RANDOMIZATION TESTS IN RANDOMIZED CLINICAL TRIALS

by

Yanying Wang
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

Dr. William F. Rosenberger, Dissertation Director

Dr. Guoqing Diao, Committee Member

Dr. Daniel B. Carr, Committee Member

Dr. Diane Uschner, Committee Member

Dr. William F. Rosenberger, Department Chair

Dr. Kenneth S. Ball, Dean,
Volgenau School of Engineering

Date: 19 July 2019

Summer Semester 2019
George Mason University
Fairfax, VA

Randomization Tests in Randomized Clinical Trials

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Yanying Wang
Master of Science
The George Washington University, 2016
Bachelor of Science
China Agricultural University, 2013

Director: Dr. William F. Rosenberger, University Professor
Department of Statistics

Summer Semester 2019
George Mason University
Fairfax, VA

Dedication

To my teacher and friend. May vigor, knowledge, and happiness be their constant companions.

For my own education, I have endeavored to unpretentiously present the principle and method of randomization tests abiding by the teachings of many great statisticians, even though I am incapable of effectively doing so. Meanwhile, if some sentences or words speak to the readers, I will consider the effort to be extraordinarily successful.

Acknowledgments

It is common to attain comforts. It is common to be wealthy. It is common to be knowledgeable. It is common to be renowned. But it is rare, indeed, to have an opportunity to study with dedicated professors and classmates, in a cooperative department, under the guidance of a sagacious teacher and a like-minded friend, on a topic that is thoughtful and engaging. The three-year journey at George Mason University (GMU) Department of Statistics as a graduate student is adventurous. By the kindnesses of many, I was gifted with the rare fortune to not only associate with great people but also receive advice from them at an enormous amount. The education enabled me to cultivate, or at least try to cultivate, academic ability, leadership character, healthy relationships (be respectful to the seniors, be friendly to the peers, and be caring towards the juniors), acceptance of authority, and service attitude–the necessary qualities to achieve anything into which one put effort.

My first thanks go to Dr William Fisher Rosenberger, who enriched my appreciation for the randomization-based inference and who reviewed this dissertation meticulously. The amount of gratitude I owe to Professor Rosenberger cannot be overstated. Repeatedly, I was marveled by his ability to mentor students from diverse background with various levels of knowledge and steer great dissertations while at the same time, as the department chair, keeping piles of responsibilities in order. In the beginning, in my mind, he was like a towering mountain, standing in a distance that I could hardly look up. But then, he walked down from the mountain top, encouraging and accompanying me to climb the mountain of life. I remember how he patiently helped me exploring what I mean to express and speaking up with courage and confidence, how he added numerous seemingly inadvertent yet master revisions to my writings, how he, through instructive words and exemplary acts, transformed my idea of academic research from the recreation of reclusive speculation to the formation of practical wisdom—It is easy to criticize the darkness, but it is more effective to light a candle. In the moments of confusion, in the situation of difficulty, I was accompanied by his benevolence. Because of his sharing his invaluable time and energy, and giving his unwavering trust and support, I was able to proceed. It is said that the debt owed to one's teacher cannot be easily repaid, for the teachers devote what they have for the students. Personally educating the students in different subjects, watching their progress, giving guidance, the teachers offer their attentions and considerations at every moment. I am proud of being one of his many students.

My thanks go to Dr Diane Uschner, who came to the department from RWTH Aachen Germany in August 2017 and has been my supportive friend ever since. I remember with vividness my first meeting with her on the second floor of the engineering building conversing about a testable hypothesis for the randomization test, which turned out to be the first watershed in my understanding of the research question. Her help to me ranges from LaTeX code, programming algorithms, basic statistical concepts, to poster design, manuscript review, general research ideas, as well as forming positive, rewarding relations with the fellow students and the faculties. I have been constantly enlivened by her attentive attitude, open-mindedness, truthful and empathetic nature. It is impossible to describe my odyssey without describing her company and inspiration, which helps me keep direction along the way. I am humbled by her friendship.

My thanks go to Dr Anand N. Vidyashankar, an intelligent and vastly learned scholar who taught my first class at GMU and who is capable of making difficult concepts, methods, theories easily understandable. During Fall 2016 and Spring 2017 semesters, I systematically studied Mathematical Statistics with him. He instilled critical thinking and prepared students to inquire and to write, which has been benefiting my study in all respects. My thanks go to Dr Guoqing Diao, Dr Daniel B. Carr, Dr John M. Lachin, who kindly served as my committee members. My thanks go to Dr Linda Davis, Dr David Holmes, Dr Brett Hunter, Dr Kenneth Strazzari, Dr Ilhan Izmirli, Dr Elizabeth Johnson, who welcomed me as a Graduate Teaching Assistant and helped me do better in communicating statistical ideas with college students. I thank Dr Wanli Qiao and Dr Scott Bruce for enlivening and insightful chats. I thank Ms Verronica Mitchell, who is the office manager in the department administrative office, for carefully responding to all my "Meeting with Professor Rosenberger" emails and for many delightful warm conversations, as well as Ms Carroll Barbour, who is the office assistant, for offering her support from time to time. I thank my classmates, Ms Jing Lei and Mr Zhantao Lin, for laughing and weeping together when taking our eight courses the many exams. I thank all the faculties, staffs, students, who have extended their helps to me directly and indirectly and made this adventure possible. I seek their inspirations and guidance perpetually.

Table of Contents

List of Tables

List of Figures

Abstract

RANDOMIZATION TESTS IN RANDOMIZED CLINICAL TRIALS

Yanying Wang, PhD

George Mason University, 2019

Dissertation Director: Dr. William F. Rosenberger

A clinical trial is a medical experiment using human volunteers. It is a highly controlled process required by the U.S. Food and Drug Administration in the research and development of medical innovations. From development to approval, an innovative therapy needs to go through up to four phases of clinical trial, which might take a considerable amount of human resources and investment. The key component of a phase III clinical trial is *randomization*, or the use of probability to assign treatments to patients. Randomization assists in mitigating certain biases and is the basis for valid statistical inference.

In this dissertation, we examine the randomization test, an inferential approach which integrates and utilizes the experimental randomization in the evaluation of the treatment difference. Randomization-based inference was introduced as the method of analyzing randomized experiments since the formal introduction of the logic of experimentation, pioneered by Sir R. A. Fisher in the 1920s. The utility of the randomization test lies in the non-circumstantial statistical validity and the connection of statistical properties to the randomization. However, the computational limitations rendered the method infeasible in the early days, and statistical analysis was mostly formulated on the basis of the normal distribution and random sampling as a matter of approximation. Because it has been largely ignored in practice, other inferential methods which may not possess the same statistical properties have been mistaken for the randomization test, including the permutation test.

Today, it has become a convention to present the study conclusions using statistical inference based on the invocation of a (parametric) population distribution function.

We will develop (i) a theoretical framework of randomization tests in terms of the hypothesis, the random mechanism, the reference set, (ii) an exploration of the statistical properties including the statistical validity and the power of the test under various models of variability in the patient responses, and (iii) a solution to the computational complexity, particularly in the analysis of multi-armed clinical trials. Further, we will discuss the randomization-based interval estimation. We will contextualize the definition of a confidence interval for the treatment difference and examine efficient algorithms for computing an interval estimate. We conclude that randomization-based inference is adaptable to nearly any primary outcome analysis, and should be used as a matter of course.

Preface

The Pursuit of Knowledge and the Scope of Science

On 5 July 1687, Sir Isaac Newton published work that is considered to be one of the most important works in the history of science. Interestingly, he named it "Mathematical Principles of Natural Philosophy" (i.e., *Philosophia Naturalis Principia Mathematica*). In contemporary English usage, the word *philosophy* may have the connotation of theoretical lucubrations with no substantial basis. Nevertheless, in this context, *philosophy* constitutes definitive conclusions and insights based on knowledge acquired from verifiable sources. The word *science* comes from Latin word *scientia* which means *knowledge* or *to know*, whereas *philosophy* comes from Latin word *philosophia* which means *love of knowledge*. It was not until 1833 that the word *scientist* came into being.

Science is essentially a quest for objective and meaningful knowledge, an endeavor in making sense of the phenomena happen in the nature. *Science* magazine, in its 125th anniversary issue, published a special feature exploring 125 big questions that face scientific inquiry over the next quarter-century. Among them, the top two questions are "What is the universe made of?" and "What is the biological basis for consciousness?" If we understand what the universe is made of and where it comes from, that will provide a sense of meaning. And if we understand the basis for consciousness and our position in the universe, that will provide a sense of purpose. Before starting any scientific inquiry (e.g., what makes an apple fall to the ground?), the implicit suppositions made a prior are three. First, events do not take place without order; nature follows laws. Second, these laws can be expressed mathematically. Third, we conceive mathematics in the way that we experience nature. It is mysterious that why there is order in nature and, further, that why observable phenomenon within nature are understandable through the constructions of mathematics, which essentially are abstract symbols in human mind and have no direct correlation with natural

phenomenon. As Albert Einstein remarks, “The most incomprehensible thing about the universe is that it is comprehensible.”

Empirical science looks for natural explanations for natural phenomena. Its primary object is the measurable properties of nature, such as height, weight, speed, and so forth. And those that are non-measurable are considered to be secondary. By this focus, science can phrase things in the language of mathematics, come up with equations, and make manipulations. However, what science takes as the primary does not comprise our primary experience of the world. For example, medical science is, at one level, meant to free people from pain and provide health. Despite its phenomenal advancements, an objective measure for pain has not been defined. Doctors can characterize a fracture by measuring the degree to which the bone bends. But there is no clue for how to put a measure with a defined unit on the amount of pain a patient with a fractured bone undergoes. The same conundrum is encountered when it comes to health. While there are stacks of measurements associated with health to make indirect inferences, health itself is not something quantifiable. In fact, an indispensable part of human experience is beyond the scope of measurements to quantify. It is not possible to introduce to a friend how delicious the food is by enumerating the amount of salt, sugar, and oil, or how extraordinary a person is by listing his height, race, and occupation.

The perspective science took looking at nature leaves out the non-measurable parameters in nature. And by focusing on the measurable parameters, we have been able to achieve external control at an extraordinary level. But the internal platform, in terms of the non-measurable parameters that are subjected to direct experience, have become mismanaged, just as a smart phone equipped with state-of-the-art hardware yet corrupted software cannot be enjoyed by the user. We observe that soldiers fighting on the battlefields embrace the scarcities in health and food, whereas people living in luxury become irritated if they miss even one meal. It is a sense of enduring practical meaning and purpose that raises the attention of the soldiers above the physical conditions, grants the liberty to adjust what to be considered as pleasant or unpleasant, and give rise to momentum and wholesomeness. In summary, the importance of empirical science to social achievement is undeniable: Clinical trials are scientific research for the evaluation of innovations in medical care with respect to efficacy and safety, and have been the seminal element of the development of modern medicine.

Understanding the spontaneous need for knowledge, enlightenment, and well-being as well as the scope and focus of empirical scientific researches, we hope to contextualize and bring out what empirical researches can offer, analyze, and do to help in the pursuit.

Chapter 1: Introduction and Literature Review

1.1 Randomization as a Basis for Inference

1.1.1 Randomization in Clinical Trials

Perhaps the first recorded clinical trial was reported in the Book of Daniel, where four subjects were compared with respect to the benefit of a vegetarian diet on health. Such a study today would be inadequate in developing scientific conclusions. In the realm of clinical trials, it is desired that evidence-based inference derived from one study yield a reliable standard for future treatment decisions in serving the best care of patients. Being comparative experiments, clinical trials move from expanding the quantity of evidence to enhancing the quality of evidence. The highest level of scientific evidence in evidence-based inference of treatment effectiveness is generated by the randomized clinical trial, first designed by Sir Bradford Hill in 1946 in the streptomycin trial (Armitage, 2003), and is currently acknowledged golden standard in medical disciplines. Randomization in clinical trials entails randomized allocation of patients to the treatments being studied. It is the implementation of randomization that makes the evidence the golden standard.

The idea of randomization as an imperative experimental principle can be traced to Sir R. A. Fisher at the Rothamsted Agricultural Experiment Station in the early 20th century. The “patients" were plots of agricultural land planted with crops, or vegetables, or grass, and the “treatments” were varieties of agricultural interventions (Kempthorne, 1992). Fisher writes in his *The Design of Experiments* about the purpose of randomization with regard to the principles of experimentation (quoting from Kempthorne (1966)) :

> In the foregoing paragraphs the subject-matter of this book [*the principles of experimentation*] has been regarded from the point of view of an experimenter, who wishes to carry out his work competently, and having done so wishes to safeguard his results, so

> far as they are validly established, from ignorant criticism by different sorts of superior persons.
>
> The element in the experimental procedure which contains the essential safeguard, is that the two modifications of the test beverage are to be prepared 'in random order.'

As it is pointed out by Fisher, there are two purposes of incorporating the randomization principle: for the experimenters to carry out the experiment competently and to "safeguard his results from ignorant criticism by different sorts of superior persons."

Later, when Sir Bradford Hill expanded the randomization principle from agricultural experiments to clinical trials, his advocacy to physicians who were eager to conduct reliable research emphasized the mitigation of selection bias by inducing unpredictability in treatment allotment. Selection bias refers to the bias that is introduced by investigators who may predict the future treatment assignment based on assignments that have been allotted and thus, intentionally or unintentionally, choose patients based on their personal idiosyncrasies. In the presence of selection bias, the objectivity of a clinical trial can be seriously compromised. We cite from Armitage (2003) the justifications for randomization given by Hill in 1952:

> It ensures that neither our personal idiosyncrasies (our likes or dislikes consciously or unwittingly applied) nor our lack of balanced judgement has entered into the construction of the different treatment groups-the allocation has been outside our control and the groups are therefore unbiased;
>
> . . . it removes the danger, inherent in an allocation based on personal judgement, that believing we may be biased in our judgements we endeavour to allow for that bias, to exclude it, and that in doing so we may overcompensate and by thus 'leaning over backward' introduce a lack of balance from the other direction;
>
> . . . and, having used a random allocation, the sternest critic is unable to say when we eventually dash into print that quite probably the groups were differentially biased through our predilections or through our stupidity.

Thus, randomization, as an intentional and systematic endeavor to design clinical trials in attainment of the best possible scientific evidence, helps to provide scientific validity. Further, if there is a need to safeguard the validly established results from "ignorant criticism", randomization accounts for a coherent basis for making statistical inference. As it is explained by Fisher, for analyzing or interpreting data from a randomized experiment, it is necessary to incorporate the experimental randomization (quoting from Kempthorne (1966)):

> In these discussions it seems to have escaped recognition that the physical act of randomization, which, as has been shown, is necessary for the validity of any test of significance, affords the means in respect of any particular body of data, of examining the wider hypothesis in which no normality is implied.
>
> This [*the randomization*], in fact, is the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced.
>
> The validity of our estimate of error for this purpose is guaranteed by the provision that any two plots, not in the same block, shall have the same probability of being treated alike, and the same probability of being treated differently in each of the ways in which this is possible. The purpose of randomization in this, as in the previous experiments exemplified, is to guarantee the validity of the test of significance...

