(\mathbf{X})$. This approach is similar to that espoused by Berger (1985). Thus the idea is to try and control the loss incurred from either decision, by simultaneously controlling the Type I and Type II errors. The loss function used is the usual discrete, $0 - K$ loss function. Sahu uses two conceptually different priors: a fitting prior and a sampling prior. This is in the same vein of Wang and Gelfand (2002). The idea is based mainly on mathematical convenience covertly, but overtly it is stated that the fitting prior, which is usually improper, allows for the analysis of the data in such a way that the experimental evidence will be more influential than the prior information. Thus the fitting prior is used for formulating the posterior distribution. This is implicit in the choice of cut-off, $k^\pi(n)$. The sampling prior, which is required to be proper, is used to generate the sample. This appears explicitly in the formulation of the sample size problem. This prior is usually formed

by limiting the range of the parameter. The use of a non-intrinsic discrete loss function, as well as two separate priors, make this approach sub-optimal in our opinion. Furthermore, the method excludes testing of precise null hypothesis directly, which is an important testing problem in many situations, in particular the survival data analysis that we are interested in investigating.

An interesting extension to the usual Bayesian decision theoretic approach is given in Walker (2003). Walker takes a non-parametric approach to the sample size problem, thus assigning a prior distribution over all density functions. The action considered is not hypothesis testing, however; rather, the decision to be optimized is stating the posterior distribution, and the optimal sample size is sought for this end. This view is similar to Lindley (1997) and Bernardo (1997). Since we are particularly interested in hypothesis testing, however, we direct the interested reader to that work.

As can be seen, there has not been many fully Bayesian work on sample size determination for hypothesis testing. While we do not wish to speculate why this is so, we believe that hypothesis testing, while overused, misused, and abused, still has a prominent role in statistical analysis. In the next section we introduce our procedure for hypothesis testing, which rivals the frequentist approach in simplicity and facility.

2.3 New Method

Our method is based on equation (2.6). The crux of our proposal is in the choice of the loss function.

2.3.1 Continuous Intrinsic Loss Functions

We eschew discrete loss functions because they require special, non-continuous prior functions on the parameter for testing precise null hypothesis. This leads to the well-known Lindley's Paradox (Lindley (1957)) when testing the normal mean. As a direct result of this, the same prior cannot be used for hypothesis testing and estimation in Bayesian analysis. This is important, since it has long been a source of discomfort for the Bayesian methodology that different priors must be used to answer different inferential questions based on the same model and evidence, see Bernardo (2011). Furthermore, discrete loss functions lead to the requirement of a prior distribution on the hypotheses. This is notoriously difficult to specify in a systematic way, and hence is usually rather arbitrary. The usual recourse is the equi-probable prior distribution on the null and alternative hypotheses, but such an assignment may be more due to convenience than any real principle.

Continuous loss functions could alleviate these problems, but then there is the matter of which loss function to choose. Robert (1996) argues that the quadratic loss has been very popular, but is not suitable as a universal loss function. The reasons for its popularity are its tractability, and a confounding

of the ideas of unbiasedness with those of losses. Furthermore, loss functions based on the discrepancy between parameter values are not invariant to re-parametrization. These considerations led Robert to define intrinsic loss functions: loss functions defined in terms of distance between distributions indexed by parameters, rather than differences in parameters. Intrinsic loss functions are derived from the sampling distribution in an automatic manner, thereby giving them some amount of objectivity and generality. They are also parametrization free, in that if the problem is re-parameterized, then the loss is the same. The choice of an appropriate discrepancy measure between distributions is not obvious, and is usually based on convenience and ease of calculation.

A well-known distance measure for distributions is the Hellinger distance:

$$H^2(p,q) = \frac{1}{2}\int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mathbf{x}$$

The disadvantages of this loss function are intractable calculations in general, and difficulty of interpretation. Another attractive distance measure, introduced by Matusita (1967), is the negative log affinity:

$$\rho(p,q) = -ln\left[\int \sqrt{pq} d\mathbf{x}\right]$$

This measure has many attractive properties for use as an intrinsic loss function:

i $\rho(p,q) > 0$

ii $\rho(p,q) < \infty$

iii $\rho\left(\prod_{i=1}^{n} p_i, \prod_{i=1}^{n} q_i\right) = \sum_{i=1}^{n} \rho(p_i, q_i)$ for any dependency structure.

The last property in particular is extremely useful, as it simplifies the calculation of the loss function for sampling data even in cases of non-independence. However, the interpretation of this loss function is not intuitive, and thus it would be difficult to set a cutoff for hypothesis testing. The ease of interpretation of the intrinsic discrepancy loss function, even at the expense of more arduous calculations, is worthwhile for calibration purposes. There are many other options for the choice of intrinsic loss function; we will use the intrinsic discrepancy loss function introduced by Bernardo and Rueda (2002).

The Bernardo Intrinsic Discrepancy Loss

Assume the general "true" model for the data is:

$$\mathcal{M}_{\mathcal{X}} = p(\mathbf{x} \mid \theta, \lambda), \mathbf{x} \in \mathcal{X}, \theta \in \Theta, \lambda \in \Lambda.$$

Let the model labelled by the null hypothesis $H_0 : \theta = \theta_0$ be:

$$\mathcal{M}_0 = p(\mathbf{x} \mid \theta_0, \lambda_0), \mathbf{x} \in \mathcal{X}, \lambda_0 \in \Lambda.$$

Then the loss function is defined as

$$l\left(\theta_0(\theta, \lambda)\right) = \inf_{\lambda_0 \in \Lambda} \delta\left[p_{\mathcal{X}}(\cdot \mid \theta, \lambda), p_{\mathcal{X}}(\cdot \mid \theta_0, \lambda_0)\right]$$

where $\delta[p, q] = min\left[\int q ln\frac{q}{p}, \int p ln\frac{p}{q}\right]$ is the minimum Kullback-Leibler directed divergence between the two distributions. Notice that the loss function does not take into account the cost of sampling, as is done in Berger (1985),

Lindley (1997) and Sahu (2006), for example. That can be easily added if deemed important; however, we avoid using the cost of sampling as it would add an additional layer of assumptions that would be best left to a case by case analysis. This loss function satisfies all the desiderata stated above, and in addition has a few more desirable properties. It is additive for conditionally independent data, which makes for easier calculation. Furthermore, and perhaps most appealing, is its ease of interpretation: the minimum average log-likelihood ratio needed to distinguish between the two models. The fact that it incorporates the log of the likelihood ratio makes this loss function apprehensible and quite familiar to statisticians and practitioners alike. It should also be noted that with this loss function, we do not assume that the nuisance parameter has the same value in both models. The use of the infimum to deal with the nuisance parameter is defended in Bernardo (2011). The other ingredient necessary for the implementation of (2.6) is the prior distribution.

2.3.2 The Prior Distribution

The class of priors that we espouse for sample size calculation is the conjugate prior. We determine the sample size as a function of the hyper-parameters, and then investigate how much of an effect the hyper-parameters have on the sample size. We consider conjugate priors for a variety of reasons, not least of which are their mathematical tractability, and modeling diversity. While some decry them as not being truly subjective priors, we find this to

be a strength. In fact, the family of conjugate prior for a given problem is determined by the sampling distribution/likelihood function, which gives it a certain amount of objectivity. The determination of the hyper-parameters are subjective, and thus it is possible to interweave expert opinion in the formulation of the prior. It seems that conjugate priors strike a balance between objective and subjective prior distribution.

Samaniego (2010) underscores the utility of conjugate priors by pointing to the fact that they improve the mixing properties and convergence rates of numerical methods like the Gibbs sampler. It is also the case that, in the exponential family of distributions, the posterior mean is a convex combination of the conjugate prior mean and the frequentist's uniform minimum variance unbiased estimator. This allows for an elegant interpretation in terms of prior knowledge and knowledge obtained from the data. A more theoretical justification for using conjugate priors is that any exponential family prior can be expressed as a mixture of conjugate priors (Robert (2007), Bernardo and Smith (1994)). Though not constructive, this fact gives a theoretical underpinning to the choice of conjugate priors.

2.3.3 The Proposed Procedure

As in Bernardo and Rueda (2002), and Bernardo (2011), let us consider the following testing situation: $H_0 : \theta = \theta_0$ verses $H_1 : \theta \neq \theta_0$. In this set up the alternative model is generally accepted as true, however significant simplification and perhaps an increase in explanatory power can be gained if

the null is true, that is, if the null is deemed compatible with the data. Or it could be that the null model is embedded in a more complex alternative model, as is the case in survival analysis, and it is required to determine, based on the sample, whether or not it is worthwhile to depart from the null model.

The best decision in the hypothesis testing problem is to reject the null if and only if the expected posterior loss under the null hypothesis is greater than some suitably chosen value:

$$E_{\theta|\mathbf{x}}\left[l\left(\theta_0,(\theta,\lambda)\right)\right]=\bar{l}>l_0.$$

This position is expounded upon in Bernardo and Rueda (2002), and Bernardo (2011). The constant, l_0, is context dependent, and represents the minimum average log-likelihood against the null. For example, if $l_0 = ln(10)$, then this indicates that we will be willing to reject the null if the minimum average likelihood ratio against the null is expected to be larger than 10, i.e., the observed sample is 10 times more likely, on average, under the alternative than the null. This type of interpretation may be more intuitive as a testing criterion, as likelihood ratios are very easy to explain and understand. Compare this to explaining why we choose the cutoff of $\alpha = 0.05$ for the frequentist procedure.

Main Result

Once we have the optimal decision, the sample size is determined from equation (2.6), which results in the following:

$$\underset{n}{\operatorname{argmin}} \left[E_{\mathbf{x}} E_{\theta|\mathbf{x}}[l] > l_0 \right].$$

Since the intrinsic discrepancy loss function is a function of the parameter only, and using Fubini's Theorem, we can reformulate our sample size method as $\underset{n}{\operatorname{argmin}} \left[E_{\theta} E_{\mathbf{x}|\theta}[l] > l_0 \right]$, which implies

$$\boxed{\underset{n}{\operatorname{argmin}} \left[E_{\theta}[l] > l_0 \right]}. \tag{2.7}$$

Equation (2.7) can be used quite generally to calculate the optimal sample size for testing a precise null hypothesis. This is the main result of this chapter, and its simplicity, elegance, and coherence should be compared to other Bayesian methods. We can determine the sample size for any type of sampling distribution, if not explicitly, then by Monte Carlo integration or some other simple numerical method. Compare this with the results of Weiss (1997), Reyes and Ghosh (2013), Rubin and Stern (1998), Wang and Gelfand (2002), and others. In fact, we have not seen a Bayesian sample size formula that is simpler to implement on any problem, the only restriction being the test should be precise.

However, the result will hold for composite hypotheses as well, as can be seen in the following theorem:

Theorem 2.3.1 *If the sampling distribution is from a regular exponential family, then result equation* (2.7) *will hold for composite hypotheses of the form* $H_0 : \theta \leq \theta_0$ *verses* $H_1 : \theta > \theta_0$.

Proof: Let us fix a $\theta_0^{'} \in \theta \leq \theta_0$ and $\theta_1 \in \theta > \theta_0$. Let us assume, without loss of generality, that there is no nuisance parameter, since the loss function, and hence the decision rule, will be taken over the infimum of the nuisance parameter, and so it does not play a role in our sample size calculations. Then our decision rule remains the same: reject $H_0 : \theta = \theta_0^{'}$ if $E_{\theta|\mathbf{x}}\left[l\left(\theta_0^{'}, \theta_1\right)\right] > l_0$ for some given l_0. Now $l\left(\theta_0^{'}, \theta_1\right) = \min\left[\int p(x \mid \theta_1) \ln\left(\frac{p(x \mid \theta_1)}{p(x \mid \theta_0^{'})}\right), \int p(x \mid \theta_0^{'}) \ln\left(\frac{p(x \mid \theta_0^{'})}{p(x \mid \theta_1)}\right)\right] \Rightarrow \int p(x \mid \theta_0^{'}) \ln\left(\frac{p(x \mid \theta_0^{'})}{p(x \mid \theta_1)}\right)$, since the sampling distribution is from the exponential family, and thus possesses the MLR property, and $\theta_1 > \theta_0^{'}$ always. If we can show that this loss, as a function of any θ in the null parameter space, is monotone decreasing, then our decision rule will be the same for the composite hypothesis and thus for our sample size determination. Let us fix another $\theta_0^{''} \in \theta \leq \theta_0 \mid \theta_0^{''} > \theta_0^{'}$. Then $\int p(x \mid \theta_0^{''}) \ln\left(\frac{p(x \mid \theta_0^{''})}{p(x \mid \theta_1)}\right) - \int p(x \mid \theta_0^{'}) \ln\left(\frac{p(x \mid \theta_0^{'})}{p(x \mid \theta_1)}\right) < 0$ for all such $\theta_0^{'}$ and $\theta_0^{''}$ in the null space, due to the fact that the integrals are Kullback-Leibler divergence measures of $p(x \mid \theta_1)$ from $p(x \mid \theta_0^{''})$ and $p(x \mid \theta_0^{'})$ respectively.

We will now turn our attention to fixed sample size determination for hypothesis testing of equality of survival functions in particular.

2.4 Application of Result to Survival Analysis

2.4.1 The Problem Statement

Let us assume that we have two independent groups of subjects, one called control (group 1), the other treatment (group 2). Let us further assume independence among subjects within each group. We are interested in the survival functions, or correspondingly the hazard functions, of each group. Specifically, we would like to determine the fixed optimal sample size required to test the hypothesis that there is no difference between the survival functions for each group after some pre-specified length of time. Similar to Lachin (2009), we assume the following about the data:

- the hazard functions, $\lambda_i(t)$, are constant for $\forall t$; this implies the exponential distribution for the event probability distribution, and
- the censoring distribution is independent of the event distribution, and is the same for both groups.

