Robust Inference for Randomized Play the Winner Design

by

An-Lin Cheng

(Under the direction of Anand Vidyashankar)

Abstract

This dissertation develops statistical methodologies for analysis of design parameters and response variables in a randomized play the winner design (RPWD). RPWD is an allocation procedure that changes the allocation probability during the experiment so as to allocate more subjects to a better performing treatment. This causes an imbalance in the number of subjects allocated to various treatment arms. Standard statistical methodologies that do not account for the adaptiveness in the allocation lead to procedures that are inefficient and inaccurate.

In this thesis, several novel methodologies using bootstrap and Hellinger distance are developed to mitigate the problems of adaptive allocation. These methodologies have potential for application in other ares of science and technology.

Index words: randomized play the winner design, allocation proportion, success probability, Hellinger distance, bootstrap, influence function, efficiency, phase transition, response adaptive design, multitype branching process, embedding theorem.

Robust Inference for Randomized Play the Winner Design

by

An-Lin Cheng

B.S. National Cheng-Kung University, Taiwan, 1999

M.S. University of Georgia, USA, 2002

A Dissertation Submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2004

Robust Inference for Randomized Play the Winner Design

by

An-Lin Cheng

Approved:

Major Professor:	Prof. Anand Vidyashankar
Committee:	Prof. Jaxk Reeves
	Prof. Paul Schliekelman
	Prof. Tharuvai N. Sriram
	Prof. XiangRong Yin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2004

Acknowledgements

I am deeply thankful to my major professor, Dr. Anand Vidyashankar. From inspire the research topics to complete this dissertation, he is such a great teacher. For his discipline, constant encouragement and inspiring guidance enrich the days of doing research.

I also want to express my sincere gratitude to the other members of my Ph.D. committee, Prof. Jaxk Reeves, Prof. Paul Schliekelman, Prof. Tharuvai N. Sriram and Prof. XiangRong Yin for their valuable comments and suggestions throughout the process of completing this dissertation and the early courses at the department of statistics in the University of Georgia.

A special thanks to my friends in Athens, Chia-Pei, David, Masa and Vivian for their support and encouragement.

Finally, I would like to dedicate this dissertation to my mon and dad for their love and unconditional support at any time; I would also like to dedicate this dissertation to my brother and sisters.

Table of Contents

Chapter 1

Randomization in Clinical Trials

1.1 Introduction

A clinical trial is an experiment that is designed to compare the efficacy and value of medical treatments (or interventions) against a control in human beings. Successful implementation of a clinical trial depends on the following: an efficient randomization scheme, evaluation of confounding factors, identification of factors that affect the primary outcome, modeling of the dependence between the primary outcome and the covariates and ethical issues concerning the study. It has been argued that a properly planned and executed clinical trial is a powerful experimental technique for assessing the effectiveness of an intervention.

Drug development and marketing of a drug within the U.S. are controlled by the federal government rules and regulations; polices and procedures are stipulated by the Food and Drug Administration (FDA) so as to ensure maximum safety for the consumers. A typical drug development process involves four phases and are popularly termed as *Phases* of a clinical trial. *Phase I* trials are intended to identify the maximum tolerable dose (MTD) and typically consist of very few subjects. Several popular designs, like up and down designs (Durham and Flournoy (1995)) are used to evaluate the MTD. The *Phase I* trials are preceeded by trials on animals to study if the drug has the potential to cure a related symptom. *Phase II* trials are intended to evaluate the toxicity and the efficacy of the drug and are typically carried out on healthy volunteers. This however is not the case in cancer trials where new drugs are evaluated on subjects to decrease the high death rates on existing therapies. *Phase III* trials are typical large scale clinical trials and are usually multi-center (several centers of study for the same intervention), double-blind (both the patient and the

investigator are blinded to the randomization), and placebo controlled (one of the treatment arms is a placebo and acts as a control) fully randomized experiments. *Phase IV* trials deal with long-term evaluation of the treatment effects on study subjects. This dissertation will focus on *Phase III* clinical trials only.

The following are the oft-touted advantages of a fully randomized experiment:

(1) *Freedom from selection bias.* If the experimenter knows for certain that the next assignment will be a treatment, or a control, he(she) may consciously(or unconsciously) bias the experiment by such decisions as to who is(or is not) a suitable experimental subject, in which category the subject belongs, etc. Such a bias is called the *selection bias*, a term coined by Blackwell and Hodges (1957). It is obvious that complete randomization eliminates selection bias, while a systematic design maximizes it. Blackwell and Hodges advocate using complete randomization, but continuing the experiment until a certain minimum number of subjects are allocated to both treatments and controls. In practice this can be very difficult to achieve, particularly in an experiment such as the Hodgkin's disease(where there are many categories of subjects). Selection bias is not a factor in blind experiments where admission to the study and related decisions are made by someone ignorant of the past assignment of treatments and controls.

(2) *Freedom from accidental bias.* In several experimental studies, systematic factors occur without the prior knowledge of the experimenter. Typical examples include time trends, sex-linked differences and differing experimental conditions. Complete randomization tends to balance out such factors and thus protect the significance level of the usual hypothesis tests. The systematic designs are quite vulnerable to accidental bias.

(3) *Randomization as a basis for inference.* Probability statements, such as the observed significance level of the experiment, can be based entirely on the randomness induced by the complete randomization between treatments and controls. This eliminates the need for probability assumptions on the responses of the individual experimental units and guarantees the validity of the stated significance level.

We now move to discuss some of the disadvantages of fully randomized experiments. These are:

(1) Randomization can cause serious imbalances in the number of subjects allocated to different treatment groups, especially if the sample size is small. Even though the *Phase III* trials are large, the number of subjects involved in a particular center is moderate in size and allocating completely at random can cause serious imbalance.

(2) The second disadvantage of a fully randomized clinical trial is that in certain situations it could be unethical to allocate subjects to treatments without taking into account the efficacy of an intervention thus far.

The alternatives that account for these disadvantages are the so called adaptive trials or restricted randomization designs. We now describe several randomization schemes that are typically used in the context of a *Phase III* clinical trials.

1.2 **Permuted Block Design**

Permuted Block Design (PBD) , first introduced by Hill (1951), describes a randomization scheme that attempts to balance the number of subjects allocated to treatments groups. For the sake of illustration let A and B denote two treatments and assume that we have a block of size $2m$. In each block, m subjects are randomly assigned to treatment A and treatment B, respectively. In particular if $m = 2$, one of the six possible combinations

$$AABB,\ ABAB,\ BAAB,\ BABA,\ BBAA,\ ABBA$$

is chosen randomly and assigned to the subjects. If n is the number of subjects in the study, then the maximum imbalance in the number (of subjects) allocated to the two treatment is m. m is usually refereed to as the maximum tolerable imbalance (MTI) parameter. Indeed, investigators using PBD can ensure balanced allocation after the $(2m)^{th}$ assignment.

PBD induces a selection bias, since the treatment assignment of the very last subject in any block is known if the length of block is known. To avoid this bias, one can choose the

block size to be random, so that the investigator will not know when the block will end and fail to guess the treatment assignment.

If the condition of the subjects do not change across time, this design can ensure more equally sized groups. But from the analysis point of view, the data resulting from this design are not independent and identically distributed (i.i.d.) data. Ignoring the design in the analysis will lead to decrease in power or exaggerated treatment effects (Matt and McHugh (1978) and Wei *et. al.*(1990))

1.3 **Stratified Randomization**

As mentioned in the introduction randomization helps alleviate the accidental bias. If the accidental bias is caused due to covariate imbalance, sometimes it is more convenient to address it at the design stage. Complete randomization performed in different strata is called stratified randomization. For stratified randomization, subjects are grouped according to different prognostic factors (for instance, sex and age). Within each stratum, independent randomization scheme is applied separately. Permuted randomization can also be applied within each strata. As an example, consider a clinical trial studying the effects of an intervention on smoking. The effect of intervention could depend on the smoking history (current smoker, ex-smoker and never smoked) and the gender (female and male). In such a trial, there are total of six strata and each subject is allocated to one of the treatment groups within each stratum separately. Permuted block randomization can be employed within each stratum.

In a multicenter clinical trial, one does not typically randomize the subjects to treatments across all clinics. Indeed, separate permuted block designs are carried out within each center separately. This allows the experimenter to evaluate the effects of unknown covariates like geography, quality and type of care in a center, and the availability of the clinical expertise. The last phenomenon is especially useful when studying a new surgical method.

In some situations, stratification can be used posthumously during the analysis stages. It has been shown that stratified analysis improves the precision of the estimators, since it improves the efficiency of the estimators and testing procedures. This phenomenon is more evident in smaller trials than the larger trials.

An increase in the number of strata due to the inclusion of several covariates leads to the requirement of a large sample size within each stratum. In a smaller study, this is infeasible and the disadvantages of including several covariates for stratified randomization are described in Pocock and Simon (1975).

1.4 Adaptive Randomization

In this section we describe designs that change the probability of allocation to treatments during the trial. These allocation probabilities could depend on a set of covariates, and the allocation history thus far, but **do not** depend on the responses of the allocated subjects. We begin with the biased coin design due to Efron (1971).

Biased Coin Design

The biased coin design is a randomization scheme for achieving balance in allocation between treatment groups. Assume that we have two treatments (1 and 2). Let $D_i = N_1(i) - N_2(i)$, $1 \leq i \leq N$ where $N_1(i)$ represents the number of subjects assigned to treatment 1 after the ith subject's assignment. Efron's biased coin design can be described as follows: Let T_{i+1} denote the treatment indicator for the $(i+1)^{\text{st}}$ subject; then

$$P(T_{i+1} = 1 | D_i) = \begin{cases} 1/2 & \text{if } D_{i-1} = 0 \\ p & \text{if } D_{i-1} < 0 \\ 1-p & \text{if } D_{i-1} > 0 \end{cases}$$

where $p \in (0.5, 1]$. Note that $D_i < 0$ implies that more number of subjects have been allocated to treatment 2 while if $D_i > 0$ implies that more number of subjects have been allocated to treatment 1. Note that the design is adaptive since the probability of allocation

to a treatment changes at different stages of the trial, unlike the completely randomized case, where it remain fixed throughout the entire trial.

Let $D_i^* = |D_i|$, then $D_i^*,\ i \geq 1$ forms a Markov chain on the state space $\{0,\ 1,\ 2, \cdots\}$. The one-step transition probability matrix given by

$$P = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdots \\ p & 0 & 1-p & 0 & 0 & \cdots \\ 0 & p & 0 & 1-p & 0 & \cdots \\ 0 & 0 & p & 0 & 1-p & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

It is possible to introduce a tolerance level into the Efron's biased coin design and make the biasing probability depend on the stage of the clinical trial and prognostic factors. Even though such a generalized design has not been studied in the literature, biased coin design that account for covariates have been studied by Pocock and Simon (1975). Further properties of the biased coin design have been studied in Chen (1999), Heckman (1985) and Smith (1984); comparsions of the biased coin design with other designs have been investigated in Chen (2000).

Wei's Urn Design

Wei's urn design, as the name suggests, also changes the allocation probability during the course of the trial but uses an urn model instead of a biased coin. This design can be described as follows: Assume that we have two treatments 1 and 2, start with an urn containing α balls of types 1 and α balls of types 2 representing the two treatments. When a subject is available for randomization, a ball is drawn at random and returned to the urn. Assignment is made according to the type of the drawn ball. If it is a type 1 ball, then β ($\beta \geq 1$) type 2 balls are added to the urn. If it is a type 2 ball, then β type 1 balls will be added to the urn. The process is repeated until all subjects have been assigned to treatment.

It was shown by Wei (1978) that this design tends to keep the probability of allocation to two treatments close to equal. An analogue of Wei's design accounting for covariates has been developed by Wei (1979).

1.5 Response Adaptive Designs

We now describe designs that are driven by ethical considerations and take into account responses of subjects allocated to treatments.

Randomized Play the Winner Design

Randomized play the winner design (RPWD) is a method of assigning subjects to intervention in clinical trials that takes into account the responses of the patients. The method inspired by the Play the Winner (PW) rule of Zelen (1969), accounts for a delayed response. The PW rule of Zelen can be described as follows: A success on a particular treatment generates a future trial on the same treatment with a new subject. A failure on a treatment generates a future trial on the alternate treatment with a new subject. If the response is unavailable, the subject is allocated using equal probability amongst all treatments. This design assigns subjects to a better performing treatment.

The main difficulty of this design is that when the subject accrual is rapid and the response is delayed, the design allocates subjects to treatments with equal probability, therefore making the adaptation irrelevant. To overcome these difficulties, Wei and Durham (1978) introduced the RPWD. This design can be described as follows: Suppose we want to assign subjects to two different treatments (say 1 and 2). We would start this procedure with an urn containing $N_1(0)$ balls of type 1 (representing treatment 1) and $N_2(0)$ balls of type 2 (representing treatment 2), corresponding to treatments 1 and 2 respectively. Once the subject is available for treatment assignment, a ball is drawn at random from the urn and returned to the urn. The subject is assigned to a treatment according to the type of ball. When the subject's response is available the urn is updated as follows: if the response is a

success on treatment 1 or a failure on treatment 2, then α type 1 balls are added to the urn; however if the response is failure on treatment 1 or a success on treatment 2, then α type 2 balls are added to the urn. The process is repeated until all subjects have been assigned to a treatment. This design tends to favor a better performing treatment by increasing the probability of allocation towards a better performing treatment thus far.

We will propose one extension of the RPWD to the r treatments case. For a clinical trial with r treatments, we would start with an urn containing $N_1(0), \cdots, N_r(0)$ balls of r different types representing r different treatments. Once a subject is available for randomization, draw a ball at random from the urn and replace it. Assign the subject to a treatment according to the type of the ball. If the treatment turns out to be a success, then r balls with the same type are added to the urn. Otherwise, r balls for each type other than the treatment are added to the urn. The process is repeated until all the subjects have been assigned a treatment.

Generalized Pólay's Urn Design

There are several other response adaptive designs that have been studied in the literature. As we mentioned earlier, the RPWD was a precursor to the various other response adaptive designs. Wei (1979) presented the following generalized Pólay urn design(GPUD). Begin with an urn contains $N_1(0), \cdots, N_r(0)$ balls of r different types representing r different treatments. Once a subject is available for treatment assignment, a ball is drawn at random from the urn and then returned to the urn. Assign the subject to a treatment according to the type of ball. When the response of a subject to treatment i turns out to be a success, then α balls of the same type are added to the urn. Otherwise, β balls for each type other than the treatment are added to the urn. Repeat this process until all the subjects have received treatment assignment. This treatment assignment rule is denoted by GPUD($\mathbf{N}, \alpha, \beta$), where $\mathbf{N}$= $(N_1(0), \cdots, N_r(0))'$. (RPWD for two treatments described above corresponds to the case $r = 2$, $\beta = \alpha$).

More recently, Li, Durham and Flournoy (1997) studied a variant of the GPUD for $r(\geq 2)$ treatments called the RPUD. In their design, while $\alpha > 0,\ \beta \equiv 0$. This design can be described as follows: start with an urn contains $N_1(0), \cdots, N_r(0)$ balls of r different types representing r different treatments. Once a subject is available for treatment, a ball is drawn at random and the subject is assigned to the treatment indicated by the ball. If the treatment is success, α balls ($\alpha > 0$) of the same type are added to the urn; else the ball is returned to the urn. The main difference between a GPUD and a RPUD is that the composition of balls in the urn will not change when there is a failure in the response. In other words, RPUD favors treatments that lead to success.

Birth and Death Urn Model

It has been noted in several simulation studies (for instance Flournoy and Rosengerber (1992) P.65 and P.23) that RPWD and GPUD introduce much variablity in the design making it harder to implement in areas such as toxicology and cancer. To overcome these difficulties, Ivanova, Rosenberger, Durham and Flournoy (2000) presented a design called the Birth and Death Urn Design (BDUD). In this design the urn starts with r types of balls representing r different treatments. When a subject is available for a treatment assignment, a ball is drawn at random. The subject is assigned to a treatment according to the type of the ball. If the outcome is a failure, the ball is not replaced. If the treatment is a success, two balls of the same type are added to the urn. In this design, if the success probability of certain treatment, say 1, is less than 0.5, then the type 1 ball will become extinct at some point and no more subjects will be assigned to treatment 1. To avoid this scenario, the authors developed a supplemented birth and death urn design called the Birth and Death Urn Design with Immigration(BDUDI).

In this design, the urn starts with r types of balls representing r different treatments and ar immigration balls (a ≥ 0). The parameter, a, is called the rate of immigration. When a subject is available for treatment assignment, a ball is drawn at random and replaced. If it is an immigration ball, choose one ball from the r different types with equal probability and

add to the urn. If it is a treatment ball, then assign the subject to a treatment according to the type of the ball. If the outcome is a failure, one ball representing the treatment assigned is removed from the urn. If the treatment is a success, one ball of the same type is added to the urn.

Ivanova *et. al* have shown that the equal allocation is not always the optimal solution to problems in a clinical trial. Simulations (in their work) suggest that when there are three treatments and success probability of these treatments have the following relationship, *viz.*, $p_1 > p_2 = p_3$, the urn design is more powerful for testing the equality of the p_i's than any other fixed allocation design.

Drop the Loser Rule

More recently, Ivanova (2003) developed a new rule called Drop the Loser Rule for assigning subjects to treatments. This is also an urn model. It starts with $r + 1$ types of balls. The first r types represent r treatments, while the balls of type $r + 1$ are called immigration balls. When a subject arrives for treatment, a ball is drawn at random. If it is an immigration ball, then no treatment assignment is made and the ball is returned to the urn together with r additional balls, one corresponding to each of the treatments. If a treatment ball is drawn (*i.e.* one of the first r types), the subject will be assigned to the corresponding treatment. If the outcome is a failure, the ball is not replaced. If the outcome is a success, the ball is replaced. Several important feature of the design have been studied by Ivanova (2003). More recently, Hu and Rosenberger (2003) show that this design has properties similar to that of PWD of Zelen but is less variable and has more adaptivness

Randomized Play the Winner design that account for covariates have been studied by Rosenberger, Vidyashankar and Agarwal (2001). The feasibility and logistics of conducting a response-adaptive, double-blind, placebo-controlled study was investigated by Eli-Lilly and company (Tamura *et. al* (1994)).

Example - Fluoxetine Trial

We now describe the clinical trial conducted by Eli-Lilly and company which motivated this dissertation. This is a multi-center clinical trial comparing fluoxetine to placebo in patients with depressive disorder. It is believed (Kupfer (1976)) that shortened rapid eye movement latency is a marker for endogenous depression. In this trial, patients were stratified into two groups: Patients with normal rapid eye movement latency(REML) and patients with shortened REML. The first six patients within each stratum were assigned by a randomized block design to either fluoxetine or placebo. The trial used two independent urns (for two different strata) to assign the patients. Both urns started with one ball for each type, representing the two treatments. Independent randomized play the winner rules were initiated with the seventh patient within each stratum. There are two primary outcomes: (1) the percentage of patients who exhibited a 50 percent or greater reduction in Hamilton Depression Scale ($HAMD_{17}$) between baseline and final active visit after a minimum of three weeks of therapy, and (2) the reduction in $HAMD_{17}$ between baseline and the final visit. Patients receiving therapy for at least 3 weeks who exhibited a 50 percent or greater reduction in $HAMD_{17}$ were defined to be responders (success in treatment). The time from baseline to final measurement was approximately 8 weeks. The time delay, along with a rapid patient arrival, did not allow an adaptive trial based on the response from final visit. Thus adaptive allocation was based on a surrogate marker to update the urn. The surrogate responder was defined as a patient exhibiting a reduction greater than 50 percent in ($HAMD_{17}$) in two consecutive visits after at least three weeks of therapy. The trial was stopped after 61 patients had responded according to the surrogate criterion. No further surrogate response was obtained for the remaining patients. There were total 89 patients in this trial.

The data related to this trial is included in the Table 1.1 where for shortened REML patients belong to strata 1, normal REML patients belongs to strata 0, treatment is denoted by 1 if the patients is treated with Fluoxetine and 0 if is placebo. There are 83 patients have final response been recorded.

1.6 **Conclusions**

In this chapter we described several basic randomization procedures that are used in a typical $phaseIII$ clinical trial. We described the RPWD and other related response adaptive designs.

In a typical clinical trial, several variables are collected; some represent the primary outcome while others represent secondary variables. For example, consider a clinical trial investigating the efficacy of drug A in lowering the cholesterol levels. In such a trial the primary variables would typically be cholesterol levels, blood pressure, and body weight. There are certain secondary variables on which the information is also collected. These could include information on the life style, diet, and genetic components. This dissertation develops robust and efficient procedures for the analysis of primary and secondary outcomes from a randomized play the winner design.

The remainder of the dissertation is structured as follows : Chapter 2 describes the basic limit theory for RPWD and develops new methodology for obtaining confidence intervals for the design parameters, Chapter 3 developes minimum Hellinger distance methodology for the analysis of outcomes from a response adaptive design, Chapter 4 is devoted to bootstrap methodology for i.i.d. and RPWD data, while Chapter 5 contains future research directions.