1.1.2 Randomization-based Inference in Clinical Trials

Consider a trial that compares the effectiveness of treatment A with treatment B. From the randomization, n_A and n_B patients are allocated to their respective treatments. In analyzing data from the trial, we observe variability in the measurements. The variability has two aspects; treatment difference and error. Error is the variability that is not attributed to the effect of treatments; it may due to the act of measuring and the individuality of patients. This aspect makes the experimental results, even validly established, subject to "ignorant criticism". Because a skeptical judge can insist that the variability is caused, not by treatments, but by error, irrespective of the apparent experimental evidence, which leads to an absurd conclusion that treatment effects cannot be examined by

experiments (Kempthorne, 1992). To safeguard the results, a plausible way is, by referring to the randomizing mechanism, to calculate the chance of such a statement being true. The smaller the chance is, the more likely that the data will be inconsistent with the statement that the variability in data is caused by error alone.

It is essential to contrast the random mechanism in randomization tests with classical likelihood-based tests, such as the t-test, z-test, permutation tests, and so forth. In the randomization test, the data are understood to be arithmetic numbers. The only random mechanism implemented by the investigators in a randomized clinical trial is the act of randomization; patients are carefully selected first through rigorous screening and a consent process so that they have similar relevant characteristics. The classical likelihood-based tests, on the other hand, assume a different random mechanism (Kempthorne, 1977):

> A classical type pf analysis consists of an examination of the data in search of a "good" model. In the simple case of data consisting of an unstructured set of arithmetic numbers, the first steps consist of constructing histograms, then recognizing that the set of numbers is like a random sample from some simple mathematical distribution (e.g. Gaussian, χ^2, Cauchy), and then applying procedures developed for random sampling from the chosen distribution. If the data do not conform, in some senses, to one of these simple distributions, one will consider transformations of the data to achieve an acceptable status... one then proceed by estimating and constructing intervals on the parameters of that Gaussian distribution.

However, as it is stated by Fisher, the basis for making statistical inference in this context is randomization. The normal theory test can be useful only if the resultant p-value provides a good approximation to that given by the randomization test (quoting from Kempthorne (1966)):

> The vital principle has often been overlooked that the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation.
>
> ...conclusion [*from the normal theory test*] have no justification beyond the fact that they agree with those which could have been arrived at by this elementary method

[*randomization test*].

As with the t-test, its appropriateness to any particular body of data may be verified arithmetically.

Kempthorne (1966) comments further:

> In this context, at least, Fisher says the only way to enable the laws of chance to be used in the interpretation of the experiment, is to introduce a chance mechanism in the design and interpret the results in terms of the same chance mechanism. The use of randomization in the design induces a conceptual population of possible results and the observed result is considered in relation to this population.

To recapitulate, first, "tests of significance in the randomized experiment have frequently been presented by way of normal law theory, whereas their validity stems from randomization theory" (Kempthorne, 1955). Second, the act of randomization does not guarantee the validity of population-based approaches. Third, the reliability of the inductive inference follows from an unbiased experiment; otherwise, any invoked statistical formality that is meant to facilitate us to look into the future can only be nominal.

1.2 Permutation Tests and Randomization Tests

In this section, we will present, from the perspective of mathematical theory, why incorporating the randomization procedure is necessary in analyzing data from randomized clinical trials. It is important to distinguish randomization tests from permutation tests.

1.2.1 Permutation Tests

A defect with regard to the random sampling viewpoint is that the population model for the data in terms of parameters is unknown. Further, the validity of the statistical conclusions to future patients such as estimates is conditioned on the model assumptions. "Permutation tests are tests for the comparison of random samples from unspecified distributions" (Kempthorne, 1969). The

feature that renders permutation tests attractive, especially when it is difficult to assume a proper population model, is that they are *distribution-free*; that is, the distribution of a test statistic under the null hypothesis is unrelated to the population distribution of patient responses.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the observed patient responses. Let $\mathcal{Y}^n_{/\boldsymbol{y}}$ be the set of all possible permutations of patient responses conditioning on $\boldsymbol{y}$. Here $\mathcal{Y}^n_{/\boldsymbol{y}}$ is the *reference set for conditional inference* (Pesarin, 2001), where $\mathcal{Y}$ is the sample space. Assume that y_i's, $i = 1, \ldots, n$, are independently sampled from a distribution P, where P is from a nonparametric family of distributions denoted by $\mathcal{P}$. Then the probability density of $\boldsymbol{y}$ (with respect to some dominating measure ξ), denoted by $f_P^{(n)}(\boldsymbol{y}) = \prod_{i=1}^n f_P(y_i)$, is invariant with respect to permutation of the argument of $\boldsymbol{y}$; that is,

$$f_P^{(n)}(y_1, \ldots, y_n) = f_P^{(n)}(y_{u_1}, \ldots, y_{u_n}), \tag{1.1}$$

where $(u_1, \ldots, u_n)$ is a permutation of $(1, \ldots, n)$. However, if y_i's are not viewed as independent and identically distributed, then we need to assume that property (1.1) holds true. This assumption is termed as *exchangeability of observed data with respect to groups* (Pesarin, 2001).

Now, since all the permutations have a common likelihood, the likelihood ratio of any two permutations is a constant regardless of the probability distribution of y_i's; that is,

$$f_P^{(n)}(\boldsymbol{y}')/f_P^{(n)}(\boldsymbol{y}'') = 1, \forall P \in \mathcal{P},$$

given that $\boldsymbol{y}', \boldsymbol{y}'' \in \mathcal{Y}^n_{/\boldsymbol{y}}$. Therefore, conditioning on $\mathcal{Y}^n_{/\boldsymbol{y}}$, $\boldsymbol{y}$ is uniformly distributed regardless of P, provided that there are not tied values in $\boldsymbol{y}$:

$$P(\boldsymbol{y} = \boldsymbol{y}^* \mid \mathcal{Y}^n_{/\boldsymbol{y}}) = \frac{f_P^{(n)}(\boldsymbol{y}^*) \cdot d\xi}{\sum_{\boldsymbol{y} \in \mathcal{Y}^n_{/\boldsymbol{y}}} f_P^{(n)}(\boldsymbol{y}) \cdot d\xi} = \frac{1}{\#[\boldsymbol{y} \in \mathcal{Y}^n_{/\boldsymbol{y}}]} = 1/n!.$$

Based on this distribution, for a given treatment difference metric, the p-value of the null hypothesis

is obtained by counting the number of permutations that result in equal or more extreme test statistic values then the observed test. Since $\boldsymbol{y}$ is a set of sufficient statistics, conditioning on $\boldsymbol{y}$ allows for making inferences about their distribution arising from family $\mathcal{P}$. A rejection of the null hypothesis that the probability distributions of the two treatment groups are identical implies a treatment difference.

Therefore, assuming independent and identically distributed patient responses and conditioning on the set of sufficient statistics $\boldsymbol{y}$, the *reference set for conditional inference* is comprised of equiprobable elements. Consequently, the distribution-free property and the flexibility in choosing a test statistic is obvious. Yet, the permutation test ignores the distribution of the treatment assignment sequence derived from the randomization procedure. The only function of the observed treatment assignment sequence is to partition the patient responses into two treatment groups. Any discussion on the type I error rate and the statistical power has no connection to the randomization.

1.2.2 Randomization Tests

In randomization tests, the collection of outcome data are viewed as arithmetic numbers, as opposed to as if a realization of random samples from a population distribution (Kempthorne, 1977, 1992). The structure of the randomization test consists of three components: the hypothesis, the reference set, and the distribution of the test statistic. The null and alternative hypotheses in the randomization test are simple; the null states that the treatments are independent of patient responses. The alternative hypothesis, on the other hand, is not simply a complement of the null hypothesis, but is a specification of the treatment difference on an individual level. The reference set holds a prominent position in the testing structure, because the distribution of the treatment assignment sequences is derived from the set, and, more importantly, the distribution test statistic is derived from the set. We illustrate the structure through the following examples.

Consider a two-armed randomized clinical trial comparing the treatment difference among four patients. At the end of the trial, the data are collected and organized in the table below (Table 1.1). From the table, the first patient was assigned to treatment A and gave a response of y_1, the second patient was assigned to treatment A and gave a response of y_2, and so forth. The third column

Table 1.1: *Data structure of a two-armed randomized clinical trial with four patients.*

Patients responses $\boldsymbol{y}$	Treatment assignment sequence $\boldsymbol{t}$	Probability $P_\phi\{\boldsymbol{T}=\boldsymbol{t}\}$
y_1, y_2, y_3, y_4	A, A, A, B	p_1

in the table is the probability of the treatment assignment sequence (A, A, A, B) being sampled using the randomization procedure. To determine the statistical significance of a hypothesis based on the experimental data, we need to find out what the patient responses are *when duplicating the completed trial by re-randomization*. This is possible with the help of a hypothesis.

The null hypothesis assumes that patient responses are unrelated to the treatments. In other words, under the null, patient responses for any treatment assignment sequences in the reference set are copied from the original responses without any alteration (Table 1.2). The second column in the table lists the treatment assignment sequences that could have been generated by the randomization procedure. The third column contains their corresponding probabilities given the randomization procedure. From this table, the probability distribution of the test statistic can be derived. For instance, consider the difference in means statistic. From the patient responses and the their treatment assignments, the values of the difference in means can be obtained, and their probabilities are determined by the randomization procedure (Table 1.3).

Under an alternative hypothesis, the probability distribution of the difference in means is similarly derived. Suppose an alternative hypothesis states that treatment A has a constant additive effect (denoted by Δ) on the patient responses in comparison to treatment B. That is to say, if a patient from treatment A were re-randomized to treatment B, the response would be decreased by Δ. Likewise, if a patient from treatment B were re-randomized to treatment A, the response would be increased by Δ. Therefore, under the alternative, a new set of patient responses can be determined for any treatment assignment sequence in the reference set (Table 1.4). Thus, the distribution of the

Table 1.2: *Reference set of a randomized clinical trial with four patients and randomization procedure ϕ and the patients responses under the null hypothesis.*

Patients responses $\boldsymbol{y}$	Treatment assignment sequence $\boldsymbol{t}$	Probability $P_\phi\{\boldsymbol{T}=\boldsymbol{t}\}$
y_1, y_2, y_3, y_4	A, A, A, B	p_1
y_1, y_2, y_3, y_4	A, A, B, A	p_2
y_1, y_2, y_3, y_4	A, A, B, B	p_3
y_1, y_2, y_3, y_4	A, B, A, A	p_4
y_1, y_2, y_3, y_4	A, B, A, B	p_5
y_1, y_2, y_3, y_4	A, B, B, A	p_6
y_1, y_2, y_3, y_4	A, B, B, B	p_7
$\dots$	$\dots$	$\dots$

Table 1.3: *The probability distribution of difference in means statistic under the null hypothesis.*

Patients responses $\boldsymbol{y}$	Treatments $\boldsymbol{t}$	Difference in means $\bar{y}_A - \bar{y}_B$	Probability $P_\phi\{\boldsymbol{T}=\boldsymbol{t}\}$
y_1, y_2, y_3, y_4	A, A, A, B	$(y_1+y_2+y_3)/3 - y_4$	p_1
y_1, y_2, y_3, y_4	A, A, B, A	$(y_1+y_2+y_4)/3 - y_3$	p_2
y_1, y_2, y_3, y_4	A, A, B, B	$(y_1+y_2)/2 - (y_3+y_4)/2$	p_3
y_1, y_2, y_3, y_4	A, B, A, A	$(y_1+y_3+y_4)/3 - y_2$	p_4
y_1, y_2, y_3, y_4	A, B, A, B	$(y_2+y_4)/2 - (y_1+y_3)/2$	p_5
y_1, y_2, y_3, y_4	A, B, B, A	$(y_1+y_4)/2 - (y_2+y_3)/2$	p_6
y_1, y_2, y_3, y_4	A, B, B, B	$(y_2+y_3+y_4)/3 - y_1$	p_7
$\dots$	$\dots$	$\dots$	$\dots$

difference in means under the alternative hypothesis is determined (Table 1.5).

The p-value of the hypothesis is constructed as follows. For a given test statistic $S(\cdot)$, it represents the likelihood of observing a test statistic that equals or exceeds the observed value $s_{\text{obs.}}$ when duplicating the already complemented trial by re-randomization. The value is determined as

$$p = \sum_{\boldsymbol{t}\in\Omega} I_{(S(\boldsymbol{t},\boldsymbol{y})\geq s_{\text{obs.}}|H)} P_\phi\{\boldsymbol{T}=\boldsymbol{t}\},$$

Table 1.4: *Reference set of a randomized clinical trial with four patients and randomization procedure ϕ and the patients responses under the alternative hypothesis*

Patients responses $\boldsymbol{y}$	Treatment assignment sequence $\boldsymbol{t}$	Probability $P_\phi\{\boldsymbol{T}=\boldsymbol{t}\}$
y_1, y_2, y_3, y_4	A, A, A, B	p_1
$y_1, y_2, y_3-\Delta, y_4+\Delta$	A, A, B, A	p_2
$y_1, y_2, y_3-\Delta, y_4$	A, A, B, B	p_3
$y_1, y_2-\Delta, y_3, y_4+\Delta$	A, B, A, A	p_4
$y_1, y_2-\Delta, y_3, y_4$	A, B, A, B	p_5
$y_1, y_2-\Delta, y_3-\Delta, y_4+\Delta$	A, B, B, A	p_6
$y_1, y_2-\Delta, y_3-\Delta, y_4$	A, B, B, B	p_7
...	...	...

Table 1.5: *The probability distribution of difference in means statistic under the alternative hypothesis (color blue denotes the responses come from treatment B).*

Patients responses $\boldsymbol{y}$	Difference in means $\bar{y}_A - \bar{y}_B$	Probability $P_\phi\{\boldsymbol{T}=\boldsymbol{t}\}$
y_1, y_2, y_3, y_4	$(y_1+y_2+y_3)/3 - y_4$	p_1
$y_1, y_2, y_3-\Delta, y_4+\Delta$	$(y_1+y_2+y_4+\Delta)/3 - (y_3-\Delta)$	p_2
$y_1, y_2, y_3-\Delta, y_4$	$(y_1+y_2)/2 - (y_3-\Delta+y_4)/2$	p_3
$y_1, y_2-\Delta, y_3, y_4+\Delta$	$(y_1+y_3+y_4+\Delta)/3 - (y_2-\Delta)$	p_4
$y_1, y_2-\Delta, y_3, y_4$	$(y_2-\Delta+y_4)/2 - (y_1+y_3)/2$	p_5
$y_1, y_2-\Delta, y_3-\Delta, y_4+\Delta$	$(y_1+y_4+\Delta)/2 - (y_2-\Delta+y_3)/2$	p_6
$y_1, y_2-\Delta, y_3-\Delta, y_4$	$(y_2-\Delta+y_3-\Delta+y_4)/3 - y_1$	p_7
...	...	...

where Ω denotes the reference set and H denote the hypothesis tested. If H is the null hypothesis, the p-value is used as the standard comparator to the type I error rate to judge statistical power. However, the construction under the an alternative hypothesis can be used to determine interval

estimate for Δ. In the randomization test, the experiment itself forms a population and the patient responses does not represent a random sample from a patient population.