Our method will be applied to the situations of no censoring, and administrative censoring but no random losses to follow-up.

2.4.2 Sample Size Determination for the Exponential Model with no Censoring

The null hypothesis is $H_0 : \theta = 0$ verses $H_0 : \theta \neq 0$, where $\theta = ln\left(\frac{\lambda_1}{\lambda_2}\right)$, and λ_i represents the hazard rate of group $i, i = 1, 2$. The likelihood for the i-th patient is:

$$L_i = \left[\lambda_1^{\delta_i} e^{-\lambda_1 t_i}\right]^{1-Z_i} \left[\lambda_2^{\delta_i} e^{-\lambda_2 t_i}\right]^{Z_i}$$

where $\delta_i = \begin{cases} 1, & \text{if event occurs} \\ 0, & \text{otherwise} \end{cases}$, and $Z_i = \begin{cases} 1, & \text{if treatment} \\ 0, & \text{if control} \end{cases}$. The likelihood for all patients is:

$$L = {\lambda_1}^{n_1} e^{-\lambda_1 \sum_{i=1}^{n_1} t_i} {\lambda_2}^{n_2} e^{-\lambda_2 \sum_{j=1}^{n_2} t_j}$$

where t_i is time of observation for control, t_j is time of observation for treatment, n_1 is the number of subjects assigned to control,and n_2 is the number of subjects assigned to treatment, and $n_1 + n_2 = n$, since we assume no censoring. The maximum likelihood estimators for λ_1 and λ_2, which are also sufficient statistics, are: $\hat{\lambda_1} = \frac{n_1}{\sum_{i=1}^{n_1} t_i}$, and $\hat{\lambda_2} = \frac{n_2}{\sum_{j=1}^{n_2} t_j}$. Asymptotically, $\hat{\lambda} \mid \lambda \sim N\left(\lambda, \frac{\lambda^2}{n}\right)$, and thus $ln\left(\hat{\lambda}\right) \mid ln(\lambda) \sim N\left(ln(\lambda), \frac{1}{n}\right)$. This implies that:

$$\hat{\theta} \mid \theta \sim N\left(\theta, \frac{1}{n_2} + \frac{1}{n_1}\right)$$

where $\hat{\theta} = ln\left(\frac{\hat{\lambda_1}}{\hat{\lambda_2}}\right)$. Let $Pr(Z_i = 1) = p$. Then $n_1 = (1-p)n$, and $n_2 = pn$. Thus:

$$\hat{\theta} \mid \theta, \sigma^2, n \sim N\left(\theta, \frac{\sigma^2}{n}\right)$$

where $\sigma^2 = \dfrac{1}{p(1-p)}$, which is assumed to be given.

We will use the following conjugate prior:

$$\theta \mid \mu, \sigma^2, n_0 \sim N\left(\mu, \frac{\sigma^2}{n_0}\right), n_0 > 0$$

where n_0 represents the "effective sample size" used to determine our prior. The larger n_0 is, the more certain we are in our prior belief about the parameter of interest.

The intrinsic discrepancy loss function, using the definition above, is:

$$l(\theta_0, \theta) = \frac{n(\theta - \theta_0)^2}{2\sigma^2} \Rightarrow l(\theta) = \frac{n\theta^2}{2\sigma^2} \tag{2.8}$$

the Mahalanobis distance between θ and θ_0, a very elegant and satisfying statistical result. These are the only two elements needed for our sample size determination. However, for the sake of comparison to the frequentist method, and for the scenario in which Bayesian statistical analysis is done after the sample is actually collected, we will have need for the following quantities. The resulting posterior, based on standard results, is:

$$\theta \mid \hat{\theta} \sim N\left(\frac{n\hat{\theta} + n_0\mu}{n + n_0}, \frac{\sigma^2}{n + n_0}\right).$$

The expected loss is:

$$\bar{l} = E_{\theta|\hat{\theta}}\left[\frac{n\theta^2}{2\sigma^2}\right] = \frac{n}{2\sigma^2}\left[\frac{\sigma^2}{n + n_0} + \left(\frac{n\hat{\theta} + n_0\mu}{n + n_0}\right)^2\right].$$

The marginal/prior predictive is:

$$\hat{\theta} \mid \mu, \sigma^2, n_0, n \sim N\left(\mu, \sigma^2\left(\frac{1}{n} + \frac{1}{n_0}\right)\right).$$

It is readily seen that:

$$E_\theta[l] = E\left[\frac{n\theta^2}{2\sigma^2}\right] = \frac{n}{2n_0} + \frac{n\mu^2}{2\sigma^2}.$$

Plugging this into equation (2.7) gives the minimum sample size as:

$$n_B > \frac{l_0}{\frac{1}{2n_0} + \frac{\mu^2}{2\sigma^2}}. \tag{2.9}$$

Note that as $n_0 \to 0$, $n_B \to 0$, and as $n_0 \to \infty$, $n_B \to \frac{\sigma^2(2l_0)}{\mu^2}$. These results can be interpreted as follows: if you are in a situations where you have no prior knowledge about θ, then the hypothesis testing problem is ill-posed; you must have some belief about θ, or else you would not entertain a test of hypothesis. In the situation where you are dogmatic in your prior opinion about θ — this is in essence the frequentist position — then the resulting sample size is similar to the frequentist sample size in equation (2.3), where $\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 = (2l_0)$.

Dependence of Sample Size on the Parameters

Figure 2.1 presents the results for sensitivity to parameters. σ, μ, and l_0 will be uncontroversial, based simply on the nature of the decision problem. The most interesting thing to note is the dependence on n_0. The resulting sample

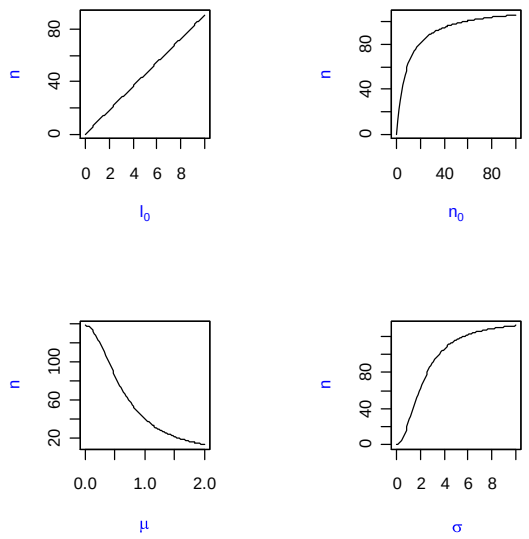

Figure 2.1: Dependence of n_B on parameters: from top-left, clockwise $\sigma = 2$, $n_0 = 10$, $\mu = \ln 2$; $\sigma = 2$, $l_0 = \ln 1000$, $\mu = \ln 2$; $\sigma = 2$, $l_0 = \ln 1000$, $n_0 = 10$; $\mu = \ln 2$, $l_0 = \ln 1000$, $n_0 = 10$.

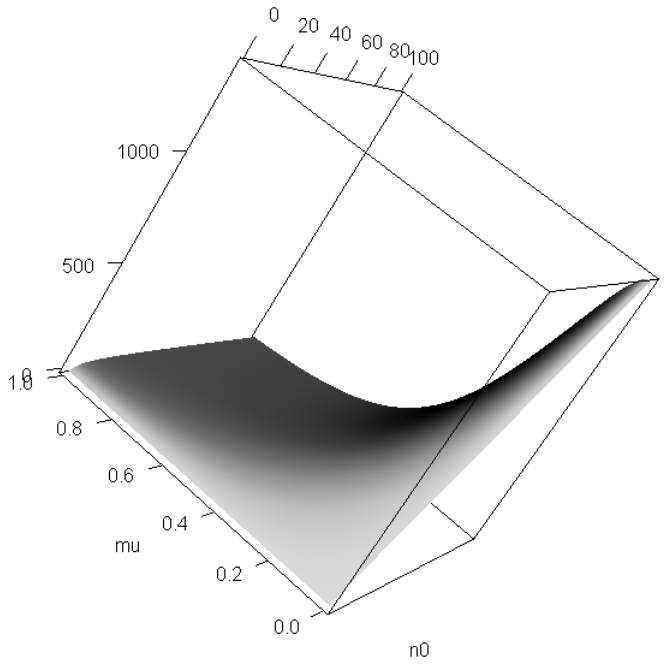

Figure 2.2: Dependence of n_B on μ and n_0 simultaneously.

size changes rapidly for values of $0 < n_0 < 30$, so choices of n_0 should be clearly stated and justified in the analysis.

Figure 2.2 illustrates the notion that if the expected value of the log hazard ratio is close to the null value, and this prior opinion is very dogmatic, then it will be very difficult to reject the null hypothesis, so the sample size shoots up towards infinity. This phenomenon is very similar to the frequentist situation where $\theta \to 0$. However, if I am not dogmatic about my prior, and so n_0 is not very large, then even if the expected prior value of the parameter is equal to the null value, the sample size required to reject the null hypothesis is not infinite. This is a marked improvement over the frequentist sample size formula, where the sample size required to reject the null becomes unbounded as the alternative hypothesised value of θ goes to 0.

Consistency of Procedure

The decision rule is $\bar{l} > l_0$, which implies $\frac{n}{2\sigma^2}\left[\frac{\sigma^2}{n+n_0} + \left(\frac{n\hat{\theta}+n_0\mu}{n+n_0}\right)^2\right] > l_0$. This can be rewritten in a familiar form: $|T| > \sigma\sqrt{\frac{2l_0}{n} - \frac{1}{n+n_0}}$, where $T = \frac{n\hat{\theta}+n_0\mu}{n+n_0}$. Since $T \mid \theta \sim N\left(\frac{n\theta+n_0\mu}{n+n_0}, \frac{n\sigma^2}{(n+n_0)^2}\right)$, then:

$$Pr(reject) = Pr\left(|T| > \sigma\sqrt{\frac{2l_0}{n} - \frac{1}{n+n_0}}\right) = 1 - \Phi\left(\frac{a-b}{\sqrt{c}}\right) + \Phi\left(\frac{-a-b}{\sqrt{c}}\right) \quad (2.10)$$

where $a = \sigma\sqrt{\frac{2l_0}{n} - \frac{1}{n+n_0}}$, $b = \frac{n\theta + n_0\mu}{n+n_0}$, and $c = \frac{n\sigma^2}{(n+n_0)^2}$. It can be seen that as $n \to \infty$, $Pr(reject) \to 0$ when $\theta = 0$, and $Pr(reject) \to 1$ otherwise.

In fact, this is a more general result, in that our decision procedure will always be consistent if we use a continuous function of the maximum likelihood-ratio estimator (MLE) to summarize our data, as can be seen in the following theorem:

Theorem 2.4.1 *If the data is summarized by a continuous function of the MLE, then our decision procedure for testing $H_0 : \theta = \theta_0$ will be consistent, that is, will reject the null if the null is false, and accept the null if the null is true, as our sample size increases.*

Proof: Let T be the MLE for θ; then standard results show that $T \sim N\left(\theta_*, \frac{\mathrm{I}^{-1}}{n}\right)$, where θ_* is true parameter value, and I is the Fisher information. Since the posterior expected loss will be a function of the MLE, say $g(T)$, then applying the delta method, we have $g(T) \sim N\left(g(\theta_*), \frac{\mathrm{I}^{-1}}{n}(g'(\theta_*))^2\right)$. The power of the procedure is $Pr\left(g(T) > l_0 \mid \theta\right) \Rightarrow Pr\left(Z > \frac{\sqrt{n}\left(l_0 - g(\theta_*)\right)}{\sqrt{\mathrm{I}^{-1}(g'(\theta_*))^2}} \mid \theta\right)$, where Z is distributed as a standard normal variable. It is clear that when $\theta = \theta_*$, and as $n \to \infty$, then the probability that we reject goes to 0.