Table 1.1: The Fluoxetine trial data

Strata	Strata=1 Treatment=1	Strata=1 Treatment=0	Strata=0 Treatment=1	Strata=0 Treatment=0
Success Proportion	12/20	7/21	13/21	10/21
Change	-12	4	-2	-7
in	-11	2	-12	0
$HAMD_{17}$	-17	-16	-10	-3
	-5	3	-21	-9
	-7	0	-4	-20
	-8	-6	2	-3
	-20	-11	-14	-3
	-8	-21	-1	2
	-15	-3	-16	-16
	-13	-16	-15	-6
	-16	3	-22	0
	-16	-2	-6	-15
	-2	2	-12	-10
	-1	-9	-5	-13
	-6	-8	-4	-13
	-3	-3	-12	-7
	-16	-4	-14	-10
	-11	-4	-14	-17
	-16	1	-17	-15
	-21	-17	-5	-18
		-15	-23	2

References:

1. Blackwell, D. and Hodges, J.L. (1957). Design for the control of selection bias. Ann. Math. Statist., 28, 449-460.

2. Chen, Y.P. (1999). Biased coin design with imbalance tolerance. Comm. Statist. Stochastic Models., 15, 953-975.

3. Chen, Y.P. (2000). Which design is better? Ehrenfest urn versus biased coin. Adv. in Appl. Probab., 32, 738-749.

4. Durham, SD, Flournoy, N (1995). Up-and-down designs I: Stationary treatment distributions. Adaptive Designs. Lecture-Notes Monograph Series 25 139-157. Flournoy, N, Rosenberger, WF (eds.). Hayward, CA: Institute of Mathematical Statistics.

5. Efron, B(1971). Forcing a sequential experiment to be balanced. Biometrika, 58, 403-417.

6. Flournoy, N. and Rosenberger, W.F. (1992). Lecture Notes-Mongraph Series Vol. 25 Institute of Mathematical Statistics.

7. Heckman, N.E. (1985). A sequential probability ratio test using a biased coin design. Ann. Statist., 13 , 789-794.

8. Hill AB (1951). The clinical trial. British Medical Bulletin, 71, 278-282.

9. Hu, F. and Rosenberger,W.F. (2003)Optimality, variability, power: evaluating response-adapative randomization procedures for treatment comparisons. J. Amer. Statist. Assoc., 98, 671-678.

10. Ivanova, Anastasia. (2003). A play-the-winner type urn design with reduced variability. Metrika, 58, 1-13.

11. Ivanova, Anastasia, Rosenberger, William F. Durham S. and Flournoy, Nancy.(2000). A Birth and Death Urn for randomized clinical trials: asymptotic methods, Biostatistics, 62, 104-118.

12. Kupfer, D.J. (1976). REM Latency: A Psychobiological Marker for Primary Depressive Disease. Biological Psychiatry, 11,159-174.

13. Li, W, Durham, SD and Flournoy, N (1997). Randomized Polya urn designs. 1996 Proceedings of the Biometrics Section of the American Statistical Association, 166-170.

14. Matts JP and McHugh RB (1978). Analysis of accrual randomized clinical trials with balanced groups in strata. J. Chronic Dis. 31, 725-740.

15. Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. Biometrics, 31, 103-115.

16. Rosenberger, W. F., Vidyashankar, A. N., and Agarwal, D. K. (2001). Covariate-Adjusted Response-Adaptive Designs for Binary Response. Journal of Biopharmaceutical Statistics, 11, 227-236.

17. Smith, R.L.(1984) Sequential treatment allocation using biased coin designs. Journal of the Royal Statistical Society: Series B, 46, 519-543.

18. Tamura, R.N., Faries, D.E., Andersen, J.S. and Heiligenstein J.H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depression disorder. Journal of the American Statistical Association, 89, 768-776.

19. Wei, L.J. (1978). An application of an urn model to the design of sequential controlled clinical trials. Journal of the American Statistical Association, 73, 559-563.

20. Wei, L.J. (1979). The generalized Pólya's urn design for sequential medical trials. The Annals of statistics, 7, 291-296.

21. Wei, L.J. and Durham S.. (1978). The randomized play-the-winner rule in medical trials. Journal of the American Statistical Association, 73, 840-843.

22. Wei, L.J., Smythe, D.Y., Lin, D.Y. and Park, T.S. (1990). Statistical inference with data-dependent treatment allocation rules. Journal of the American Statistical Association, 85, 156-162.

23. Zelen, M. (1969). Play the Winner Rule and the Controlled Clinical Trial. American Statistical Association Journal, 64, 131-146.

Chapter 2

Confidence Intervals for the Design Parameters of RPWD.

2.1 Introduction

In Chapter 1, we introduced several randomization methods in clinical trials; the focus of this dissertation is on randomized play the winner design. Analysis of the design parameters and inference for secondary and primary variables of interest is complex. In spite of the availability of enormous computational resources, statistically sound inferential methods that work in general situations is as yet unavailable.

Frequently, experimenters are interested not only in proportions allocated to various treatments but are also interested in several qualitative and quantitative characteristics of the design, response variables and their interactions. In this chapter, we collect several known technical results and use this opportunity to develop notations and terminology that will be used in the entire thesis. We also develop new computational tools for inference concerning the design parameters.

2.2 Randomized Play the Winner Design

We begin by recalling the randomized play the winner design. The following 2-treatment design will be used throughout our theoretical descriptions in this thesis. Consider an urn containing $\mathbf{R} = (N_1(0),\ N_2(0))$ balls corresponding to two treatments, treatment 1 and treatment 2. When a subject is available for randomization, a ball is drawn at random (type i say) and returned to the urn and the subject is assigned to the treatment represented by the ball. When the response from the subject is available, the urn is updated as follows: if

response was a success then α balls of the same type are added to the urn; else α balls of the other type are added to the urn. The process is repeated. The design will be represented as RPWD(α).

We now turn to discuss various properties of the design. A natural first question concerns the proportion of subjects allocated to each of the treatments after n subjects have been randomized. We denote by $N_1(n)$ and $N_2(n)$, the number of subjects assigned to treatment 1 and treatment 2 respectively. Note that $N_1(n)+N_2(n) = n$. A key tool to study the behavior of $N_1(n)$ and $N_2(n)$ is an embedding rule introduced by Athreya and Karlin (1967). To make precise statements concerning these results, we make a detour into multitype continuous time Branching processes.

2.3 Multitype Continuous Time Branching Processes

Consider a population containing two types of particles evolving over time. Let $\mathbf{R} = (N_1(0),\ N_2(0))$ represent the number of type 1 and type 2 particles at time 0. All type i $(i = 1, 2)$ particles live an exponential length of time with parameter λ_i and reproduce according to the offspring distribution $P_i(j_1,\ j_2)$, where $P_i(j_1,\ j_2)$ represents the probability that a type i parent produces j_1 particles of type 1 and j_2 particles of type 2. Let us denote by $X = (x_1,\ x_2)$, the generic random variable representing the offspring distribution. Then

$$P_i(X = (j_1,\ j_2)) = P_i(j_1,\ j_2) \tag{2.3.1}$$

and let $m_{ij} = E_i X_j$. Let

$$M = \begin{pmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{pmatrix}. \tag{2.3.2}$$

denote the mean matrix. Let us denote by $\boldsymbol{Z}(t) = (Z_1(t),\ Z_2(t))$ the number of type 1 and type 2 particles at time t. Let $M(t) = E\boldsymbol{Z}(t)$. Note that

$$M(t) = \begin{pmatrix} m_{11}(t) & m_{12}(t) \\ m_{21}(t) & m_{22}(t) \end{pmatrix} \tag{2.3.3}$$

where $m_{ij}(t)$ represents the expected number of the type j offspring produced by a single type i parent. We will assume that there exists t_0 and $0 < t_0 < \infty$ such that

$$M(t_0) \text{ is irreducible and positively regular} \tag{2.3.4}$$

i.e. there exist t_0 such that $m_{ij}(t_0) > 0$, for any i and j. Under this assumption, Perron-Frobenius Theory of positive matrices ensures the existence of a strictly positive maximum eigenvalue $\lambda_1(t_0)$ of $M(t_0)$ such that

$$(i) \quad \text{for every other eigenvalue } \lambda(t_0),\ |\lambda(t_0)| < \lambda_1(t_0)$$

and

$$(ii) \text{ the algebraic and geometric multiplier of } \lambda_1(t_0) \text{ are unity.}$$

Furthermore,

$$\lim_{t \to 0} M(t) = I \tag{2.3.5}$$

where I is the identity matrix of order 2, and

$$M(t+u) = M(t)M(u). \tag{2.3.6}$$

These conditions ensure that $M(t)$ has the following representation, *viz*,

$$M(t) = \exp(At). \tag{2.3.7}$$

Thus, the eigenvalues of $M(t)$ can be expressed as $e^{\lambda_i t}$, where λ_i are the eigenvalues of A and $M(t)$ and $A(t)$ have the same eigenvectors. This implies that we can arrange the eigenvalues of A as

$$\lambda_1 > Re(\lambda_2)$$

and the left and right eigenvectors of λ_1 are $\boldsymbol{u}$ and $\boldsymbol{v}$ and are normalized such that

$$\boldsymbol{u}' \cdot \boldsymbol{v} = 1 \text{ and } \boldsymbol{u}' \cdot \mathbf{1} = 1 \tag{2.3.8}$$

where $\mathbf{1} = (1,\ 1)'$.

2.4 Imbedding Urn Schemes into a Continuous Time Markov Branching Processes

To describe the composition of the urn, we consider a Markov branching processes with $N_1(0) + N_2(0)$ particles and the life time of all particles are exponentially distributed with parameter 1. Group the particles into two groups. Type 1 represents branching processes initialed by $N_1(0)$ treatment 1 balls and type 2 represents branching processes initialed by $N_2(0)$ treatment 2 balls. Let $\boldsymbol{Z}(t) = (Z_1(t),\ Z_2(t))$ represent the population size at time t. Let $\{\tau_n,\ n \geq 1\}$ denote the times at which a split or a branching occurs. Then $\boldsymbol{Z}(\tau_n) = (Z_1(\tau_n),\ Z_2(\tau_n))$ represents the stochastic process of population size at time of split. The next theorem, first proved by Athreya and Karlin (1967), shows that as a stochastic process it is equivalent to the composition of the urn.

Theorem 2.4.1. The stochastic processes

$$\{(N_1(n),\ N_2(n),\ n \geq B)\} \text{ and } \{\boldsymbol{Z}(\tau_n),\ n \geq 1\}$$

are equivalent. ∎

We now describe how the imbedding helps in deriving the asymptotic urn composition. Indeed a immediate Corollary of the above Theorem 2.4.1 using the results from branching process is the following, *viz.*

Corollary 2.4.2.

$$\lim_{n\to\infty} \frac{N_i(n)}{n} = v_i \quad a.s. \tag{2.4.1}$$

where v_i is as in (2.3.8). ∎

We now return to the RPWD and express (2.4.1) in terms of the parameters of the design. To this end, let $p_1 = P\{\text{success on treatment 1} \mid \text{subject was treated with 1}\}$ and $p_2 = P\{\text{success on treatment 2}| \text{ subject was treated with 2}\}$. Let $q_1 = 1-p_1$ and $q_2 = 1-p_2$.

From the discussion above, the matrix M is given by

$$M = \begin{pmatrix} p_1 & q_2 \\ q_2 & p_2 \end{pmatrix}. \tag{2.4.2}$$

The eigenvalues of M are $\lambda_1 = 1$ and $\lambda_2 = p_1 + p_2 - 1$. The corresponding left eigenvector of λ_1 is $v = (q_1 + q_2)^{-1}(q_2,\ q_1)$. Thus, (2.4.1) reduces to:

$$\frac{N_i}{n} \to Q_i \ as\ n \to \infty, \quad a.s. \tag{2.4.3}$$

where $Q_i = (q_{3-i})(q_1 + q_2)^{-1}$. In other words, the asymptotic urn composition is given by $(q_1+q_2)^{-1}(q_2,\ q_1)$. Note that if $q_1 = q_2$ the asymptotic proportion is $(\frac{1}{2},\ \frac{1}{2})$ which is identical to complete randomization. If $q_1 > q_2$ *i.e.* the probability of success on treatment 2 is greater than that of treatment 1, then more subjects will be allocated to treatment 1.

We next focus on the second order results concerning the urn composition. The results for Markov branching process (Athreya and Karlin (1968)) show that they crucially depend on the difference of $2\lambda_2 - \lambda_1$ being positive, negative or 0. Our next theorem uses the results from Markov Branching process and states the results in terms of RPWD. Let $\delta = (p_1 + p_2 - 1)$.

Theorem 2.4.3. The follows are true:

(i) If $\delta < 1/2$ then, as $n \to \infty$

$$\sqrt{n}(\frac{N_1}{n} - Q_1) \xrightarrow{d} N(0, \sigma^2) \tag{2.4.4}$$

where $\sigma^2 = (3 + 2\delta)(1 - 2\delta)^{-1}(Q_1(1 - Q_1))$.

(ii) If $\delta = 1/2$ then, as $n \to \infty$

$$\sqrt{\frac{n}{\log n}}(\frac{N_1}{n} - Q_1) \xrightarrow{d} N(0, \sigma^2) \tag{2.4.5}$$

where $\sigma^2 = 4Q_1(1 - Q_1)$.

(iii)If $\delta > 1/2$ then, as $n \to \infty$

$$n^{(\frac{1}{2}-\delta)}(\frac{N_1}{n} - Q_1) \xrightarrow{d} W \tag{2.4.6}$$

where W is a non-degenerate random variable. ∎

Remark 2.4.4. When $\delta = 1/2$, the limiting variance was identified by Matthews and Rosenberger (1997) while if $\delta < 1/2$ Smythe and Rosenberger (1995) identified the limiting variance. ∎

Remark 2.4.5. A rigorous proof for the identification of the limiting variance for $\delta \geq 1/2$ is as yet unavailable. ∎

2.4.1 Large Sample Theory

In this section, we describe the basic statistical techniques associated with estimation of success probabilities on treatment 1 and 2, *viz.*, p_1 and p_2. Let us denote by T_j the treatment indicator, *i.e.* $T_j = i$ if the j^{th} subject was assigned to treatment i, $i = 1, 2$. Let $I_{j,i} = 1$ if $T_j = i$. Let X_j denote the response variable for the j^{th} patient. Then

$$p_1 = P\{X_j = 1 | T_i = 1\} \tag{2.4.7}$$

and

$$p_2 = P\{X_j = 1 | T_i = 2\}. \tag{2.4.8}$$

Let $\boldsymbol{p} = (p_1,\ p_2)$. The likelihood $\mathcal{L}_n$ of the data is

$$\mathcal{L}_n(\boldsymbol{p}) = \prod_{i=1}^{2} \prod_{j=1}^{n} p_i^{X_j I_{j,i}} q_i^{(1-X_j) I_{j,i}}. \tag{2.4.9}$$

A simple differentiation with respect to p_1 and p_2 shows that

$$\hat{p}_1(n) = \frac{\sum_{j=1}^{n} X_j I_{j,1}}{\sum_{j=1}^{n} I_{j,1}}.$$

Now, $\hat{p}_1$ can be expressed as

$$\hat{p}_1(n) = \frac{1}{N_1(n)} \sum_{j=1}^{n} X_j I_{j,1}. \tag{2.4.10}$$

In a similar vein,

$$\hat{p}_2(n) = \frac{1}{N_2(n)} \sum_{j=1}^{n} X_j I_{j,2}. \tag{2.4.11}$$

Rosenberger, Flournoy and Durham (1997) have established the consistency and asymptotic normality of $\hat{p}_1(n)$ and $\hat{p}_2(n)$.

Theorem 2.4.6. Under the population model described in (2.4.7) and (2.4.8), $\hat{p}_1(n)$ and $\hat{p}_2(n)$ are strongly consistent and asymptotic normally distributed; *i.e.*

$$\lim_{n\to\infty} (\hat{p}_1(n),\ \hat{p}_2(n)) = (p_1,\ p_2) \qquad a.s. \tag{2.4.12}$$

and

$$\lim_{n\to\infty} P\left[(\sqrt{N_1}(\hat{p}_1 - p_1),\ \sqrt{N_2}(\hat{p}_2 - p_2)) \leq (x_1,\ x_2)\right] = \Phi\left(\frac{x_1}{\sqrt{p_1 q_1}}\right)\Phi\left(\frac{x_2}{\sqrt{p_2 q_2}}\right). \tag{2.4.13}$$

Remark 2.4.7. Since $\frac{N_1}{n} \to q_2(q_1+q_2)^{-1}$ (by (2.4.3)), we can re express the above limit result as

$$\lim_{n\to\infty} P\left[(\sqrt{n}(\hat{p}_1 - p_1),\ \sqrt{n}(\hat{p}_2 - p_2)) \leq (x_1,\ x_2)\right] = \Phi\left(\frac{x_1}{\sqrt{p_1 q_1 Q_1}}\right)\Phi\left(\frac{x_2}{\sqrt{p_2 q_2 Q_2}}\right). \tag{2.4.14}$$

2.5 Inference for Success Probability

In this section, we describe various techniques for constructing confidence intervals for the success probabilities. Wei(1988) was the first to initiate a test for the equality of success probabilities. Indeed , he developed a permutation test of

$$H_0 :\ p_1 = p_2 \tag{2.5.1}$$

and showed that ignoring the design in the analysis leads to an exaggerated treatment effect. Further work by Begg (1990) discusses the alone mentioned work of Wei (1988).

Confidence intervals for p_1 and p_2 were developed by Rosenberger *et. al* (1999) using asymptotic theory. They showed, using extensive simulations, that in small samples, ignoring the design in the analysis is anti conservative and hence the coverage probability based on asymptotic theory could have serious drawbacks (in terms of converge) in small samples.

Rosenberger and Hu (1999)investigated resampling procedures for constructing confidence intervals for the success probability. They considered situations where $\delta < 1/2$, even though theorem 2.4.6 holds without any restrictions on δ.

In the next section, we describe the basic resampling technique and develop a kernel smoothing approach to improve the performance of bootstrap confidence intervals. Using

the kernel smoothing to improve the accuracy of bootstrap confidence interval have been investigated in the content of i.i.d. data by Efron (1985), Polansky and Schucany (1997), Silverman and Young (1987) and Fisher and Hall (1991).

2.5.1 The Bootstrap Algorithm for RPWD

Let $\mathcal{X}_n = (X_j,\ T_{j\cdot},\ j = 1,\ \cdots,\ n)$ denote the response from the RPWD and $\hat{p_1}(n)$ and $\hat{p_2}(n)$ be as defined in (2.4.10) and (2.4.11). Using $\hat{p_1}(n)$ and $\hat{p_2}(n)$ as the success probability, we generate B randomized play the winner designs yielding the data

$$(\mathcal{X}_n^*(k),\ k = 1,\ \cdots,\ B) \tag{2.5.2}$$

where

$$\mathcal{X}_n^*(k) = \{(X_j^*,\ T_j^*),\ j = 1,\ 2,\ \cdots, n\}. \tag{2.5.3}$$

Now define the bootstrap version of (2.4.10) and (2.4.11) (for the k^{th} bootstrap sample) as

$$\hat{p}_1^*(k) = (N_1^*(n))^{-1} \sum_{j=1}^{n} \mathcal{X}_j^* T_{j1}^* \tag{2.5.4}$$

$$\hat{p}_2^*(k) = (N_2^*(n))^{-1} \sum_{j=1}^{n} \mathcal{X}_j^* T_{j2}^*. \tag{2.5.5}$$

The $100(1-\alpha)\%$ confidence interval based on the bootstrap samples is given by

$$CI - 1 = (\hat{p}_i^{*(B\alpha/2)}, \hat{p}_i^{*(B(1-\alpha)/2)}) \tag{2.5.6}$$

where $\hat{p}_i^{*(j)}$ is the j^{th} order statistic of $\{\hat{p}_i^{*(k)},\ 1 \le k \le B\}$.

We can also obtain the $100(1-\alpha)\%$ confidence interval for the estimates, $\hat{P}$, based on large sample theory, *viz.*,

$$CI - 2 = \hat{p}_i \pm z_{\alpha/2} \sqrt{\frac{\hat{p}_i \hat{q}_i}{N_i}} \tag{2.5.7}$$

where $z_{\alpha/2}$ represents the critical points from the standard normal distribution.

Since the variance is unknown, we replace the critical points from the standard normal distribution to the critical points from a t distribution. This yields

$$CI-3=\hat{p}_i \pm t_{\alpha/2,N_i-1}\sqrt{\frac{\hat{p}_i\hat{q}_i}{N_i}}. \tag{2.5.8}$$

Note that the above interval is conditioned on the design.

By approximating $\hat{p}_i - p_i$ by $\hat{p}_i^* - \hat{p}_i$, where $\hat{p}_i^*$ is an individual bootstrap estimate, $i = 1, 2$. Rosengerber and Hu(1999) studied the confidence interval given by

$$CI-4=(2\hat{p}_i-\hat{p}_i^{*(B(1-\alpha)/2)}, 2\hat{p}_i-\hat{p}_i^{*(B\alpha/2)}). \tag{2.5.9}$$

Using an adhoc approximation to constructing confidence intervals for the success probability. Rosenberger and Hu (1999) suggest the following confidence interval, *viz.*,

$$CI-5=(\hat{p}_i-Z_i^{*B(1-\alpha/2)}, \hat{p}_i-Z_i^{*B(\alpha/2)}) \tag{2.5.10}$$

where

$$Z_i^*(j)=\sqrt{\left(\frac{N_i^*(j)\hat{p}_i\hat{q}_i}{N_i\hat{p}_i^*(j)\hat{q}_i^*(j)}\right)}(\hat{p}_i^*(j)-\hat{p}_i)$$

and $Z_i^{*(j)}$ is the j^{th} order statistics of $\{Z_i^*(j),\ 1 \le j \le B\}$.