Even if inference been placed in the framework of patient population and random sampling, the randomization test still has its validity as a likelihood ratio test, and the p-value is determined by the reference set alone. We demonstrate this point via a test of the null hypothesis. In a permutation test, patient outcomes $\boldsymbol{y}$ are regarded as a realization of random variables while the probability distribution of the treatment assignments $\boldsymbol{t}$ is not considered. A more holistic perspective for analyzing data from clinical trials is to view both $\boldsymbol{y}$ and $\boldsymbol{t}$ as random variables. Let $\boldsymbol{X} = (\boldsymbol{T}, \boldsymbol{Y})$ be a random vector, where $\boldsymbol{T} = (T_1, \ldots, T_n)$ is the treatment assignment sequence and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is the patient responses. The distribution of $\boldsymbol{T}$, denoted by P_ϕ, is derived from the randomization procedure. The distribution of $\boldsymbol{Y}$, denoted by probability measure P with density $f_P(\boldsymbol{y})$, is unknown.

Let $f_X(\boldsymbol{x})$ be the probability density of $\boldsymbol{X}$ with respect to a dominating measure ξ. Since $\boldsymbol{x}$ is a set of sufficient statistics, conditioning on $\boldsymbol{x}$ allows for making inferences about the distribution of $\boldsymbol{X}$ arising from a nonparametric family of distributions (Pesarin, 2001). Note that this argument alone does not provide a substantial basis for a valid generalization of the experimental results to a large context. Under the null hypothesis that the treatments are unrelated to patient outcomes, $\boldsymbol{T}$ is stochastically independent of $\boldsymbol{Y}$ (i.e., $\boldsymbol{T} \perp\!\!\!\perp \boldsymbol{Y}$). Therefore, we have $f_X(\boldsymbol{x}) = P_\phi(\boldsymbol{T} = \boldsymbol{t}) \cdot f_P(\boldsymbol{y})$, and the likelihood ratio of two observations with identical patient outcomes $\boldsymbol{y}$ (e.g., $\boldsymbol{x} = (\boldsymbol{t}, \boldsymbol{y})$, $\boldsymbol{x}' = (\boldsymbol{t}', \boldsymbol{y})$) is decided only by the randomization, and is not dependent on the distribution of $\boldsymbol{X}$:

$$\frac{f_X(\boldsymbol{x})}{f_X(\boldsymbol{x}')} = \frac{P_\phi(\boldsymbol{T} = \boldsymbol{t}) \cdot f_P(\boldsymbol{y})}{P_\phi(\boldsymbol{T} = \boldsymbol{t}') \cdot f_P(\boldsymbol{y})} = \frac{P_\phi(\boldsymbol{T} = \boldsymbol{t})}{P_\phi(\boldsymbol{T} = \boldsymbol{t}')},$$

where $\boldsymbol{t}, \boldsymbol{t}' \in \Omega$, the reference set induced by the randomization procedure. Note that the permutation of data is no longer the key element. The reference set does not have to be equiprobable and the data do not have to be exchangeable. These concepts are now replaced by the reference set and computation of the distribution of the test statistic with respect to that reference set.

Let $S(\cdot)$ be a measure of treatment difference. Let $\boldsymbol{x}_{\text{obs.}}$ be the observation from a trial, p-value

is thereby computed as

$$p = \sum_{\boldsymbol{t} \in \Omega} I_{(S(\boldsymbol{x}) \geq S(\boldsymbol{x}_{\text{obs.}}) | \boldsymbol{Y} = \boldsymbol{y})} P_{\phi}\{\boldsymbol{T} = \boldsymbol{t}\},$$

which is equivalent to the equation we have derived previously without the notion of patient population, conditioning, and sufficient statistics.

Edgington and Onghena (2007) discuss applying randomization tests in designed experiments where the use of nonrandom samples is prevalent. Because the experimental designs result in equiprobable treatment assignment sequences, they suggest using the *set of data permutations* as a replacement for the reference set to reduce the computation in calculating the p-value. The set of data permutations is comprised of all possible permutations of patients outcomes while fixing the treatment assignment sequence. The p-value is computed as the proportion of data permutations that result in equally or more extreme test statistic values than the observation.

Table 1.6: *The analogy of the set of data permutations and the reference set when treatment assignment sequences are equiprobable.*

Set of data permutations				Reference set				Probability	
Group *A*		Group *B*		y_1	y_2	y_3	y_4	$P_{\phi_1}\{\boldsymbol{T} = \boldsymbol{t}\}$	$P_{\phi_2}\{\boldsymbol{T} = \boldsymbol{t}\}$
y_1	y_2	y_3	y_4	A	A	B	B	1/6	1/4
y_1	y_3	y_2	y_4	A	B	A	B	1/6	1/8
y_1	y_4	y_2	y_3	A	B	B	A	1/6	1/8
y_2	y_3	y_1	y_4	B	A	B	A	1/6	1/8
y_2	y_4	y_1	y_3	B	A	A	B	1/6	1/8
y_3	y_4	y_1	y_2	B	B	A	A	1/6	1/8

The analogy of the set of data permutations and the reference set when treatment assignment sequences are equally likely is presented in Table 1.6. Consider an experiment which assigns four patients to two A treatments and two B treatments. The design produces six equiprobable treatment assignment sequences (see the second column in table 1.6). Suppose patient outcomes are y_1, y_2, y_3, y_4. Then six permutations of patient responses can be obtained by fixing the treatment

assignment sequence (see the first column in table 1.6). Comparing the two columns, we see a one-to-one relationship between each data permutation and each treatment assignment sequence. Thus, enumerating the number of data permutations is arithmetically equivalent to using the reference set in obtaining the p-value. However, the equivalence no longer exists when the sequences are not equiprobable (see the last column in the table for an example). More essentially, assumptions from population-based testing theory would be needed for permuting the observed patient outcomes among treatment groups, which alters the nature of randomization-based inference.

So, a randomization test is different from a permutation test, in that, under the null, it depends only on the randomization distribution, which may not be exchangeable.

1.2.3 Inferential Procedures Overview

An overview of the inferential procedures applied in randomized clinical trials is summarized in Figure 2.1. Here the term parametric tests refers to the hypothesis tests of population parameters under the Neyman-Pearson paradigm (Rosenberger and Lachin, 2016). Although aiming at a common research question, population-based and randomization-based inferences diverge as they condition on different component of the data: randomization tests condition on the observed patient outcomes $\boldsymbol{y}$ while permutation tests and parametric tests condition on the observed treatment assignment sequences $\boldsymbol{t}$.

Kempthorne (1977) discusses three types of empirical investigations in the statistical profession. They are: (1) an observational study; (2) a survey in which population of interest is totally defined and accessible; (3) a comparative experiment (i.e., the comparing treatments are chosen by the investigators and are imposed on the patients). What distinguishes the comparative experiment from the other two is the imposition of interventions on the experiment units as well as the concept of population. Kempthorne (1977) states that different investigations are distinctive in purposes and scopes, and it is crucial that the subsequent approach to inference should conform to the type of investigation; only in a survey of a definite and accessible population, is making inference about attributes of the population reasonable.

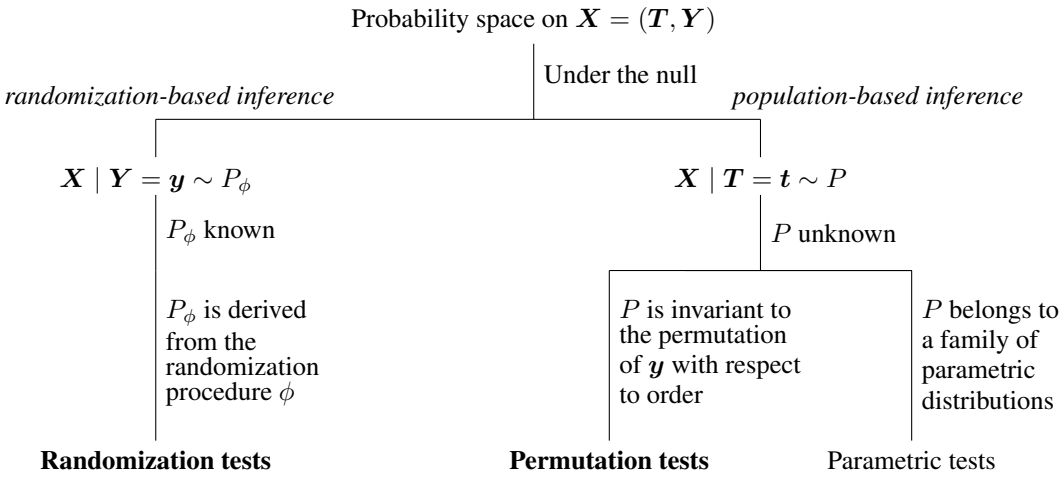

Figure 1.1: *Overview of the inferential procedures applied in randomized clinical trials*

1.3 Validity, Statistical Power, Test Statistic, and Generalization

The idea of using a hypothesis test as an *accept-reject* rule can be seen in a paper by Neyman and Pearson (1933). In standard Neyman-Pearson hypothesis testing theory (Lehmann and Romano, 2006), patient responses are regarded as a realization of a random variable which has a distribution function indexed by a parameter (or a set of parameters). The goal is to determine a decision rule to ascertain if a particular value of the parameter is plausible in order that the probabilities of committing type I and type II errors are minimized.

Statistical power is the probability of correctly rejecting the null hypothesis (of independence) under repeated sampling. As in population-based testing approaches, the power of the randomization test is simulated by re-sampling patient responses from a population distribution and re-randomizing treatment assignments. It is interpreted as how reproducible the statistical decision would be in a population of repetitions under a certain randomization procedure. Investigating the power is helpful in improving the randomization if some background knowledge of the nature of treatment effects, or the potential confounding factors, or the heterogeneity in patient responses in

general (e.g., time trend, outliers, skewness) is known beforehand. We will explore this topic in later chapters.

In addition, the randomization test is statistically valid by construction provided that the significance level α is an achievable size and the comparability among treatment groups is not violated by, for example, selection bias. This is because, for each time the data set being regenerated, the randomization test updates the distribution of the test statistic with respect to the newly generated outcomes and the reference set. Therefore, the distribution is always correct for the data, and thus the p-value is uniformly distributed. In this regard, an inflated type I error rate can be taken as an indication of a flawed study or an improper reference set (i.e., inconsistent with the randomization procedure that produces the original treatment allocations).

Because of the built-in statistical validity, the choice of the test statistic in a randomization test is no longer confined by the distributional assumptions of the patient population, nor is influenced by any misspecification in that population model. As we have shown in Section 1.2.2, even from the perspective that patient outcomes are random samples from a population, the randomization test has its rationale as a valid likelihood ratio test on the basis of sufficiency and conditioning. Consequently, randomization tests can be adapted to nearly every type of primary outcome analysis, including covariate-adjusted analysis (Parhat et al., 2014), sequential monitoring (Plamadeala and Rosenberger, 2012), and stratified analysis (Rosenberger and Lachin, 2016). Any type of primary outcome variable can be analyzed by an appropriate test statistic, capturing categorical, ordinal, continuous, time-to-event outcomes (Rosenberger et al., 2019).

It is sometimes believed that the generalizability of the experimental conclusions to a relevant population of interest is enabled and can be estimated by invoking statistical formality via the notion of random sampling , or via the formulation of the testing hypothesis. For example, Proschan and Dodd (2019) indicate that preservation of type I error rate is an implication for valid generalization from the specific trial participants to a larger context. They show this using a conditioning argument as support. However, such a connotation cannot be invoked by the preservation of type I error rate alone; rather, a statistically valid conclusion may not be scientifically objective due to a biased experiment. The validity of the generalization of trial results to future similar patients relies on the

design and proper conduct of the trial rather than on the accuracy of a statistical model applied in the inference. The entire population cannot be unambiguously characterized on the strength of a finite number of experiments; how could it be possible to sample and experiment with patients who will appear in future and whose diversity in attributes can scarcely be understood? After all, randomized clinical trials do not involve the design and execution of a random sampling procedure. ß

The rationale or validity underlying randomization tests follows from the principle that statistical inference is subordinate to experimentation. Thus, the testing method should be consistent with the experimental design and conduct. While the p-value of a hypothesis targets only at the given set of experiment data, this limited purview does not impede the generalizability of the experiment conclusions. For randomization-based inference, the rationale for generalization comes descendingly: because a treatment effect exists, the experiment provides the gold standard level of evidence, and the statistical test is pursuant to the experimental design. Therefore, the inferential conclusion can be compatible with the actual treatment effect, and thus can be applied to future clinical decisions assuming that future patients have similar characteristics. On the other hand, if the researchers have no faith in the validity of experimentation or treatment effects, and if making valid generalization can be single-handedly addressed by invoking a patient population and applying decision theory, then why bother to conduct a trial in the first place?

1.4 Interval Estimation

Randomization-based interval estimation is closely connected with the treatment effect model specified in the alternative hypothesis. We found in Kempthorne (1969, 1977, 1979, 1982) and Edgington (2007, Section 13.6) some suggestions on how a confidence interval of an additive treatment effect can be interpreted and constructed. We recall the test of constant additive treatment effect in Section 1.2.2. Let H_Δ be the hypothesis that the difference between treatment A and B is Δ for each and every patient, where Δ is a real number. Under H_Δ, by the construction of randomization distribution, the probability of rejecting H_Δ when H_Δ is true,

$$P(\text{rejecting } H_\Delta | H_\Delta \text{ is true}),$$

should be controlled at the prescribed significance level. Therefore, the set of Δ values for which H_Δ will not be rejected in a level α test can be regarded as a $(1-\alpha)\%$ confidence set for the treatment difference.

This definition of confidence level differentiates randomization-based interval estimation from population-based interval estimation: for example, let $y_i, i = 1, \ldots, n$ be a random sample from the population distribution indexed by an unknown parameter μ. A confidence interval for μ is an interval calculated from the random sample, such that it captures the true value of μ with a specified probability under repeated sampling (Bhattacharyya and Johnson, 1977). If the construction is based on inverting a test statistic, then the level of confidence associated with the interval corresponds to the significance level of a test of the null hypothesis, and the interval is basically a point estimate with an error bound. Kempthorne (1992) comments that interval estimation in the randomization-based framework should be named *consonance region*, because the values of the treatment difference specified by the regions are *consonant with the given data* at a chosen significance level, in order to distinguish it from the population-based confidence interval.

The confidence interval of a treatment shift is specified by performing a series of significance tests over a range of Δ. To determine the upper and lower limits of the confidence interval thus requires efficient algorithms. There are numerous in carrying out this search approach. For example, will the confidence region be continuous (i.e., an interval)? How will the confidence interval be impacted by the choice of test statistic? What do we do if the subtraction of treatment shift Δ from patient responses results in negative numbers when negative values have no practical meaning in the study? Is it applicable to derive a confidence interval for other types of treatment effect, such as the ratio? These questions will be examined in Chapter 6.