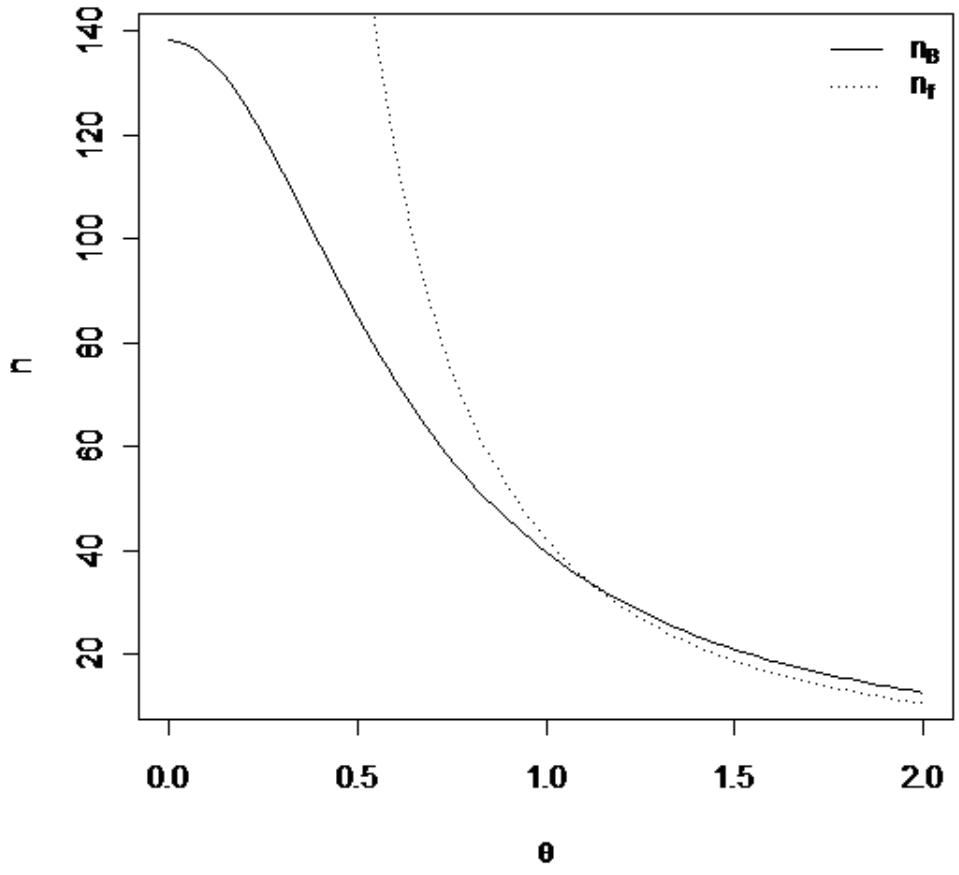

Figure 2.3: Comparison of n_B to n_f: $\sigma = 2$, $\mu = \ln 2$, $n_0 = 10$, $l_0 = \ln 1000$, $\alpha = 0.05$, $\beta = 0.1$.

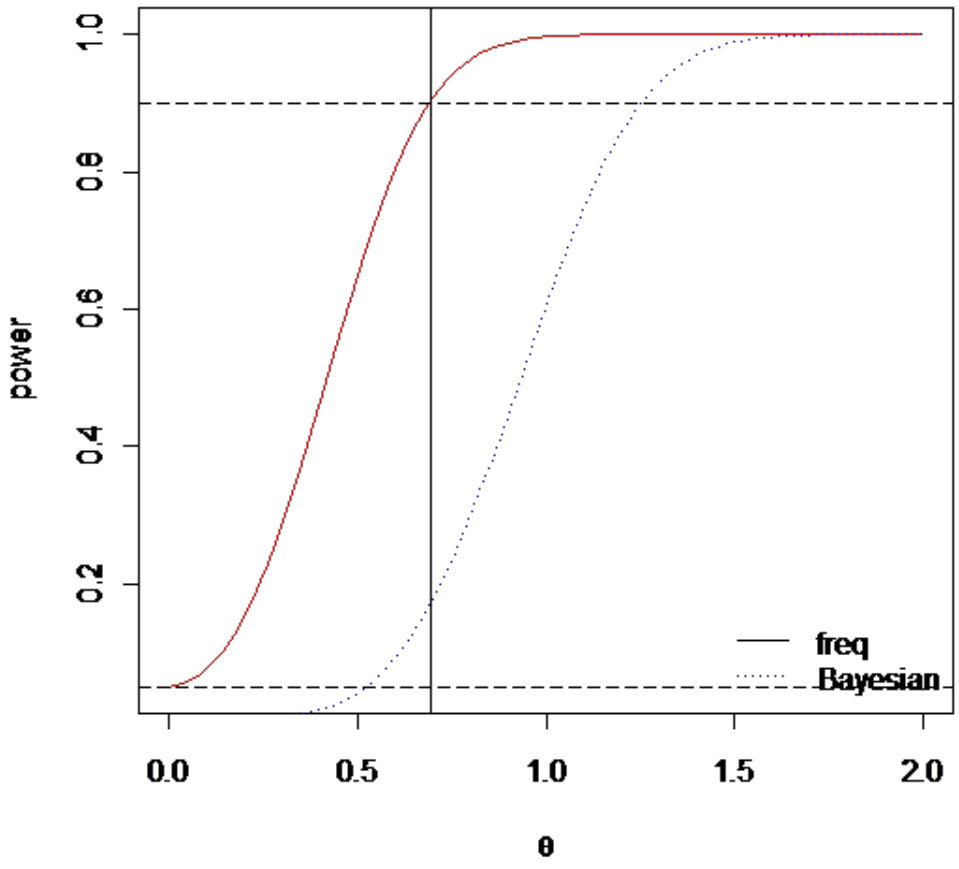

Figure 2.4: Comparison of power: $n_B = 64$, $n_f = 88$,$\sigma = 2$, $\mu = \ln 2$, $n_0 = 10$, $l_0 = \ln 1000$, $\alpha = 0.05$.

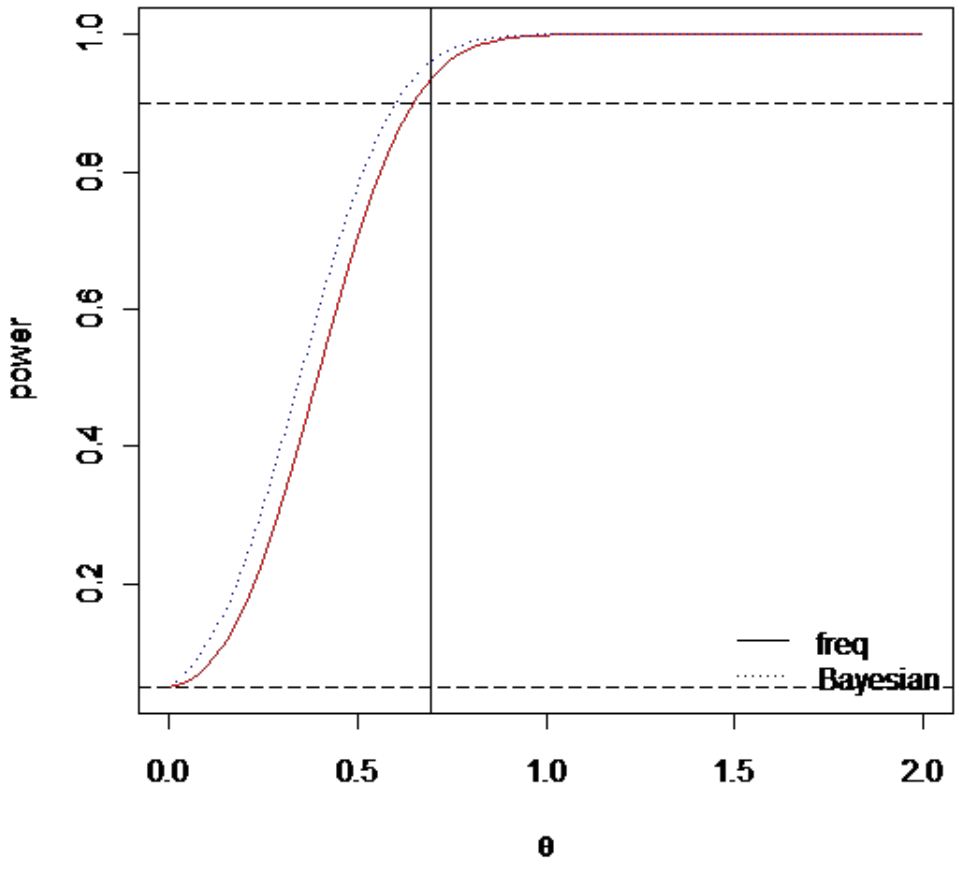

Figure 2.5: Comparison of power: $n_B = 100$, $n_f = 100$,$\sigma = 2$, $\mu = \ln 2$, $n_0 = 10$, $l_0 = 2.228843$, $\alpha = 0.05$.

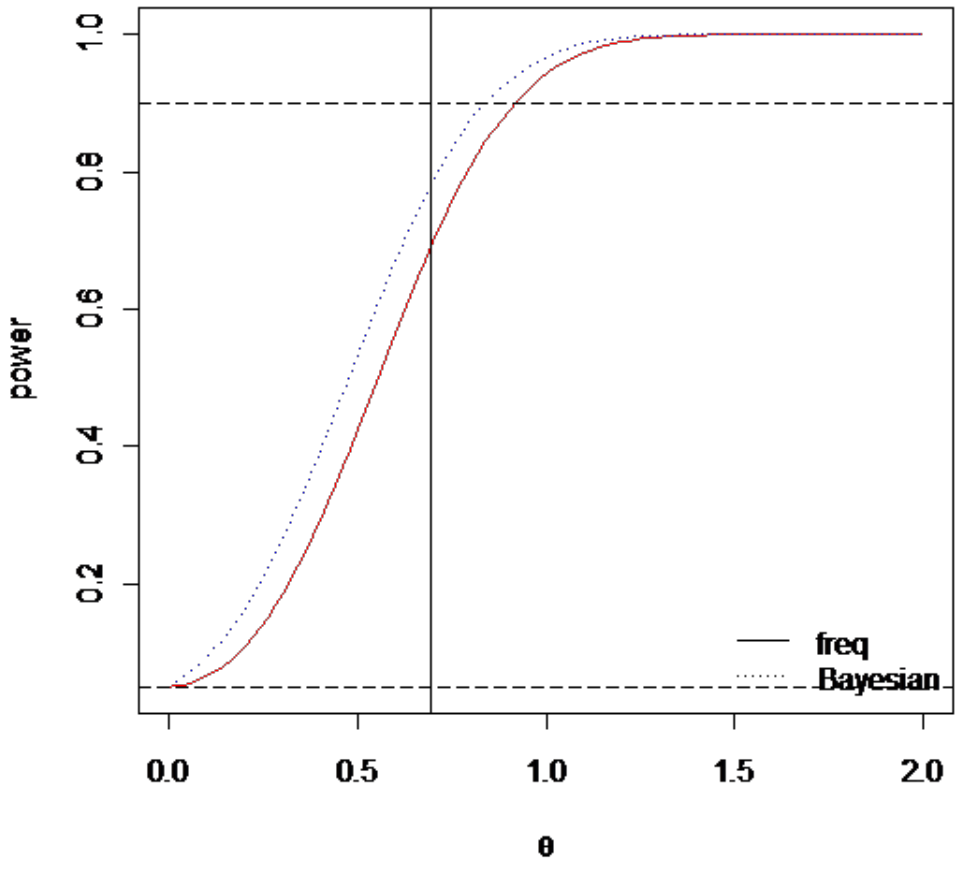

Figure 2.6: Comparison of power: $n_B = 50$, $n_f = 50$,$\sigma = 2$, $\mu = \ln 2$, $n_0 = 10$, $l_0 = 2.05722$, $\alpha = 0.05$.

Comparison with Frequentist Sample Size

While our testing criterion is different from that used in the frequentist method, it may be interesting to see what the results are for each in a given situation. For this example we use the Lupus Nephritis study from Lewis, Hunsicker, Lan, et al. (1992), which is reintroduced in Lachin (2009). We would like to determine the minimum sample size necessary to test the aforementioned null hypothesis, where treatment is plasmapheresis plus standard immune-suppressive therapy and the control is standard therapy alone. Let us assume that the trial is designed to have a power of 0.9 to detect a 50% reduction in risk. The size of the test is $\alpha = 0.05$. It should be noted that Lachin did the sample size determination under the assumption of a one-tailed test, but we use a two-tailed test since this is the scientific and unbiased alternative.

Under the assumption of no losses to censoring, that is, everyone is followed until an event is observed, the frequentist sample size is given by equation (2.3). For $\sigma^2 = 4$ (assumption of equal assignment to treatment or control), $\alpha = 0.05$, $\beta = 0.1$, $\theta = ln(2)$ (50% reduction in risk), then $n_f = 88$. For $\sigma^2 = 4$, $n_0 = 10$ (a standard choice representing reasonable amount of uncertainty), $\mu = ln(2)$ (representing a prior belief that is consistent with the frequentist minimal significant difference), and $l_0 = ln(1000)$ (a rather stringent criteria for rejecting the null), then $n_B = 63$. This comparison may be extended for various choices of the hypothesised log hazard ratio (μ for our procedure). The results are shown in figure 2.3. Two things are note-

worthy: our procedure gives smaller sample sizes for all values of θ up to 1.12, which translates to a hazard ratio of about 3. The second thing is that the frequentist procedure becomes unstable as $\theta \to 0$, requiring increasingly larger sample sizes to reject the null hypothesis. Our procedure, conversely, approaches a limiting size of $n_B = 2 * n_0 * l_0$ as $\mu \to 0$. This satisfying result is due simply to the fact that we account for variability in our opinion of the true value of the log hazard ratio.

While our subjective choices for μ and l_0 above may seem reasonable under the given setup, the choice of n_0 may require more justification. This is especially true given the dependence of sample size on this parameter, as shown in figure 2.1. One way is to ask the investigators of the range expected for the log hazard ratio θ . Dividing the resulting range by 6 would give a rough estimate of the prior standard deviation of θ. Then, given that $\sigma^2 = 4$, the value of n_0 could be calculated. This procedure of eliciting the mean and standard deviation for θ only makes sense because we assume the normal distribution. In fact, Samaniego (2010) argues that, for the exponential family, all conjugate priors can be characterized by their mean and variance. Otherwise, eliciting a prior is indeed a difficult undertaking, and we do not wish to hand-wave this critical aspect of our analysis. Much more work needs to be done in this area.