2.5.2 Kernel Smoothing of Bootstrap Samples

In this section, we develop the kernel smoothing of bootstrap samples to generate confidence intervals for the success probabilities. We define:

$$\hat{f}_B(x)=\frac{1}{B}\sum_{j=1}^{B}K_{h_B}(x-\hat{p}_i^{*j}) \tag{2.5.11}$$

where

$$K_{h_B}(x)=\frac{1}{h_B}K\left(\frac{x}{h_B}\right)$$

and h_B is the so-called bandwidth. Then $\hat{f}_B(\cdot)$ is an approximation to $h_{N_i}(x)$, the sampling distribution of $\hat{p}$. The lower bound and the upper bound for the $100(1-\alpha)\%$ confidence interval for p_i is given by

$$CI-6=(c_1^*,\ c_2^*), \tag{2.5.12}$$

where c_1^* and c_2^* satisfy

$$\int_{-\infty}^{c_1^*} \hat{f}_{B_i}(x)dx = \alpha/2 \tag{2.5.13}$$

and

$$\int_{c_2^*}^{\infty} \hat{f}_{B_i}(x)dx = \alpha/2. \tag{2.5.14}$$

The choice of h_B is important. The correctness of the confidence intervals depends on the rate of convergence of h_B to 0. Using Theorem 1 of Devroye (1987) we have the following Theorem.

Theorem 2.5.1. Assume $h_B \to 0$ and $Bh_B \to \infty$ as $B \to \infty$. Then

$$\lim_{B\to\infty} \hat{f}_{B_i}(x) = h_{N_i}(x), \quad a.s. \tag{2.5.15}$$

and

$$\lim_{B\to\infty} \int_{\Re} |\hat{f}_{B_i}(x) - h_{N_i}(x)|dx = 0. \tag{2.5.16}$$

Remark 2.5.2. In the above Theorem, n remains fixed and $B \to \infty$ (independent of n).

Remark 2.5.3. The above Theorem implies that as $B \to \infty$, $\hat{f}^*_{B_i}(x)$ approximate the sampling distribution of $\hat{p}_i(n)$.

2.5.3 Simulation Results

All simulations were carried out using 5000 simulations with 2000 bootstrap samples per simulation. The initial urn composition is $N_1(0) = N_2(0) = 1$ and $\alpha = 1$. The simulation was done in the Fortran language with eight parallel processors. For all simulations, we assume $r = 2$. If $\hat{p}_i,\ i = 1, 2$, was 0 or 1 or if N_1 and N_2 were 0, that replication was discarded. The kernel density was chosen to be the Gaussian kernel.

Table 2.1 contain the results of simulation for the randomized play the winner design with sample size equals to 30. Coverage probability for true p_1 and p_2 were computed for each of the six confidence intervals CI-1 to CI-6 with signlficant level equals to 0.05 and the average length of the confidence intervals have also been obtained.

We can tell that CI-3 requires longer length than CI-2 to achieve the same coverage rate. CI-1, CI-2 and CI-4 are anticonservative (coverage < 0.95), but the coverage rate for CI-1 increases as $p_1 + p_2$ increases. The coverage rate for CI-5 decreases as $p_1 + p_2$ increases. Furthermore CI-5 requires longer length than CI-6 for the same coverage rate. CI-6 (based on the kernel density estimates) can always achieve the desired probability (0.95) with moderately shorter length compared to the other procedures. Therefore, we conclude that the confidence interval obtained by kernel smoothing of bootstrap samples yields an optimal coverage with shortest length amongst the competing methodologies.

Table 2.1: Results for the RPWD. Simulated coverage probabilities (P) for true success probabilities and the average length (L) for the confidence intervals. Significant level=0.05, $N_1(0) = N_2(0) = 1$, $\alpha = 1$, $n = 30$, 5000 simulations and B=2000.

p_1		CI-1	CI-2	CI-3	CI-4	CI-5	CI-6
p_2							
0.5	P	0.9152	0.9136	0.9354	0.8602	0.9504	0.9432
	L	0.4935	0.4994	0.5502	0.4935	0.5668	0.5070
0.5	P	0.9092	0.9096	0.9668	0.8496	0.9478	0.9390
	L	0.4928	0.5002	0.5511	0.4928	0.5675	0.5061
0.7	P	0.9364	0.9252	0.9426	0.8882	0.9586	0.9454
	L	0.4459	0.4318	0.4691	0.4459	0.4914	0.4577
0.5	P	0.9126	0.9060	0.9692	0.8078	0.9222	0.9450
	L	0.5281	0.5573	0.6397	0.5281	0.6169	0.5430
0.75	P	0.9316	0.9194	0.9446	0.8232	0.9188	0.9510
	L	0.4763	0.4649	0.5315	0.4763	0.5165	0.4898
0.75	P	0.9330	0.9222	0.9896	0.8250	0.9203	0.9544
	L	0.4754	0.4643	0.5295	0.4754	0.5158	0.4893

Table 2.1 continuous

0.76	P	0.9360	0.9158	0.9352	0.8318	0.9120	0.9466
	L	0.4682	0.4580	0.5250	0.4682	0.5065	0.4818
0.75	P	0.9306	0.9186	0.9878	0.8210	0.9204	0.9510
	L	0.4787	0.4684	0.5378	0.4787	0.5196	0.4922
0.77	P	0.9524	0.9304	0.9454	0.8264	0.9152	0.9608
	L	0.4602	0.4464	0.5072	0.4602	0.4917	0.4738
0.75	P	0.933	0.9194	0.98686	0.8198	0.9152	0.9494
	L	0.4835	0.4747	0.5473	0.4835	0.5256	0.4972
0.8	P	0.9412	0.9362	0.9504	0.8160	0.9038	0.9536
	L	0.4566	0.4440	0.5128	0.4566	0.4774	0.4708
0.8	P	0.9386	0.9354	0.9896	0.8042	0.8992	0.9520
	L	0.4534	0.4392	0.5067	0.4534	0.4712	0.4677

2.6 Confidence Intervals for Allocation Parameters

In this section we develop the confidence intervals for the allocation proportion. This problem has not been investigated in the literature and we present first results for the problem. We recall from Theorem 2.4.3, the behavior of

$$(\frac{N_1}{n} - Q_1,\ \frac{N_2}{n} - Q_2)$$

depends on the value of $\delta = p_1 + p_2 - 1$. Indeed if $\delta < 1/2$, classical central limit type normalization, *viz* $\sqrt{n}$ holds and we have an asymptotic normal distribution. If $\delta = 1/2$, the normalization changes to $\sqrt{\frac{n}{\log n}}$ and limiting distribution is still normal. The most interesting case is when $\delta > 1/2$. In this case the normalization is $n^{1/2-\delta}$ and not much is known about the limit random variable.

Note further that if $\delta < 1/2$, the variance of the limiting normal distribution, which is given by

$$\sigma^2(\delta) = (3 + 2\delta)(1 - 2\delta)^{-1}(Q_i(1 - Q_i)) \tag{2.6.1}$$

has a singularity at $\delta = 1/2$. These issues complicate the construction of confidence intervals for Q_i.

We now describe the bootstrap technique developed in the previous section to construct confidence intervals for Q_i. We will use the notations from the previous sections. Note that using the data $\{\mathcal{X}^*(k),\ k = 1,\ \cdots,\ B\}$ we obtain for the k^{th} bootstrap sample

$$N_i^*(k) = \sum_{j=1}^{n} I^*_{\{T_j=i\}}. \tag{2.6.2}$$

Our first methodology for constructing confidence interval for Q_i is

$$CI_{1,1} = (\frac{N_i^{*(B\alpha/2)}}{n},\ \frac{N_i^{*(B(1-\alpha)/2)}}{n}) \tag{2.6.3}$$

where $N_i^{*(j)}$ is the j^{th} order statistic of

$$\{N_i^*(1),\ \cdots,\ N_i^*(B)\}.$$

The second and third confidence intervals for Q_i are based on the large sample theory and are given by

$$CI_{1,2} = \left(\frac{N_i}{n}\right) \pm \frac{z_{\alpha/2}}{\sqrt{n}} \left((3+2\delta)\left(\frac{Q_i(1-Q_i)}{1-2\delta}\right)\right)^{1/2}$$

and

$$CI_{1,3} = \left(\frac{N_i}{n}\right) \pm \frac{t_{\alpha/2,N_i-1}}{\sqrt{n}} \left((3+2\delta)\left(\frac{Q_i(1-Q_i)}{1-2\delta}\right)\right)^{1/2}.$$

If $\delta = 1/2$ then the corresponding intervals are given by

$$CI_{2,2} = \left(\frac{N_i}{n}\right) \pm z_{\alpha/2}(\frac{\log n}{n})^{1/2}(4Q_i(1-Q_i))^{1/2}$$

and

$$CI_{2,3} = \left(\frac{N_i}{n}\right) \pm t_{\alpha/2,N_i-1}(\frac{\log n}{n})^{1/2}(4Q_i(1-Q_i))^{1/2}.$$

If $\delta > 1/2$, then the corresponding interval is given by

$$CI_{3,2} = \left(\frac{N_i}{n}\right) \pm c_\alpha n^{\delta-1/2}\hat{\sigma}^2$$

where c_α is the percentage point from the distribution of W and $\hat{\sigma}^2$ is an estimate of the variance of $\frac{N_1}{n}$. We can also compare the above confidence intervals with their bootstrap versions', for example analogous to $CI_{1,2}$, we can define $CI^*_{1,2}$ as follows:

$$CI^*_{1,2} = \left(\frac{N_i}{n}\right) \pm \frac{z_{\alpha/2}}{\sqrt{n}} \left((3+2\delta^*)\left(\frac{Q^*_i(1-Q^*_i)}{1-2\delta^*}\right)\right)^{1/2}$$

where $\delta^* = p^*_1 + p^*_2 - 1$ and $p^*_i = \frac{1}{B}\sum_{k=1}^{B} p^*_i(k)$. Similar substitutes for δ and p_i for $CI_{1,3},\ CI_{2,2},\ CI_{2,3}$ and $CI_{3,2}$ will yield the corresponding bootstrap versions which we denote by $CI^*_{1,3},\ CI^*_{2,2},\ CI^*_{2,3}$ and $CI^*_{3,2}$.

Finally, we adopt the kernel smoothing technique from the previous section. Again using the data $\{\mathcal{X}^*_n(k),\ k = 1,\ \cdots, B\}$ we construct the smooth bootstrap density using

$$\hat{f}^*_{B_i}(x) = \frac{1}{B}\sum_{j=1}^{B} K_{h_B}(x - \frac{N^*_i(j)}{n}). \tag{2.6.4}$$

The lower bound and the upper bound for the $100(1-\alpha)\%$ confidence interval for Q_i is given by $(c^*_1,\ c^*_2)$, where c^*_1 and c^*_2 satisfy

$$\int_{-\infty}^{c_1^*} \hat{f}_{B_i}^*(x)dx = \alpha/2 \tag{2.6.5}$$

and

$$\int_{c_2^*}^{\infty} \hat{f}_{B_i}^*(x)dx = \alpha/2. \tag{2.6.6}$$

The following section will contain all the simulation results including the coverage probability and the length of the confidence intervals.

2.6.1 **Simulation Results**

The simulations were performed under the same scenario as mentioned in section 2.5.1. Table 2.2, 2.3 and 2.4 contain the results of the coverage probability for the asymptotic allocation proportion defined in (2.4.3) regarding to the above confidence intervals with signlficant level equals to 0.05 and the average length of the confidence intervals have also been obtained.

Table 2.2 contains the simulation results associated with $\delta < 1/2$. Comparing $CI_{1,2}$ and $CI_{1,3}$ with $CI_{1,2}^*$ and $CI_{1,3}^*$ respectively, see that by using the bootstrap correction for δ yields shorter confidence interval with same coverage. $CI_{1,1}$ has almost the same coverage rate as the confidence interval based on the kernel smoothing. We should note that the length of the confidence interval based on the kernel smoothing of the bootstrap samples is significantly shorter, the coverage rate is higher than these confidence intervals based on the asymptotic theory.

Table 2.3 contain the simulation results associate with $\delta = 1/2$. In this case, the confidence intervals based on the asymptotic theory have longer confidence intervals compared to these based on kernel smoothing.

Table 2.4 contain the simulation results accociate with $\delta > 1/2$. In this case the confidence based on the kernel smoothing techniqure and percertial from the bootstrap sample are much surprior comparing to the confidence intervals that using the asymptotic distribution's critial value.

Table 2.2 Results for the RPWD. Simulated coverage probabilities ($N_i,\ i = 1,\ 2$) for asymptotic allocation proportions and the average length (L) for the confidence intervals. Significant level=0.05, $N_1(0) = N_2(0) = 1$, $\alpha = 1$, $n = 30$, 5000 simulations, B=2000 and $\delta < 1/2$.

p_1		$CI_{1,1}$	$CI_{1,2}$	$CI_{1,3}$	$CI^*_{1,2}$	$CI^*_{1,3}$	CI-Ker
p_2							
0.5	N_1	0.9986	0.9665	0.9825	0.9648	0.9828	0.9998
	L	0.4768	0.6356	0.6972	0.6140	0.6736	0.4918
0.5	N_2	0.9986	0.9665	0.9819	0.9648	0.9822	0.9998
	L	0.4769	0.6356	0.6972	0.6140	0.6735	0.4918
0.7	N_1	0.9968	0.9810	0.9901	0.9594	0.9690	0.9968
	L	0.5319	0.8856	0.9577	0.8507	0.9198	0.5487
0.5	N_2	0.9968	0.9810	0.9862	0.9594	0.9644	0.9968
	L	0.5317	0.8856	1.0017	0.8507	0.9619	0.5487

2.7 **Conclusions**

In this chapter we developed confidence intervals for the design parameters and success probabilities. We introduced a novel kernel smoothing method that provided optimum coverage and smaller length compared to other existing methodologies, when dealing with success probability

We also developed a method for constructing confidence interval for the allocation proportions. This is a challenging problem due to the differential behavior of $\frac{N_i}{n}$ across the values of δ. The problem of providing unified inference for all δ without prior knowledge of δ is an interesting open problem and we plan to pursue it in the future.

Table 2.3 Results for the RPWD. Simulated coverage probabilities ($N_i,\ i = 1,\ 2$) for asymptotic allocation proportions and the average length (L) for the confidence intervals. Significant level=0.05, $N_1(0) = N_2(0) = 1$, $\alpha = 1$, $n = 30$, 5000 simulations, B=2000 and $\delta = 1/2$.

p_1		$CI_{2,1}$	$CI_{2,2}$	$CI_{2,3}$	$CI^*_{2,2}$	$CI^*_{2,3}$	CI-Ker
p_2							
0.75	N_1	0.9942	0.9976	0.9994	0.9996	1.0000	0.9972
	L	0.6447	1.2477	1.3898	1.2589	1.4024	0.6662
0.75	N_2	0.9944	0.9976	1.0000	0.9996	1.0000	0.9972
	L	0.6448	1.2478	1.3888	1.2589	1.4015	0.6662

Table 2.4 Results for the RPWD. Simulated coverage probabilities ($N_i,\ i = 1,\ 2$) for asymptotic allocation proportions and the average length (L) for the confidence intervals. Significant level=0.05, $N_1(0) = N_2(0) = 1$, $\alpha = 1$, $n = 30$, 5000 simulations, B=2000 and $\delta > 1/2$.

p_1		$CI_{3,1}$	$CI_{3,2}$	$CI^*_{3,2}$	CI-Ker
p_2					
0.8	N_1	0.9948	0.8734	0.8372	0.9964
	L	0.6747	0.9197	0.7255	0.6980
0.8	N_2	0.9944	0.8734	0.8374	0.9964
	L	0.6749	0.9200	0.7258	0.6980

References:

1. Athreya, K. B. and Karlin, S. (1967). Limit theorems for the split times of branching processes. Journal of Mathematical Mechanics, 17, 257-277.

2. Athreya, K. B. and Karlin, S. (1968). Branching processes and related limit theorem. The Annals of Mathematical Statistics, 39, 1801-1817.

3. Begg, C. B. (1990) On inferences fromWeis biased coin design for clinical trials. With discussion and a reply by the author. Biometrika, 77, 467-484.

4. Devroye, L. (1987). A Course in Density Estimation, Birkhauser Boston.

5. Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. Biometrika, 72, 45-58.

6. Fisher, N.I. and Hall, P. (1991) Bootstrap algorithms for small sample size. Journal of Statistical Planning and Inference. 27, 157-169.

7. Matthews, P. C. and Rosenberger, W. F. (1997). Variance in Randomized Play-the-Winner Clinical Trials. Statistics and Probability Letters, 35, 233-240.

8. Polansky, A.M. and Schucany, W.R.(1997) Kernel smoothing to improve bootstrap confidence intervals. (English summary) Journal of the Royal Statistical Society: Series B. 59, 821-838.

9. Rosenberger, W. F., Flournoy, N., and Durham, S. D. (1997). Asymptotic Normality of Maximum Likelihood Estimators from Multiparameter Response-Driven Designs. Journal of Statistical Planning and Inference, 60, 69-76.

10. Rosenberger, W. F. and Hu, F. (1999). Bootstrap Methods for Adaptive Designs. Statistics in Medicine, 18, 1757-1767.

11. Silverman, B.W. and Young, G.A.(1987) The bootstrap: to smooth or not to smooth. Biometrika, 74, 469-479.

12. Smythe R. T. and Rosenberger W. F. (1995). Play-the-winner design, generalized Pólya Urns, and markov branching processes. Adaptive Designs, Hayward, CA: Institute of Mathematical Statistics, 13-22.

13. Wei, L.J. (1988). Exact two-sample permutation tests based on the randomized play-the-winner rule. Biometrika , 75, 603-606.

Chapter 3

Minimum Hellinger Distance Estimators

3.1 Introduction

In Chapters 1 and 2 we described various randomization schemes adopted in $PhaseIII$ clinical trials and developed inferential methodology for the analysis of the design related parameters. However in a typical clinical trial, information on several variables are collected. For example, consider a clinical trial investigating the efficacy of drug A in lowering the cholesterol levels. In such a trial information on cholesterol levels, blood pressure, and body weight would be collected. Further, information on the lifestyle, diet and genetic components may also be collected. In this chapter, we develop robust methodology for the analysis of these continuous variables from the RPWD.

Minimum Hellinger distance procedure (MHDP) for analysis of independent and identically distributed (i.i.d.) data has been studied in the literature. Beran (1977) investigated the MHDP for continuous data and showed that the minimum Hellinger distance estimator (MHDE) of a finite dimensional parameter (in a parametric model) is as efficient as the MLE (Maximum Likelihood Estimator) under the true model assumption; furthermore, MHDE possess a "stability property" when the true distribution is in the neighborhood of the assumed parametric model. RPWD methodology naturally leads to situations where fewer subjects are sometimes allocated to one of the treatment arms. In these situations, it is difficult to identify the true distribution of the data under consideration. Drawing on the results from the i.i.d. literature, it is conceivable that the methodology based on MHDE would be more robust than the MLE and perhaps just as efficient as the MLE.

In this chapter, we will introduce the minimum Hellinger distance estimators for the randomized play the winner design and study it's asymptotic properties. We will also develop confidence intervals for these estimators using the asymptotic theory. Finally, we describe a new computational methodology (using Monte Carlo techniques) for obtaining the estimators. A SAS macro that implements these methodologies for Gaussian model is included. We begin with a brief introduction to the MHDP for the i.i.d. data.