1.5 Conditional Reference Set

The use of the conditional reference set is suggested by Cox (1982) as a method of *analysis* for the Efron's bias coin design (BCD). In his treatise, Cox first advocates the use of BCD in clinical trials:

> [*The BCD*] ensures with high probability near balance at all stages of the experiment

> and avoids the objection to any form of block design that towards the end of a block, treatment allocation can be successful predicted [*which makes room for selection bias*].

He mentions that in small experiments, unlike in a long term experiment, the BCD can lead to the presence of treatment allocations that provides no information about the treatment effect (e.g., all As). To test H_0 in this context, Cox proposes that:

> We should take the randomization distribution not over all designs [*the original reference set*] but only over those arrangement with the same or nearly the same terminal lack of balance [*the observed value of* $N_A(n) - N_B(n)$].

The justification for applying the conditioning method given by Cox is that $N_A(n) - N_B(n)$ is an ancillary statistic under H_0. Additionally, Cox proposes that the conditioning can also include other aspects of the treatment assignments that are ancillary and if there is a reason to think relevant:

> For instance, if the background model has a roughly linear trend, we should condition the randomization on the realized value of $\sum r\Delta_r$ [$\Delta_r = N_A(r) - N_B(r)$].

Further, Cox comments that

> Such conditioning of the randomization is closely related to the argument for rejecting unsatisfactory designs in a more conventional context. Restricted randomization is essentially a way of achieving the necessary conditioning without modification of the usual analysis, and with the avoidance altogether of specified 'bad' designs.

The word "designs" in the above quotation refers to a treatment assignment sequence, not a randomization procedure.

The conditional reference set can be introduced more naturally in the setting of multiple comparisons, or when the randomization procedure is covariate-adaptive or response-adaptive. For example, consider a randomized clinical trial comparing three treatments A, B, C among six patients. The treatment assignments used in the trial is (A, C, A, A, C, B). If we want to compare the difference between treatment A and B, then we need a subset of the reference set that randomizes only

treatment A and B while holds the C assignments fixed. Such a subset is known as the conditional reference set (for the pairwise comparison).

As another example, consider a randomization procedure as follows (adapted from Proschan and Dodd (2019)): recruit six patients and allocate them to one of the two treatments by flipping a fair coin. If all of them are assigned to a same treatment, recruit another six patients and allocate them to the opposite treatment. If the first six patients are not all assigned to the same treatment, then recruit another six patients and allocate them to one of the two treatments by flipping a fair coin. The reference set used in the test is supposed to be conditioned on the observed treatment assignments of the first six patients. Like the reference set, the conditional reference set provides an objective basis for deriving the probability distribution of the test statistic, and thus tends to preserve the type I error rate of the test.

1.6 Monte Carlo Re-Randomization Tests

The p-value of an exact randomization test can be estimated by the Monte-Carlo re-randomization test. The idea is to estimate the distribution of the test statistic by Monte-Carlo simulation. For a given set of patient responses, the treatment assignment sequence is regenerated L times, and the test statistic is re-computed for each time. Then the p-value is determined as the proportion of the L simulations that results in a test statistic value that equals or exceeds the observed statistic. The two-sided Monte Carlo p-value estimator is defined as

$$\hat{p} = \frac{\sum_{l=1}^{L} I(|S_l| \geq |S_{obs.}|)}{L}. \tag{1.2}$$

The value of Monte Carlo re-randomization tests can be appreciated better if the historical setting is understood. We cite from Rosenberger and Lachin (2016)

> The great founders of the randomization clinical trial understood the importance of randomization as a basis for inference, but were limited in their ability to perform it, due to computational limitations of the day...While in the past, much of the literature was

> focused on finding the asymptotic distribution of the randomization test, such approximations were often inaccurate for moderate sample sized, and the accuracy was highly dependent on the type pf the randomization procedure employed...The computation of randomization tests using Monte Carlo is now the preferred technique...this method is simple and relatively foolproof...such tests can be performed using standard software in seconds.

From the above quotation, we see that an impediment of popularizing randomization-based inference in the early days is the limitation in computation; an exact randomization test is not computationally efficient as the sample size gets larger than around fifteen even with the modern computational facilities (Rosenberger and Lachin, 2016). With regard to the limitation, the solutions provided during the 1980's focused on approximating the exact randomization distribution when the number of patients goes infinity. Now with the advancement of computers, the exact randomization distribution can be simulated with efficiency by Monte-Carlo simulation. The accuracy of the p-value estimate is guaranteed by convergence theory. But, somehow, statistical inference turns astray from its original basis in randomization.

Some necessary concerns for carrying out the re-randomization tests includes the choice of L (Rosenberger and Lachin, 2016). The choice of L depends on how large the p-value is expected to be. With an *MSE* argument (1.3), it is straight forward that the smaller the expected p-value is, the larger the L will be. A more advanced measurement for finding L is suggested by Plamadeala and Rosenberger (2012). Later, Galbete and Rosenberger (2015) demonstrated through simulations that taking L to be 15,000 will be accurate enough to estimate a small p-values, for instance, no greater than 0.05.

$$\text{MSE}(\hat{p}) = \frac{p(1-p)}{L} \leq \frac{1}{4L}. \tag{1.3}$$

To apply the Monte-Carlo method to estimating a conditional randomization test (i.e., a test uses the condition reference set), it is required to have an efficient algorithm for simulating the conditional reference set, particularly in multiple comparisons. Here we give some examples when

the test is conditioned on the observed number of A treatments assigned in a (two-armed) trial (denoted by n_A). Zhang and Rosenberger (2011) examine a naive approach for sampling from the conditional reference set. Generate a large number of treatment assignment sequences, say K, using the unconditional sampling scheme. Next, keep only the sequences satisfying the condition, and the p-value estimate $\hat{p}_c$ is computed from them by the following equation,

$$\hat{p}_c = \frac{\sum_{l=1}^{K} I(|S_l| \geq |S_{obs.}|, N_A(n) = n_A)}{\sum_{l=1}^{K} I(N_A(n) = n_A)}, \tag{1.4}$$

where $N_A(n)$ denotes the number of A assignments in a sequence of length n.

In the above approach, the number of the sequences should be large enough to approximate the conditional randomization distribution. Plamadeala and Rosenberger (2012) show that it gets computationally unfeasible and complicated when n_A deviates from 0.5. They suggest sampling directly from the conditional reference set as an alternative approach. For instance, for complete randomization design, jth treatment assignment can be sampled by using the formula:

$$p_j = \frac{n_A - N_A(j-1)}{n - (j-1)}. \tag{1.5}$$

They also provide a formula to enable conditional sampling for Efron's biased coin design. For more complex randomization procedures, the formula for direct sampling can be difficult to derive (Rosenberger and Lachin, 2016).

1.7 Contributions of the Thesis

Thus far we have presented the application of randomization in clinical trials and the use of the randomization test as a valid inferential method with regard to the exposition by many statisticians, including Fisher (1935b), Anscombe (1948), Kempthorne (1952a), Armitage (2003). Noticing that the concepts and use of the notion of population, statistical significance, validity, statistical power, and confidence interval in randomization-based inference have not been comprehensively addressed in past research, particularly in the context of randomized clinical trials; this chapter explores them

afresh.

The contributions of the dissertation encompasses four aspects, covering tests of significance as well as estimations of treatment effects. We develop (i) a holistic theoretical framework of randomization tests in terms of the hypothesis, the random mechanism, the reference set, (ii) an exploration of the statistical properties including the statistical validity and the power of the test under various models of variability in the patient responses, and (iii) a solution to the computational complexity, especially in the analysis of multi-armed clinical trials. Further, we discuss (iv) the estimation of the treatment difference in the randomized-based framework. We contextualize the definition of confidence interval for the treatment difference and examine an efficient algorithm for computing an interval estimate. We conclude that randomization-based inference is adaptable to nearly any primary outcome analysis, and should be used as a matter of course.

1.8 Outline of the Thesis

The outline of the thesis is as follows. In Chapter 2, we explore the statistical properties of randomization tests in two-armed clinical trials through simulations, in terms of the type I error rate and statistical power under three scenarios of heterogeneity in patient responses. In each of the scenario, we compare two test statistics and eight randomization procedures. Also, we contrast the randomization test with the permutation test and the t-test. In Chapter 3, we explain the rationale of randomization-based inference in handling multiple treatments, which begins with a discussion of the analysis of variance largely based on Kempthorne's work in a different context and, later, covers the topics of multiple comparisons as well as the analysis of factorial designs. Chapter 4 contains a collection of generalized randomization procedures for clinical trials with more than two treatments, and a description of the algorithm for the conditional Monte Carlo re-randomization test is included. The statistical properties of the randomization test in handling data from multi-armed clinical trials are also explored in comparison to the population-based tests in this chapter, which includes the re-analysis of two three-armed randomized clinical trials using the randomization test. In Chapter 5, we define a confidence interval for the treatment difference in randomization-based estimation and examine an efficient algorithm for computing an interval estimate. Application to

estimating confidence intervals for data from randomized clinical trials as well as a comparison of performance of randomization-based with population-based interval estimation are also incorporated. Finally, future work and concluding remarks of randomization-based and population-based inference in the context of randomized clinical trials are presented in Chapter 6.

Chapter 2: Statistical Properties of Randomization Tests in Two-armed Randomized Clinical Trials

2.1 Randomization Procedures for Two-armed Randomized Clinical Trials

We briefly introduce the randomization procedures that will be compared in the following simulation study. They are (i) complete randomization (CR), (ii) the random allocation rule (RAR), (iii) the truncated binomial design (TBD), (iv) permuted blocked design (PBD), (v) random block design (RBD), (vi) the urn design (UD) (Wei, 1977, 1978), (vii) Efron's biased coin design (BCD) (Efron, 1971), and (viii) the big stick design (BSD) (Soares and Wu, 1982).

In *complete randomization*, all the treatment assignments are independent Bernoulli random variables with success probability 1/2. The *random allocation rule* and *truncated binomial design* are forced balanced procedures, which lead to exactly $n/2$ patients assigned to each of the two treatments. The *random allocation rule* produces $\binom{n}{n/2}$ equiprobable sequences. As we can find in Chapter 3 of Rosenberger and Lachin (2016), "One can think of the *random allocation rule* in terms of an urn model. Suppose an urn contains $n/2$ balls of type A and $n/2$ balls of type B. Each time a patient is ready to be randomized, a ball is drawn and not replaced, and the corresponding treatment is assigned. This continues until the urn is depleted." In the *truncated binomial design*, each assignment is decided by a Bernoulli random variable with success probability 1/2 until one treatment has been assigned $n/2$ times. Then the rest of the patients are assigned to the opposite treatment.

The *permuted block design* and the *random block design* are forced balance designs within blocks. The *permuted blocked design* divides the n treatment assignments into blocks. Each block contains m assignments. Note that the size of the last block may be smaller than m. Within each

block, a forced balance design, usually a *random allocation rule*, is used. Thus the maximum imbalance at any time during the trial is $m/2$. The *random block design* assumes a random size for each block to reduce the chance of selection bias. In our model (Rosenberger and Lachin, 2016), the block sizes are sampled from a set of values (2, 4, ..., $2B_{\text{max}}$) uniformly at random (except the last block), where the maximal block size $B_{\text{max}} \geq 1$.

Other designs that are developed to balance treatment assignments are Efron's *biased coin design*, the *big stick design*, and the *urn design*. Efron's *biased coin design* allocates treatment assignments according to a probability model

$$P(T_i = \text{treatment } A) = \begin{cases} \frac{1}{2}, \text{if } N_A(i) - N_B(i) = 0, \\ p, \text{if } N_A(i) - N_B(i) < 0, \\ 1-p, \text{if } N_A(i) - N_B(i) > 0, \end{cases}$$

where $N_A(i)$ and $N_B(i)$ denote the number of A assignments and the number of B assignment in i allocations, and the parameter $p \in (0.5, 1]$. The *big stick design* is a modification of Efron's *biased coin design* by imposing an *imbalance intolerance* parameter b (Rosenberger and Lachin, 2016). The probability model is given by

$$P(T_i = \text{treatment } A) = \begin{cases} \frac{1}{2}, \text{if } |N_A(i) - N_B(i)| < b, \\ 0, \text{if } N_A(i) - N_B(i) = b, \\ 1, \text{if } N_A(i) - N_B(i) = -b. \end{cases}$$

The *urn design* is an adaptive *biased coin design*, where the probabilities of assignment adapt according to the degree of imbalance (Wei, 1977, 1978). For an *urn design* with parameter α and β, the urn starts with α balls of each of two types (*A* and *B*). If a type A ball is drawn, it is replaced and β type B ball is added to the urn, and vice versa. In the simulation, we set $\alpha = 0$ and $\beta = 1$.

The urn is so designed as to increase the probability of assignment to the treatment that has been selected least often thus far.

2.2 Randomization Tests, Permutation Tests, and the t-test

We first verify the validity of randomization tests in analyzing data from randomized clinical trials in comparison to permutation tests and the *t*-test via simulations. Type I error rate and power for testing the null hypothesis that patient responses are independent of treatment assignments are simulated under a linear time trend model using difference in means statistic under two randomization procedures (TBD, and PBD with RAR within each block). Time trends are systematic temporal changes in measurements. In clinical trials, they can appear with the sequential recruitment of patients as a result of maturation or deterioration of physiological conditions, and is often not perceived by the investigators in advance of the trial. The impact of time trends in biasing the treatment effect appears to be highly dependent on the randomization (Tamm and Hilgers, 2014).

In the simulation, patient responses are sampled from $N(\Delta, 1)$, $\Delta \in \{0, 0.1, \ldots, 2\}$, plus a time trend ranging linearly on the interval (-2.2]. Type I error rate and power of the test correspond to parameter values $\Delta = 0$ and $\Delta > 0$ respectively. The rejection rates are estimated by averaging the rejection indicator over 10,000 simulated data sets with sample size $n = 50$. In each data set, both patient responses and treatment assignment sequence are regenerated. Monte Carlo simulation is applied to compute permutation tests and randomization tests. The p-value is estimated by the Monte Carlo re-randomization test with the number of re-generated treatment assignment sequences $L = 15,000$.