Notice that our procedure does not operate under the Type I and Type II constraints of the frequentist procedure, and thus comparing sample sizes may seem fallacious. However, one of our objectives is to present an alter-

native procedure to the frequentist framework, one that may be attractive to investigators in certain circumstances. The procedure we present severely tests the null, and is guaranteed regardless of the actual unknown value of θ. We do not try to maximize the power of the test for a given value of θ, and so a test based on our sample size calculations cannot be as powerful as the frequentist's. See, for example, the comparison of the power curves in figure 2.4. The Bayesian procedure, with its stricter cutoff, over-controls type I error, and has very low power in comparison to the frequentist approach. However, if we ignore the sample size determination problem and just investigate power for a given sample size, then some power comparison can be done.

For a given sample size, our testing rule given above can be shown to be a convex combination of the frequentist estimator and the prior mean; this is a direct consequence of using a conjugate prior. We can determine the value of l_0 that will control the Type I error of the Bayesian decision rule for the test for given μ and n_0. Using the power function for our decision rule in equation (2.10) given above, and using the following parameters: $\sigma^2 = 4$, $n_0 = 10$, $\mu = ln(2)$, $n = 88$, and $\theta = 0$ (this would give the corresponding Type I error), then the corresponding cutoff is $l_0 = 2.204321 \approx ln(9)$, hardly a stringent condition on the null hypothesis by our standard. This means that we are willing to reject the null if, regardless of the true value of θ, the observed data is 9 times more likely under the alternative than the null hypothesises model. At any rate, using the same sample size for both the

Bayesian decision rule and the usual frequentist test, and determining the corresponding value of l_0 that will control the Type I error, we find that the Bayesian procedure is in fact more powerful than the frequentist procedure. This is illustrated in both 2.5 and 2.6. These calculations are valid, once again, only if the sample size is already given; this may be the case if no experimental design was done, but observations were collected simply based on time, money, or some other limited resource. Notice that the Bayesian procedure is uniformly more powerful than the frequentist procedure. This is the power of using prior information. Note that the results are the same regardless of the choice of n_0, so it does not depend on how sure we are in our prior. Also, as long as $\mu > 0$, the Bayesian test be more powerful than the frequentist test. A similar result is shown by Samaniego (2010, chapter 5) with regards to estimation of a parameter using the squared error loss function. There he shows theoretically that the Bayesian estimate will beat the best frequentist estimate (UMVUE) for a wide range of priors, if performance is measured by the frequentist risk function. It can be argued that one can unscrupulously choose the prior to obtain a more powerful test. However, the choice of hyper-parameters are clearly stated, and if they seem unreasonable then the procedure can be contested. Also, the result is surprisingly robust to the choice of n_0, and this hyper-parameter is the most controversial one. At any rate, as long as all assumptions are in the open then critiquing the choices of hyper-parameters are a welcome part of the statistical discourse.

In general, we can always choose a value for l_0 such that the testing procedure we use have a Type I error of α, as long as we use a function of the MLE to summarize our data. This is an immediate consequence of theorem 2, which we state as follows:

Corollary 2.4.1.1 *For the conditions given in theorem 2, and the power function provided in equation* (2.10)*, then choosing* $\frac{\sqrt{n}\,(l_0 - g(\theta_0))}{\sqrt{I^{-1}(g'(\theta_0))^2}} = Z_{1-\frac{\alpha}{2}} \Rightarrow$ $l_0 = \frac{Z_{1-\frac{\alpha}{2}}\sqrt{I^{-1}(g'(\theta_0))^2}}{\sqrt{n}} + g(\theta_0)$ *will ensure that the sample size chosen by our procedure will control Type I error at level* α*.*

Comparison with Semi-Bayesian Methods

As shown in Adcock (1997) and Pezeshk (2003), the ACC can be adopted for hypothesis testing by using the following: $H_0 : \mu^{'} = \mu_0^{'}$ verses $H_1 : \mu^{'} = \mu_0^{'} + e$, where $\mu^{'}$ is the posterior mean of θ and e is the half length of the highest posterior density (HPD) interval, and is assumed given. The sample size is $n_{ACC} = \frac{\sigma^2\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{e^2} - n_0$. This would give a sample size, under the conditions above, of $n_{ACC} = 78$.

As can be seen in Weiss (1997), given the following specification of the test $H_0 : \theta = \theta_0$ verses $H_1 : \theta \sim N\left(\theta_0 + e, \frac{\sigma^2}{n_0}\right)$, a minimal sample size can be determined numerically, but not with an explicit solution. However, Adcock (1997) showed that as $n_0 \to \infty$, implying that the alternative becomes a point alternative, then the sample size is $n_W = \frac{\sigma^2\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{e^2}$. This is equivalent to the frequentist sample size.

2.4.3 Sample Size Determination for Administrative Censoring Only

Let us assume, as in Lachin (2009), that T_S = maximum length of study, T_R = recruitment period, we have uniform entry during the recruitment period, and we have no random losses to follow-up. These assumptions imply $E(\delta|\lambda) = \left[1 - \frac{e^{-\lambda(T_S - T_R)} - e^{-\lambda T_S}}{\lambda T_R}\right]$. Thus, asymptotically, $\hat{\theta} \mid \theta, \sigma^2, n \sim N\left(\theta, \frac{\sigma^2}{n}\right)$, where $\sigma^2 = \left(\frac{1}{(1-p)E[\delta \mid \lambda_1]} + \frac{1}{pE[\delta \mid \lambda_2]}\right)$.

For the prior, we choose $\pi(\lambda_1, \lambda_2) = \pi(\lambda_1)\pi(\lambda_2) = \frac{{\beta_1}^{\alpha_1}{\beta_2}^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}{\lambda_1}^{\alpha_1 - 1}e^{-\beta_1\lambda_1}{\lambda_2}^{\alpha_2 - 1}e^{-\beta_2\lambda_2}$ This implies $\pi(\theta, \lambda_2) = \frac{{\beta_1}^{\alpha_1}{\beta_2}^{\alpha_2}}{\Gamma(\alpha_1)\Gamma(\alpha_2)}e^{\alpha_1\theta}{\lambda_2}^{\alpha_1 + \alpha_2 - 1}e^{-\lambda_2\left(\beta_1 e^{\theta} + \beta_2\right)}$. For computation facility, let us assume $\alpha_1 = \frac{\beta_1}{2}$, and $\alpha_2 = \frac{\beta_2}{2}$. Then

$$\lambda_2 \mid \theta \sim Gamma\left(\frac{\beta_1 + \beta_2}{2}, \beta_1 e^{\theta} + \beta_2\right) \tag{2.11}$$

$$e^{\theta} \sim \mathcal{F}(\beta_1, \beta_2) \tag{2.12}$$

Calculations from Bernardo (2011) show that $l = \frac{n}{2} ln\left(1 + \frac{\theta^2}{\sigma^2}\right)$. Thus $E_\theta[l] = \frac{n}{2} E_\theta\left[ln\left(1 + \frac{\theta^2}{\sigma^2}\right)\right]$. This expectation can be done by the Monte Carlo method, for example. Thus, using inequality (2.7), the minimum sample size can be found.

Comparison to Frequentist Method

An example, once again taken from the Lupus Nephritis example in Lachin (2009), is a study in which $T_R = 4$, $T_S = 6$, $\theta = ln(2)$ (treatment will result in a 50% reduction in hazard rate compared to control), and $p = 0.5$. In addition to these parameters, we would require the investigators to give us an estimate of the variances for λ_1 and λ_2, for the sake of prior specification. In the Nephritis study, it is assumed that we are interested in patients in the control with hazard $\lambda_1 = 0.3$ for the purpose of sample size determination. Let us say that the investigators think that the standard deviation of both hazards is 0.2, which expresses a fair amount of uncertainty. This implies that $\beta_1 = \beta_2 = 12.5$. These, along with our usual cutoff of $l_0 = ln(1000)$, gives the following approximate range estimate for sample size:

$$n_B = (131, 136).$$

We give an interval estimate to take into account the variability of the Monte Carlo method. The above estimate is based on $50,000$ repetitions in our Monte Carlo simulation. If it is also assumed that $\alpha = 0.05$, $\beta = 0.1$, and that the test is two-tailed, then the resulting frequentist sample size, using equation (2.2) is:

$$n_f = 156.$$

The point being that, once again, sample size is reduced by using our methods. Our assumptions regarding the prior distributions can be easily elicited

from the investigators, but clearly the resulting sample size will depend on the hyper-parameters.

Dependence of Sample Size on Parameters

From figure 2.7, it can be seen that the sample size for exponential survival analysis with administrative censoring only is quite dependent on the hyper-parameters. Small values of β_1 and β_2 correspond to relatively large variances for our hazard ratio prior beliefs, but also comparatively large values of our expected prior belief, hence smaller sample sizes are needed to reject the null. Large values of β_1 and β_2 correspond to small variances for our hazard ratio priors, and also small expected values, meaning that we are really dogmatic about our beliefs, hence a large sample size is be needed to reject the null. What is clear is that this is a situation in which we would especially insist on publishing the values of the hyper-parameters, and the reasons behind selecting them, so that anyone can see and dispute as they see fit.

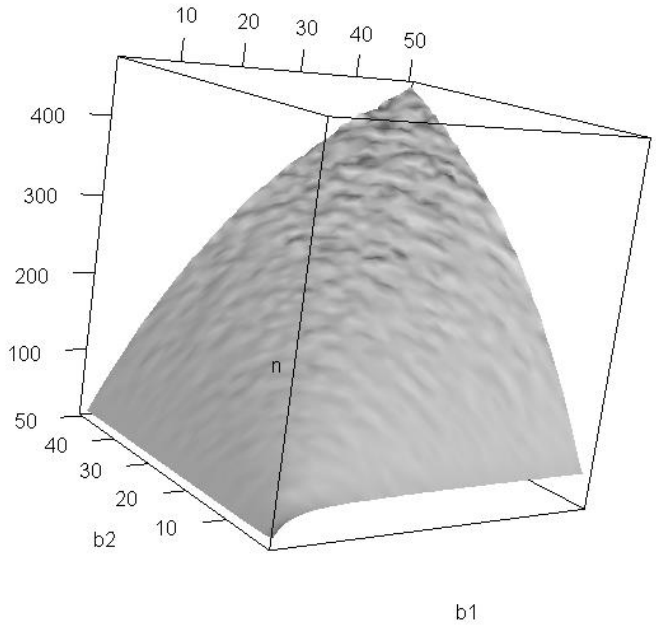

Figure 2.7: Dependence of sample size on the hyper-parameters β_1 and β_2

Chapter 3

BLINDED SAMPLE SIZE RE-ESTIMATION

Abstract

In this chapter, we investigate blinded sample size re-estimation/adjustment (SSR) for testing the hypothesis of equal survival functions for randomized, parallel group, double-blind clinical trials. There is currently a dearth of methods for the aforementioned problem. This is probably because blinded re-estimation methods usually focus on re-estimating the nuisance parameter, like the variance parameter for testing normal means, in the sample size formula. For survival data, either the variance parameter is assumed to be known, as is the case in no censoring, or it is a function of the treatment effect, as is the case in administrative censoring only. Attempts to re-estimate the treatment effect in the sample size formula have not been fruitful.

We introduce a Bayesian method for blinded SSR for survival data, using the prior predictive distribution of the observations to assign each observation to either group in a blinded fashion, and then using a Bayesian decision theoretic approach to re-calculate sample size based on the available data. Parenthetically, it should be noted that it is not important what method is used to recalculate the sample size; the pivotal issue here is the blinded estimation of the interim test statistic. This approach has the distinct advantage of taking uncertainty of the parameter values into account. The method can be used for any type of sampling distribution assumption, and for both precise and composite hypothesis testing. However, we illustrate our method by assuming constant hazard functions for both group, and apply the method under the assumption of no censoring only. We show that, under certain circumstances, our blinded estimate of the treatment effect performs as well as the unblinded estimate. This blinded estimate of the treatment effect can be used directly in the frequentist sample size formula, or indirectly in our procedure to update the sample size.

3.1 Introduction

Sample size re-estimation/adjustment (SSR) is an extension of the fixed sample size design approach. In SSR, the fixed sample size that was determined before the experimental data are collected is re-calculated based on an interim analysis of the parameters that are used to determine the fixed sam-

ple size. The central idea behind SSR is to ensure that the study is not under-powered due to miss-specification of the parameters in the sample size formula. For example, testing the difference in means in two normal populations that are assumed to have the same variance, the sample size formula is: $n = \frac{\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2 \sigma^2}{\theta^2}$, where θ is the alternative hypothesised difference in means. The variance parameter, σ, is considered a nuisance parameter in this case. If σ was under-estimated, then the sample size would be too small to detect the hypothesised signal, in other words the experiment would be under-powered. Both σ and θ are a priori point estimates based on knowledge of similar trials and expert experience. There is no way to account for the uncertainty in these estimates using the frequentist framework, but it is recognized that they may be incorrect, and so we re-estimate σ based on information collected in the ongoing trial. The reason for this is that the ongoing trial should give the most appropriate information about the parameter. The parameter of interest, θ in the normal means case, is not re-estimated, since that is usually set by the alternative hypothesis.

In the case of hypothesis testing of equality of survival functions, the parameters are the log hazard ratio and the probability of observing an event, and since they are functionally related then estimating one implies estimating the other. It is not surprising then that very little work has been done in using SSR for survival data from the frequentist perspective. The usual adaptive design for survival analysis is a group sequential design. Thus we are trying to extend the purview of the usual SSR to include time-to-event analysis.