3.2 Minimum Hellinger Distance Estimators (i.i.d. case)

We begin by describing the minimum Hellinger distance estimation for continuous i.i.d. data. Let $f(x)$ and $g(x)$ be any two densities; the Hellinger distance between $f(x)$ and $g(x)$ is defined as the L_2-norm of the difference between square root of density functions, *viz.*, $\sqrt{f(x)}$ and $\sqrt{g(x)}$, *i.e.*

$$\begin{aligned} HD^2(f,g) &= ||f(x)^{1/2} - g(x)^{1/2}||_2^2 \\ &= \int [(f(x))^{1/2} - (g(x))^{1/2}]^2 dx \\ &= 2 - 2\int f^{1/2}(x) g^{1/2}(x) dx. \end{aligned} \tag{3.2.1}$$

Let $X_1, \ldots, X_n$ be i.i.d. real valued random variables with density belonging to a specified parametric family $\{f(\cdot|\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta\}$, where $\Theta \subset \Re^p$ is the parameter space. Let $\mathcal{G}$ be a class of density functions. We will assume that $\mathcal{G}$ is metrized by L_1 convergence, *i.e.* $g_n \to g$ if and only if $g_n \xrightarrow{L_1} g \Leftrightarrow (\int |g_n(x) - g(x)|dx \to 0$ as $n \to \infty$.) The minimum Hellinger Distance functional (MHDF) of θ is defined to be a mapping (possibly multivalued). $T : \mathcal{G} \longrightarrow \Theta$ such that

$$T(g) = \arg\min_{\boldsymbol{\theta} \in \Theta} HD^2(f(\cdot|\boldsymbol{\theta}), g) \tag{3.2.2}$$

$$= \arg\max_{\boldsymbol{\theta} \in \Theta} \gamma(\boldsymbol{\theta}|g) \tag{3.2.3}$$

where

$$\gamma(\boldsymbol{\theta}|g) = \int_{\Re} f^{1/2}(x|\boldsymbol{\theta}) g^{1/2}(x) dx. \tag{3.2.4}$$

Given the observations $(X_1, \ldots, X_n)$, let $g_n(\cdot)$ denote any estimator of the density g of X_1. Then the MHDE of $\boldsymbol{\theta}$ is given by $T(g_n)$, *i.e.*

$$T(g_n) = \arg\min_{\boldsymbol{\theta}\in\Theta} HD_n^2(f(\cdot|\boldsymbol{\theta}), g_n), \tag{3.2.5}$$

where

$$HD_n^2(f(\cdot|\boldsymbol{\theta}, g_n) = 2 - 2\int_{\Re} f^{1/2}(x|\boldsymbol{\theta}) g_n^{1/2}(x)dx. \tag{3.2.6}$$

Equivalently

$$T(g_n) = \arg\max_{\boldsymbol{\theta}\in\Theta} \gamma_n(\boldsymbol{\theta}|g_n). \tag{3.2.7}$$

When there is no scope for confusion we will use $HD_n^2(f(\cdot|\boldsymbol{\theta}, g_n)$ or $HD_n^2(\boldsymbol{\theta})$ based on the notational convenience. In the same spirit, we will use $\gamma(\boldsymbol{\theta}, g_n)$ and $\gamma(\boldsymbol{\theta})$ interchangeably. A choice for $g_n(\cdot)$ is the kernel density estimator

$$g_n(x) = \frac{1}{nc_n}\sum_{j=1}^{n} K\{\frac{x - X_j}{c_n}\}. \tag{3.2.8}$$

where $K(\cdot)$ is a kernel density. It is known that (see Devroye (1987)) if $c_n \to 0$, and $nc_n \to \infty$ then $g_n \xrightarrow{L_1} f$. Now using Cauchy-Schwartz inequality, it follows that

$$HD_n^2(f(\cdot|\boldsymbol{\theta}), g_n) \to 0. \tag{3.2.9}$$

Beran (1977) has shown that the MHDE is "robust" compared to the maximum likelihood estimator when data contaminations are present. Furthermore MHDE is known to be asymptotically efficient under a specified parametric family of densities and is minimax robust in a small Hellinger metric neighborhood of the given family (Beran 1977).

Tamura and Boos (1986) have studied the MHDE when the data are vector valued. They established the affine-invariance property of the MHDE for multivariate location and scale. They also established that the breakdown point (*i.e.* the smallest fraction of contamination that can cause the estimator to take arbitrary large values) of the estimator to be at least $\frac{1}{4}$. The breakdown point of an affine-invariant M-estimator is at most $\frac{1}{d+1}$, where d is the dimension of parameter space. The MHDE for the multivariate case is superior since it is independent of the dimension.

In the context of discrete data, Simpson (1987) showed that the MHDE has 50% breakdown point. Simpson (1989) also developed a deviance test using Hellinger distance and studied the impact of contamination on the test size and power. More recently, Lindsay (1994) studied a generalized disparity method for estimation in parametric models with discrete support. Indeed, a particular choice of the disparity measure yields the MHDE. Sriram and Vidyashankar (2000) studied MHDE for supercritical branching process. Other weighted likelihood based approaches for robustness are based on the work of Markatou, Lindsay and Basu (1997).

3.3 Minimum Hellinger Distance Estimators for Randomized Play The Winner Design

In this section, we develop the MHDP for analysizing the continuous outcomes from a RPWD. We begin with notations and terminology. Let T_i denote the treatment indicator $T_i = 1$ if the i^{th} subject received treatment 1 and $T_i = 2$ if the i^{th} subject receiving treatment 2. Let

$$\mathcal{A}_n^{(i)} = \{1 \leq j \leq n : T(j) = i\}$$

denote all those subjects (amongst the first n subjects) randomized to treatment i. Then the cardinality of $\mathcal{A}_n$ is

$$|\mathcal{A}_n^{(i)}| = N_i. \tag{3.3.1}$$

Let

$$\mathcal{A}(i) = \{j \geq 1 : T(j) = i\}$$

denote the set of all subjects randomized to treatment i. The data can be represented by $\{X_{i,\ \nu(i,j)},\ j \in \mathcal{A}_n,\ i = 1,\ 2\}$. Note that given the treatment assignment, the responses of subjects receiving a particular treatment are i.i.d. and given the treatment assignment, the responses of subjects receiving treatment 1 are independent of the responses of the subjects receiving treatment 2.

While such a conditional independence is useful one can actually extract a "full" independence of the allocated sequences. This is made precise in the following theorem of Melfi and Page (2000).

Theorem 3.3.1. Let $\{(X_{1,j}, X_{2,j}),\ j \geq 1\}$ be a collection of i.i.d. random variables and let F_{X_1} denote the distribution of $\{X_{1,j}, j \geq 1\}$ and G_{X_2} denote the distribution of $\{X_{2,j}, j \geq 1\}$. Let $\mathcal{F}_n = \sigma < T_1, \ldots, T_n >$. Assume that $(X_{1,n}, X_{2,n})$ is independent of $\mathcal{F}_{n-1}$. Let $\{\nu(i,j),\ i = 1, 2,\ j = 1, 2, \ldots\}$ be a collection of positive, increasing in j (for all i), integer-valued, almost surely finite random variables such that

$$\{\nu(i,j) = k\} \in \mathcal{F}_{j-1}$$

for all $i = 1, 2$. Assume that $p(\nu(1,j) = \nu(2,k)) = 0$ for all integers j and k. Then

(i) $\{X_{1,\nu(1,j)},\ j \in \mathcal{A}(1)\}$ are i.i.d. with common distribution F_{X_1};

(ii) $\{X_{1,\nu(2,j)},\ \ j \in \mathcal{A}(2)\}$ are i.i.d. with common distribution G_{X_2}; and

(iii) the two sequences $\{X_{1,\nu(1,j)},\ j \geq 1\}$ and $\{X_{2,\nu(2,j)},\ j \geq 1\}$ are independent of one another. ∎

In order to establish the criterion function for estimating the parameters, we make the following assumption, *viz*,

(E1) $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are not functionally dependent.

We note that this condition can easily be removed by modeling the dependence between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. This introduces more complex notations and technical issues and hence is not pursued in this work. Let

$$\boldsymbol{F}(\cdot|\ \Xi) = \begin{pmatrix} f(\cdot|\boldsymbol{\theta}) \\ g(\cdot|\boldsymbol{\eta}) \end{pmatrix} \text{ and } \boldsymbol{H} = \begin{pmatrix} h_1 \\ h_2 \end{pmatrix} \tag{3.3.2}$$

where $\boldsymbol{\Xi} = (\boldsymbol{\theta},\ \boldsymbol{\eta})'$. Let $\Theta_1 \in \Re^p$ denote the parameter space corresponding to $\boldsymbol{\theta}$, $\Theta_2 \in \Re^p$ denote the parameter space corresponding to $\boldsymbol{\eta}$ and $\boldsymbol{\Theta} = \Theta_1 \times \Theta_2$ denote the parameter

space corresponding $\boldsymbol{\Xi}$. Let

$$VHD(\boldsymbol{F},\ \boldsymbol{H}) = \begin{pmatrix} HD^2(f(\cdot|\boldsymbol{\theta}), h_1) \\ HD^2(g(\cdot|\boldsymbol{\eta}), h_2) \end{pmatrix} \tag{3.3.3}$$

denote the vector of squares of Hellinger distances between the components of $\boldsymbol{F}$ and $\boldsymbol{H}$.

Now analogous to (3.2.2), we define the MHDF to be the functional (possibly multivalued) $T : \mathcal{G} \times \mathcal{G} \to \boldsymbol{\theta}$ such that

$$\begin{aligned} T(\boldsymbol{H}) &= \arg\min_{\boldsymbol{\Xi}\in\boldsymbol{\Theta}} \{VHD(\boldsymbol{F}(\cdot|\ \Xi),\ \boldsymbol{H}\} & (3.3.4)\\ &= \arg\max_{\boldsymbol{\Xi}\in\boldsymbol{\Theta}} \{\boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H})\} & (3.3.5) \end{aligned}$$

where

$$\boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H}) = \left(\int_{\Re} f^{1/2}(x|\boldsymbol{\theta}) h_1^{1/2}(x)dx,\ \int_{\Re} g^{1/2}(x|\boldsymbol{\eta}) h_2^{1/2}(x)dx\right)'.$$

Now, if $\boldsymbol{H}_{\mathrm{n}}$ are the estimators of $\boldsymbol{H}$ based on the data (to be described below) then the MHDE of $\boldsymbol{\Xi}$ is given by

$$T(\boldsymbol{H}_{\mathrm{n}}) = \arg\min_{\boldsymbol{\Xi}\in\boldsymbol{\Theta}} \{VHD(\boldsymbol{F}(\cdot|\ \Xi),\ \boldsymbol{H}_{\mathrm{n}}\} \tag{3.3.6}$$

We choose for $\boldsymbol{H}_{\mathrm{n}}$

$$\boldsymbol{H}_{\mathrm{n}} = (h_{1,n},\ h_{2,n})'$$

the following kernel density estimates, given by

$$h_{i,n}(x) = \frac{1}{N_i} \sum_{j\in\mathcal{A}_n(i)} K\left(\frac{x - X_{i,\nu(i,j)}}{c_n}\right), \quad i = 1, 2 \tag{3.3.7}$$

Now using the sample version for VHD we get

$$\begin{aligned} VHD_n(\boldsymbol{F},\ \boldsymbol{H}) &= (HD_n^2(f(\cdot|\boldsymbol{\theta}), h_{1,n}),\ HD_n^2(g(\cdot|\boldsymbol{\eta}), h_{2,n})) \\ &= 2\cdot\mathbf{1} - \boldsymbol{\gamma}_{\mathrm{n}}(\boldsymbol{\Xi},\ \boldsymbol{H}) & (3.3.8) \end{aligned}$$

where

$$\boldsymbol{\gamma}_{\mathrm{n}}(\boldsymbol{\Xi},\ \boldsymbol{H}) = (\gamma_{n,1}(\boldsymbol{\theta}),\ \gamma_{n,2}(\boldsymbol{\eta})).$$

3.4 Existence and Uniqueness

In this section, we will establish the existence and uniqueness of the MHDE, defined through a minimization of (3.3.6). Recall that

$$\boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H}) \ = \ \left(\int_{\Re} f^{1/2}(x|\boldsymbol{\theta})h_1^{1/2}(x)dx,\ \int_{\Re} g^{1/2}(x|\boldsymbol{\eta})h_2^{1/2}(x)dx\right)'$$

We will make the following assumptions through out this chapter.

(E2) The parameter spaces Θ_1 and Θ_2 are locally compact.

(E3) $f(\cdot|\boldsymbol{\theta})$ and $g(\cdot|\boldsymbol{\eta})$ are upper semi-continuous.

Our first theorem shows that under a further weak regularity condition (3.3.6) exists.

Theorem 3.4.1. Assume (E1)-(E3). Let $\boldsymbol{\Theta}_K = K_1 \times K_2$, where $K_i \subset \Theta_i$ is compact for all $i = 1, 2$.

(i) Assume that

$$\sup_{\boldsymbol{\Xi}\in\boldsymbol{\Theta}^c} \boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H}) < \sup_{\boldsymbol{\Xi}\in\boldsymbol{\Theta}_K} \boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H})$$

where $\boldsymbol{\Theta}^c = \Theta_1^c \times \Theta_2^c$, $\Theta_1^c = K_1^c \cap \Theta_1$, and $\Theta_2^c = K_2^c \cap \Theta_2$.

(ii) If $\boldsymbol{\Xi}_1 \neq \boldsymbol{\Xi}_2$ then $\boldsymbol{H}(\cdot|\boldsymbol{\Xi}_1) \neq \boldsymbol{H}(\cdot|\boldsymbol{\Xi}_2)$ on a set of positive Lebesgue measure.

Under the above condition (3.3.6) exists. Furthermore $T(\boldsymbol{F})$ is unique.

Proof. The method of proof involves two steps.

(1) We will show that $VHD(\boldsymbol{F},\ \boldsymbol{H})$ is lower semi-continuous.

(2) We will then use (i) along with condition (ii) of the theorem to establish the existence of the minimizers of (3.3.6).

We begin with (1). Note that

$$VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}),\ \boldsymbol{H}) = 2\cdot\mathbf{1} - 2\boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H}). \tag{3.4.1}$$

Under (E3), $\boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H})$ is upper semi-continuous function. Hence $VHD(\boldsymbol{F},\ \boldsymbol{H})$ is lower semi-continuous. Now, since K_1 and K_2 are compact subsets of Θ_1 and Θ_2 respectively,

$\boldsymbol{\Theta}_K = K_1 \times K_2$ is also a compact subset of $\boldsymbol{\Theta}$. Hence from the lower semi-continuity of VHD there exists a $\boldsymbol{\Xi}^*$ such that

$$(m_1, m_2)' = VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}^*),\ \boldsymbol{H}) = \inf_{\boldsymbol{\Xi} \in \boldsymbol{\Theta}_K} VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}),\ \boldsymbol{H}). \tag{3.4.2}$$

Hence using (3.4.1)

$$\boldsymbol{\gamma}(\boldsymbol{\Xi}^*,\ \boldsymbol{H}) = \frac{1}{2}(2 - m_1,\ 2 - m_2)'.$$

Now, using condition (i), we have that for all $\boldsymbol{\Xi} \notin \boldsymbol{\Theta}_K$, we have $\boldsymbol{\gamma}(\boldsymbol{\Xi},\ \boldsymbol{H}) < \frac{1}{2}(2-m_1,\ 2-m_2)'$. Hence $\boldsymbol{\Xi}^*$ minimizes $VHD(\cdot|\boldsymbol{\Xi})$ on $\boldsymbol{\Theta}$. We next prove the uniqueness of $T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}))$. Note that

$$VHD(\boldsymbol{F}(\boldsymbol{\Xi}),\ \boldsymbol{F}(\boldsymbol{\Xi}_0)) = \begin{pmatrix} HD^2(f(\cdot|\boldsymbol{\theta}), f(\cdot|\boldsymbol{\theta_0})) \\ HD^2(g(\cdot|\boldsymbol{\eta}), g(\cdot|\boldsymbol{\eta_0})) \end{pmatrix}$$

and is minimized at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ uniquely by Beran's (1977) Theorem 1. by the identifiability assumption implied by (ii). ∎

3.5 L_1- Convergence of Kernel Density Estimators for RPWD

In this section we deal with the L_1 convergence of the kernel density estimators that are required in the proof of our consistency results. We recall that the kernel density estimators of h_1 and h_2 (the densities of responses for treatment 1 and treatment 2 respectively) are given by

$$h_{i,n}(x) = (N_i c_n)^{-1} \sum_{j \in \mathcal{A}_n(i)} K\left(\frac{x - X_{i,\nu(i,j)}}{c_n}\right), \quad i = 1, 2. \tag{3.5.1}$$

Our first theorem establishes the strong pointwise consistency and strong L_1 consistency of $h_{i,n}(\cdot)$ and $E(h_{i,n}(\cdot))$.

Theorem 3.5.1. Assume that $c_n \to 0$ and $nc_n \to \infty$ as $n \to \infty$. Then for almost all x (with respect to the Lebesgue Measure)

$$\lim_{n \to \infty} h_{i,n}(x) = h_i(x) \ \text{ a.s.}, \tag{3.5.2}$$

and

$$\lim_{n\to\infty} E(h_{i,n}(x)) = h_i(x) \text{ a.s..} \tag{3.5.3}$$

Furthermore,

$$\lim_{n\to\infty} \int_{\Re} |h_{i,n}(x) - h_i(x)| = 0, \tag{3.5.4}$$

and

$$\lim_{n\to\infty} \int_{\Re} |E(h_{i,n}(x)) - h_i(x)| = 0. \tag{3.5.5}$$

Proof. By Melfi's Theorem, $\{X_{i,\nu(i,j)}, j \in \mathcal{A}(i)\}$ are i.i.d. random variables. Hence, by Theorem 1 of Devroye (1987), (3.5.2) follows. (3.5.3) is now a consequence of Glick's Theorem. We next calculate

$$\begin{aligned} E\left(K\left(\frac{x - X_{i,\nu(i,j)}}{c_n}\right)\right) &= \int K\left(\frac{x-y}{c_n}\right) h_i(y)dy \\ &= c_n \int K(t) h_i(x + tc_n)dt \end{aligned} \tag{3.5.6}$$

Now, conditioning on the treatment assignment and using (3.5.6)

$$E(h_{i,n}(x)) = \int K(t) h_i(x + tc_n)dt. \tag{3.5.7}$$

Thus, to complete the proof we need to show that (3.5.7) converges to $h_i(x)$. Now

$$|E(h_{i,n}(x)) - h_i(x)| \leq \int K(t)|h_i(x + t(c_n) - h_i(x)|dt. \tag{3.5.8}$$

By the bounded convergence theorem, right hand side of (3.5.8) converges to 0 as $n \to \infty$ yielding (3.5.4). Finally, by integrating (3.5.8) and interchanging the order of integration (using Tonelli's theorem), it follows again by the bounded convergence theorem that

$$\lim_{n\to\infty} \int |E(h_{i,n}(x)) - h_i(x)| = 0$$

yielding (3.5.5). ∎

3.6 Continuity and Consistency of the MHDF

In this section, we study the consistency of the MHDE *via* the continuity of the MHDF defined in (3.3.6). Recall that $\mathcal{G}$ is the class of densities and $T : \mathcal{G} \times \mathcal{G} \to \boldsymbol{\Theta}$ defined by

$$T(\boldsymbol{H}) = \arg \max_{\boldsymbol{\Xi} \in \boldsymbol{\Theta}} \boldsymbol{\gamma}(\boldsymbol{\Xi}|\boldsymbol{H}).$$

Our first result establishes the continuity of T. Assume that (E1-E3) and conditions of Theorem 3.4.1. hold.

Theorem 3.6.1. Assume further (E1)-(E3) that $T(\boldsymbol{H})$ is unique. Then T is continuous, *i.e.* if $h_{1,n} \xrightarrow{L_1} h$, $h_{2,n} \xrightarrow{L_1} h_2$, then

$$\lim_{n\to\infty} T(\boldsymbol{H}_n) = T(\boldsymbol{H}). \tag{3.6.1}$$

Proof. Let $h_{1,n} \xrightarrow{L_1} h_1$ and $h_{2,n} \xrightarrow{L_1} h_2$. By Theorem 3.4.1, there exists $\boldsymbol{\Xi}_n \in \boldsymbol{\Theta}$ such that $T(\boldsymbol{H}_n) = \boldsymbol{\Xi}_n$. Also, by Theorem 3.4.1, there exist $\boldsymbol{\Xi} \in \boldsymbol{\Theta}$ such that

$$T(\boldsymbol{H}) = \boldsymbol{\Xi}.$$

Thus, to prove (3.6.1) it is enough to show that

$$\boldsymbol{\Xi}_{\mathrm{n}} \to \boldsymbol{\Xi}_0. \tag{3.6.2}$$

We now show that is sufficient to prove that

$$\lim_{n\to\infty} \sup_{\boldsymbol{\Xi} \in \boldsymbol{\Theta}} |VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}_n),\ \boldsymbol{H}_{\mathrm{n}}) - VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}),\ \boldsymbol{H})| = \mathbf{0}. \tag{3.6.3}$$

To this end, suppose (3.6.3) holds and (3.6.2) does not hold. By compactness of $\boldsymbol{\Theta}_K$, we have that there exists $\boldsymbol{\Xi}_* \neq \boldsymbol{\Xi}_0$ and a subsequence n_k such that

$$\boldsymbol{\Xi}_{n_k} \to \boldsymbol{\Xi}_*. \tag{3.6.4}$$

Hence by (3.6.3)

$$VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}_{n_k}),\ \boldsymbol{H_{n_k}}) \to VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}_*),\ \boldsymbol{H}). \tag{3.6.5}$$

This implies that

$$VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}_*),\ \boldsymbol{H}_*) = VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}),\ \boldsymbol{H})$$

contradicting the uniqueness of $T(\boldsymbol{H})$. Now we show that (3.6.3) holds. Note that

$$\begin{aligned} &\sup_{\boldsymbol{\Xi}\in\Theta} |VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}_n),\ \boldsymbol{H}_{\mathrm{n}}) - VHD(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}_0),\ \boldsymbol{H})| \\ \leq &\begin{pmatrix} \sup_{\boldsymbol{\theta}\in K_1} |HD^2(f(x|\boldsymbol{\theta}_0), h_{1,n}) - HD^2(f(x|\boldsymbol{\theta}_0), h_1)| \\ \sup_{\boldsymbol{\eta}\in K_2} |HD^2(g(x|\boldsymbol{\eta}_0), h_{2,n}) - HD^2(g(x|\boldsymbol{\eta}_0), h_2)| \end{pmatrix}. \end{aligned}$$

By Theorem 1 (ii) of Beran (1977), each of the components on the RHS of the above converges to 0, proving (3.6.3). ∎

Now, using the compactness of K_1 and K_2 and using Theorem 3.5.1 and 3.6.1 we get strong consistency of the MHDE. We state this as a Theorem.