The presence of a time trend invalidates the assumptions of the permutation test (exchangeability (Pesarin, 2001)) and the *t*-test (normality and homogeneity of variance). In addition, when the TBD is employed, the treatment assignment sequences are no longer equally likely, which invalidates the assumptions of the permutation test even further. In Figure 2.1 we see that only the randomization test preserves the nominal type I error rate consistently. For the permutation test and the *t*-test, however, type I error rates are deflated under the PBD and inflated under the TBD. Note also that under the PBD, the power of the randomization test is highest among the three. For instance, the

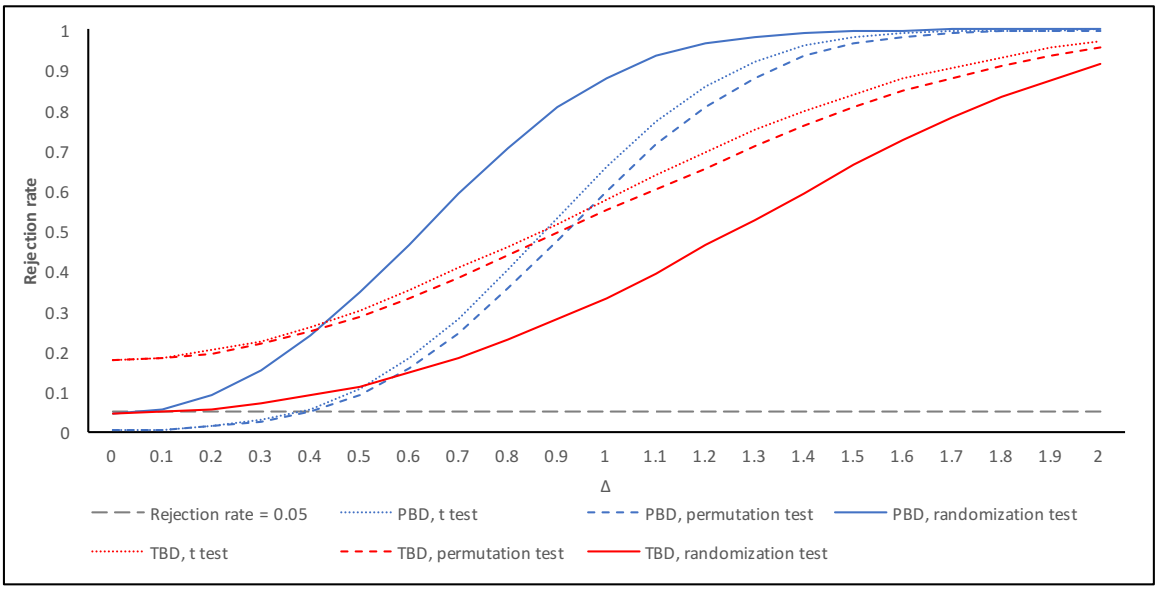

Figure 2.1: *Power curves of the randomization test, the permutation test, and the t-test under a linear time trend model and two randomization procedures.*

power, given by the randomization test, is 0.82 at $\Delta = 0.9$, whereas those given the t-test and the permutation test are 0.52 and 0.49 respectively.

It is usually believed that nonparametric tests have less power than parametric tests if there is no misspecification. If the randomization test is classified as nonparametric, then the results obviously contradict the conventional assertion. It can be seen further that the performance of the permutation test is close to the t-test, yet dissimilar to the randomization test, which emphases the gap between population-based and randomization-based inference in the analysis of randomized clinical trials. The normal-law test is invalid in the presence of the linear drift; it cannot be used to approximate a randomization test.

2.3 Power of Randomization Tests and the Choice of Randomization Procedure and Test Statistic

Though randomization tests are statistically valid by construction, which randomization procedure or test statistic would be most useful for detecting the treatment difference with desired sensitivity depends on the nature of variability in patient responses as well as how large the treatment differences are likely to be (Anscombe, 1948). In this section, we explore the impact of the randomization procedure and test statistic on the power of the randomization test under three models of heterogeneity in patient responses: time trend, outliers, and heavy-tailed outcome data.

2.3.1 Time Trend

The time trend model follows from the previous section. The simulation results are as follows. The lowest power occurs when the TBD is employed, whereas the PBD and RBD achieve the highest power at all Δ values. For example, when the mean shift Δ is 1, the power for TBD, PBD, and RBD are 0.35, 0.88, and 0.90, respectively. Recall that block designs ensure balanced assignment within every block of patients and thereby balance treatment assignments throughout the course of a trial. On the contrary, the TBD can result in a serious imbalance at some point in the trial, for treatment assignments can end with either A's or B's with non-negligible probability. It is also observed that the type I error rate of the test is preserved under all randomization procedures. However, comparing the power curves of the test using the difference in means statistic to those using the simple rank statistics, the power appears to not be sensitive to the change of the test statistic.

Since the PBD and RBD have the highest power, we further examine the performance of the PBD under a series of block sizes $m, m = 2, 4, 8, \ldots, 44, 50$. Note that when $m = 50$, the PBD is the same as the RAR. The result (Figure 2.3) displays a consistent increase of power with the decrease of block size for both test statistics, which implies that the confounding influence of a time trend can be alleviated by adjusting block size in relation to the expected amount of heterogeneity in patient responses over time.

2.3.2 Outliers

The presence of outliers is modeled by a Cauchy distribution $Cauchy(x_0, \gamma)$, where x_0, γ are the location and scale parameter respectively. Under H_0, patient outcomes are sampled from $Cauchy(0, 1)$. Under H_A, patient responses to treatment A are sampled from $Cauchy(\Delta, 1)$. Figure 2.3 shows overlapped power curves under both the difference in means statistic and the simple rank statistic, indicating that the influence of the randomization procedures compared on power is negligible. But the test of the difference in means has relatively low power in comparison to the test based on the linear rank statistic. For example, if $\Delta = 2$, the rejection rates under difference in mean statistic are approximately 0.29, while those under simple rank statistic are approximately 0.88. Given that the Cauchy distribution does not has a finite mean, it is not unexpected that the difference in means statistic appears to be insensitive in detecting the treatment group difference.

We further study a less extreme model. Under H_0, patient outcomes are sampled from $N(0, 1)$ with 10% random contamination sampled from $N(5, 1)$. Under H_A, patient responses to treatment A are sampled from $N(\Delta, 1)$ with 10% random contamination sampled from $N(5 + \Delta, 1)$. Figure 2.4 shows that, when the simple rank statistic is applied, the distinction between power curves under different randomization procedures is negligible. When the difference in means statistic is used, the impact of the choice of randomization procedure is partially observed: the power given by UD(0,1) and CR are lower those given by other procedures. Since balancing treatment assignments is not considered in the design of CR, and the UD tends to CR asymptotically (Rosenberger and Lachin, 2016), this observation is sensible. The result implies that the simple rank statistic is capable of mitigating the influence of unbalanced treatment assignments on power. Moreover, it is again observed that, for all randomization procedures, the test of difference in means produces low power comparing to the test using simple rank statistic. It is therefore concluded that, under the assumed outliers model, the choice of test statistic has a greater impact on power than the choice of randomization procedure.

2.3.3 Heavy-tailed Distribution

Another source of variability in patients is heavy-tailed outcome data. We model this by the exponential distribution $Exp(\Delta)$, where Δ is the distribution mean. Under H_0, patients outcomes are sampled from $Exp(1)$. Under H_A, patient responses to treatment *A* are sampled from $Exp(1+\Delta)$. The results summarized in Figure 2.5 indicate that the influence of the randomization procedure on power is dependent on the choice of test statistic. When using the simple rank statistic, the power of the test is not affected by the choice of randomization procedures, indicated by overlapped power curves for all randomization procedure compared. However, the impact of the randomization procedure is observed when using the difference in means statistic. Specifically, the powers curves of UD and CR are much lower than those given by other randomization procedures. For instance, at $\Delta = 1$, the power under CR, UD(0,1), and RBD is 0.34, 0.50, and 0.64, respectively. This phenomenon is similarly observed under the mixed-normal model in the previous section.

2.3.4 Conclusion

The three extreme examples above demonstrate that choosing a suitable randomization procedure with regard to the research outcome (test statistic) and the clinical circumstance is a dynamic process which can be facilitated by studying the power of randomization tests. We have found that periodic balance of treatment allocation helps mitigate the confounding influence of a time trend on the analysis of the treatment effect. When variability in patient responses is unrelated to a sequential order (e.g, outliers, heavy-tailed), however, the power appears less contingent on the randomization procedures compared, and the choice of test statistic has more significant impact. Other factors may also be considered in the selection of a randomization procedure, including selection bias, balance, ethics, and covariates. More details can be found in Rosenberger and Lachin (2016).

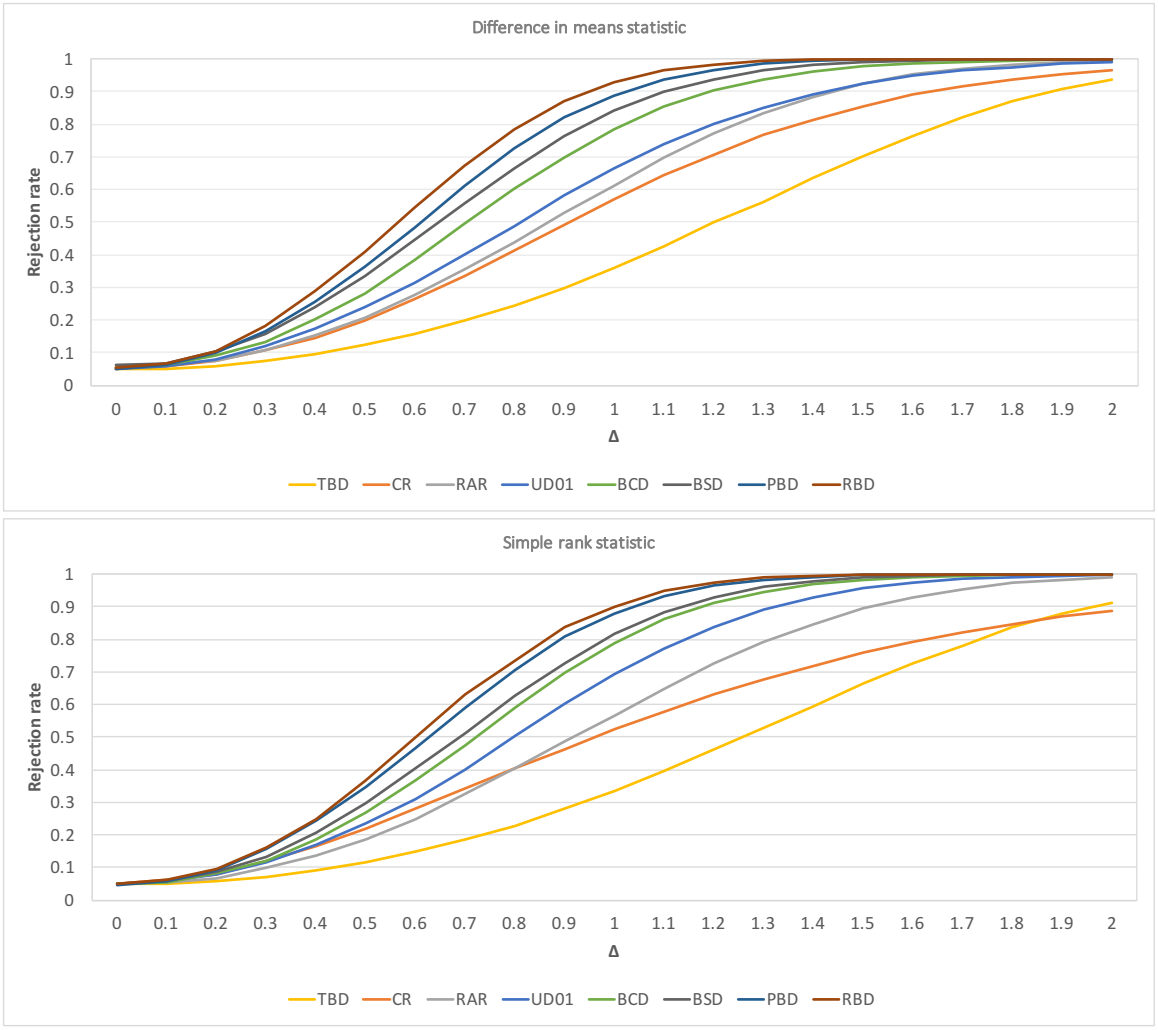

Figure 2.2: *Power curves of randomization tests under a linear drift and eight randomization procedures.*

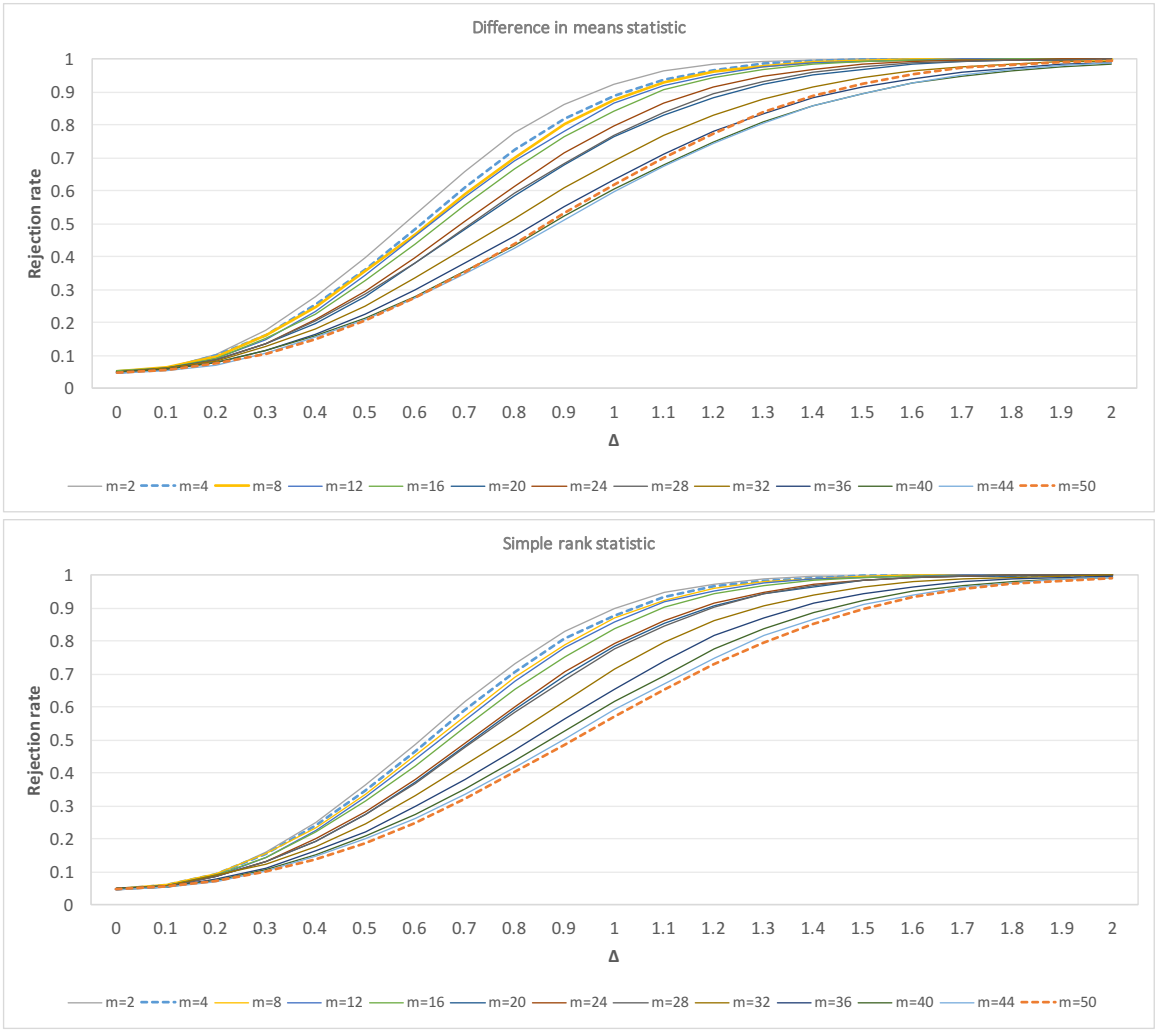

Figure 2.3: *Power curves of randomization tests under a linear drift and PBDs with different block size m.*