SSR is done to determine whether or not it is necessary to increase the sample size. The procedure does not involve an interim hypothesis test. While blinded SSR does employ an interim analysis, which makes the design adaptive, it is not a member of the sequential design approaches. Sequential designs, as defined in Shih (2006), entails one or more interim analyses to determine whether or not the study should continue, hence the final sample size is not fixed. The decision in sequential designs is based on an hypothesis test, and so there will be as many hypothesis tests as there are interim analyses. If only one interim analysis is planned, then the design is usually called a two stage adaptive design. In this particular type of sequential design, before data are collected the timing of the interim analysis is determined, and the final number of observations will be based on the result of that analysis. Whereas for SSR, the timing of the interim analysis is not predetermined; the interim analysis does not involve any hypothesis testing, and the purpose of the interim analysis is to recalculate the fixed sample size based on updated information from the data, not to determine the efficacy or futility of the experimental design. Thus, other than the fact that both procedures employ an interim analysis, SSR and two-stage adaptive designs have nothing else in common, and SSR should not be confounded with sequential designs.

SSR can be done in a blinded or un-blinded manner. Blinded SSR involves re-calculating estimates of the parameters needed for the sample size without knowing group assignment, using statistics calculated from the sample based on the interim observations. All that is available are the data values, for

example event and censored times, but there is no knowledge of whether or not that observation belonged to a person that received treatment. According to ICH E9 guideline documentation, section 4.4, sample size adjustment is allowed as long as the blind is preserved, and any effect on Type I error is taken into account. All studies on blinded SSR done so far has not found any significant change in the nominal Type I error. The main theoretical and practical issue that has nettled frequentist approaches to the problem is how to calculate statistics without knowledge of group assignment, especially if these statistics estimate the parameter of interest. This is necessary, since these blinded statistics will be substituted for the parameter values used in the sample size calculation.

There has been very little work done on blinded SSR for survival data, due to the aforementioned fact that there is no nuisance parameter to re-estimate if no censoring is assumed, and the nuisance parameter is a function of the treatment parameter otherwise. Furthermore, almost no work has been done on re-estimating the parameter of interest, and what has been done is unsuccessful. It may be the case that, unlike the variance parameter, the treatment parameter is considered so important to the problem that it would not do to re-estimate it. We do not hold to that belief, however, as all parameters are uncertain, and if there is a way to decrease, or at least take into account, this uncertainty then it should be done. There has not been, to the best of our knowledge, any investigation done on blinded SSR for hypothesis testing of survival data from a fully Bayesian perspective. We

will not speculate as to why that is, but we believe that it is an important problem, and that the Bayesian approach may provide some theoretical and practical results that may not be easily accomplished within the frequentist framework. We believe that the Bayesian approach is capable of estimating the treatment effect, and any other parameter, in a more coherent way than previous attempts. This is in contradistinction to the frequentist approach which focuses on recalculating a nuisance parameter that is independent of the treatment parameter. The main contribution here is in the method of estimating any and all parameters in a blinded fashion based on the interim data.

The rest of this chapter will be organised as follows. We will do a literature review on the existing methods for blinded SSR. We then introduce our new method, and study its performance via simulation studies.

3.2 Literature Review

Adaptive design procedures like two-stage adaptive design, or other such group sequential design procedures, are usually viewed as more appropriate for survival data. This is due to the fact that survival data usually involves mortality, and these group sequential designs allow for early stopping of the study due to futility. These types of adaptive designs are very flexible, but due to the effect of Type I error inflation and power considerations, they are still viewed suspiciously by regulatory bodies, see Berry (2004, 2006)

and Shih (2006). However, a cursory glance at the current literature reveals that the scientific community is headed in the direction of adaptive designs for clinical trials, especially those designs done from a Bayesian framework. At any rate, at least for now, blinded SSR procedures play an important role in adaptive designs of clinical trials. The investigations on blinded SSR have been mostly performed from a frequentist perspective. Most studies on blinded SSR can be bifurcated into two groups: those that are concerned with binary endpoints, and those that involve continuous endpoints. There have been very few studies on survival endpoints in particular. We will focus our review on continuous endpoints, with the usual assumption that the endpoint of interest is normally distributed.

Gould and Shih (1992) investigated blinded SSR for normally distributed data, by estimating the common within-group variance, σ^2, in equation (2.1). They first show theoretically that the effect of blinded SSR on the Type I error rate for normally distributed end point is in most cases negligible. They then present two methods for estimating the unknown variance in a blinded fashion. The first is a so-called simple adjustment based on the pooled estimate of the sample variance, using the hypothesised treatment difference. To wit, $(n-1)s^2 = (n-2)\hat{\sigma}^2 + np(1-p)(\bar{x}_1 - \bar{x}_2)^2$, where s is overall sample variance, $\hat{\sigma}$ is the variance estimate that is required, and p is unknown proportion of subjects in group 1. The idea is, if the alternative hypothesised difference Δ is true, and if n is large enough, then $\hat{\sigma}^2 \approx \frac{n-1}{n-2}\left(s^2 - \frac{\Delta^2}{4}\right)$, provided that $\Theta = 0.5$, where Θ is known overall proportion of subjects assigned to group

1. Note that this does not take into account the uncertainty in the assumed treatment difference given by the alternative hypothesis. This approach also assumes that the interim sample size is large enough so that the hypothesised treatment difference will be a good estimator of the sample treatment difference, and that the proportion of observations collected at the interim stage from either group will be close to 0.5. We leave the reader to judge the validity of these assumptions. However, the result is simple and elegant. This would not work for survival data, however, since either it is the case that the variance is known, or the variance is unknown and depends on the treatment difference, thus requiring an estimation of the parameter of interest. The second suggested method is to use the Expectation-Maximization (EM) algorithm, assuming that the group assignments are missing at random. They concluded that the algorithm did a good job of estimating the unknown variance, but at the same time did not do well at estimating the treatment difference. This, in itself, is suggestive. The point is moot, however, since the EM method was later discredited by Friede and Kieser (2002).

Friede and Kieser (2002) show that the EM algorithm used by Gould and Shih (1992) to re-estimate the variance parameter in a blinded fashion is inappropriate for three main reasons. Firstly, estimates from the EM procedure are too dependent on the initialization parameters of the EM algorithm. Secondly, the stopping rule used does not guarantee convergence of the algorithm. These two problems are not surprising, since the procedure is not based on any theoretical consideration, but based on intuition and simulation

studies. Lastly, the EM procedure depends on simple randomization pattern for the clinical trial. It is not obvious how it can be extended for other types of randomization schemes, like block randomization. The severity of the first two issues cannot be ignored, and hence the EM procedure for re-estimating the variance parameter is no longer considered as a valid approach.

Kieser and Friede (2003) suggest their own sample size re-estimation procedures for normally distributed end-points via re-estimation of the variance parameter. The first procedure is to simply use the pooled data, ignore group assignment, and calculate the grouped sample variance and use that as an estimate for the variance parameter. This procedure has a recognized positive bias, which they suggest can be reduced by using a procedure similar to the simple adjustment procedure suggested by Gould and Shih (1992). They show, numerically, that the effect on the Type I error rate is minimal for either procedure, at least in the situation where the t-test is used. Once again, these procedures are not useful for survival data, since either the variance parameter is known, or else it depends on the treatment parameter.

Xing and Ganju (2005) use enrollment order of subjects and randomization block size to re-estimate the variance parameter for continuous endpoints. Their method can be used for normal and non-normal endpoints. They calculate an estimate of the variance parameter that is based on the known block endpoint sums. They show that their estimator is unbiased analytically. Also, for certain block sizes, it is close to the unblinded estimator of the variance parameter on average; the variability of the estimator can

only be studied numerically. This method, once again, is unsuitable for survival endpoints as the variance parameter that they study does not depend on the parameter of interest. Also noteworthy is the fact that, for all the frequentist methods reviewed thus far, there is no theoretical justification for the development of the estimators. All these procedures depend on instincts, and all are focused on finding an unbiased estimator of the variance parameter. Placing such importance on unbiasedness is a hallmark feature of most frequentist statistical methods, even though unbiasedness is not an optimal feature associated with any known loss function. That is to say, if loss is taken into account in deriving an optimal procedure, unbiasedness is not the result of minimizing any known loss function. It has more of an intuitive appeal than a clear statistical underpinning.

In contrast to the types of experimental situations described hitherto, namely the t-test, survival data analysis does not have a nuisance parameter that is independent of the parameter of interest. Todd, Valdes-Marquez, and West (2012) try to re-estimate the sample size in survival data analysis by re-estimating the survival rates, given by the alternative hypothesis, in a blinded manner. There seems to be a few inconsistencies in their approach. Namely, they give the following formula for number of events: $D = \frac{4\left(Z_{1-\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{\theta^2}$, where $\theta = ln\left(\frac{\lambda_2}{\lambda_1}\right)$. This equation assumes testing the equality of two survival functions for independent groups, one control, the other experimental treatment, that each subject is randomly assigned to a group with probability 0.5, and furthermore that the hazard rates are constant. It also crucially

assumes that the number of events are equal to the number of observations, that is, all subjects are followed until an event is observed, see Lachin (2009, section 9.5). They then go on to give two equations for converting number of events, D, to number of patients needed, n, based on the aforementioned equation. However, that is inconsistent, since the equation for D assumes that the number of events is equal to the number of observations. At any rate, they look at two distinct situations: the fixed follow-up situation, where each patient is followed up for the same length of time, regardless of when they were signed up, and the variable follow-up situation, in which the trial is designed to last for a fixed time, and each patient's follow-up depends on when they were recruited. The sample size needed for the former case is given by: $n = \dfrac{D}{1 - \frac{1}{2}\left[S_{CR}(F) + S_{ER}(F)\right]}$, where F is the length of time that each patient will be followed up for, and S_{CR} and S_{ER} are the reference survival functions for the control group and experimental group respectively. For the latter case, what they term variable follow-up, the sample size equation is slightly more complicated, but also uses the reference survival functions for both groups given by the alternative hypothesis. They claim that these reference survival functions are nuisance parameters, but they obviously depend on the reference hazard functions, so re-estimating the reference survival functions without re-estimating the log hazard ratio is inconsistent. Their idea is to calculate an average survival rate based on the data collected using the Kaplan-Meier method on the pooled data, then use that to update the old survival rate average which is based solely on the alternative hypothesis.

Why the use of the average, and what are the theoretical justifications of such a technique are not disclosed to us. Once again, we see intuitive, ad hoc procedures, likely proposed due to simulation results, without any attempt at theoretical justification. Note also that if these reference survival functions can be recalculated from the data, then one is justified in wondering why not update the alternative hypothesis as well. If it is recognized that these initial estimates of the survival function may not be accurate, then it follows that the alternative hypothesised value of the log hazard ratio used in the sample size calculation must also not be accurate.

Recognizing that, for survival analysis at least, due to the dependence of all the parameters involved, it is inconsistent to try and re-estimate survival probabilities without re-estimating the treatment difference, Xie, Quan, and Zhang (2012), try to remedy this situation. They at least acknowledge that updating the treatment effect parameter – the alternative hypothesised log hazard ratio – is the most consistent approach for survival data, since the so-called nuisance parameters are also a function this parameter of interest. They opine that no reliable estimate of the treatment effect can be obtained in a blinded manner using only the survival data of the ongoing trial, and thus try an alternate approach of using surrogate end-points as a way of re-estimating the treatment effect in a blinded manner. The gist is as follows. Assume that previous trials, similar to the one being conducted, have demonstrated a treatment effect on a surrogate endpoint. The surrogate endpoint should be such that the information is available to all involved in the clinical

trial, like weight or blood pressure. They then propose three methods for re-estimating the treatment effect in the current trial: a classification approach, and two EM algorithm approaches. It is not made clear what the guidelines are for choosing the surrogate endpoint. For the classification approach in particular, they further assume that the surrogate endpoint has higher values for the treatment group, on average. The classification approach then is to classify observations if their surrogate endpoint values are larger than the median value of the pooled data. They give no theoretical justification for such a classification rule. At any rate, the conclusions they come to are that the EM algorithm approaches are too dependent on initialization values, and the use of surrogate endpoints also prove unreliable. All three methods result in huge, unquantifiable variation in the blinded treatment effect estimate, and are thus unsuitable for blinded SSR.

Hartley (2012) offers a "mostly" Bayesian solution to re-estimating the treatment effect in a trial with normally distributed endpoints. His method pertains mainly to effect size, that is the ratio of treatment effect and standard deviation, for comparing difference between two normal means. His approach would not be suitable for survival data, as he explicitly assumes independence of the treatment effect parameter and the variance parameter. He also assumes that the interim sample size at which the re-estimation is done is very large, an unlikely situation for survival data. His approach is simply to average the conditional power used for traditional sample size calculations over all possible values of the parameters, given the pooled variance

estimate from the interim data.

As far as we know, no one hitherto has investigated a blinded SSR method based on the Bayesian framework for survival data.

3.3 New Method

Our method can be limned as follows. Assume that the initial sample size, n, is determined by some method. We of course suggest our method, given in detail in chapter 2, equation (2.9), which involves minimization of expected loss. At any rate, the method of initial sample size determination is not important. Given the interim data on $n_* < n$ number of subjects, a blinded estimate of the treatment parameter is determined. This is done by using the Bayesian classification method of assigning observations to either group based on their prior predictive probabilities: if a certain observation has a higher prior predictive probability of being in group 1, say, than in group 2, then the observation is assigned as a group 1 measurement. Once we have an estimate of the treatment parameter, we will recalculate the sample size needed, N, based on minimization of expected loss. However, we will use the updated distribution of the parameter of interest in order to calculate sample size, this update due to the observations collected at the interim. Another important difference from the method used in chapter 2 is that the cutoff for our decision rule will be based on controlling Type I error, rather than thinking about the average likelihood of the data under the alternative verses the null. This

is done to ensure that Type I error is controlled, as is required by regulatory guidelines. The new sample size is $n_b = n_* + N$. If $n < n_b < 1.5n$, then it will be used as the new sample size estimate, and the appropriate number of additional subjects will be recruited and randomized for the trial. Notice that the factor 1.5 in the above is rather arbitrary. Here we assume that the investigators would set some upper limit on what the maximum sample size can be. Whatever that limit is in actuality can be substituted for 1.5. Also, note that the important thing is the blinded estimate of the test statistic; this can be used to update the sample size formula for any method, including the frequentist sample size formula, equation (2.3).