Theorem 3.6.2. Assume that the $T(\boldsymbol{H})$ is unique. Then, the sequence of MHDE defined in (3.3.6) converges a.s. to $T(\boldsymbol{H})$. ∎

3.7 Joint Asymptotic Normality of MHDE of $\boldsymbol{\Xi}_{\mathrm{n}}$

In this section, we deal with the joint asymptotic normality $\boldsymbol{\Xi}_{\mathrm{n}}$. We will assume throughout this section that the conditions (E1-E3) and the conditions of Theorem 3.4.1, 3.5.1 and 3.6.1 hold. We need the following regularity conditions on $\{f(\cdot|\boldsymbol{\theta}),\ \theta \in \Theta_1\}$ and $\{g(\cdot|\boldsymbol{\eta}),\ \boldsymbol{\eta} \in \Theta_2\}$.

(D1) $f(\cdot|\boldsymbol{\theta})$ and $g(\cdot|\boldsymbol{\eta})$ are twice continuously differentiable functions of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$.

(D2) Assume further that $||\boldsymbol{\nabla} f^{1/2}(\cdot|\boldsymbol{\theta})||_2$ and $||\boldsymbol{\nabla} g^{1/2}(\cdot|\boldsymbol{\eta})||_2$ are continuous and bounded.

Using the (**D1**) and (**D2**) and partially differentiating with respect to $\boldsymbol{\Xi}$ we get

$$\boldsymbol{\nabla} VHD(\boldsymbol{\Xi}) = 0 \tag{3.7.1}$$

Let $\boldsymbol{\Xi}_n$ be the solution to (3.7.1). Now applying one term Taylor expansion of (3.7.1) we get

$$\boldsymbol{\nabla} VHD_n(\boldsymbol{\Xi}_0) = \boldsymbol{\nabla} VHD_n(\boldsymbol{\Xi}_n) + (\boldsymbol{\Xi}_n - \boldsymbol{\Xi}_0)' D_n(\boldsymbol{\Xi}_n^*) \tag{3.7.2}$$

where $\boldsymbol{\Xi}_n^* = (\boldsymbol{\theta}_n^*, \boldsymbol{\eta}_n^*)' \in U_n(\boldsymbol{\theta}_0) \times V_n(\boldsymbol{\eta}_0)$ and

$$U_n(\boldsymbol{\theta}_0) = \{\boldsymbol{\theta} | \boldsymbol{\theta} = t\boldsymbol{\theta}_0 + (1-t)\boldsymbol{\theta}_n\} \tag{3.7.3}$$

$$V_n(\boldsymbol{\eta}_0) = \{\boldsymbol{\eta} | \boldsymbol{\eta} = t\boldsymbol{\eta}_0 + (1-t)\boldsymbol{\eta}_n\} \tag{3.7.4}$$

Thus,

$$(\boldsymbol{\Xi}_n - \boldsymbol{\Xi}_0)' = \boldsymbol{\nabla} V H D_n(\boldsymbol{\Xi}_0) D_n^{-1}(\boldsymbol{\Xi}_n^*). \tag{3.7.5}$$

Hence,

$$\begin{aligned} \sqrt{n}(\boldsymbol{\Xi}_n - \boldsymbol{\Xi}_0)' &= \sqrt{n}\boldsymbol{\nabla} V H D_n(\boldsymbol{\Xi}_0)(D_n^{-1}(\boldsymbol{\Xi}_n^*) - D_n^{-1}(\boldsymbol{\Xi}_0)) \\ &\quad + \sqrt{n}\boldsymbol{\nabla} V H D_n(\boldsymbol{\Xi}_0) D_n^{-1}(\boldsymbol{\Xi}_0) \\ &= T_{n,1} + T_{n,2}. \end{aligned} \tag{3.7.6}$$

The following regularity conditions will be needed for proving our theorm.

R1. The kernel $K(\cdot)$ density is symmetric with compact support (K).

R2. Let $\{\alpha_n,\ n \geq 1\}$ be a sequence diverging to infinity. Assume that

$$\lim_{n\to\infty} n \sup_{t \in \text{supp}(K)} P_g(|x - c_n t| > \alpha_n) = 0$$

$$\lim_{n\to\infty} n \sup_{t \in \text{supp}(K)} P_f(|x - c_n t| > \alpha_n) = 0$$

where $\text{supp}(K)$ is the support of the kernel density $K(\cdot)$.

R3. Let

$$M_n(1) = \sup_{|x| \leq \alpha_n} \sup_{t \in \text{supp}(K)} \left| f^{-1}(x) f(x + tc_n | \boldsymbol{\theta}) \right|$$

$$M_n(2) = \sup_{|x| \leq \alpha_n} \sup_{t \in \text{supp}(K)} \left| g^{-1}(x) g(x + tc_n | \boldsymbol{\eta}) \right|.$$

Assume

$$\sup_{n \geq 1} M_n(i) < \infty \text{ for } i = 1, 2.$$

R4. $nc_n^2 \to \infty$.

Lemma 3.7.1. Assume (R1)-(R4), (D1)-(D2) hold. Assume all the conditions from the previous section hold. Then

$$\lim_{n\to\infty}(D_n^{-1}(\boldsymbol{\Xi}_n^*) - D_n^{-1}(\boldsymbol{\Xi}_0)) = 0. \tag{3.7.7}$$

Proof. Note that D_n is a block diagonal matrix with diagonals $D_{n,1}$ and $D_{n,2}$ (say). Tamura and Boos (1984) have shown that $D_{n,i} \to \infty, \quad i = 1, 2$ as $n \to \infty$. Their result is true for fixed sample size. However, using Renyi type theorem for random sample sizes, (3.7.7) follows. ∎

Our next lemma studies the behavior of $T_{n,2}$ in (3.7.6)

Lemma 3.7.2 Assume (R1)-(R4) and (D1)-(D2) hols. Assume further that the conditions of section 3.6 hold. Then

$$T_{n,2} = \sqrt{n}\left(S^1_{N_1}, S^2_{N_2}\right)' + o_p(1) \tag{3.7.8}$$

where

$$o_p(1) \to 0 \text{ as } n \to \infty$$

$$S^i_{N_i} = \frac{1}{N_i}\sum_{j=1}^{N_i}\psi_i(X_{1,\nu(i,j)}),\ i = 1, 2 \quad ,$$

$$\psi_1(x) = (f(x|\boldsymbol{\theta}))^{-1}\boldsymbol{\nabla} f(x|\boldsymbol{\theta}),$$

$$\psi_2(x) = (g(x|\boldsymbol{\eta}))^{-1}\boldsymbol{\nabla} g(x|\boldsymbol{\eta}).$$

Proof. The proof follows along the same lines as in Tamura and Boos(1984). Convergence to 0 of the $o_P(1)$ term follows from the arguments as in Lemma 3.7.1. ∎

Theorem 3.7.3. Assume that the conditions (R1)-(R4), (E1)-(E3), (D1)-(D2) hold. Then, as $n \to \infty$

$$\sqrt{n}(\boldsymbol{\Xi}_n - \boldsymbol{\Xi}_0) \xrightarrow{d} N_2(\mathbf{0}, \Sigma)$$

where

$$\Sigma = \begin{pmatrix} Q_1 I_1^{-1}(\boldsymbol{\theta}_0) & 0 \\ 0 & Q_2 I_2^{-1}(\boldsymbol{\eta}_0) \end{pmatrix}. \tag{3.7.9}$$

where

$$I_1^{-1}(\boldsymbol{\theta}_0) = 4[\int (\boldsymbol{\nabla} f^{1/2}(x|\boldsymbol{\theta}))(\boldsymbol{\nabla} f^{1/2}(x|\boldsymbol{\theta})' dx]$$

and

$$I_2^{-1}(\boldsymbol{\eta}_0) = 4[\int (\boldsymbol{\nabla} g^{1/2}(x|\boldsymbol{\eta}))(\boldsymbol{\nabla} g^{1/2}(x|\boldsymbol{\eta})') dx].$$

Proof. Using (3.7.6) and lemma 3.7.1 and lemma 3.7.2, it is enough to show that $\sqrt{n}T_{n,2}$ converges to a bivariate normal distribution. We will use the Cramer-Wold device. Let l_1 and l_2 be any column vectors of constants. Now, using the lemma 3.7.2, we can express the linear comination of $S^1_{N_1}$ and $S^2_{N_2}$ as

$$\sqrt{n}(\sum_{i=1}^{2} \frac{1}{N_i} \sum_{j=1}^{N_i} l_i \psi_i(X_{i,\nu(i,j)})). \tag{3.7.10}$$

Thus to complete the proof we need to show that the term

$$\sqrt{n} \sum_{i=1}^{2} \frac{1}{N_i} \sum_{j=1}^{N_i} l_i \psi_i(X_{i,\nu(i,j)})$$

converges a normal distribution. Now

$$\sqrt{n} \sum_{i=1}^{2} \frac{1}{N_i} \sum_{j=1}^{N_i} l_i \psi_i(X_{i,\nu(i,j)}) \quad = \quad G_{n,1} + G_{n,2} + G_{n,3} + G_{n,4} \tag{3.7.11}$$

where

$$\begin{aligned}
G_{n,1} &= \frac{1}{N_1} \sum_{j \in \mathcal{A}_n(1)} l_1' \psi_1(X_{1,\nu(i,j)}) = \frac{1}{N_1} \sum_{j=1}^{[n \cdot Q_1]} l_1' \psi_1(X_{1,\nu(i,j)}) \\
G_{n,2} &= \frac{1}{N_2} \sum_{j=1}^{[nQ_2]} l_2' \psi_2(X_{1,\nu(2,j)}) \\
G_{n,3} &= \frac{1}{N_1} \sum_{[Q_1 n] \wedge N_1 \le j \le N_1 \vee |Q_1 n]} l_i' \psi_1(X_{1,\nu(1,j)}) \\
G_{n,4} &= \frac{1}{N_2} \sum_{[Q_2 n] \wedge N_2 \le j \le N_2 \vee |Q_2 n]} l_2' \psi_2(X_{2,\nu(2,j)}).
\end{aligned}$$

By Theorem 3.3.1 and central limit theorem for i.i.d. random variables $G_{n,1} + G_{n,2}$ converges to a linear combination of Gaussian random vectors. Furthermore,

$$D_{n,1}^{-1}(\boldsymbol{\theta}_0) \to D_1^{-1}(\boldsymbol{\theta}_0) \tag{3.7.12}$$

and

$$D_{n,2}^{-1}(\boldsymbol{\eta}_0) \to D_2^{-1}(\boldsymbol{\eta}_0). \tag{3.7.13}$$

We will now show that $G_{n,3} + G_{n,4}$ converges to 0 in probability. Now, for any $\delta > 0$,

$$\begin{aligned} P(|G_{n,3}| > \epsilon) &= P(|G_{n,3}| > \epsilon : |\frac{N_1}{n} - Q_1| < \delta) + P(|G_{n,3}| > \epsilon : |\frac{N_1}{n} - Q_1| > \delta) \\ &= \text{(i)} \qquad\qquad\qquad\qquad + \text{(ii)} \end{aligned}$$

Let $E(n,\delta) = \{\omega : |\frac{N_1(w)}{n} - Q_1| < \delta\}$.

Note that

$$\begin{aligned} \text{(i)} &= P\left\{ \left(\frac{1}{N_1} \sum_{j=N_1 \wedge [nQ_1]}^{N_1 \vee [nQ_1]} \psi_1(X_{1,\nu(1,j)}) \right) > \epsilon : \quad E(n,\delta) \right\} \\ &\leq P\left(\frac{1}{n(Q_1+\delta)} \sum_{j=[n(Q_1-\delta)]}^{n(Q_1+\delta)} \psi_1(X_{1,\nu(1,j)}) > \epsilon \right) \\ &\leq P\left(\max_{1 \leq k \leq n(Q_1+\delta)} \sum_{j=[n(Q_1-\delta)]}^{k} \psi_1(X_{1,\nu(1,j)}) \geq n(Q_1+\delta)\epsilon \right) \\ &\leq \frac{C}{n^2} \quad \to 0 \ \text{ as } \ n \to \infty. \end{aligned} \tag{3.7.14}$$

where (3.7.14) follows from Kolmogorov's maximal inequality (Chung (1974) P. 116.) Similar argument shows that $G_{n,4} \xrightarrow{P} 0$. As for (ii)

$$\text{(ii)} \leq P(|\frac{N_1}{n} - Q_1| > \delta) \to 0 \ \text{ as } \ n \to \infty.$$

Thus combining , (3.7.12) and (3.7.13) the theorem follows since $\frac{N_i}{n} \to Q_i$ by Corollary 2.4.2. ∎

3.8 **Robustness of MHDE**

In this section we deal with the robustness of the MHDE. We describe the robustness properties through a study of the $\boldsymbol{\alpha}$-influence function and the breakdown point. We begin with the $\boldsymbol{\alpha}$-influence function. We will denote by $\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z})$ the contaminated model, *i.e.*

$$\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z}) = (\mathbf{1} - \boldsymbol{\alpha})\boldsymbol{F}(\cdot|\boldsymbol{\Xi}) + \boldsymbol{\alpha}\boldsymbol{U}_{\boldsymbol{z}} \tag{3.8.1}$$

where

$$\boldsymbol{U_z} = \begin{pmatrix} U_{Z_1} \\ U_{Z_2} \end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}.$$

U_{Z_i} are uniform densities on the interval $(Z_i - \epsilon, Z_i + \epsilon)$ where $\epsilon > 0$. Note that $f(\cdot|\boldsymbol{\theta}\ ,\ \alpha_1,\ Z_1)$ represents a $(1-\alpha_1)\%$ contamination with distant "outliers". Similarly, $g(\cdot|\boldsymbol{\eta}\ ,\ \alpha_2,\ Z_2)$ represents a $(1-\alpha_2)\%$ contamination with distant "outliers". Our first main result of this section is contained in the following theorem.

Theorem 3.8.1. Assume that the conditions of Theorem 3.4.1. hold. If $T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z})$ is unique for all $\boldsymbol{z}$, then

(i) $T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z})$ is a bounded continuous function of $\boldsymbol{z}$ such that

$$\lim_{\boldsymbol{z}\to\infty} T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z}) = \boldsymbol{\Xi}. \tag{3.8.2}$$

Furthermore,

(ii)

$$\lim_{\boldsymbol{\alpha}\to 0}(T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z}) - \boldsymbol{\Xi})\boldsymbol{\alpha}^{-1} = RF_T(\boldsymbol{z})$$

where

$$RF_T(\boldsymbol{z}) = \begin{pmatrix} (I_1(\boldsymbol{\theta}))^{-1}[\int_{\Re} U_{Z_1}(x)\psi_1(x|\boldsymbol{\theta})dx] \\ (I_2(\boldsymbol{\eta}))^{-1}[\int_{\Re} U_{Z_2}(x)\psi_2(x|\boldsymbol{\eta})dx] \end{pmatrix}.$$

Proof. Since $\boldsymbol{\alpha}$ are fixed, let us denote, using the uniqueness of $T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi},\ \boldsymbol{\alpha},\ \boldsymbol{z})$, the MHDF by $\boldsymbol{\Xi_z}$. To establish continuity, we need to show that

$$\lim_{\boldsymbol{z}\to\infty}(\boldsymbol{\Xi_z}) = \boldsymbol{\Xi}. \tag{3.8.3}$$

Suppose not then without loss of generality, by going to a subsequence if necessary, we may assume

$$\lim_{Z\to\infty} \boldsymbol{\Xi_z} = \boldsymbol{\Xi}_1. \tag{3.8.4}$$

Define

$$VHD_{\boldsymbol{z}}(\boldsymbol{t}) = \begin{pmatrix} 2 - 2\int f^{1/2}(x|\boldsymbol{t}_1) f^{1/2}(x|\boldsymbol{\theta}, \alpha_1, z_1)dx \\ 2 - 2\int g^{1/2}(x|\boldsymbol{t}_2) g^{1/2}(x|\boldsymbol{\eta}, \alpha_2, z_2)dx \end{pmatrix}. \tag{3.8.5}$$

Define

$$R_{\boldsymbol{z}}(\boldsymbol{t}) = \begin{pmatrix} (2 - 2\int f^{1/2}(x|\boldsymbol{t}_1)[(1-\alpha_1)^{1/2} f^{1/2}(x|\boldsymbol{\theta}) + \alpha_1^{1/2} U_{z_1}^{1/2}(x)]dx \\ (2 - 2\int g^{1/2}(x|\boldsymbol{t}_2)[(1-\alpha_2)^{1/2} g^{1/2}(x|\boldsymbol{\eta}) + \alpha_2^{1/2} U_{z_2}^{1/2}(x)]dx. \end{pmatrix}. \tag{3.8.6}$$

Now, using our arguments in Theorem 3.5.1 we can conclude that

$$\lim_{\boldsymbol{z}\to\infty} \sup_{\boldsymbol{t}\in\boldsymbol{\Theta}} |R_{\boldsymbol{z}}(\boldsymbol{t}) - VHD_{\boldsymbol{z}}(\boldsymbol{t})| = 0. \tag{3.8.7}$$

Now

$$\begin{aligned} \lim_{\boldsymbol{z}\to\infty} VHD_{\boldsymbol{z}}(\boldsymbol{\Xi}) &= \lim_{\boldsymbol{z}\to\infty} R_{\boldsymbol{z}}(\boldsymbol{\Xi}) \\ &= \begin{pmatrix} (1-\alpha_1)^{1/2} \int f^{1/2}(x|\boldsymbol{\theta}_1) f^{1/2}(x|\boldsymbol{\theta})dx \\ (1-\alpha_2)^{1/2} \int g^{1/2}(x|\boldsymbol{\eta}_1) g^{1/2}(x|\boldsymbol{\eta})dx \end{pmatrix} \\ &< \begin{pmatrix} (1-\alpha_1)^{1/2} \\ (1-\alpha_2)^{1/2} \end{pmatrix} \\ &= \lim_{\boldsymbol{z}\to\infty} VHD_{\boldsymbol{z}}(\boldsymbol{\Xi}). \end{aligned} \tag{3.8.8}$$

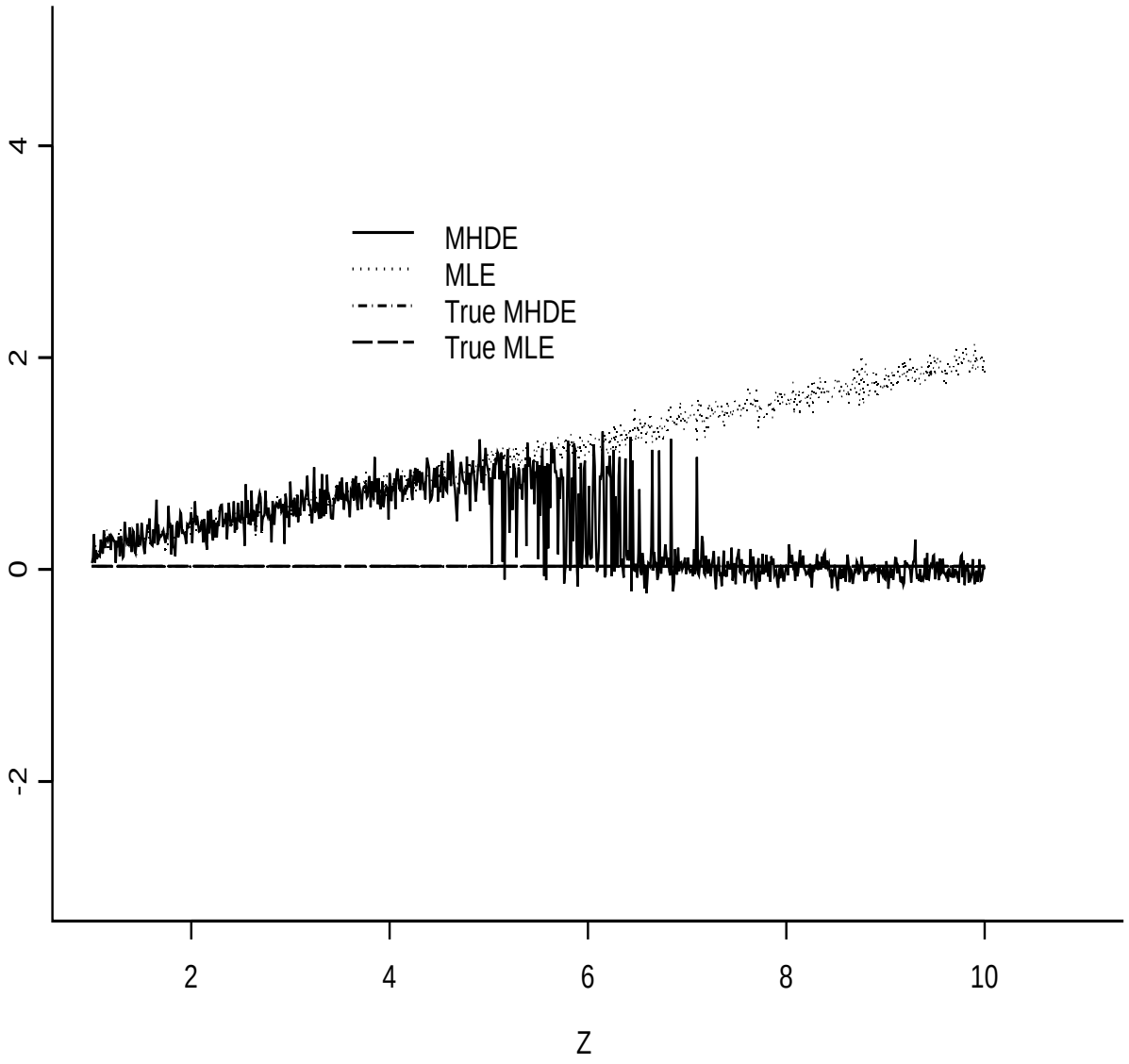
Furthermore, since $\boldsymbol{\Xi}_{\boldsymbol{z}}$ minimizes $VHD_{\boldsymbol{z}}(\boldsymbol{\Xi})$

$$\lim_{\boldsymbol{z}\to\infty} VHD_{\boldsymbol{z}}(\boldsymbol{\Xi}_{\boldsymbol{z}}) \leq \lim_{\boldsymbol{z}\to\infty} VHD_{\boldsymbol{z}}(\boldsymbol{\Xi}) \tag{3.8.9}$$

which is a contradiction to (3.8.2). Thus, $\boldsymbol{\Xi}_1 = \boldsymbol{\Xi}$.