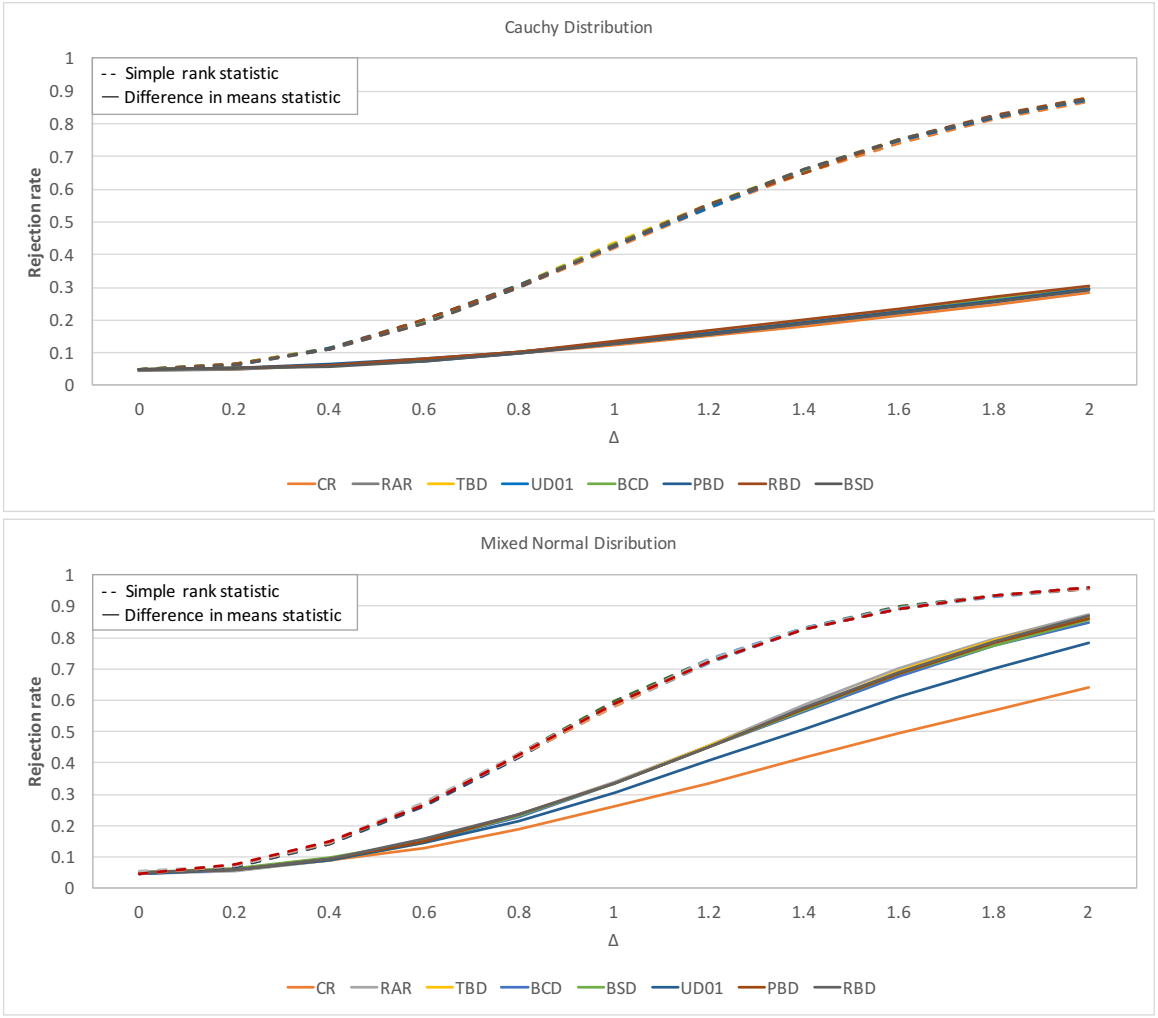

Figure 2.4: *Power curves of randomization tests under two outliers model and eight randomization procedures.*

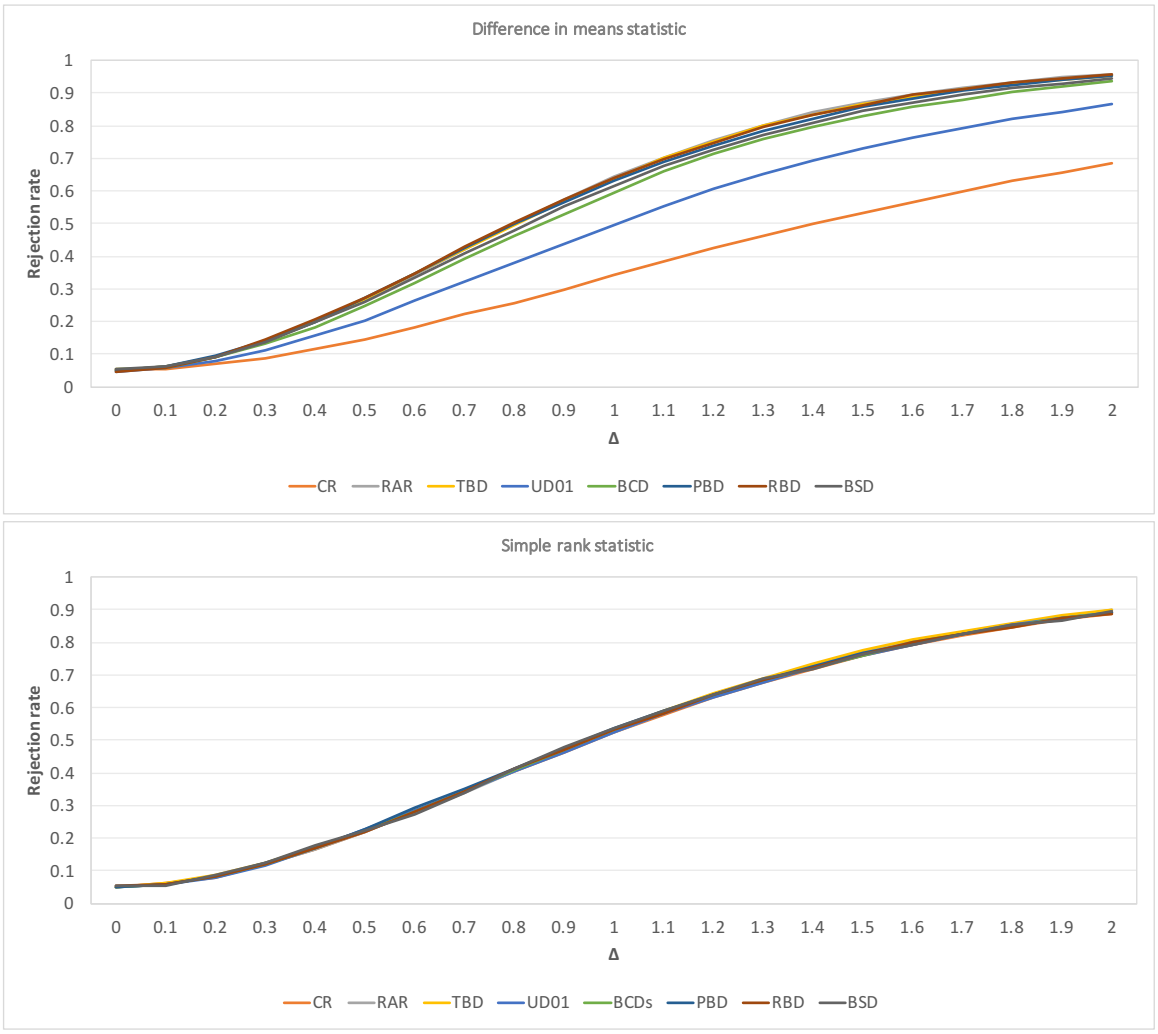

Figure 2.5: *Power curves of randomization tests under a heavy-tailed model and eight randomizations.*

Chapter 3: Randomization Tests for Multi-armed Randomized Clinical Trials

While an adequate amount of literature have focused on randomization tests in two-armed randomized clinical trials, in this chapter, we re-examine the rationale of randomization tests afresh in the context of multi-treatment randomized clinical trials in testing global treatment difference, multiple comparisons, and factorial designs. To develop an illustrative theoretical framework, we refer to the analysis of variance formulation (Kempthorne, 1955).

3.1 The Analysis of Variance

The popularity of analysis of variance (ANOVA) in the 20th century and its applications developed thereof are considered to be attributed to the brilliance of Sir R. A. Fisher: "The first published paper with an analysis of variance was the analysis by Fisher and Mackenzie of the results of a $2 \times 12 \times 3$ factorial experiment on potatoes, which appeared in 1923" (Cochran, 1980).

The ANOVA is a modeling tool for investigating treatment comparisons. Consider a randomized clinical trial evaluating K treatments among n patients, $K > 2$. From the data, an one-way ANOVA table can be computed (Table 3.1). The *treatment mean square* $SS_B/(K-1)$ and the *residual mean square* $SS_W/(n-K)$ are regarded as the ANOVA estimates of the effect variability per treatment group and the residual variability per patient, respectively. The ratio of two mean squares thus gives a meaningful measure of the overall heterogeneity among the treatment groups. Although nowadays ANOVA is usually presented as a way to linearly decompose the variance for given set of realizations of random variables, it was not originally proposed in such a manner. Fisher writes in his discussion to a paper by J. Wishart (Fisher, 1935b) that "the analysis of variance is not a mathematical theorem, but rather a convenient method of arranging the arithmetic." We find in Kempthorne (Kempthorne, 1987) some explanations of the quotation:

Table 3.1: *ANOVA table for comparing k treatments*

Source of Variation	Sum of Squares	d.f.	Mean Square	Ratio of Mean Squares
between treatments	$SS_B = \sum_j n_j(\bar{y}_j - \bar{y})^2$	$K-1$	$\frac{SS_B}{K-1}$	$\frac{SS_B/(K-1)}{SS_W/(n-K)}$
within treatments	$SS_W = \sum_{ij}(y_{ij} - \bar{y}_j)^2$	$n-K$	$\frac{SS_W}{n-K}$	
total	$SS_T = \sum_{ij}(y_{ij} - \bar{y})^2$	$n-1$		

> "Analysis of variance" is not analysis of 'variance' [of random variables] but is analysis of variability and covariability of given *data*. ...analysis of variance is, indeed, a species of "arithmetic" that is related to linear models [to explain the observations] without any conception of random variable.

> ...analysis of variance is not necessarily a decomposition of a sum of squares into quadratic forms, as we can see in model: $y_{ij} = \mu a_i b_j$, first explored by Fisher and Mackenzie (1923) and developed over the past, say, 20 years, by various workers.

We see from these comments that, first, the analysis of variance does not have to be related to random variables and, second, the analysis does not have to be done linearly. Another interesting implication is that linear models of treatment effect on patients does not have to be formulated as conditional expectations of random samples from population distribution function: "the Gaussian error model [and, indeed, the notion that our data (the vector y) are a realization of random variable] is an idea that is a huge leap from the data and the arithmetic" (Kempthorne, 1987).

For example, the response of ith patient in treatment group j can be decomposed as the combination of deviations of treatment means $\bar{y}_j$ from the grand mean $\bar{y}$ and the deviations of patient

responses from the corresponding treatment group mean

$$y_{ij} = \bar{y} + (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j),\ i = 1, \ldots, n_j,\ j = 1, \ldots, K,$$

or

$$(y_{ij} - \bar{y}) = (\bar{y}_j - \bar{y}) + (y_{ij} - \bar{y}_j),\ i = 1, \ldots, n_j,\ j = 1, \ldots, K.$$

The decomposition equation given above does not have apparent connection with the Gaussian error model that $\{y_{ij}\}$'s are population expectations with errors that are normally and independently distributed around zero with constant variance. The identity

$$\sum_{ij} y_{ij}^2 = \sum_{ij} \bar{y}^2 + \sum_{ij} (\bar{y}_j - \bar{y})^2 + \sum_{ij} (y_{ij} - \bar{y}_j)^2$$

or

$$\sum_{ij} (y_{ij} - \bar{y})^2 = \sum_{ij} (\bar{y}_j - \bar{y})^2 + \sum_{ij} (y_{ij} - \bar{y}_j)^2$$

holds true simply as the result of arithmetic properties regardless of the nature of $\{y_{ij}\}$'s, be they random samples or not. In other words, conjoining the notion that data are random samples from population distribution and the basic arithmetic properties is but an additional component. From this arithmetic aspect of ANOVA, we develop the framework for randomization-based multiple treatment comparisons.

3.2 Test of Overall Treatment Difference Significance

Here we briefly discuss two test statistics in evaluating the overall treatment difference that are related to the ANOVA structure: the ratio of mean squares and the Kruskal Wallis H-statistic.

3.2.1 Ratio of Mean Squares and Additive Model

We first review an additive model on the basis of normal theory that is widely used and thoroughly studied. The term "additive" emphasizes that there is no patient-by-treatment interaction. Patient responses in jth treatment group, $j = 1, \ldots, K$, are assumed to constitute a random sample from a normal population with mean β_j and common variance σ^2 (Bhattacharyya and Johnson, 1977). Under the additive model, the population null hypothesis is specified as $\beta_1 = \ldots = \beta_K = 0$. If the assumption of common variance holds true, the ratio of treatment mean square over residual mean square would follow a F distribution with $df = (K-1, n-K)$. If it does not, however, the p-value thus obtained would underestimate the level of significance to be attached to the null (Kempthorne, 1952a). Note that, initially, the F-test is meant for examining the equality of variances of two normal populations. The F statistic, whose distribution is called a F distribution with $(n_1 - 1, n_2 - 1)$ degrees of freedom is given as

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{\sum_i (X_i - \bar{X})^2/(n_1 - 1)\sigma_1^2}{\sum_i (Y_i - \bar{Y})^2/(n_2 - 1)\sigma_2^2},$$

where X_i's and Y_i's are independent random samples from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ (Bhattacharyya and Johnson, 1977).

For randomization-based inference, an additive model that facilitates the interpretation of descriptive statistics is proposed as follows. Suppose uth patient is assigned to treatment j, the response y_u is regarded as the sum of x_u and β_j, where x_u is the response of patient u in the trial under some basic condition (Kempthorne, 1955):

$$\begin{aligned} y_u &= x_u + \beta_j \\ &= \bar{x} + \beta_j + (x_u - \bar{x}),\ u = 1, \ldots, n. \end{aligned}$$

Let δ_u^j be the indicator random variable that patient u is assigned to treatment j. The group mean of

treatment j is thereby expressed as

$$\bar{y}_j = \bar{x} + \beta_j + \frac{1}{n_j}\sum_{u=1}^{n} \delta_u^j (x_u - \bar{x}).$$

Unlikely the normal theory model, the quantity $(x_u - \bar{x})$ is no longer a realization of a normally distributed random variable, but as a fixed *unknown* value that is attached to $\bar{y}_j$ according the distribution of δ_u^j, which is derived from the randomization employed. The significance of the null hypothesis is obtained with reference to the randomization null distribution of the ratio of mean squares statistic. The variance estimator, which is a concern in the population-based model, is irrelevant.