Now for the details. Let us assume that we are interested in testing the equality of two survival functions for a double blind, parallel arm clinical trials study, where one group, group 1 say, is considered control, and the other the treatment group. For the purpose of sample size determination, let us assume that the hazard rate for both groups are constant. Let us further assume that there are no censored times due to loss to follow up, that is, all patients are followed until an event is observed. The null hypothesis is $H_0 : \theta = 0$ verses $H_0 : \theta \neq 0$, where $\theta = ln\left(\frac{\lambda_1}{\lambda_2}\right)$, and λ_i represents the hazard rate of group $i, i = 1, 2$. The initial sample size estimate, n, has already been determined by some method. We further assume that there is a fixed recruitment period, t, and recruitment/entry times are uniformly distributed over that period. The sample size re-estimation is scheduled at the recruitment of the ${n_*}^{\text{th}}$ patient, perhaps when 75% of patients have been recruited, or some-

thing like that. Since some observations will be administratively censored at the time of the re-estimation, say t_*, then the appropriate likelihood at the interim analysis is:

$$\begin{aligned} L &= \left({\lambda_1}^{\delta_i} e^{-\lambda_1 t_i}\right)^{n_1} \left({\lambda_2}^{\delta_k} e^{-\lambda_2 t_k}\right)^{n_2} \\ &= {\lambda_1}^{d_1} e^{-\lambda_1 \left(\sum_{i=1}^{d_1} t_i + \sum_{j=1}^{n_1-d_1} t_j \right)} {\lambda_2}^{d_2} e^{-\lambda_2 \left(\sum_{k=1}^{d_2} t_k + \sum_{l=1}^{n_2-d_2} t_l \right)} \end{aligned}$$

where $\delta_i = \begin{cases} 1, & \text{if event occurs} \\ 0, & \text{otherwise} \end{cases}$, n_1 and n_2 are the number of patients in each group ($n_1 + n_2 = n_*$), and d_1 and d_2 are the number of observed events in each group. It can easily be deduced that the maximum likelihood estimates (MLE) of the interim hazard rates are: $\hat{\lambda}_{1*} = \dfrac{d_1}{\sum t_i + \sum t_j}$ and $\hat{\lambda}_{2*} = \dfrac{d_2}{\sum t_k + \sum t_l}$. The interim MLE for θ is:

$$\hat{\theta}_* = \ln \frac{\hat{\lambda}_{1*}}{\hat{\lambda}_{2*}}. \tag{3.1}$$

This is how we will summarize our interim data, and we will use this statistic to update our prior belief about θ. The asymptotic distribution of the interim test statistic is:

$$\hat{\theta}_* \mid \theta \sim N\left(\theta, \frac{1}{n_*}\left[\frac{1}{pE(\delta \mid \lambda_1)} + \frac{1}{(1-p)E(\delta \mid \lambda_2)}\right]\right).$$

It can be shown that $E(\delta \mid \lambda) = 1 - \dfrac{1}{\lambda t_*} + \dfrac{e^{-\lambda t_*}}{\lambda t_*}$, provided recruitment is uniform and censoring is due to no further observations after t_* (administra-

tive censoring) only. Thus, assuming $p = 0.5$ – this represents the proportion assigned to control – then: $\hat{\theta}_* \mid \theta \sim N\left(\theta, \frac{2\sigma_*^2}{n_*}\right)$, where

$$\sigma_*^2 = \frac{1}{1 - \frac{1}{\lambda_1 t_*} + \frac{e^{-\lambda_1 t_*}}{\lambda_1 t_*}} + \frac{1}{1 - \frac{1}{\lambda_2 t_*} + \frac{e^{-\lambda_2 t_*}}{\lambda_2 t_*}}. \tag{3.2}$$

Our conjugate prior is: $\theta \sim N\left(\mu, \frac{\sigma^2}{n_0}\right)$, where $\sigma^2 = \frac{1}{p(1-p)}$ for exponential survival data with no censoring, and n_0 is an "effective" sample size for my prior. The resulting posterior is:

$$\theta \mid \hat{\theta}_* \sim N\left(\frac{n_*\sigma^2\hat{\theta}_* + 2n_0\sigma_*^2\mu}{n_*\sigma^2 + 2n_0\sigma_*^2}, \frac{2\sigma_*^2\sigma^2}{n_*\sigma^2 + 2n_0\sigma_*^2}\right).$$

Now, using the Bayesian decision theoretic procedure presented in chapter 2, equation (2.7), we determine the optimal additional sample size needed, N, given the data, by taking expectation of the intrinsic loss function, equation (2.8). This implies

$$E_{\theta|\hat{\theta}_*}(l) = \frac{N}{2\sigma^2} E_{\theta|\hat{\theta}_*}\left(\theta^2\right) = \frac{N}{2\sigma^2}\left[\frac{2\sigma^2\sigma_*^2}{n_*\sigma^2 + 2n_0\sigma_*^2} + \left(\frac{n_*\sigma^2\hat{\theta}_* + 2n_0\sigma_*^2\mu}{n_*\sigma^2 + 2n_0\sigma_*^2}\right)^2\right] > l_0$$

$$\Rightarrow N > \frac{l_0}{\frac{\sigma_*^2}{n_*\sigma^2 + 2n_0\sigma_*^2} + \frac{1}{2\sigma^2}\left(\frac{n_*\sigma^2\hat{\theta}_* + 2n_0\sigma_*^2\mu}{n_*\sigma^2 + 2n_0\sigma_*^2}\right)^2}$$

Note that the above development assumes that group membership is known, thus our estimate for $\hat{\theta}_*$ is unblinded. Also, the parameter l_0 used in the decision rule and sample size determination is determined in a different way

than in chapter 2. Whereas we suggested that $l_0 = \ln 1000$, implying that we will reject the null hypothesis when the minimum average likelihood ratio against the null is expected to be larger than 1000, we now suggest that l_0 be chosen such that the Type I error is controlled for a sample of size $1.5 * n$. This ensures that our blinded procedure will always control the Type I error, as is required by regulatory bodies. We can do this by using the power function for our decision rule, (2.10). The next step is now to determine $\hat{\theta}_*$ in a blinded manner, that is, without knowledge of group assignment for each observation. The method we propose is to use prior information on λ_1 and λ_2 to classify each observation. Let us expound this idea.

We assume that each time has an event distribution $f(t_i \mid \lambda_j) = \lambda_j e^{-\lambda_j t_i}$, where $j = 1, \text{or } 2$. Using the Bayesian classification rule, Johnson and Wichern (1992, chapter 11.2), and assume the cost of misclassification is the same, and that the prior probabilities of being in either group are equal, then the optimal classification decision is to assign the observation to group 1 if $f(t_i \mid \lambda_1) > f(t_i \mid \lambda_2)$. This is optimal in that this rule minimizes the total probability of misclassification. Another way of thinking about it is that this rule minimizes my posterior expected loss, if I use a discrete loss function for misclassification, and equal prior probabilities. Since we do not know the values of $\lambda_1$and λ_2, we use the prior predictive distributions to determine the optimal classification rule. Assume a conjugate gamma prior for each parameter as follows: $\pi(\lambda \mid \alpha, \beta) = \dfrac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$. The gamma family allows for a wide range of models for prior belief. The prior predictive is then:

$f(t_i \mid \alpha, \beta) = \dfrac{\alpha\beta^{\alpha}}{(t_i+\beta)^{\alpha+1}}$. Thus the classification rule is given by assigning an event observation to group 1 if:

$$\frac{\alpha_1\beta_1^{\alpha_1}}{(t_i+\beta_1)^{\alpha_1+1}} > \frac{\alpha_2\beta_2^{\alpha_2}}{(t_i+\beta_2)^{\alpha_2+1}}. \tag{3.3}$$

Similarly, the prior predictive for the censored observations, equivalently those subjects who have survived up to the time of interim re-estimation, is given by: $1 - F(t_i \mid \alpha, \beta) = \dfrac{\beta^{\alpha}}{(t_i+\beta)^{\alpha}}$. From this we find the optimal classification of survival times is to assign a censored observation to group 1 if:

$$\frac{\beta_1^{\alpha_1}}{(t_i+\beta_1)^{\alpha_1}} > \frac{\beta_2^{\alpha_2}}{(t_i+\beta_2)^{\alpha_2}}. \tag{3.4}$$

Once we have assigned all observations up to time t_* using these aforementioned decision rules, we can then use equation (3.1) to determine our interim test statistic. This estimate of θ will be our blinded estimator, and as such we will label it $\hat{\theta}_{*b}$.

There are, it seems to us, three outstanding issues to discuss. Firstly, when to conduct the interim re-estimation. This is investigation dependent; however, based on simulation studies, we suggest when 75% of the observations or more have been collected. This should give enough time for useful information to be collected, while giving the investigators enough time to plan an extension of recruitment if needed. One of the main advantages of blinded sample size re-estimation is that we can choose when to do the interim calculations after we see the data. This is due to the fact that everyone

remains blind, so there is no concern about introducing bias, and the Type I error will be controlled, due to our choice of l_0 which will be set before data collection begins. Secondly, how to choose the hyper-parameters for the gamma distribution of θ may be problematic. There are two ways to go about this: showing distributions of θ for various values of $\alpha$and β, or equivalently showing the prior predictive distribution of t_i to the investigators/experts. We prefer the latter, as it seems to us that it would be easier to think of the actual observations and what their distributions may look like, than an abstraction such as λ. Either way, we believe that as long as the hyper-parameters are explicitly stated, then once again it is up to the reader of the report to determine whether or not these values are reasonable, and furthermore, a range of sample sizes can be estimated for different possible hyper-parameters. The third concern is the utility of sample size re-estimation. The hazard rates have to be fairly high, so that enough event times are observed before the interim analysis, or else there can not be a useful sample size re-estimation. Therefore the procedure we describe above is only pertinent to some survival analyses, specifically those with a high enough hazard rate.

We will now investigate the statistical properties of our blinded estimator, and our procedure, via numerical analysis.

Table 3.1: Type I error when interim at 50^{th} percentile

	$\lambda = 0.0008$	$\lambda = 0.008$	$\lambda = 0.01$
Type I error	0	0.034	0.058
mean(n_b)	103	110	108
sd(n_b)	11	13	13
mean($\hat{\theta}_{*b}$)	1.946	-0.245	-0.804
sd($\hat{\theta}_{*b}$)	0.686	0.720	0.772
mean($\hat{\theta}_{*ub}$)	-0.003	0.004	0.000
sd($\hat{\theta}_{*ub}$)	0.824	0.376	0.357

Table 3.2: Type I error when interim at 75^{th} percentile

	$\lambda = 0.0008$	$\lambda = 0.008$	$\lambda = 0.01$
Type I error	0	0.038	0.051
mean(n_b)	95	109	109
sd(n_b)	NA	12	13
mean($\hat{\theta}_{*b}$)	2.537	-0.050	-0.610
sd($\hat{\theta}_{*b}$)	0.619	0.527	0.596
mean($\hat{\theta}_{*ub}$)	-0.038	-.006	-0.006
sd($\hat{\theta}_{*ub}$)	0.609	0.279	0.274

Table 3.3: Type I error when interim at 90^{th} percentile

	$\lambda = 0.0008$	$\lambda = 0.008$	$\lambda = 0.01$
Type I error	0	0.025	0.057
mean(n_b)	93	114	108
sd(n_b)	10	11	12
mean($\hat{\theta}_{*b}$)	2.489	-0.024	-0.574
sd($\hat{\theta}_{*b}$)	0.504	0.446	0.523
mean($\hat{\theta}_{*ub}$)	-0.007	0.013	0.010
sd($\hat{\theta}_{*ub}$)	0.510	0.251	0.241

Table 3.4: Comparison of blinded and unblinded estimates for $\lambda_1 = 0.008$ when interim at 50^{th} percentile

	$\theta = 0.693$	$\theta = 0.5108$	$\theta = 0.3567$
mean($\hat{\theta}_{*b}$)	0.455	0.304	0.170
sd($\hat{\theta}_{*b}$)	0.649	0.647	0.680
mean($\hat{\theta}_{*ub}$)	0.714	0.522	0.352
sd($\hat{\theta}_{*ub}$)	0.434	0.410	0.407
mean(n_b)	106	109	108
sd(n_b)	12	13	13

Table 3.5: Comparison of blinded and unblinded estimates for $\lambda_1 = 0.008$ when interim at 75^{th} percentile

	$\theta = 0.693$	$\theta = 0.5108$	$\theta = 0.3567$
mean($\hat{\theta}_{*b}$)	0.712	0.543	0.361
sd($\hat{\theta}_{*b}$)	0.414	0.429	0.468
mean($\hat{\theta}_{*ub}$)	0.704	0.529	0.359
sd($\hat{\theta}_{*ub}$)	0.296	0.286	0.308
mean(n_b)	104	107	109
sd(n_b)	12	12	13

Table 3.6: Comparison of blinded and unblinded estimates for $\lambda_1 = 0.008$ when interim at 90^{th} percentile