Continuity of $\boldsymbol{\Xi}_{\boldsymbol{z}}$ follows from the Hellinger continuity of the functional T while the boundedness of $\boldsymbol{\Xi}_{\boldsymbol{z}}$ follows from (3.8.2). ∎

Remark 3.8.2. The functional T viewed as a function of z is called the α-influence curve.



Graph 3.1: The following graph represents the α-influence curve with 20% contamination.

Remark 3.8.3. Note that $\boldsymbol{F}(\cdot|\boldsymbol{\Xi}, \boldsymbol{\alpha},\ \boldsymbol{z})$ models an experiment where the observations are mixed with approximately $\boldsymbol{\alpha}\%$ gross errors located near $\boldsymbol{z}$. The above theorem compare $T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi}, \boldsymbol{\alpha},\ \boldsymbol{z}))$ with $T(\boldsymbol{F}(\cdot|\boldsymbol{\Xi})) = \boldsymbol{\Xi}$. ∎

Remark 3.8.4. The Graph 1 on the last page described the influence of z for the Hellinger distance estimator. Note also that the graphs for various $\boldsymbol{\alpha}$'s change dramatically, implying that the convergence of the $\boldsymbol{\alpha}$-influence curve need not be uniform in $\boldsymbol{z}$. To contrast our results with the MLE, we note that the α-influence curve of the MLE is unbounded since

$$|\hat{\boldsymbol{\theta}}_{MLE,\boldsymbol{z}}| \to \infty$$

as $\boldsymbol{z} \to \infty$. This can be seen from the Graph 2. ∎

We now move on to describe the breakdown point of MHDE.

Theorem 3.8.5. Define

$$\hat{\boldsymbol{B}} = \sup \boldsymbol{\gamma}(\boldsymbol{\Xi}|\boldsymbol{H}). \tag{3.8.10}$$

Assume that the MHDE lies to the interior of $\boldsymbol{\Theta}$. Let

$$\boldsymbol{B}^* = \lim_{\boldsymbol{z}^* \to \infty} \sup_{\boldsymbol{z} \geq \boldsymbol{z}^*} \boldsymbol{\gamma}(\boldsymbol{\Xi}_{\boldsymbol{z}}|\boldsymbol{H}). \tag{3.8.11}$$

If

$$\boldsymbol{\alpha} < (\hat{\boldsymbol{B}} - \boldsymbol{B}^*) \cdot (\hat{\boldsymbol{B}} - \boldsymbol{B}^*)' \cdot (\mathbf{1} + (\hat{\boldsymbol{B}} - \boldsymbol{B}^*) \cdot (\hat{\boldsymbol{B}} - \boldsymbol{B}^*)')^{-1} \tag{3.8.12}$$

then there is no sequence of contaminated families of distributions such that

$$|T(\boldsymbol{\Xi}_{\boldsymbol{z}}) - T(\boldsymbol{\Xi})| \to \infty \tag{3.8.13}$$

as $\boldsymbol{z} \to \infty$.

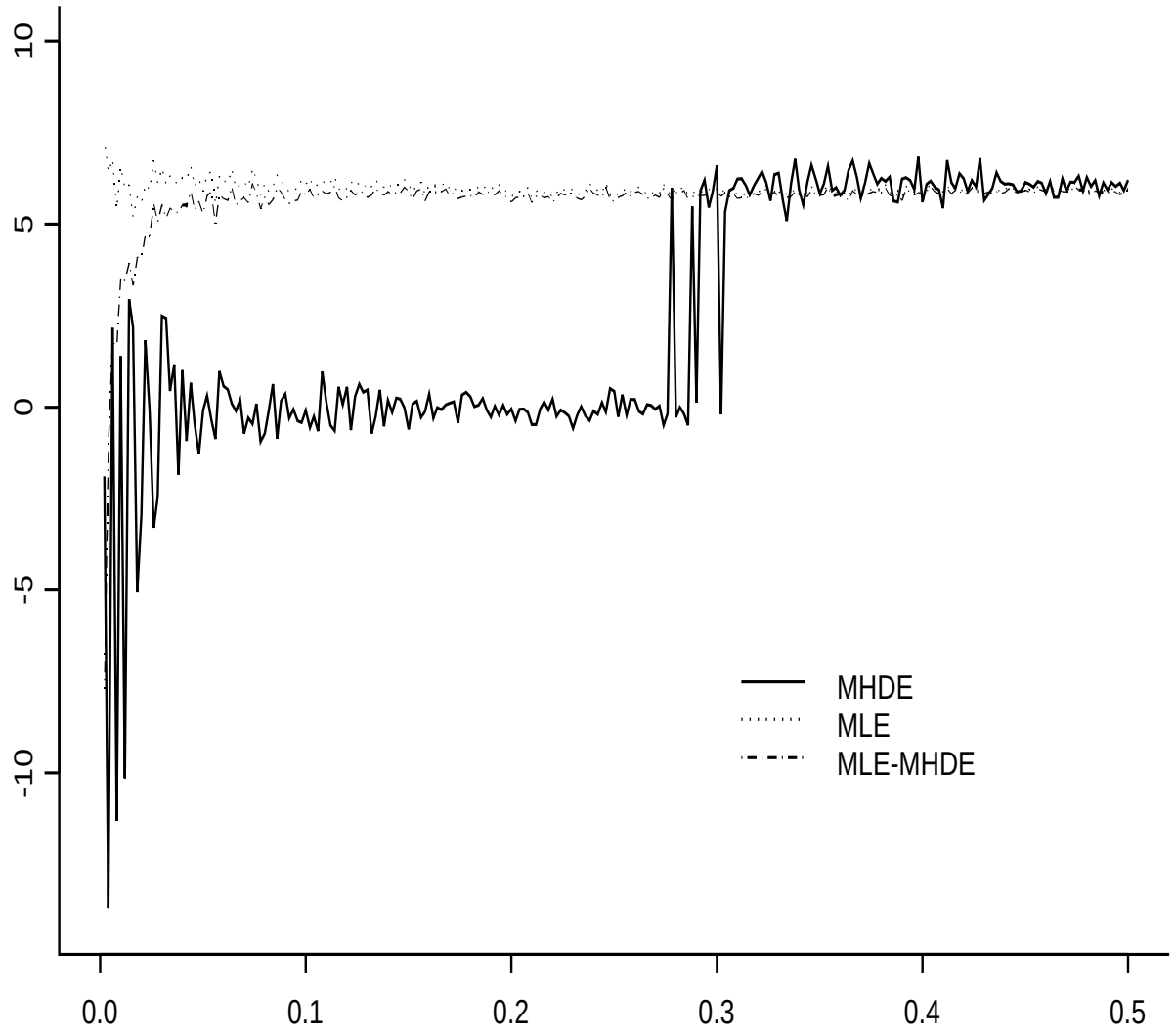
Proof. Assume that

$$|T(\boldsymbol{\Xi}_{\boldsymbol{z}}) - T(\boldsymbol{\Xi})| \to \infty$$

as $\boldsymbol{z} \to \infty$. Then there exist sequences $\boldsymbol{\Xi}_n$ such that $||\boldsymbol{\Xi}_n|| \to \infty$ and

$$\boldsymbol{\gamma}(\boldsymbol{\Xi}_n|\boldsymbol{H}_n) > \boldsymbol{\gamma}(\boldsymbol{\Xi}_n|\boldsymbol{H})$$

for infinitely many n. Now



Graph 3.2: The following graph represents the relative change in the estimator due to the change in the contamination proportion. α is the percentage of contamination.

$$\boldsymbol{\gamma}(\boldsymbol{\Xi}_n|\boldsymbol{H}) \;\geq\; ((\mathbf{1}-\boldsymbol{\alpha})^{1/2})\cdot\hat{\boldsymbol{B}}^{'} \tag{3.8.14}$$

and

$$\boldsymbol{\gamma}(\boldsymbol{\Xi}_n|\boldsymbol{H}_n) \leq ((\mathbf{1}-\boldsymbol{\alpha})^{1/2}) + \boldsymbol{\alpha}^{1/2}. \tag{3.8.15}$$

Furthermore,

$$\boldsymbol{\gamma}(\boldsymbol{\Xi}_n|\boldsymbol{H}_n) \leq \boldsymbol{B}^* + \boldsymbol{\delta} \tag{3.8.16}$$

for some $\delta > 0$. Hence it follows that

$$((\mathbf{1}-\boldsymbol{\alpha})^{1/2})\cdot\hat{\boldsymbol{B}}^{'} \leq ((\mathbf{1}-\boldsymbol{\alpha})^{1/2})\cdot\boldsymbol{B}^* + \boldsymbol{\alpha}^{1/2} \;\blacksquare \tag{3.8.17}$$

Remark 3.8.6. The above theorem provides the maximum amount of contamination that is possible without making the MHDE to breakdown. ∎

3.9 Simulation Results

In this section we deal with computation. Under the regularity condition for the Gaussian data $VHD(\cdot)$ function is differentiable with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Hence by partially differentiating equation (3.3.6) with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, we obtain the estimating equations for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ and setting them equal to 0. This leads to

$$\int \frac{\nabla f(x|\boldsymbol{\theta})}{\sqrt{f(x|\boldsymbol{\theta})}}\sqrt{h_{1,n}(x)}dx = 0 \tag{3.9.1}$$

and

$$\int \frac{\nabla g(x|\boldsymbol{\eta})}{\sqrt{g(x|\boldsymbol{\eta})}}\sqrt{h_{2,n}(x)}dx = 0, \tag{3.9.2}$$

where $\nabla f(x|\boldsymbol{\theta})$ represents the partial derivative with respect to $\boldsymbol{\theta}$ and $\nabla g(x|\boldsymbol{\eta})$ represents the partial derivative with respect to $\boldsymbol{\eta}$. Let us denote by $s_{1,\boldsymbol{\theta}}(x) = (f(x|\boldsymbol{\theta}))^{1/2}$ and $s_{2,\boldsymbol{\eta}}(x) = (g(x|\boldsymbol{\eta}))^{1/2}$. The estimation functions can be rewritten as:

$$\int \nabla s_{1,\boldsymbol{\theta}}(x)(h_{1,n}(x))^{1/2}dx = 0 \tag{3.9.3}$$

and

$$\int \nabla s_{2,\boldsymbol{\eta}}(x)(h_{2,n}(x))^{1/2}dx = 0, \tag{3.9.4}$$

where $\nabla s_{1,\boldsymbol{\theta}}(x)$ represents the partial derivative with respect to $\boldsymbol{\theta}$ and $\nabla s_{2,\boldsymbol{\eta}}(x)$ represents the partial derivative with respect to $\boldsymbol{\eta}$. Evaluating the solution to (3.9.3) and (3.9.4) yields MHDE. We will describe a new algorithm called one step Monte Carlo approximation for obtaining the MHDE in the following section.

3.9.1 Estimation of MHDE for the Normal Model

In this section we will introduce a new numerical method to solve equations (3.9.3) and (3.9.4) for $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Assuming $f_{\boldsymbol{\theta}}(x)$ to be the Normal distribution with mean μ and variance σ^2, we will describe the Newton-Raphson method and introduce a new algorithm "one-step Monte Carlo approximation method".

Newton-Raphson Method

Beran (1977) applied the Newton-Raphson method for solving MHDE. Let us describe the algorithm he carried out briefly. Let us focus on solving equation (3.9.3) now. In our case $\boldsymbol{\theta} = (\mu, \sigma)'$, so we can rewrite (3.9.3) as the following two equations:

$$\int \nabla s^{(1)}_{\mu,\sigma}(x)(h_{1,n}(x))^{1/2}dx = 0 \tag{3.9.5}$$

and

$$\int \nabla s^{(2)}_{\mu,\sigma}(x)(h_{1,n}(x))^{1/2}dx = 0. \tag{3.9.6}$$

We have two equations (3.9.5) and (3.9.6) and have to solve for two unknown parameters $(\mu, \sigma)'$. We can apply multi-variate Newton-Raphson method to obtain an iterative algorithm as follows:

$$\begin{pmatrix} \hat{\mu}^{(k+1)} \\ \hat{\sigma}^{(k+1)} \end{pmatrix} = \begin{pmatrix} \hat{\mu}^{(k)} \\ \hat{\sigma}^{(k)} \end{pmatrix} - D_N \begin{pmatrix} f_1(\boldsymbol{\theta}^{(k)}) \\ f_2(\boldsymbol{\theta}^{(k)}) \end{pmatrix}. \tag{3.9.7}$$

Where D_N is a matrix defined as follows:

$$D_N = \begin{pmatrix} \int \nabla s^{(1,1)}_{\mu,\sigma}(x)(h_{1,n}(x))^{1/2}dx & \int \nabla s^{(1,2)}_{\mu,\sigma}(x)(h_{1,n}(x))^{1/2}dx \\ \int \nabla s^{(2,1)}_{\mu,\sigma}(x)(h_{1,n}(x))^{1/2}dx & \int \nabla s^{(2,2)}_{\mu,\sigma}(x)(h_{1,n}(x))^{1/2}dx \end{pmatrix}^{-1}.$$

Following the standard procedure for iterative algorithm, we can use update formula (3.9.7) to obtain the MHDE.

One Step Monte Carlo Approximation Method

Recall that from (3.2.7), finding the MHDE of $\boldsymbol{\theta}$ is equivalent to finding the $\boldsymbol{\theta}$ that maximizes the following:

$$\int (f_{\boldsymbol{\theta}}(x))^{1/2}(h_{i,n}(x))^{1/2}dx = \int \{\frac{(f_{\boldsymbol{\theta}}(x))^{1/2}}{(h_{i,n}(x))^{1/2}}\}(h_{i,n}(x))dx.$$

using strong law of large numbers, the above integral can be approximated by

$$\frac{1}{M}\sum_{j=1}^{M}(\frac{f_{\boldsymbol{\theta}}(y_{i,j})}{h_{i,n}(y_{i,j})})^{1/2}, \tag{3.9.8}$$

where $y_{i,j} \sim h_{i,n}$ and M is the number of the Monte Carlo samples . We need to find the value of $\boldsymbol{\theta}$ that maximizes (3.9.8). When the underlying distribution of f_θ is $N(\mu, \sigma^2)$, (3.9.8) becomes the following :

$$\frac{1}{M}\sum_{j=1}^{M}\frac{w_{i,j}}{\sqrt[4]{2\pi\sigma^2}}\exp(-\frac{1}{4\sigma^2}(y_{i,j}-\mu)^2), \quad w_{i,j} = \frac{1}{\sqrt{h_{i,n}(y_{i,j})}}. \tag{3.9.9}$$

Taking the partial derivative of (3.9.9) with respect to μ and σ^2 and setting them to 0, we obtain the following recursive equations for μ and σ^2, *viz.*,

$$\hat{\mu}_{(m+1)} = \frac{\sum_{j=1}^{M} w_{i,j}\exp(-\frac{1}{4\hat{\sigma}^2_{(m)}}(y_{i,j}-\hat{\mu}_{(m)})^2)y_{i,j}}{\sum_{j=1}^{M} w_{i,j}\exp(-\frac{1}{4\hat{\sigma}^2_{(m)}}(y_{i,j}-\hat{\mu}_{(m)})^2)} \tag{3.9.10}$$

and

$$\hat{\sigma}^2_{(m+1)} = \frac{\sum_{j=1}^{M} w_{i,j}\exp(-\frac{1}{4\hat{\sigma}^2_{(m)}}(y_{i,j}-\hat{\mu}_{(m)})^2)(y_{i,j}-\hat{\mu}_{(m)})^2}{\sum_{j=1}^{M} w_{j}\exp(-\frac{1}{4\hat{\sigma}^2_{(m)}}(y_{i,j}-\hat{\mu}_{(m)})^2)}. \tag{3.9.11}$$

If the kernel K is a standard normal density, we have

$$h_{i,N_i}(x) \;=\; \frac{1}{N_i c_n}\sum_{l=1}^{N_i} K\left\{\frac{x - X_{i,\nu(i,l)}}{c_n}\right\}$$

$$
\begin{aligned}
&= \frac{1}{N_i c_n}\sum_{l=1}^{N_i}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{x-X_{i,\nu(i,l)}}{c_n})^2)\\
&= \frac{1}{N_i}\sum_{l=1}^{N_i}\frac{1}{\sqrt{2\pi c_n^2}}\exp(-\frac{1}{2}(\frac{x-X_{i,\nu(i,l)}}{c_n})^2)\\
&= \frac{1}{N_i}\sum_{l=1}^{N_i}\phi(X_{i,\nu(i,l)}, c_n^2),
\end{aligned}
$$

where ϕ is the normal density with mean equal to $X_{i,\nu(i,l)}$ and variance equal to c_n^2. Thus, $h_{i,n}(x)$ is a mixture of normal densities with mixing proportion $\frac{1}{N_i}$. Therefore, in the first step of the algorithm, we generate a random variable $y_{i,j}$ which has the distribution $h_{i,n}(x)$. Note that in the update formulas (3.9.10) and (3.9.11), $w_{i,j} = \frac{1}{\sqrt{h_{i,N_i}(y_{i,j})}}$ which depends on the choice of the kernel density K. When K is chosen to be standard normal density, we have

$$
w_{i,j} = \left[\frac{1}{N_i c_n}\sum_{l=1}^{N_i}\frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{y_{i,j}-X_{i,\nu(i,l)}}{c_n})^2)\right]^{-1/2}.
$$

If $K(\cdot)$ were Epanechnikov kernel the weight reduces to

$$
w_{i,j} = \left[\frac{1}{N_i c_n}\sum_{l=1}^{N_i}0.75(1-(\frac{y_{i,j}-X_{i,\nu(i,l)}}{c_n})^2)\right]^{-1/2}, \quad \left|\frac{y_{i,j}-X_{i,\nu(i,l)}}{c_n}\right| \le 1.
$$

Evaluating w_{ij} for Epanechnikov Kernel:

We begin with the following claim that will help us evaluate Epanechnikov kernel.

Claim: Let U_1, U_2, U_3 and U_4 are i.i.d. uniform $[0,1]$ random variables, then for $a > 1$, $U_1^{1/a}U_2$ has density $\frac{a}{a-1}(1-x^{a-1})I_{[0\le x\le 1]}$ also $(-U_3^{1/a})U_4$ has density

$$
\frac{a}{a-1}(1-x^{a-1})I_{[-1\le x\le 0]}.
$$

Proof. We first calculate the density function of $U_1^{1/a}U_2$. Consider the following:

$$
\begin{aligned}
P(U_1^{1/a}U_2 \le x) &= E[P(U_1^{1/a}U_2 \le x|U_2)]\\
&= E\{\frac{x^a}{U_2^a}I_{(U_2\ge x)} + I_{(x>U_2}\}\\
&= x^a E[\frac{1}{U_2^a}I_{(U_2\ge x)}] + P(U_2 \le x)\\
&= x^a[\frac{1}{1-a} - \frac{x^{1-a}}{1-a}] + x.
\end{aligned}
$$

Differentiating the above equation with respect to x, we get the density for the above claim. Similar argument will give the density function of $(-U_3^{1/a})U_4$, and hence the claim follows.

Remark 3.9.1 If $a = 3$ and randomly choosing between $U^{1/3}U_2$ and $(-U_3^{1/3})U_4$ with equal probability, we obtain a random variable whose density is the Epanechikov density, *i.e.* if X is a random variable defined by

$$X = \begin{cases} U^{1/3}U_2 & \text{with prob } 1/2 \\ -U_3^{1/3}U_4 & \text{with prob } 1/2 \end{cases}$$

then the density of X is given by the Epanechikov density. ∎

The one step Monte-Carlo approximation algorithm can be described as follows:

1. Generate random variables for each data point from the kernel density with mean $X_{i,\nu(i,k)}$ and variance c_n^2. Choose one of then with equal probability $(\frac{1}{N_i})$ and retain it. Repeat M times. Using the initial values for μ and σ, *viz.*, $\hat{\mu}^{(0)}$ =median$\{X_{i,\nu(i,l)}\}$ and $\hat{\sigma}^{(0)} = (0.674^{-1})$median$\{|X_{i,\nu(i,l)} - \hat{\mu}^{(0)}|\}$.