3.2.2 Kruskal-Wallis H statistic

The Kruskal-Wallis H test (Kruskal and Wallis, 1952) is designed to evaluate the hypothesis that the samples are from the same population. It is generally known as a non-parametric test proposed under the random sampling model using the rank statistics of the original measurements. As suggested by Kruskal and Wallis (1952), "one of the most important applications of the test is in detecting differences among the population means."

The Kruskal-Wallis H statistics is formulated as

$$H = \sum_{j=1}^{k} \frac{(\bar{r}_j - \bar{r})^2}{\sigma^2/n_j},$$

where r_i denotes the rank of y_i in the n observations. If there is no ties, then σ^2 reduces to

$$\frac{1}{n-1}\sum_{i=1}^{n}(r_i - \bar{r})^2 = \frac{n(n+1)}{12}, \bar{r} = \frac{n+1}{2},$$

and thus the statistic reduces to

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{k} n_j \bar{r}_j^2 - 3(n+1).$$

The statistics has its relevance to the χ^2 distribution (Kruskal and Wallis, 1952): Let $X_1, \ldots, X_n$ be a random sample from a normal distribution $N(\mu, \sigma^2)$. Then the distribution of statistic

$$\frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} = \frac{s^2}{\sigma^2/(n-1)}$$

is a χ^2 distribution with $n-1$ degree of freedom (Bhattacharyya and Johnson, 1977). For determining p-value, Kruskal and Wallis (1952) state that if the samples y_i's come from identical continuous populations and the group size n_j's are not too small, then H statistic is distributed (approximately) as $\chi^2_{(k-1)}$. In randomization-based inference, the observed test statistic is compared to its randomization null distribution, and the normalization in the above equation is unnecessary.

3.3 Multiple Comparisons

The motivation of developing multiple comparisons is found in Tukey (1949):

> The practitioner of the analysis of variance often wants to draw as many conclusions as are reasonable about the relation of the true means for individual "treatments," and a statement by the F-test (or the z-test) that they are not all alike leaves him thoroughly unsatisfied. The problem of breaking up the treatment means into distinguishable groups has not been discussed at much length, the solutions given in the various textbooks differ and, what is more important, seem solely based on intuition.

The analytical question usually centers around the control of the *familywise error rate* (FWER) (Benjamini and Hochberg, 1995), which is the probability of committing at least one type I error under simultaneous comparisons. First mentioned by Fisher (1935a), the procedure for weak control

was is a two-step process (Hinkelmann and Kempthorne, 2008): at the first step, test the overall null hypothesis by the size α F-test, then, at the second step, test each pairwise comparison by the size α t-test. For achieving strong control, a number of approaches have been proposed, among which the method that makes use of the Bonferroni inequality (Dunn, 1961) is one of the more popular.

An additional element that adds to the complexity of the issue is that whether the type I error rate of a test is truly controlled at the designated level. For randomization-based inference, this, however, is not a concern, provided that the trial is an valid experiment. In the randomization model, the distribution for testing the overall null hypothesis is generated from the entire reference set. For comparing a subset of the treatments, such as a pairwise comparison, the null hypothesis is restricted to the difference between certain treatments. Thus the randomization distribution is to be generated from the corresponding subset of the entire reference set (known as the conditional reference set) that only randomizes the treatments compared in the null (Edgington and Onghena, 2007, Rosenberger and Lachin, 2016). In either case, the distribution is objectively given, although a conditional test may be difficult to compute depending on the complexity of the randomization procedure.

3.4 The Analysis of Factorial Design

In addition to examining the effect of each treatment, a factorial design enables the evaluation of treatment interaction: "Fisher conducted experiments using factorial design to increase the precision of experimental outcomes and to further study how key factors may jointly modify the outcomes" (Berger, 2018). Suppose the researchers investigate four combinations of two treatments each at two levels (Table 3.2). Treatment interaction is understood as follows. If the two treatments were acting independently, then the effect of treatment A would be unaffected by the level of B. The difference between the effects at the two levels is thus a possible measure of the extend to which the treatments interact (under an additive model).

In this situation, the patients are randomized as if entering a four-armed trial, but the analysis is targeted at comparing three treatments (including the placebo). The test of the overall treatment difference can be incorporated easily in the ANOVA using the ratio of mean squares statistic, since

	B	$Placebo$
A	(A, B)	$(A, Placebo)$
$Placebo$	$(B, Placebo)$	$(Placebo, Placebo)$

Table 3.2: A 2$\times$ 2 factorial design

the sum of squares of the three treatment contrasts, $\bar{A}, \bar{B}, \bar{AB}$, provide a partitioning of the total sum of squares into three single df sums of squares (see Hinkelmann and Kempthorne (2008) Section 7.2.3 for details). Alternatively, we may use the H statistic. The ANOVA table for a trial comparing $p \times q$ combinations of two treatments with r replicates is presented in Table 3.3.

Let $\bar{y}_{PP}, \bar{y}_{AP}, \bar{y}_{BP}, \bar{y}_{AB}$ denote the treatment averages of the four groups. Citing Kempthorne (1952b), The average effects of treatment A and B over the two levels can be estimated as

$$\bar{A} = \frac{1}{2}(\bar{y}_{AP} - \bar{y}_{PP} + \bar{y}_{AB} - \bar{y}_{BP}),$$

$$\bar{B} = \frac{1}{2}(\bar{y}_{BP} - \bar{y}_{PP} + \bar{y}_{AB} - \bar{y}_{AP}),$$

and the interaction can be estimated as

$$\bar{AB} = \frac{1}{2}(\bar{y}_{AB} - \bar{y}_{AP} - \bar{y}_{BP} + \bar{y}_{PP}).$$

For the normal theory test, the significance of the null hypotheses ($\bar{A} = 0, \bar{B} = 0, \bar{AB} = 0$) can be obtained by comparing the estimated effects to a t distribution with variance estimated by MSE/r and $df = pq(r-1)$, where $p = q = 2$ and $MSE = SS_E/pq(r-1)$ (Bhattacharyya and Johnson, 1977). Note that for the randomization test, unlike the pairwise comparison, the distribution of the test statistic is derived from the reference set (not the conditional reference set).

Table 3.3: *ANOVA table for a $p \times q$ factorial design with r replicates*

Source of Variation	Sum of Squares	d.f.	Ratio of Mean Squares
factor a	$SS_a = qr\sum_{j=1}^{p}(\bar{y}_j - \bar{y})^2$	$p-1$	$\frac{SS_a/(p-1)}{SS_E/pq(r-1)}$
factor b	$SS_b = pr\sum_{i=1}^{q}(\bar{y}_i - \bar{y})^2$	$q-1$	$\frac{SS_b/(q-1)}{SS_E/pq(r-1)}$
interaction $a \times b$	$SS_{ab} = r\sum_{i=1}^{p}\sum_{j=1}^{q}(\bar{y}_{ij} - \bar{y}_i - \bar{y}_j + \bar{y})^2$	$(p-1)(q-1)$	$\frac{SS_{ab}/(p-1)(q-1)}{SS_E/pq(r-1)}$
residual	$SS_E = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y}_{ij})^2$	$pq(r-1)$	
total	$SS_T = \sum_{i=1}^{p}\sum_{j=1}^{q}\sum_{k=1}^{r}(y_{ijk} - \bar{y})^2$	$pqr-1$	

Chapter 4: Randomization of More Than Two Treatments and Conditional Monte Carlo Re-Randomization Test

4.1 Randomization of More Than Two Treatments

We now discuss how to randomize in a trial with more than two treatments. A selected collection of randomization procedures generalized to the multi-armed treatment allocation are summarized. They are (i) complete randomization, (ii) the random allocation rule, (iii) the truncated binomial design, (iv) the permuted block design and (v) random block design, and (vi) the urn design. In addition, we propose generalizing methods for (vii) Eforn's biased coin design and (viii) the big stick design, which heretofore were designed for only randomizing two treatments.

As we can find in Chapter 3 of Rosenberger and Lachin (2016), "Complete randomization becomes a simple multinomial probability generator with K equally likely outcomes. Then random allocation rule can be thought of as an urn with n/K balls representing each treatment. Truncated binomial randomization becomes a multistage process whereby K-treatment complete randomization is used, and each treatment is subsequently dropped when the n/Kth patient is assigned to that treatment, until only one treatment is left. All subsequent patients are then assigned to that treatment."

The permuted block and random block designs are forced balance designs within blocks. The permuted block design divides the treatment assignments into blocks with equal size (except the last block). The random block design assumes a random size for each block to reduce the chance of selection bias. The block sizes are sampled from a set of values $(K, 2K, \ldots, B_{\max}K)$ uniformly at random, unless the last block is incomplete. For block designs, the random allocation rule or the truncated binomial design is usually applied for randomization within each block. Hence, the block designs for multiple treatment are generalized accordingly.

The urn design (Wei, 1977, 1978) is an adaptive biased coin design where the probabilities of assignment adapt according to the degree of imbalance. For a generalized urn design with parameter α and β, the urn starts with α balls for each type. If one type of ball is drawn, it is replaced and β balls for each of the other types is added to the urn. A generalized urn design with $\alpha = 0$ and $\beta = 1$ has allocation rule (Rosenberger and Lachin, 2016)

$$P\left(T_i = j \mid N_j(i-1)\right) = \frac{i - 1 - N_j(i-1)}{(i-1)(K-1)}, i \geq 2,$$

$$P(T_1 = j) = \frac{1}{K}, j = 1, \ldots, K.$$

We now generalize two procedures that heretofore have not been expressed for $K > 2$ treatments. A generalization of *Efron's biased coin design* (Efron, 1971) is proposed as follows. First, calculate the weight for allocating treatment j at ith assignment as

$$\begin{aligned} W\left(T_i = j \mid N_j(i-1)\right) =& 1/K, \text{ if } KN_j(i-1) - (i-1) = 0, \\ & 2p/K, \text{ if } KN_j(i-1) - (i-1) < 0, \\ & 2(1-p)/K, \text{ if } KN_j(i-1) - (i-1) > 0, \end{aligned}$$

where $j = 1, \ldots, K, p \in (0.5, 1)$. The term $N_j(i-1)$ denotes the number of allocations to treatment j in $(i-1)$ assignments. Hence, if the number of allocations is above the expected value (i.e., $(i-1)/K$), then the weight is decreased. If it is below the expected ratio, then the weight is increased. Denote the value of $W\left(T_i = j \mid N_j(i-1)\right)$ by w_{ij}. The allocating probability of treatment j is defined as $w_{ij} / \sum_j w_{ij}$. The rule reduces to an Efron's biased coin design with p when $K = 2$.

In the generalized *big stick design*, (Soares and Wu, 1982) we guarantee that the absolute pairwise imbalance between any two treatments at any allocation cannot exceed an imbalance intolerance parameter b (positive integer). The allocation rule is as follows. First, calculate the weigh for

allocating treatment j at ith assignment as

$$W\left(T_i = j \mid N_j(i) - N_{j'}(i), j \neq j'\right) = 1/K, \text{ if } |N_j(i) - N_{j'}(i)| < b \text{ for all } j',$$

$$0, \text{ if } N_j(i) - N_{j'}(i) = b \text{ for some } j',$$

$$1, \text{ if } N_j(i) - N_{j'}(i) = -b \text{ for some } j',$$

where $j, j' \in \{1, \ldots, K\}$. Denote the value of $W(T_i = j \mid N_j(i) - N_{j'}(i), j \neq j')$ by w_{ij}. Next, if the lower imbalance limit $-b$ is reached by one or more treatments, then the next assignment would only be chosen from the these treatments with equal probability. That is, if $w_{ij} = 1$ for some j, then for $j' \neq j$ let $w_{ij'} = 0$ if $w_{ij'} < 1$. Finally, treatment j is allocated with probability $w_{ij} / \sum_j w_{ij}$. The rule reduces to a big stick design with imbalance tolerance b when $K = 2$.

The validity of the randomization procedures above is based on the principle that the allocation ratio at every assignment be preserved (i.e., $P(T_i = j) = 1/K, i = 1, \ldots, n, j = 1, \ldots, K$). (Kuznetsova and Tymofyeyev, 2011) The proofs for the generalized Efron's biased coin design and the generalized big stick design are relegated to an appendix. The other procedures are easier to see due to their interchangeable structures.

These generalized procedures preserve the allocation ratio at every assignment (i.e., $P(T_i = j) = 1/K, i = 1, \ldots, n, j = 1, \ldots, K$), since the allocating probabilities of the treatments are interchangeably defined. We now prove this property for the generalized Efron's biased coin design and the generalized big stick design.

From the definition of the generalized Efron's biased coin design, the sum of all the weights assigned at ith allocation is

$$\sum_{j=1}^{K} w_{ij} = \frac{1}{K} \sum_{j=1}^{K} I_{\left(N_j(i) = \frac{i}{K}\right)} + \frac{2p}{K} \sum_{j=1}^{K} I_{\left(N_j(i) < \frac{i}{K}\right)} + \frac{2(1-p)}{K} \sum_{j=1}^{k} I_{\left(N_j(i) > \frac{i}{K}\right)},$$

where $N_j(i)$ is the number of treatment j in i allocations. Since there are K treatments to be

assigned at each allocation, we also have the identity

$$\sum_{j=1}^{K} I_{\left(N_j(i)=\frac{i}{K}\right)} + \sum_{j=1}^{K} I_{\left(N_j(i)<\frac{i}{K}\right)} + \sum_{j=1}^{K} I_{\left(N_j(i)>\frac{i}{K}\right)} = K.$$

Combing the two equations, we have

$$\sum_{j=1}^{K} w_{ij} = 1 + \frac{2p-1}{K}\left(\sum_{j=1}^{K} I_{\left(N_j(i)<\frac{i}{K}\right)} - \sum_{j=1}^{K} I_{\left(N_j(i)>\frac{i}{K}\right)}\right).$$

Thus the allocation rule can be expressed as, for any $i = 1, \ldots, n, j' = 1, \ldots, K$,

$$\begin{aligned} E(N_{j'}(i+1) - N_{j'}(i) \mid N_1(i), \ldots, N_K(i)) &= P(T_i = j' \mid N_1(i), \ldots, N_K(i)) \\ &= \frac{1}{K + (2p-1)\sum_{j=1}^{K}(I_{\left(N_j(i)<\frac{i}{K}\right)} - I_{\left(N_j(i)>\frac{i}{K}\right)})}, \text{if } N_{j'}(i) = \frac{i}{K}, \\ &\quad \frac{2p}{K + (2p-1)\sum_{j=1}^{K}(I_{\left(N_j(i)<\frac{i}{K}\right)} - I_{\left(N_j(i)>\frac{i}{K}\right)})}, \text{if } N_{j'}(i) < \frac{i}{K}, \\ &\quad \frac{2(1-p)}{K + (2p-1)\sum_{j=1}^{K}(I_{\left(N_j(i)<\frac{i}{K}\right)} - I_{\left(N_j(i)>\frac{i}{K}\right)})}, \text{if } N_{j'}(i) > \frac{i}{K}. \end{aligned}$$