	$\theta = 0.693$	$\theta = 0.5108$	$\theta = 0.3567$
mean($\hat{\theta}_{*b}$)	0.760	0.558	0.405
sd($\hat{\theta}_{*b}$)	0.342	0.356	0.372
mean($\hat{\theta}_{*ub}$)	0.703	0.516	0.364
sd($\hat{\theta}_{*ub}$)	0.267	0.260	0.251
mean(n_b)	107	110	113
sd(n_b)	12	12	11

3.4 Application of Result

3.4.1 Simulation Study

For the following, we assume a 2 year recruitment period, where subjects are uniformly entered into a double-blind, parallel arm study. Each subject is randomly assigned to control (group 1) or treatment (group 2) with probability 0.5. We assume constant hazard rates, so that the survival times are exponentially distributed for both groups. We further assume that all subjects will be followed until an event is observed, that is, no censoring. We desire a two-tailed test of $\theta = \ln \frac{\lambda_1}{\lambda_2} = 0$. Assume $\alpha = 0.05$, and a power of 0.9 to detect $\theta = \ln 2 \Rightarrow \lambda_2 = 0.5\lambda_1$, a 50% reduction in risk. For the calculation of σ_*^2 in equation (3.2), we use the expected values of the priors for λ_1 and λ_2. Let us assume that the initial sample size estimate, n, is calculated based on the frequentist procedure, which would yield a sample size of $n = 88$. We use the following settings throughout the simulations: $n_0 = 10$, $\mu = \ln 2$, $\sigma^2 = 4$ (see chapter 2.4.2 for justification of these parameters), and $l_0 = 2.273364$, which corresponds to the cutoff for our decision procedure that guarantees control of Type I error for up to a sample size of $1.5n = 132$, provided $\theta = 0$. Recall that we will only increase the sample size if $n < n + N \equiv n_b < 1.5n$. We assume that the investigators believe that, on average, the hazard rate for the control group is $\lambda_1 = 3$ per year ≈ 0.008 per day. Finally, the hyperparameters for the gamma functions are $\alpha_1 = 6$, $\alpha_2 = 2$, $\beta_1 = 300$, and $\beta_2 = 60$. These values are chosen based on the prior predictive distributions

of the observations.

The objectives of the simulation study are as follows:

- show that Type I error is controlled for blinded SSR for various values of θ, and
- compare the blinded and unblinded estimates of θ.

These comparisons will be done when 50%, 75%, and 90% of the subjects are recruited. Using the Law of Large Numbers, the probabilities will be estimated by averaging over 1000 simulations.

3.4.2 Discussion of Results

The results underscore the difficulties in doing a blinded interim analysis, especially when the treatment effect is being estimated. However, some useful information can be gleaned without knowledge of group assignment. From tables 3.1, 3.2, and 3.3 it can be seen that Type I error is generally controlled for the blinded SSR, and this is so for a wide range of λ values. The level of control is best when $\lambda = 0.008$; this is due to the fact that the prior specification, which determines the classification of event times and censored times, is fine-tuned for this situation. If the prior specification over-estimates λ then the procedure over-controls the Type I error, and if it under-estimates the true value of λ then it barely controls the Type I error. Due to the fact that a Bayesian procedure was used for the blinded SSR, we cannot expect

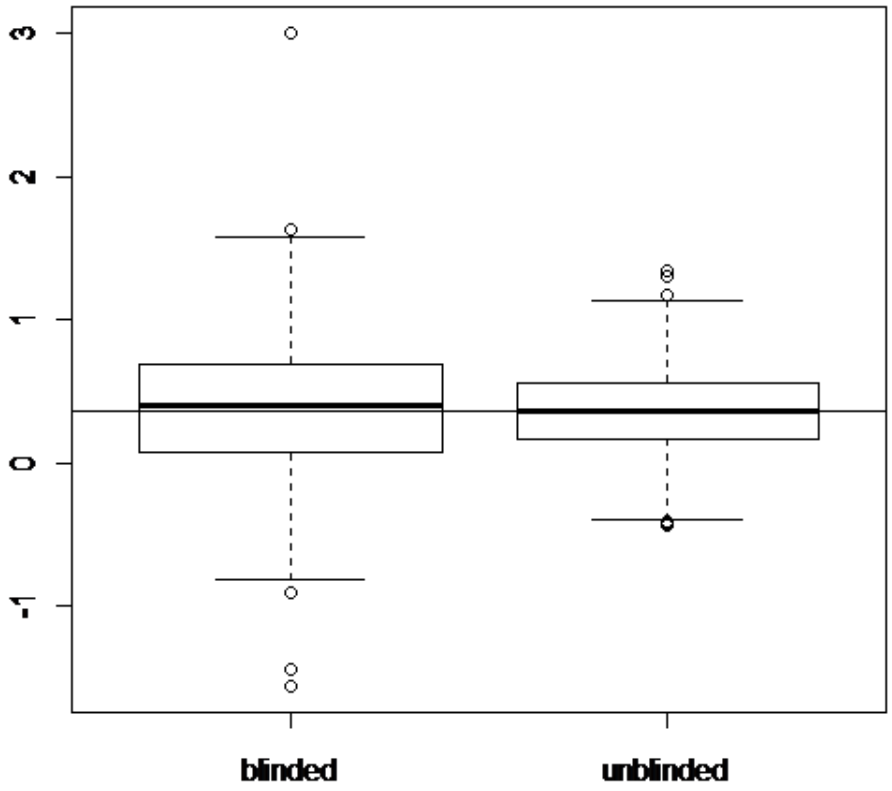

Figure 3.1: Comparison of $\hat{\theta}_{*b}$ and $\hat{\theta}_{*ub}$ at interim $75th$ percentile, $\theta = \frac{1}{0.7} = 0.3567$.

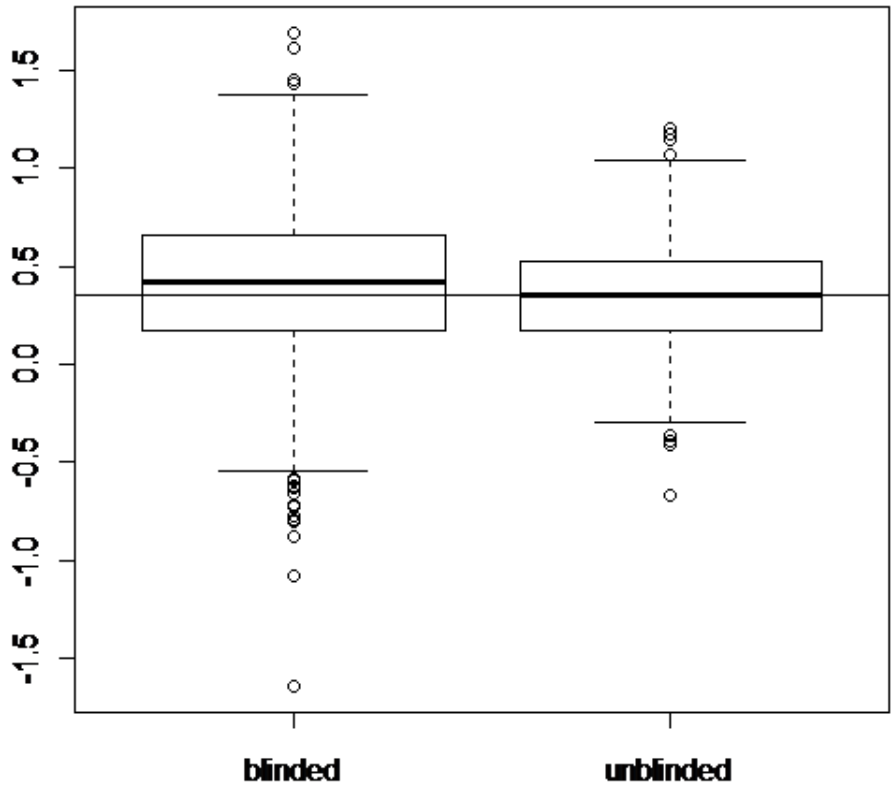

Figure 3.2: Comparison of $\hat{\theta}_{*b}$ and $\hat{\theta}_{*ub}$ at interim $90th$ percentile, $\theta = \frac{1}{0.7} = 0.3567$.

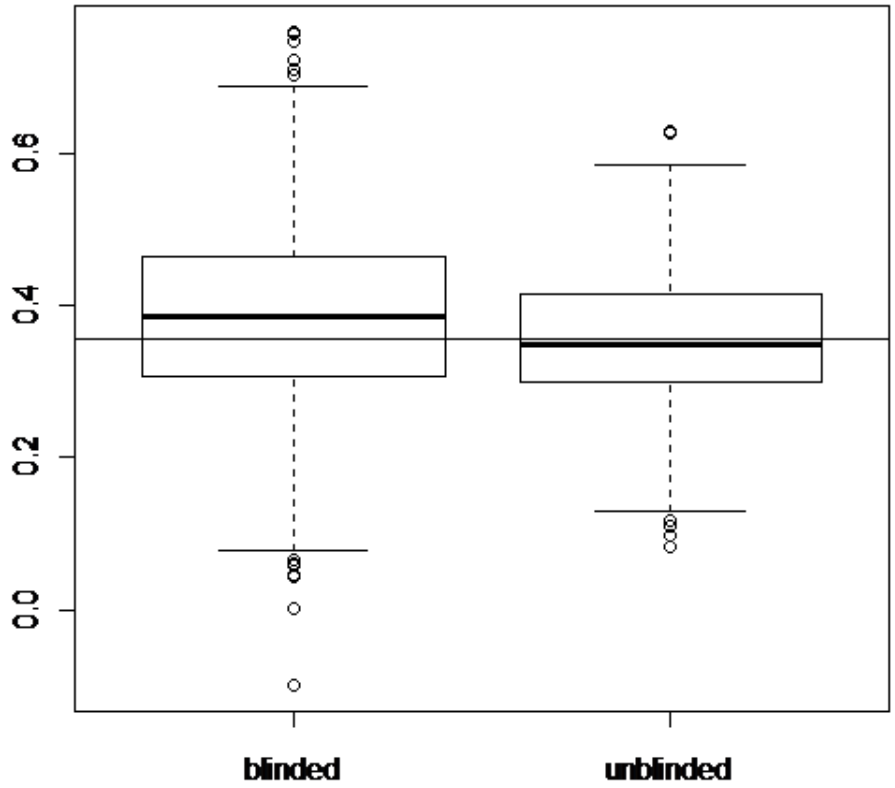

Figure 3.3: Comparison of $\hat{\theta}_{*b}$ and $\hat{\theta}_{*ub}$ at interim $75th$ percentile, $\theta = \frac{1}{0.7} = 0.3567$, $n = 1000$.

exact control of the Type I error, and we are quite satisfied with the level of control given by the procedure.

Quite predictably, the blinded estimate of the parameter of interest at the interim, $\hat{\theta}_{*b}$, improves as the timing of the interim estimation gets later in the recruitment time frame. This is to be expected, since the later we conduct the interim analysis, the more information will be available to us for classification of the times. It seems that midway through the accrual phase is too early for a reliable estimate, and so we would advise to wait until at least three quarters of the subjects have been recruited before conduction of a blinded SSR. Another factor that affects the estimate is the true hazard rate. If it is very low, then it is likely that few or no event times will be observed, and so we will have a poor estimate. This can be seen in the aforementioned three tables, for $\lambda = .0008$. Thus we would advise that a blinded SSR be done only in the case where the hazard rate is expected to be fairly high. The accuracy and precision of the estimate can also be affected by the rate of accrual. If patients are recruited too quickly, then there will not be enough event times for a useful estimation. The interplay of hazard rate, length of recruitment time, and accrual rate is quite complicated, and all factors would affect the blinded estimate of θ. This requires further study.

Tables 3.4, 3.5, and 3.6 illustrate the effect of miss-specification of θ on the blinded estimate. These are all done under the assumption that $\lambda_1 = 0.008$ per day. It was initially thought that the treatment would result in a 50% reduction in risk; that is the shown in the second column. The third and fourth

columns show what the estimates would be if the reduction in risk is in fact 40% and 30% respectively. Once again, the blinded estimates improve as we time the interim analysis later in the recruitment period. What is remarkable is that the blinded estimator does a good job, on average, of tracking the true value of θ, and is pretty close to the performance of the unblinded estimator, the unblinded estimator being the bench mark of performance here. What is unsatisfactory, however, is the variability of the blinded estimator, which is further displayed in figures 3.1 and 3.2. These graphs show the spread of the blinded estimate in comparison to the unblinded estimate for miss-specification of $\theta = \ln \frac{1}{0.7}$, that is, when the reduction in risk is only 30% instead of the expected 50%. The blinded estimate when the interim is conducted after the 90^{th} percentile of recruitment is better than that of the 75^{th} percentile, but is still more variable than the unblinded estimate. For example, the IQR (interquartile range) for the blinded estimate when the actual reduction in risk is only 30%, and the interim analysis is timed at the 90^{th} percentile, is 36% larger than the unblinded estimate. This variability is to be expected, however, since we are trying to estimate a parameter with only partial information from the sample. However, figure 3.3 shows that this variation decreases with increase in initial sample size and ipso facto increase in n_*. For $n = 1000$, the standard deviation of the blinded and unblinded estimates are 0.121 and 0.087 respectively; their IQR's are 0.156 and 0.115 respectively. Thus as n_* increases, the variability, as measured by the standard deviation and IQR, decreases for the blinded estimate,$\hat{\theta}_{*b}$. However, while it

is obvious, based on the distribution of the unblinded estimator given above, that the unblinded estimator $\hat{\theta}_*$ is a consistent estimator, we do not expect that $\hat{\theta}_{*b}$ will also be consistent. In fact, it is impossible, since there will always be a misclassification error associated with the blinded estimator that will certainly make it a biased and inconsistent estimator. We only claim here that the variation will decrease with increased initial sample size. Thus for very large studies the blinded estimator exhibits improved performance in terms of precision. Therefore, we have increased confidence that our blinded estimator can perform reasonably well under certain circumstances, and thus provide useful information for SSR.