2. Obtain the updates using (3.9.10) and (3.9.11).

3. When $|\hat{\mu}_{(m+1)} - \hat{\mu}_{(m)}| < \epsilon$ and $|\hat{\sigma}_{(m+1)} - \hat{\sigma}_{(m)}| < \epsilon$ for small ϵ, say 10^{-6} then stop; else go to step 2.

We will present simulation results which were carried out using SAS software. The simulations compares the efficiency between MHDE and MLE with outlier, and incorporating the randomized play the winner design with two treatments. We start with an urn containing 5 ball of each type and assume treatment A has success probability p_1 and treatment B has success probability p_2. Let N_1^0 denote the number of type A balls in the urn at the beginning of the trial and N_2^0 denotes the number of type B balls in the urn at the beginning of the trial. Let N_1^i and N_2^i denote the number of type A and B ball, after the i^{th} patient's response is observed and the urn has been updated. The simulation procedure works as follows:

1. Generate a uniform(0,1) random variable, say u_1.

2. If $u_1 > \frac{N_1^i}{(N_1^i+N_2^i)}$, assign patient $(i+1)$th to treatment B and generate a N(5,3) random variable, representing the secondary variable. Otherwise, assign patient to treatment A and generate a N(0,1) random variable, representing the secondary variable.

3. Generate a uniform(0,1) random variable, say u_2. If the treatment assignment in step 2 is A and $u_2 < P_1$, then call this treatment a success, and add one type A ball to the urn. Otherwise, add a type B ball. If the treatment assignment in step 2 is B, we will update the urn similarly.

4. Repeat steps 1, 2 and 3 for 30 times to represent a sample of size 30.

5. Calculate MHDE and MLE for both treatments.

6. Repeat the above steps 1000 times.

To illustrate the robustness property of the MHDE, we will change some of the treatment A's secondary variable to outliers; say a N(2,1) random variable. We obtain the proportion of times that MHDE and MLE fall in the true confidence interval; *i.e.* for treatment A, the true confidence interval is $(0 - 1.96\frac{1}{\sqrt{N_1}}, 0 + 1.96\frac{1}{\sqrt{N_1}})$, where N_1 is the number of patients allocated to treatment A ; similarly for treatment B. Table 3.1 contains the results assuming $p_1 = p_2 = 0.5$. The numbers in the bold font represent the cases that the proportion of times that MHDE falling into true confidence interval is higher than MLE. From Table 3.1, we can see that as the values of the outliers become larger, the probability of MLE falling into the true confidence interval is much smaller than MHDE. Increasing the number of outliers make the situation even worse.

Instead of assuming $p_1 = p_2 = 0.5$, the next simulation assumes $p_1 = 0.8$ and $p_2 = 0.2$, which means treatment A has a higher success probability than treatment B. Table 3.2 contains the results. From Table 3.2, we notice that the effect of outliers is similar to Table 3.1, there is a decreasing in the proportion of data which contains outliers for MLE. Due to more patients being assigned to treatment A, the results in Table 3.2 are not as dramatic as in Table 3.1.

Table 3.1: Results for the RPWD. The probability that the MHDE and MLE fall in the true confidence intervals with outlier from treatment A. The number of outlier equals 1, 2 and 3 with scale from N(2,1) to N(6,1). Significant level=0.05, $N_1(0) = N_2(0) = 5$, $\alpha = 1$, $n = 30$, 1000 simulations and $p_1 = p_2 = 0.5$

outlier		N(2,1)		N(3,1)		N(4,1)		N(5,1)		N(6,1)	
		A	B	A	B	A	B	A	B	A	B
1	MHDE	0.893	0.932	**0.895**	0.945	**0.895**	0.934	**0.912**	0.940	**0.923**	0.935
	MLE	0.923	0.936	**0.876**	0.954	**0.816**	0.944	**0.727**	0.938	**0.661**	0.940
2	MHDE	**0.795**	0.942	**0.751**	0.926	**0.740**	0.949	**0.809**	0.941	**0.878**	0.930
	MLE	**0.792**	0.947	**0.669**	0.942	**0.447**	0.953	**0.293**	0.957	**0.135**	0.935
3	MHDE	**0.671**	0.924	**0.482**	0.937	**0.391**	0.936	**0.523**	0.952	**0.621**	0.940
	MLE	**0.639**	0.934	**0.374**	0.940	**0.130**	0.943	**0.040**	0.963	**0.005**	0.953

Table 3.2: Results for the RPWD. The probability that the MHDE and MLE fall in the true confidence intervals with outlier from treatment A. The number of outlier equals 1, 2 and 3 with scale from N(2,1) to N(6,1). Significant level=0.05, $N_1(0) = N_2(0) = 5$, $\alpha = 1$, $n = 30$, 1000 simulations and $p_1 = 0.8, p_2 = 0.5$

outlier		N(2,1)		N(3,1)		N(4,1)		N(5,1)		N(6,1)	
		A	B	A	B	A	B	A	B	A	B
1	MHDE	**0.947**	0.931	0.909	0.949	**0.890**	0.932	**0.923**	0.927	**0.932**	0.935
	MLE	**0.941**0.942		0.905	0.955	**0.854**	0.948	**0.790**	0.949	**0.741**	0.943
2	MHDE	**0.877**	0.927	**0.836**	0.945	**0.816**	0.937	**0.845**	0.942	**0.908**	0.922
	MLE	**0.876**	0.953	**0.771**	0.951	**0.571**	0.954	**0.386**	0.956	**0.211**	0.943
3	MHDE	**0.747**	0.923	**0.561**	0.932	**0.585**	0.939	**0.918**	0.930	**0.925**	0.926
	MLE	**0.730**	0.933	**0.422**	0.948	**0.232**	0.954	**0.786**	0.952	**0.697**	0.948

3.10 Data Analysis

In this section, we will provide a robust and efficient data analysis for the Fluoxetine trial data that was introduced in the chapter 1. We obtain the MHDE and MLE of the primary outcome, the difference of $HAMD_{17}$ between baseline and final visit for all four strata . The data are provided in chapter 1.

As we can see the from the following table, due to the robustness property of MHDE, the standard deviation estimators are much smaller than the one from the MLE. Also we can have a better idea about the performance of the treatments in different groups by comparing the mean using the MHDE. The absolute change in $HAMD_{17}$ between the baseline and final visits seem to be substantially higher for treatment 1 than treatment 0 and the change is more pronounced in strata 1 than in strata 0.

In the next chapter we will provide robust semi-parametric confidence intervals for these parameters using the bootstrap methodology.

Table 3.3: Analysis results for the Fluoxetine trial data

Strata	Strata=1 Treatment=1	Strata=1 Treatment=0	Strata=0 Treatment=1	Strata=0 Treatment=0
MHDE	-10.88 (5.04)	-3.92 (5.73)	-10.14 (6.51)	-9.59 (5.96)
MLE	-11.20 (5.97)	-5.71 (7.68)	-10.81 (7.13)	-8.62 (6.88)

3.11 Concluding Comments

In this chapter we introduced MHDE for RPWD. We studied the robustness and asymptotic properties of our estimates. Our estimates were fully efficient at the true model and possessed a stability property. The breakdown point of out estimates were found to be 1/2.

Further work on developing robust goodness of fit procedures is being investigated.

SAS Macro

```
**********************************************************************
This SAS macro implementing the one step Monte-Carlo approximation
method for calculating the Minimum Hellinger Distance Estimator.
One can just creat a SAS data set with name one and variable name
x to prosecute this program and obtain the Minimum Hellinger Distance
Estimator of mean and standard deviation for variable x.
**********************************************************************;
options nodate pageno=1 formdlim=-;
%let cn=0.7; *the bandwidth;
data yvar;
x=.;
xx=.;
run;
proc means data=one noprint;
var x;
output out=oneout n=n;
run;
data oneout;
set oneout;
call symput('n',n);
run;
%macro one;
%do i=1 %to 500;
data two;
set one;
u1=ranuni(0);
```

```
u2=ranuni(0);
u3=ranuni(0);
u4=ranuni(0);
a=u1**(1/3)*u2;
b=-(u3**(1/3)*u4);
u=ranuni(0);
if u > 0.5 then y=a;
else y=b;
xx=&cn*y+x;
keep x xx;
run;
data three;
choice=ceil(ranuni(0)*&n);
set two point=choice;
output;
stop;
run;
proc append base=yvar data=three;
run;
%end;
%mend;
%one;
data one1;
set one;
do i=1 to 500;
xxx=x;
output;
```

```
end;
keep i x;
run;
proc sort data=one1;
by i;
run;
data yvar;
set yvar ;
if xx=. then delete;
run;
data yvar1;
set yvar;
do j=1 to &n;
y=xx;
output;
end;
keep y;
run;
data one2;
merge yvar1 one1;
run;
data one3;
set one2;
k=((y-x)/&cn);
g=(3/4)*(1-k**2);
if abs(k)> 1 then g=0;
else g=g;
```

```
run;
proc means data=one3 noprint;
by i;
var g;
output out=one4 sum=sum;
run;
data one5;
merge yvar one4;
keep xx sum;
run;
data one5;
set one5;
gn=sum/(&n*&cn);
run;
proc means data=one noprint;
var x;
output out=oone mean=mean median=mu max=max std=std min=min;
run;
proc transpose data=oone out=otwo;
run;
data three;
set otwo;
do i=1 to &n;
z=_name_;
end;
run;
data three;
```

```
set three;
if z=”mu”;
keep col1 z;
run;
data three;
set three;
do j=1 to &n;
y=col1;
output;
end;
run;
data four;
merge one three;
keep x y;
run;
data four;
set four;
z=abs(x-y)/0.674;
run;
proc means data=four noprint;
var z;
output out=ofour median=sn;
run;
%macro two;
data one6;
set one5;
num1=(1/sqrt(gn))*exp(-(1/(4*&sn**2))*(xx-&mu)**2)*xx;
```

```
dem=(1/sqrt(gn))*exp(-(1/(4*&sn**2))*(xx-&mu)**2);
num2=(1/sqrt(gn))*exp(-(1/(4*&sn**2))*(xx-&mu)**2)*(xx-&mu)**2;
run;
proc means data=one6 noprint;
var num1 dem num2;
output out=one7 sum=num1 dem num2;
run;
data one8;
set one7;
muhat=num1/dem;
snhat=sqrt(num2/dem);
run;
data start;
merge one8 start;
delta1=&mu-muhat;
delta2=&sn-snhat;
keep delta1 delta2 muhat snhat o;
run;
data start;
set start;
mu=muhat;
sn=snhat;
keep delta1 delta2 mu sn o;
run;
data start;
set start;
call symput('sn',sn);
```

```
call symput('mu',mu);
output;
run;
%mend;
%macro update;
%two;
data start;
set start;
if (abs(delta1) le 0.000001 and abs(delta2) le 0.000001) then o=1;
else o=0;
call symput('o',o);
run;
%mend update;
%macro three;
data start ;
merge oone ofour;
keep mu sn;
run;
data start ;
set start;
call symput('sn',sn);
call symput('mu',mu);
run;
%update;
%do i=1 %to 30;
%if (&o=1)%then %goto exit;
%if (&o=0) %then %goto update;
```

```
%update:
%update;
%end;
data start;
set start;
o=symget(&o);
run;
%exit;
%mend;
%three;
proc print data=start;
run;
```

References:

1. Beran, R. (1977). Minimum Hellinger distance estimates for parametric models. The Annals of Statistics, 5, 445-463.

2. Chung, K. L. (1974). A course in probability theory. Academic Press.

3. Devroye, Luc. (1987). A Course in Density Estimation. Birkhauser Boston.

4. Lindsay, B. G. (1994). Efficiency verses robustness: the case for minimum Hellinger distance and related methods. The Annals of Statistics, 22, 1081-1114.

5. Markatou, M., Basu, A. and Lindsay, B.. (1997). Weighted likelihood estimating equations: the discrete case with applications to logistic regression. Robust statistics and data analysis, II. Journal of Statistical Planing and Inference, 57, 215-232.

6. Melfi, V. F. and Page C.. (2000). Estimation after adaptive allocation. Journal of Statistical Planing and Inference, 87, 353-363.

7. Simpson, D. G. (1987). Minimum Hellinger distance estimation for the analysis of count data. Journal of the American Statistical Association, 82, 802-807.

8. Simpson, D. G. (1989). Hellinger deviance tests: efficiency breakdown points, and examples. Journal of the American Statistical Association, 84, 107-113.

9. Sriram, T. N. and Vidyashankar, A. N. (2000). Minimum Hellinger distance estimation for supercritical Galton-Watson processes. Statistics and Probability Letters, 50, 331-342.

10. Tamura, R. N. and Boos, D. D..(1986). Minimum Hellinger distance estimation for multivariate location and covariate. Journal of the American Statistical Association, 81, 223-229.

Chapter 4

Bootstrap Confidence Intervals for Minimum Hellinger Distance Estimators

4.1 Introduction

In the previous chapter, we introduced the MHDE for the randomized play the winner design. We developed the asymptotic theory and showed that one can construct confidence intervals using the limiting asymptotic distribution. From our results in the previous chapter, we observed that the limiting covariance matrix is independent of the design. However from the point of view of prediction, it will be important to incorporate the design features in the analysis especially in a small sample setting. To remedy this issue, we propose a bootstrap approach for constructing the confidence intervals for the parameters. As can be seen from our simulation results, we notice that the proposed bootstrap procedure produces approximately correct sized confidence intervals with smaller length compared to the results from the large sample theory.

The bootstrap procedure suggested here is a modification of the available parametric bootstrap theory for the i.i.d. data and involves bootstrapping of the design. Bootstrapping of the design for estimating the design related parameters has been investigated by Rosenberger and Hu (1999). Incidentally, no work has been performed to study the confidence intervals for the design parameters. We address this in this chapter. The parametric bootstrap methodology for i.i.d. data has a long history, as can be seen from the works of Efron (1985) and Diciccio and Romano (1988). The robustness aspects of such a theory has not been investigated thus far. In the next section, we develop the notations and methodology parametric for the bootstrap of i.i.d. data using MHDE and study the properties of the resulting estimators. In section 3 we develop the parametric bootstrap procedure for data

from RPWD and in section 4 we study their properties. Section 5 is devoted to our simulation results. In the rest of the chapter, we will assume the appropriate conditions from chapter 3 hold.

4.2 Parametric Bootstrap for i.i.d. data using MHDE

Let $\{X_n, n \geq 1\}$ be i.i.d. with density $f(x|\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta \subset \Re^p$. Let $\boldsymbol{\theta}_n$ be the MHDE of $\boldsymbol{\theta}$ based on the estimate of the density g_n, *viz.*

$$\boldsymbol{\theta}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} HD^2(f(\cdot|\boldsymbol{\theta}), g_n). \tag{4.2.1}$$

The parametric bootstrap technique involves obtaining i.i.d. random variables $\{\mathcal{X}_i^*, 1 \leq i \leq n\}$ where

$$\mathcal{X}_i^* \sim f(x|\boldsymbol{\theta}_n). \tag{4.2.2}$$

The bootstrap version of the density estimator using $\{\mathcal{X}_1^*, \ldots, \mathcal{X}_n^*\}$ is given by

$$h_n^*(x) = \frac{1}{nc_n}\sum_{i=1}^{n} K(\frac{x - \mathcal{X}_i^*}{c_n}) \tag{4.2.3}$$

where $K(\cdot)$ is a bounded density.

Remark 4.2.1 Note that c_n is not changed into c_n^*. $h_n^*(\cdot)$ has been used previously by Hall *et. al* (1992) to find an optimal bandwidth in density estimation problem. Our goals here are however different.

The bootstrap version of the MHDE denoted by (BMHDE) is given by

$$\boldsymbol{\theta}_n^* = \arg\min_{\boldsymbol{\theta}\in\Theta} HD^2(f(x|\boldsymbol{\theta}), h_n^*).$$

Using conditions of Theorem 3.4.1, $\boldsymbol{\theta}_n^*$ exists and is unique. Using the results from Chapter 3, to prove the consistency of $\boldsymbol{\theta}_n^*$, we need to establish the L_1-consistency of $h_n^*(\cdot)$. We study various strong convergence theorems concerning $h_n^*(\cdot)$ in our next section. The proofs of the strong consistency are based on a version of strong law of large numbers due to Chung (1974) and has also been used by Bozorgnia *et. al* (1997).

4.3 L_1-Convergence of Bootstrapped versions of Kernel Density Estimators for i.i.d. Data

In this section, we establish the strong pointwise consistency and strong L_1 consistency of $h_n^*(\cdot)$ defined in (4.2.3). We first state the following law of large numbers due to Chung for array of i.i.d. random variables

Theorem 4.3.1. Let $\{x_{n,i}\}$ be an array of rowwise independent random variables. Let $\Psi : \mathcal{R} \to \mathcal{R}^+$ satisfying $\Psi(t)$ is even and continuous such that $(|t|^r\Psi(|t|)$ is non-decreasing and $t^{-r+\delta}\Psi(t)$ is non-increasing. Assume that $EX_{n,i} = 0$ for all i and n and that

$$\sum_{n\geq 1}\sum_{i=1}^{n}(\Psi(a_n))^{-1}E(\Psi(|x_{n,i}|)) < \infty$$

and

$$\sum_{n\geq 1}\left[\sum_{i=1}^{n}E\left(\left|\frac{x_{n,i}}{a_n}\right|^p\right)\right]^{kr} < \infty$$

for some $1 < p \leq 2$ and some positive integer k. Then

$$\lim_{n\to\infty} a_n^{-1}\sum_{i=1}^{n}x_{n,i} = 0 \quad a.s. \ \blacksquare$$

Our next theorem establishes the strong pointwise consistency and strong L_1-consistency.

Theorem 4.3.2. Assume $c_n \to 0$, $nc_n \to \infty$ as $n \to \infty$ Then, for almost all x (with respect to the Lebesgue measure)

$$\lim_{n\to\infty} h_n^*(x) = f(x|\boldsymbol{\theta}). \ \text{ a.s.} \tag{4.3.1}$$

Furthermore,

$$\lim_{n\to\infty}\int_{\Re}|h_n^*(x) - f(x|\boldsymbol{\theta})|dx = 0. \tag{4.3.2}$$

Proof. Fix an x in a set of positive Lebesgue measure and note that

$$|h_n^*(x) - f(x|\boldsymbol{\theta})| \leq |h_n^*(x) - h_n(x)| + |h_n(x) - f(x|\theta). \tag{4.3.3}$$

From Theorem 1 of Devroy (1987), the second term on the RHS of (4.3.3) converges to 0 as $n \to \infty$. Thus we only have to establish the convergence of the first term to 0 as $n \to \infty$. As

for the first term, note first that $K(\cdot)$ is bounded also

$$|h_n^*(x) - h_n(x)| \leq |h_n^*(x) - E^* h_n^*(x)| + |E^* h_n^*(x) - h_n(x)|.$$

Now applying theorem 4.3.1 with $\Psi(t) = |t|^5$ and $r = 5$ and $a_n = nc_n$, it follows that the first term converges to 0 $a.s.$ as $n \to \infty$. The convergence of the second last term to 0 follows from law of large numbers for double arrays of random variables. Finally (4.3.3) follows using Glick's theorem. ∎

4.4 Strong Consistency of BMHDE

In this section we prove the consistency of BMHDE.

Theorem 4.4.1. $\hat{\boldsymbol{\theta}}_{BMHDE}$ is a strongly consistent estimator of $\boldsymbol{\theta}_0$.

Proof. From Theorem 4.3.2, h_n^* is a strongly L_1-consistent estimator of $f(x|\boldsymbol{\theta}_0)$. Hence by Theorem 3.6.1, $T(h_n^*)$ converges to $\boldsymbol{\theta}_0$ with probability one as $n \to \infty$. ∎

Let $\{\boldsymbol{\theta}_n^*(j),\ \ j = 1, \ldots, B\}$ denote BMHDE based on B bootstrap samples. One can then combine these bootstrap estimates to obtain an approximation to the sampling density of $\boldsymbol{\theta}_n$, using $a_n(x)$ the following bootstrap kernel density estimator, *viz.*

$$a_{n,B}^*(x) = \frac{1}{Bc_B} \sum_{j=1}^{B} K\left(\frac{x - \boldsymbol{\theta}_n^*(j)}{c_B}\right) \tag{4.4.1}$$

where $K(\cdot)$ is a p-dimensional kernel density. Note that by Devroye's Theorem 1, as $B \to \infty$,

$$a_{n,B}^*(x) \to a_n^*(x) \tag{4.4.2}$$

and

$$\int |a_{n,B}^*(x) - a_n^*(x)| \to 0. \tag{4.4.3}$$

4.5 BMHDE for RPWD

In this section we define the BMHDE for RPWD and establish the consistency of the BMHDE. In the process we will also establish the strong L_1-consistency of our bootstrap versions of the density estimators.