Chapter 4

CONCLUSION

Fixed sample size determination has traditionally been the purview of the frequentist approach, since from the Bayesian perspective it makes sense to have a more adaptive approach for experimental design. This is because the Bayesian approach is not constrained by Type I error and power. However, we recognize that the frequentist approach to experimental design is very important and useful. It answers a different question: What should we expect if we repeat this experiment many times? Whereas the Bayesian approach is concerned with the particular experiment at hand. We are not arguing that one is superior to the other; that would be nonsensical. What we try to demonstrate is that, even in the frequentist setting, the Bayesian framework may be informative. In particular, if prior opinion is strong, and backed by much experience, then it makes sense to incorporate this into the statistical planning and analysis for a given experiment. This is likely to be

the case in medical, biological, and pharmaceutical fields. Furthermore, from a theoretical standpoint, when testing precise null hypotheses, the so-called nuisance parameter has been a problem for the frequentist approach. This may be resolved if we assign a probability distribution to our uncertainty about the parameters in a given problem. What we are striving for is to provide the practitioner with an alternative tool that is complementary to the usual approaches. It is up to the individual to determine, based on the problem, which approach to use.

In this thesis we investigated two related problems. The first, investigated in chapter 2, was how to determine a fixed sample size to test a precise null hypothesis. The second, investigated in chapter 3, was how to adjust the fixed sample size, based on blinded results of the ongoing trial. For fixed sample size determination, the Bayesian approach we espouse has two advantages: 1) we can incorporate any relevant prior knowledge into our experimental design, and 2) we can take into account the uncertainty in the parameters we use for sample size estimation, rather than using fixed point estimates. As can be seen, for example, in figure 2.6, incorporating prior information can lead to a uniformly more powerful procedure than the frequentist procedure for the same given sample size. It seems to us then that the Bayesian approach is, in some sense, a generalization of the frequentist approach. We have, in our main result in chapter 2, equation (2.7), a very simple and elegant sample size formula that rivals that of the frequentist approach for testing hypotheses in terms of ease of use. It is generated almost completely

from the sampling distribution, and the only additional assumptions necessary – the values of the hyper-parameters and the value of the cutoff – are a small price to pay for the generality it offers. The sampling distribution can be as complicated as the situation necessitates. The calculations necessary are also rather straightforward. This is in contradistinction to other Bayesian methods. The beauty of the result lies in the use of an intrinsic loss function. While there are several candidates for intrinsic loss functions that can be used, the one introduced by Bernardo and Rueda (2002) is particularly attractive due to its ease of interpretation. Note, as well, that we could include in our loss function the cost of recruiting and conducting the experiment on a new patient without any difficulty, similar to Bernardo (1997) and Lindley (1997). In reality, this factor may be more important than anything else in determining sample size. However, we think that it is hard to define such a cost in a dimensionless way. More work needs to be done on that as well.

For the particular example of exponential survival analysis with no right censoring, we subsume the frequentist sample size formula, equation (2.3) as a special case of our sample size formula, equation (2.9). It was shown that the frequentist sample size formula can be derived from the Bayesian sample size formula by assuming that our prior is known with infinite precision. While analytical results are hard to come by for more complex cases, we suspect that this will always be true. The two situations that we studied – exponential survival with no censoring, and with administrative censoring only – show that our sample size formula will generally give lower values than

the frequentist approach. However, because their criteria are so different, it is truly hard to compare the procedures in a meaningful way. We maintain that the testing criterion introduced by Bernardo and Rueda (2002), which we use, is more intuitive, and allow for easier exposition to practitioners. It is at least a viable alternative to the size and power criteria used by frequentist.

The ineluctable and conspicuous issue with the Bayesian approach is always the choice of prior distribution to characterize our uncertainty. We will not be cavalier about this serious issue; it is important, and we believe that much more work needs to be done on prior specification. Ideally, we would like our assumptions that go into the selection of a prior to be independently verifiable, similar in spirit to the program launched by Jaynes (2003), for example. The choice of prior should always be context dependent, but by employing the conjugate prior we hope to give a certain amount of objectivity in that the distribution family of our prior is based on the sampling distribution. Simultaneously, we balance this with relevant prior information in the selection of the hyper-parameters. Our analysis shows that sample size is sensitive to prior specification, and as such the choice of hyper-parameters, and the reasons behind these choices, should be clearly stated in any report for all to dispute.

The controversial topic of prior specification comes up again in our blinded sample size re-estimation (SSR) procedure. In estimating the test statistic – the log hazard ratio – at a given interim point in an ongoing trial in a blinded manner, we needed to assign observations to either group without knowledge

of group information. Xie et. al. (2012) suggest a method that seems to us to be quite difficult to implement in practice, as it requires knowledge from previous trials of a covariate whose group information is available in spite of the blind, and that is highly correlated with survival time. We suggest that we use the partial information available to us in the ongoing trial, along with the prior information the investigators have about the hazard rates of the control group. We then classify the observations based on the prior predictive distribution of the event and survival times. The accuracy and precision of our blinded estimator relies heavily on the prior information, and if it is poorly captured, or too general, then it does not make sense to even attempt a blinded SSR. What we did show is that, under certain circumstances, the blinded interim estimate of θ was very close to the unblinded estimate, and this is a promising result. The blinded estimator kept good track of the true value of θ, giving hope that the sample size can be re-estimated in a reliable fashion. Of course, for both the original and the subsequent sample size determination, we would suggest the Bayesian approach. However, we recognize that our approach may not be appropriate in some cases, and as such the frequentist approach would be used. Either way, a reliable blinded estimate of θ is necessary, and we feel that we have made some strides towards that goal.

We plan on investigating the roles of accrual rate and randomized right censoring on blinded SSR. We would also like to research blinded SSR when a Weibull distribution for survival times is assumed, or with the inclusion

of covariates via a proportional hazards model. However, ultimately our intentions are to move away from blinded SSR and extend our research to truly adaptive designs, that is, methods that allow for multiple testing during a trial, the ability to switch patients from control to treatment or vice versa, and early stopping due to efficacy or futility. This is the quintessential goal, in our minds, for statistical planning of a clinical trial: to use all the information available for that trial, including prior knowledge, in a meaningful, coherent way.

BIBLIOGRAPHY

[1] Adcock, C. J. (1997). Sample size determination: a review. The statistician, 261-283.

[2] Berger, J. O. (1985). Statistical decision theory and Bayesian inference. Springer-Verlag (New York).

[3] Bernardo, J. M. (1997). Statistical inference as a decision problem: the choice of sample size. The Statistician, 151-153.

[4] Bernardo, J. M. (2011). Integrated objective Bayesian estimation and hypothesis testing. Bayesian statistics, 9, 1-68.

[5] Bernardo, J. M., and Rueda, R. (2002). Bayesian hypothesis testing: A reference approach. International Statistical Review, 70(3), 351-372.

[6] Bernardo, J. M., and Smith, A. F. M. (1994). Bayesian Theory Wiley. New York.

[7] Berry, D. A. (2004). Bayesian statistics and the efficiency and ethics of clinical trials. Statistical Science, 175-187.

[8] Berry, D. A. (2006). Bayesian clinical trials. Nature reviews Drug discovery,5(1), 27-36.

[9] Cantor, A. B. (1992). Sample size calculations for the log rank test: a Gompertz model approach. Journal of clinical epidemiology, 45(10), 1131-1136.

[10] De Santis, F. (2004). Statistical evidence and sample size determination for Bayesian hypothesis testing. Journal of statistical planning and inference, 124(1), 121-144.

[11] Desu, M.M. and Raghavarao, D. (1990) Sample Size Methodology. New York: Academic Press.

[12] Friede, T., and Kieser, M. (2002). On the inappropriateness of an EM algorithm based procedure for blinded sample size re-estimation. Statistics in Medicine, 21(2), 165-176.

[13] Gould, A., and Shih, W. J. (1992). Sample size re-estimation without unblinding for normally distributed outcomes with unknown variance. Communications in Statistics-Theory and Methods, 21(10), 2833-2853.

[14] Hartley, A. M. (2012). Adaptive blinded sample size adjustment for comparing two normal means—a mostly Bayesian approach. Pharmaceutical statistics, 11(3), 230-240.

[15] ICH Steering Committee. (1998). Statistical Principles for Clinical Trials (E9). In International Conference on Harmonisation of Technical

Requirements for Registration of Pharmaceuticals for Human Use. International Conference on Harmonisation.

[16] Jaynes, E. T. (2003). Probability theory: the logic of science. Cambridge university press.

[17] Jeffreys, H. (1998). The theory of probability. OUP Oxford.

[18] Johnson, R. A., and Wichern, D. W. (1992). Applied multivariate statistical analysis (Vol. 4). Englewood Cliffs, NJ: Prentice hall.

[19] Julious, S. A. (2004). Sample sizes for clinical trials with normal data. Statistics in medicine, 23(12), 1921-1986.

[20] Kieser, M., and Friede, T. (2003). Simple procedures for blinded sample size adjustment that do not affect the type I error rate. Statistics in medicine, 22(23), 3571-3581.

[21] Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. Controlled clinical trials, 2(2), 93-113.

[22] Lachin, J. M. (2009). Biostatistical methods: the assessment of relative risks (Vol. 509). John Wiley and Sons.

[23] Lachin, J. M., and Foulkes, M. A. (1986). Evaluation of sample size and power for analyses of survival with allowance for nonuniform patient entry, losses to follow-up, noncompliance, and stratification. Biometrics, 507-519.

[24] Lakatos, E. (1988). Sample sizes based on the log-rank statistic in complex clinical trials. Biometrics, 229-241. Lee, S. J., and Zelen, M. (2000). Clinical trials and sample size considerations: another perspective. Statistical Science, 15(2), 95-110.

[25] Lewis, E. J., Hunsicker, L. G., Lan, S. P., Rohde, R. D., and Lachin, J. M. (1992). A controlled trial of plasmapheresis therapy in severe lupus nephritis.New England Journal of Medicine, 326(21), 1373-1379.

[26] Lindley, D. V. (1957). A statistical paradox. Biometrika, 187-192.

[27] Lindley, D. V. (1997). The choice of sample size. The Statistician, 129-138. Matusita, K. (1967). On the notion of affinity of several distributions and some of its applications. Annals of the Institute of Statistical Mathematics, 19(1), 181-192.

[28] Pezeshk, H. (2003). Bayesian techniques for sample size determination in clinical trials: a short review. Statistical Methods in Medical Research, 12(6), 489-504.

[29] Reyes, E. M., and Ghosh, S. K. (2013). Bayesian Average Error-Based Approach to Sample Size Calculations for Hypothesis Testing. Journal of biopharmaceutical statistics, 23(3), 569-588.

[30] Robert, C. P. (1996). Intrinsic losses. Theory and decision, 40(2), 191-214.

[31] Robert, C. (2007). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science and Business Media.

[32] Rubin, D. B., and Stern, H. S. (1998). Sample size determination using posterior predictive distributions. Sankhya: The Indian Journal of Statistics, Series B, 161-175.

[33] Sahu, S. K., and Smith, T. M. F. (2006). A Bayesian method of sample size determination with practical applications. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(2), 235-253.

[34] Samaniego, F. J. (2010). A comparison of the Bayesian and frequentist approaches to estimation. Springer Science and Business Media.

[35] Schlaifer, R., and Raiffa, H. (1961). Applied statistical decision theory.

[36] Shih, W. J. (2006). Group sequential, sample size re-estimation and two-stage adaptive designs in clinical trials: a comparison. Statistics in Medicine, 25(6), 933-941.

[37] Snapinn, S. M., Jiang, Q. I., and Iglewicz, B. (2005). Illustrating the impact of a time-varying covariate with an extended Kaplan-Meier estimator. The American Statistician, 59(4).

[38] Spiegelhalter, D. J., and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion.Statistics in medicine, 5(1), 1-13.

[39] Todd, S., Valdés-Márquez, E., and West, J. (2012). A practical comparison of blinded methods for sample size reviews in survival data clinical trials. Pharmaceutical statistics, 11(2), 141-148.

[40] US Food and Drug Administration. (2004). Challenges and opportunities report–March 2004. Retrieved from FDA: http://www.fda.gov/ScienceResearch/SpecialTopics/CriticalPathInitiative/CriticalPathOppo

[41] Walker, S. G. (2003). How many samples?: a Bayesian nonparametric approach. Journal of the Royal Statistical Society: Series D (The Statistician), 52(4), 475-482.

[42] Wang, F., and Gelfand, A. E. (2002). A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models. Statistical Science, 193-208.

[43] Weiss, R. (1997). Bayesian sample size calculations for hypothesis testing. The Statistician, 185-191.

[44] Xie, J., Quan, H., and Zhang, J. (2012). Blinded assessment of treatment effects for survival endpoint in an ongoing trial. Pharmaceutical statistics, 11(3), 204-213.

[45] Xing, B., and Ganju, J. (2005). A method to estimate the variance of an endpoint from an on-going blinded trial. Statistics in medicine, 24(12), 1807-1814.