Resampling of data in the context of RPWD has not been investigated in the literature. We begin with the methodology for bootstrapping data from RPWD. Let $p_i(q_i),\ i = 1, 2$ denote the estimates probability of success (failure) on treatments 1 and 2 respectively based on n observations. Note that

$$\hat{p}_i(n) = \frac{1}{N_i(n)} \sum_{j=1}^{n} X_j I_{j,i}, \quad i = 1, 2.$$

We first generate the design using the initial urn composition for the data and success probability vector $(\hat{p}_1(n), \hat{p}_2(n))$. Let $\{T_k^*(j), k \geq 1\}$ denote the collection of treatment indicators for the jth bootstrap sample. Now, given $\{T_i^*, 1 \leq i \leq n\}$ we associate x bootstrap response $\mathcal{X}_1^*$ where

$$\mathcal{X}_i^* \sim \begin{cases} f(\cdot|\boldsymbol{\theta}_n) & \text{if} \quad T_i^* = 1 \\ g(\cdot|\boldsymbol{\eta}_n) & \text{if} \quad T_i^* = 2. \end{cases} \tag{4.5.1}$$

Let us denote by $H_n^* = (f(\cdot|\boldsymbol{\theta}_n),\ g(\cdot|\boldsymbol{\eta}_n))'$.

Let us denote by $\nu^*(i, j)$ the index of the jth patient receiving the ith treatment in the bootstrap sample. Then the data from the bootstrap is given by $\{X^*_{i,\ \nu^*(i,j)}, j \in \mathcal{A}_n^*(i), i = 1,\ 2\}$, where

$$\mathcal{A}_n^*(i) = \{1 \leq j \leq n | T_j^* = i\}. \tag{4.5.2}$$

Let

$$\mathcal{A}^*(i) = \{j | T_j^* = i\}.$$

We now state an extension of Melfi's Theorem under our bootstrap setting.

Theorem 4.5.1. $\{\mathcal{X}^*_{i,\nu^*(i,j)}, \quad j \in \mathcal{A}^*_{(i)}\}$ are i.i.d. with distribution $h^*_{n,i}$ where $h^*_{n,i}$ is the ith component of H_n^*.

Proof. The proof follows exactly as in Melfi and Page(2000) by conditioning on the treatment assignment and appealing to Kolmogorov's consistency theorem. ∎

Now, following analogous arguments as in Chapter 3, we define the BMHDE of $\boldsymbol{\Xi}$ based on the j^{th} bootstrap sample to be

$$\boldsymbol{\Xi}_n^*(j) = \arg\min_{\boldsymbol{\Xi} \in \Theta} VHD(F(\cdot|\boldsymbol{\Xi}), H_n^*) \tag{4.5.3}$$

where

$$H_n^* = (h_{n,1}^*(\cdot), \quad h_{n,2}^*(\cdot)). \tag{4.5.4}$$

Note that

$$h_{n,i}^*(x) = \frac{1}{N_i^* c_n} \sum_{j \in \mathcal{A}_n^*(i)} K\left(\frac{x - \mathcal{X}_{i,\nu^*(i,j)}^*}{c_n}\right). \tag{4.5.5}$$

Our next theorem describes the existence and uniqueness of BMHDE for RPWD. The proof is similar to the proof of Theorem 3.4.1 and hence is omitted.

Theorem 4.5.2. Under the conditions of Theorem 3.4.1, $\boldsymbol{\Xi}_n^*(j)$ exists and is unique. ∎

We next turn to study the consistency of our BMHDE. From Theorem 3.6.1 we know that this depends on the strong L_1-consistency of $h_{n,i}^*(\cdot)$, $i = 1, 2$. Our next theorem addresses this issue.

Theorem 4.5.3. Assume that $c_n \to 0$ and $nc_n \to \infty$ as $n \to \infty$. Then $h_{n,i}^*(x) \xrightarrow{a.s.} h_i(x)$, $i = 1, 2$ where $h_i(x)$ is the ith component of $\boldsymbol{H}(x) = (h_1(x), h_2(x))$. Furthermore,

$$\lim_{n\to\infty} \int |h_{n,i}^*(x) - H_i(x)| dx = 0 \quad \text{a.s.}. \tag{4.5.6}$$

Proof. As in Theorem 4.3.1,

$$|h_{n,i}^*(x) - h_i(x)| \leq |h_{n,i}^*(x) - h_{n,i}(x)| + |h_{n,i}(x) - h_i(x)|. \tag{4.5.7}$$

The second term on the RHS of (4.5.7) converges to 0 by Theorem 3.6.1. The first term converges to 0 by Theorem 4.3.1 as in the proof of theorem 4.3.2.. ∎

These results along with theorem 3.6.1 yield the following consistency theorem of BMHDE. We omit the proof.

Theorem 4.5.4. Under the condition of Theorem 3.4.1, BMHDE defined in (4.5.3) is a strongly consistent for $\boldsymbol{\Xi}_0$. ∎

4.6 Simulation Results Using the Bootstrap

In chapter 3 we constructed the confidence interval for the parameters using the asymptotic theory. However, the asymptotic results do not reflect the design. This is due to the fact

that the design is ancillary to the inference. In small samples, this could lead to erroneous inference(Bai, Hu and Rosenberger (2002) and Rosenberger, Flournoy and Durham (1997)).

In this section, we will use the bootstrap methodology developed in previous section to construct accurate and robust confidence intervals. We will also construct the asymptotic confidence intervals and through simulations to compare these results. We will assume a normal model for the responses.

4.6.1 Bootstrap Procedure for Randomized Play the Winner Design

All of our results are based on 2000 simulations with 1000 bootstrap samples per simulation. All simulations were carried out on the super computer sp2 with eight processors. All the programming was preformed in Fortran 90 language. We being with the algorithm for implementing the bootstrap procedure described in the previous section.

Algorithm

1. Obtain the success probability of treatment 1 and 2 from the data set; say $\hat{p_1}$ and $\hat{p_2}$ and the estimates of the mean and variance; say $(\hat{\mu}_1, \hat{\sigma}_1^2)$ and $(\hat{\mu}_2, \hat{\sigma}_2^2)$.

2. We start the procedure with an urn containing one balls of type 1 and one balls of type 2 corresponding to treatments 1 and 2 respectively. For example, if a type i ball has been drawn, then assign a response by generating $N(\hat{\mu}_i, \hat{\sigma}_i^2)$ and simulate treatment outcome by $\hat{p_i}$. If the outcome is a success on treatment i or a failure on treatment $3-i$, then update the urn by adding 1 type i balls to the urn. Repeat this process for 30 subjects.

3. Obtain the MHDE of $(\hat{\mu}_1, \hat{\sigma}_1^2)$ and $(\hat{\mu}_2, \hat{\sigma}_2^2)$ for the data generated in step 2.

4. Repeat steps 2 and 3 for 1000 times, yielding 1000 bootstrap estimates,

$$\mu^*_{1,1},\ \mu^*_{1,2},\ \cdots,\ \mu^*_{1,1000},\ \text{ and }\ \sigma^*_{1,1},\ \sigma^*_{1,2},\ \cdots,\ \sigma^*_{1,1000}$$

$$\mu^*_{2,1},\ \mu^*_{2,2},\ \cdots,\ \mu^*_{2,1000},\ \text{ and }\ \sigma^*_{2,1},\ \sigma^*_{2,2},\ \cdots,\ \sigma^*_{2,1000}$$

5. The 95% Bootstrap confidence interval for μ_1 is (L, U), where U is the 97.5% percentile of $\mu^*_{1,i}$'s and L is the 2.5% percentile of $\mu^*_{2,i}$'s. The 95% confidence interval for μ_2 is obtained similarly

6. Perform step 4 and 5 for σ_1^2 and σ_2^2.

We will study the effect of bootstrapping using the MHDE and compare it against bootstrapping using MLE. We investigate four different cases:

I_1 MHDE-MHDE: The parameters in the model for the original data were estimated using MHDE. Parameters in the model for the bootstrap samples were estimated using MHDE.

I_2 MHDE-MLE: The parameters in the model for the original data were estimated using MHDE. Parameters in the model for the bootstrap samples were estimated using MLE.

I_3 MLE-MHDE: The parameters in the model for the original data were estimated using MLE. Parameters in the model for the bootstrap samples were estimated using MHDE.

I_4 MLE-MLE: The parameters in the model for the original data were estimated using MLE. Parameters in the model for the bootstrap samples were estimated using MLE.

Data from treatment 1 has one contamination with data from $N(6, 1)$ while treatment 2 data has no contamination. We study various choices of p_1 and p_2. From Table 4.1, it follows that MHDE-MHDE is always superior to MHDE-MLE in the sense of coverage probability. Also MLE-MHDE is always superior to MLE-MLE. When $p_1 = 0.77$ and $p_2 = 0.75$ the coverage rates are significantly smaller than $p_1 = 0.75$ and $p_2 = 0.75$ or $p_1 = 0.8$ and $p_2 = 0.7$ for all four cases, which is an interesting new phenomenon. This could possibly be attributed to theorem 2.4.3. Further analysis needed is to characterize this phenomenon.

We conclude that bootstrap method confidence interval provide a moderate coverage probability while maintain a short confidence interval length.

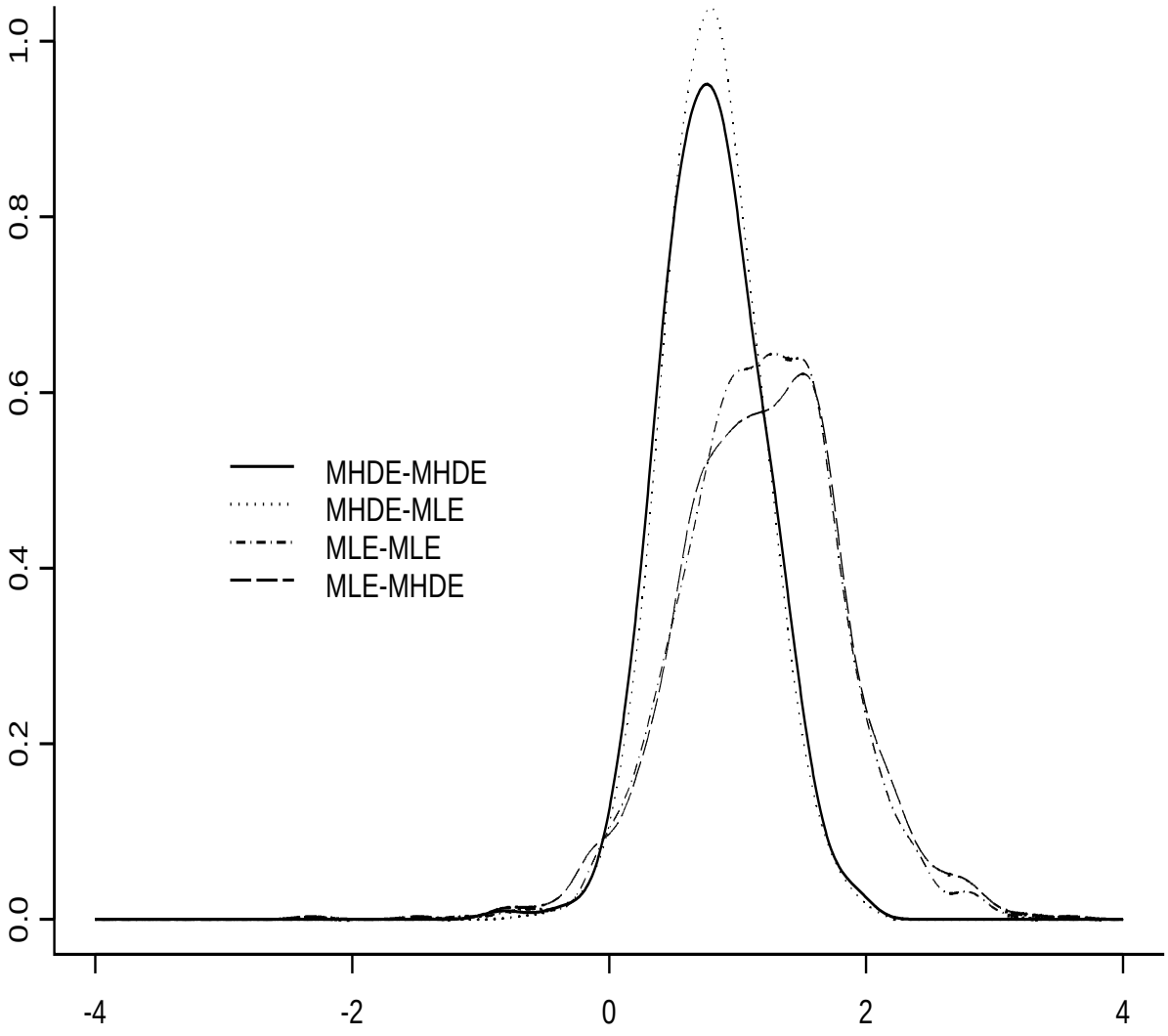
Graph 4.1 contains the bootstrap sample kernel density estimator with respect to four different cases of bootstrapping mentioned above. We can see that due to the effect of outlier, the shape of the density estimators of MLE-MHDE and MLE-MLE are no longer symmetric. Furthermore, the center of the estimates are far from the true mean 0 compare to MHDE-MHDE and MHDE-MLE. We can also see that the if we use MHDE to calculate the original data (MHDE-MHDE and MHDE-MLE), then do the bootstrapping, the estimates are almost center at 0 and have a symmetric curve. MHDE-MHDE is more robust compare to MHDE-MLE, since MHDE-MHDE has lower peak. Therefore, through this simulation study, we can conclude that the bootstrap density estimator using MHDE of continuous variable have a better approximation to the true density function of that variable.

4.7 **Computational Issues**

All the statistical algorithms that were developed in this dissertation have an a companying SAS macros. One of the difficulties that we encountered is that some of the computations were time consuming since the Monte-Carlo method we proposed in section 3.9.1 required us to generate enormous amounts of random variables. Thus to construct one bootstrap confidence interval using SAS(well optimized) in SAS macro, it takes sixteen hours. It is almost impossible to do a full simulation study using SAS for this scenario. Therefore we adopted a parallel programming approach using Fortran 90 to make this project feasible. Under the parallel computing environment, it took 2.5 minutes to construct one bootstrap confidence interval. The conversion from Fortran 90 to parallel computing environment involves understanding message-passing interface (MPI).This is well described in Gropp and Skjellum (1999).

Table 4.1. Results for the coverage probability rates for true mean (treatment 1 has true mean 0 and treatment 2 has true mean 5) (μ) and the average confidence interval length (L) for bootstrap confidence interval (1 and 2) and asymptotic confidence interval (1^* and 2^*) under 0.05 significant level. $N_1(0) = N_2(0) = 1$, $\alpha = 1$, $n = 30$, 2000 simulations and B=1000.I_1, I_2, I_3, I_4 are as defined.

p_1			1	2	1^*	2^*
p_2						
0.5	I_1	μ(L)	0.8622(1.7921)	0.8778(1.6904)	0.9233(2.8780)	0.9373(2.5283)
0.5	I_2	μ(L)	0.8371(1.6483)	0.8594(1.5641)		
	I_3	μ(L)	0.9615(2.6910)	0.9347(1.9273)	0.9994(5.9541)	0.9787(3.2315)
	I_4	μ(L)	0.9381(2.4001)	0.9219(1.7671)		
0.75	I_1	μ(L)	0.8474(1.9583)	0.8700(1.8100)	0.9088(3.3008)	0.9105(2.5812)
0.75	I_2	μ(L)	0.8174(1.7798)	0.8532(1.6599)		
	I_3	μ(L)	0.9480(2.9918)	0.9276(2.0744)	0.9920(7.0656)	0.9704(3.3820)
	I_4	μ(L)	0.9203(2.6378)	0.9169(1.8831)		
0.77	I_1	μ(L)	0.8157(1.8298)	0.7965(1.7028)	0.9208(3.2569)	0.9191(2.6600)
0.75	I_2	μ(L)	0.7930(1.6616)	0.7837(1.5589)		
	I_3	μ(L)	0.9294(2.7312)	0.8571(1.9905)	0.9982(6.6728)	0.9692(3.5447)
	I_4	μ(L)	0.9020(2.4032)	0.8426(1.7972)		
0.8	I_1	μ(L)	0.8630(1.8232)	0.8709(1.9337)	0.9214(3.0818)	0.9055(2.7013)
0.7	I_2	μ(L)	0.8323(1.6697)	0.8550(1.7642)		
	I_3	μ(L)	0.9545(2.7192)	0.9284(2.2449)	0.9960(6.3387)	0.9562(3.5628)
	I_4	μ(L)	0.9323(2.4187)	0.9147(2.0247)		
0.9	I_1	μ(L)	0.8691(1.9212)	0.8528(2.0102)	0.9114(3.4663)	0.8837(2.7354)
0.8	I_2	μ(L)	0.8406(1.7496)	0.8348(1.8297)		
	I_3	μ(L)	0.9465(2.8243)	0.9279(2.3735)	0.9970(6.7813)	0.9575(3.7917)
	I_4	μ(L)	0.9261(2.5039)	0.9133(2.1321)		



Graph 4.1: The kernel density estimators for the bootstrap samples with respect to four different methods.

4.8 **Data Analysis**

In this section, we will continue the data analysis for the Fluoxetine Trial data from Chapter 3. We will use the bootstrap method that is described in this chapter to construct confidence intervals for the data that provided in Chapter 1.

Table 4.2 presents the confidence intervals for the mean parameters of the Fluoxetine trial data. Note that in strata 1, the confidence interval obtained by MHDE do not overlap, showing that there could be a treatment effect in strata 1. This is further confirmed using the parametric bootstrap using MHDE. Note that the results from asymptotic theory could be suspect since the number of observations is less than 20 in each treatment group.

Table 4.2: Analysis results for the Fluoxetine trial data.

Strata	Strata=1 Treatment=1	Strata=1 Treatment=0	Strata=0 Treatment=1	Strata=0 Treatment=0
Asy. CI by MHDE	(-13.09, -8.67)	(-6.37, -1.47)	(-12.92, -7.36)	(-12.14, -7.04)
Length	(4.42)	(4.9)	(5.56)	(5.1)
Asy. CI by MLE	(-13.82, -8.58)	(-8.99, -2.48)	(-13.86, -7.76)	(-11.56, -5.68)
Length	(5.24)	(6.51)	(6.1)	(5.88)
Para. bootstrap(MHDE)	(-13.16, -8.63)	(-7.11, -0.1)	(-13.48, -7.04)	(-12.78, -6.31)
Length	(4.53)	(7.01)	(6.44)	(6.47)
Para. bootstrap(MLE)	(-14.13, -8.63)	(-9.99, -1.34)	(-13.66, -6.53)	(-12.63, -4.92)
Length	(5.5)	(8.65)	(7.13)	(7.71)

4.9 **Concluding Remarks**

In this chapter we developed the bootstrap based methodology using the MHDE for RPWD data. Using our methodology, we were able to detect and confirm a treatment effect in Fluoxetine Trial data which was not detected using the MLE technique. We should note

that this treatment effect was also suggested in the work of Tamura *et. al* (1994) but was not considered by FDA due to the suspicious methodology was used.

Results in our dissertation establish a robust and full efficient methodology for analyzing data from RPWD. Our methods have a potential for application to other response adaptive trials and will be studied in our future work.

References:

1. Bai, Z.D., Hu,F.F. and Rosenberger, W. F.. (2002). Asymptotic properties of adaptive designs for clinical trials with delay response, The Annals of Statistics, 30, 122-139.

2. Bozorgnia, B, Patterson, R.F. and Taylor, R.L. (1997). Chung type strong laws for arrays of random elements and bootstrapping. Stochastic Analysis and Application, 15, 651-669.

3. Chung, K. L. (1974). A course in probability theory. Academic Press.

4. Devroye, Luc. (1987). A Course in Density Estimation, Birkhauser Boston.

5. Diciccio, T.J. and Romano, J.P. (1988) A review of bootstrap confidence intervals. Journal of the Royal Statistical Society. Series B (Methodological), 50, 338-354.

6. Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems. Biometrika, 72, 45-58.

7. Hall, P (1992). The Bootstrap and Edgeworth Expansion. Springer-Verlag.

8. Gropp, W., Lusk, A. and Skjellum, A. (1999). Using MPI portable parallel programming witht the message-passing interface 2^{th} edition, Scientific and engineering computation series.

9. Rosenberger, W. F. and Hu, F. (1999). "Bootstrap Methods for Adaptive Designs," Statistics in Medicine, 18, 1757-1767.

10. Rosenberger, W. F., Flournoy, N., and Durham, S. D. (1997). "Asymptotic Normality of Maximum Likelihood Estimators from Multiparameter Response-Driven Designs," Journal of Statistical Planning and Inference, 60, 69-76.

Chapter 5

Future Research Directions.

In this chapter, we will describe some questions that arise naturally from the thesis. In Chapter 2, we described a new methodology, namely kernel smoothing of bootstrap samples, to construct confidence intervals for success probabilities. The theoretical properties of such a methodology in the context of response adaptive designs has not been previously investigated in the literature. We propose to study this problem in the future.

Obtaining confidence intervals for the Q_i(the asymptotic allocation proportion) is a very difficult and a challenging problem. A unified methodology for constructing confidence intervals for Q_i(independent of the value of δ) is an important problem.

In chapter 3, we developed the MHDE for RPWD. A key assumption that we used in our study was the independence of the response and the randomization variables. While in the Eli-Lilly trial the randomization was independent of the response variable, this may not be the case in general. A new research topic involves modeling this dependence using latent variables and then developing inferential methodology. A technique to address this problem involves modelling using mixtures.

Developing robust testing methodology along the lines of the Hellinger deviance test and extending the methods of this dissertation to other response adaptive designs such as the "drop-the-loser" rule is an important problem that is worth pursuing.