Taking expectation, we have

$$\begin{aligned} E(N_{j'}(i+1) - N_{j'}(i)) &= \frac{P(N_{j'}(i) = \frac{i}{K}) + 2pP(N_{j'}(i) < \frac{i}{K}) + 2(1-p)P(N_{j'}(i) > \frac{i}{K})}{E\left(K + (2p-1)\sum_{j=1}^{K}(I_{\left(N_j(i)<\frac{i}{K}\right)} - I_{\left(N_j(i)>\frac{i}{K}\right)})\right)} \\ &= \frac{1 + (2p-1)\left(P(N_{j'}(i) < \frac{i}{K}) - P(N_{j'}(i) > \frac{i}{K})\right)}{K(1 + (2p-1)\frac{\sum_{j=1}^{K}\left(P(N_j(i)<\frac{i}{K}) - P(N_j(i)>\frac{i}{K})\right)}{K})} = \frac{1}{K}, \end{aligned}$$

since by symmetry

$$P(N_j(i) < \frac{i}{K}) = P(N_j(i) > \frac{i}{K}) = \frac{1 - P(N_j = \frac{i}{K})}{2}, j = 1, \ldots, K.$$

Likewise, we prove the preservation of allocation ratio of the generalized big stick design. From the definition, at ith allocation, if there exists treatment j such that $N_j(i) - N_{j'}(i) = -b$ holds true for some $j', j' \neq j$, then treatment j would be assigned with probability 1 in the next allocation, and treatment j' would be not be assigned since the upper imbalance limit is reached (i.e., $N_{j'}(i) - N_j(i) = b$). Otherwise, if $|N_j(i) - N_{j'}(i)| < b$ for all $j = 1, \ldots, K, j' \neq j$, then all the treatments would be assigned with probability $1/K$ in the next allocation. Therefore, the sum of all the weights at ith allocation is

$$\begin{aligned}\sum_{j=1}^{K} w_{ij} =& (1 - \prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)}) \sum_{j=1}^{K} I_{\left(N_j(i) - N_{j'}(i) = -b, \forall j' \neq j\right)} \\ &+ \frac{1}{K} \prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)}.\end{aligned}$$

The probability of allocating treatment j'' at position i is thus rewritten as, for $i = 1, \ldots, n$,

$$\begin{aligned}E(N_{j''}(i+1) - N_{j''}(i) \mid N_{j''}(i) - N_{j'}(i), \forall j' \neq j'') &= P(T_i = j'' \mid N_{j''}(i) - N_{j'}(i), \forall j' \neq j'') \\ &= \frac{I_{\left(N_{j''}(i) - N_{j'}(i) = -b, \forall j' \neq j''\right)}}{\sum_{j=1}^{K} I_{\left(N_j(i) - N_{j'}(i) = -b, \forall j' \neq j\right)}}, \text{ if } \prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)} = 0, \\ &\quad \frac{1}{K}, \text{ if } \prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)} = 1.\end{aligned}$$

Taking expectation, we have

$$
\begin{aligned}
&E(N_{j''}(i+1) - N_{j''}(i)) \\
&\quad = E\left(\frac{I_{\left(N_{j''}(i) - N_{j'}(i) = -b, \forall j' \neq j''\right)}}{\sum_{j=1}^{K} I_{\left(N_j(i) - N_{j'}(i) = -b, \forall j' \neq j\right)}} \right) P(\prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)} = 0) \\
&\qquad + \frac{1}{K} P(\prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)} = 1) \\
&\quad = \frac{P(N_{j''}(i) - N_{j'}(i) = -b, \forall j' \neq j'')}{\sum_{j=1}^{K} P(N_j(i) - N_{j'}(i) = -b, \forall j' \neq j)} P(\prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)} = 0) \\
&\qquad + \frac{1}{K} P(\prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| < b, \forall j' \neq j\right)} = 1).
\end{aligned}
$$

Because the allocation rule is interchangeably defined among the treatments, all K treatments have equal probability to achieve a same event. By symmetry, we have

$$
P(N_j(i) - N_{j'}(i) = -b, \forall j' \neq j) = P(N_{j''}(i) - N_{j'}(i) = -b, \forall j' \neq j''), \text{ for any } j'' \neq j.
$$

Hence, the expectation reduces to

$$
\begin{aligned}
E(N_{j'}(i+1) - N_{j'}(i)) = &\frac{1}{K} P(\prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| b, \forall j' \neq j\right)} = 0) \\
&+ \frac{1}{K} P(\prod_{j=1}^{K} I_{\left(|N_j(i) - N_{j'}(i)| b, \forall j' \neq j\right)} = 1) = \frac{1}{K}.
\end{aligned}
$$

4.2 Conditional Monte Carlo Re-Randomization Test

To recap, the p-value of a randomization test is estimated by a Monte Carlo re-randomization test. The idea is to use Monte Carlo simulation to estimate the randomization distribution of a test statistic. For given patient response data, the treatment assignment sequence is regenerated L times, and the test statistic is re-computed each time. The p-value is determined as the proportion of the L simulations that gives a test statistic value S_l which is at least as extreme as the observed test statistic $s_{obs.}, 1 \leq l \leq L$. The two-sided Monte Carlo p-value estimator is calculated as (Plamadeala and Rosenberger, 2012):

$$\hat{p} = \frac{\sum_{l=1}^{L} I(|S_l| \geq |s_{obs.}|)}{L}.$$

The test statistic S is chosen to incorporate the information on the treatment effect.

In the multiple comparisons problem, the re-randomization test is carried out on the conditional reference set, which only randomizes the treatments that are included in the hypothesis while keeping the other treatment assignments fixed. For certain randomization procedures, for example, complete randomization, the random allocation rule, and the block designs with each block filled by the random allocation rule, sampling directly from the conditional reference set is equivalent to re-randomizing by the corresponding two-treatment procedures. For more complicated procedures that do not generate treatment assignment sequences with equal probability (e.g. the truncated binomial design, Efron's biased coin design, the big stick design, and Wei's urn design), the convenient approach does not apply. One may consider generating a massive number of sequences and discard those that do not satisfy the conditions necessary to make a subgroup comparison. However, the computational intensity prevents the naive approach from being feasible. In this regard, we develop a new algorithm that sequentially selects each element of the sequence from the conditional reference set, so that, without loss of efficiency, only the desired treatment would be generated at the position. We explain the method with the following example.

Consider a randomized clinical trial investigating the differences between four treatments $1, 2, 3, 4$. Let the observed treatment assignment sequence be $\boldsymbol{T}^* = (1, 3, 2, 3, 4, 2, 1, 4)$. Suppose the trial

was randomized according to randomization procedure ϕ. For a pairwise comparison of treatment 1 and treatment 4, we need to sample from a subset of the reference set that only randomizes treatment 1 and 4. In other words, we want to generate treatment assignment sequences only of the form $\boldsymbol{T} = (1 \text{ or } 4, 3, 2, 3, 1 \text{ or } 4, 2, 1 \text{ or } 4, 1 \text{ or } 4)$ with respect to P_ϕ, the probability distribution derived from procedure ϕ.

At position i, let p_{ij} be the probability of generating treatment j with respect to $P_\phi, j = 1, \ldots, 4, i = 1, \ldots, 8$. To re-generate a treatment assignment from the *unconditional* reference set, we sample a random number y from the uniform distribution ranging from 0 to $\sum_j p_{ij}$. If $p_{i1} < y < p_{i1} + p_{i2}$, for instance, then assign treatment 2 at the position i, thereby $T_i = 2$ with probability p_2. To re-generate a treatment assignment from the *conditional* reference set, we impose the following constrains. If $T_i^* = 1$ or 4 and, further, if $p_{i1} + p_{i4} > 0$, then continue sampling y until $y < p_{i1} + p_{i4}$. Next, let $T_i = 1$ if $y < p_{i1}$, and $T_i = 4$ if $p_{i1} \leq y < p_{i1} + p_{i4}$. However, if $p_{i1} + p_{i4} = 0$, then the treatment assignment sequence produced so far cannot be from the conditional reference set, because the probability of having either treatment 1 or treatment 4 at position i is zero. Thus we discard this re-randomization and start a new one. Similarly, if $T_i^* = 2$ and, further, if $p_{i2} > 0$, then assign treatment 2 to T_i. If $p_{i2} = 0$, we restart the re-randomization. In this way, the sequences are generated under P_ϕ and those that are not in the conditional reference set are not produced.

The process terminates when the re-randomization sequences reaches L, and then allows the computation of the p-value estimate. The discontinuation of re-randomization takes place only in the generalized big stick design, where the situation $p_{ij} = 0$ can be encountered. Even in this case, the algorithm is efficient and the computation of a single p-value estimate is completed in seconds with a regular laptop. Citing an example in our simulation, it takes $378,643$ re-randomization to produce $20,000$ desirable sequences when $b = 3$, Note that the expected number of discontinuations is proportionate to the cardinality of the corresponding conditional reference set.

Now we demonstrate that the algorithm is equivalent to the naive approach that samples a large amount of sequences and keeps only those satisfying the conditions. Let $\boldsymbol{t} \in \Omega_c$, the conditional reference set. Suppose L sequences are regenerated naively under P_ϕ, and $\boldsymbol{T}_l = \boldsymbol{t}$ for some $l, 1 \leq$

$l \leq L$. Then the Monte Carlo conditional p-value estimate of $P_c(\boldsymbol{T} = \boldsymbol{t})$ is calculated as

$$P_c(\boldsymbol{T} = \boldsymbol{t}) = \frac{P(\boldsymbol{T} = \boldsymbol{t}, \boldsymbol{T} \in \Omega_c)}{P(\boldsymbol{T} \in \Omega_c)} \approx \frac{\sum_{l=1}^{L} I_{(\boldsymbol{T}_l = \boldsymbol{t})} I_{(\boldsymbol{T}_l \in \Omega_c)} / L}{\sum_{l=1}^{L} I_{(\boldsymbol{T}_l \in \Omega_c)} / L}.$$

If $\boldsymbol{T}_l \in \Omega_c$ for all l, then the denominator reduces to 1, and the equation reduces to

$$\hat{P}_c(\boldsymbol{T} = \boldsymbol{t}) = \frac{\sum_{l=1}^{L} I_{(\boldsymbol{T}_l = \boldsymbol{t})}}{L},$$

which is also the formula for computing the Monte Carlo conditional p-value estimate when applying the algorithm we proposed. The algorithm produces sequences in such a way that, at each position, treatments are sampled with probability defined by P_ϕ while the undesirable treatments are discarded. Consequently, $\boldsymbol{T}_l \in \Omega_c$ for all l.

4.3 Simulation of Error Rates

In this section, we apply randomization tests to K treatment comparisons, $K = 4$, and examine the impact of the randomization procedure and the test statistic on the power of the test under two models of variability in patient responses: time trend and outliers. Under the alternative hypotheses H_A, the treatments compared have constant additive effects on patient responses denoted by $\beta_j, j = 1, \ldots, 4$. In each situation, eight randomization procedures and two test statistics (one based on the original values and one based on the order statistics of the values) are compared. The error rates are averaged across $10,000$ simulated data sets. In each simulation, both patient responses and treatment assignment sequence are regenerated. The p-value is estimated by the Monte Carlo re-randomization test with the number of re-generated sequences $L = 20,000$ and sample size $n = 100$.

The eight randomization procedures compared in the simulation study are described in Chapter 4: complete randomization (CR), the random allocation rule (RAR), the truncated binomial design (TBD), the urn design with parameters $\alpha = 0, \beta = 1$ (UD(0,1)), Efron's biased coin design (BCD)

with parameter $p = 2/3$, permuted blocked design (PBD) with block size $m = 8$, random block design (RBD) with maximal block size $KB_{\max}$, $B_{\max} = 3$, the big stick design (BSD) with imbalance intolerance parameter $b = 3/2$. For the block designs, a RAR is used within each block.

4.3.1 Time Trends

The time trend is modeled by a linear drift. Patient responses to treatment j are sampled from normal distribution $N(\beta_j, 1)$ plus a linear drift ranging on the interval $(-2, 2]$, where $\beta_1 = 0, \beta_2 = \Delta, \beta_3 = 1.5\Delta, \beta_4 = 1.75\Delta, \Delta \in \{0, 0.1, \ldots, 1\}$. The null and the alternative hypothesis correspond to the cases where $\Delta = 0$ and $\Delta > 0$, respectively.

In the first step, we examine the test of the overall null hypothesis that patient responses are independent of the treatments. We compare the performance of three tests: the randomization test using the ratio of mean squares statistic, the population-based F test using F distribution with $df = (3, 96)$, and the population-based Kruskal-Wallis H test using χ^2 distribution with $df = 3$. The simulation results show interesting information (Figure 4.1). First, the type I error rate of the (population-based) F test is highly inflated under TBD, and is deflated under other randomization procedures. In particular, the block designs give the most deflated error rate. Only the error rates under the RAR and CR are preserved at level 0.05. The nonparametric Kruskal-Wallis H test gives similar results. The type I error rates of the randomization tests are preserved for all procedures. Second, the power of test changes with the change of randomization procedures. For randomization tests, the highest power is achieved when using the block designs, which is followed by the biased coin designs. In particular, the BSD gives higher power than the BCD. It is also observed that the power curve under the urn design is below that under the RAR. The lowest power is seen when using the TBD, a procedure that can result in serious imbalanced at some stage in the trial. When complete randomization is employed, the power of the three tests (i.e., randomization test using the ratio of mean squares statistic, the population-based F test, and the Kruskal-Wallis H test) are almost the same. Under the RAR, the power of the randomization test is slightly higher then those given by the other two.

Next, we examine the performance of the conditional randomization test for subgroup comparisons. The conditional randomization test and the t-test are compared for a pairwise comparison (treatment 1 and treatment 4) using the difference in means statistic. A t distribution with $df = 96$ and variance estimated by $(n_1^{-1} + n_4^{-1})MSE$ is used for calculating the rejection rate of the t-test. The results are summarized in Figure 4.3. For the conditional randomization test, the type I error rates are preserved. A slight inflation of the type I error rate (i.e., 0.058) is seen for our generalization of the BSD. We do not see this phenomena on other occasions, for example, if K is 2. This is unexpected. Note that the our simulation can only estimate the type I error rate to two digits with accuracy given the number of replications. The power curves given by the block designs overlap, and the power given by CR and the RAR overlap. For the t-test, the type I error rate is deflated under the block designs. The power curves under CR and the RAR are close to those given by the randomization test.

Lastly, we apply the randomization test to a factorial design. Let treatment $1, 2, 3, 4$ represent $(placebo, placebo)$, $(A, placebo)$, $(B, placebo)$, (A, B), respectively, then the simulation scenario can be expanded to a factorial design provided that the randomization procedure assigns equal number of treatments to each group. The test of average effects and treatment interaction are presented in Figure 4.2. For the randomization test, the change of power with regard to the randomization procedures is again observed, and the changing pattern is consistent with that presented in the overall comparison in Figure 4.1. For the t-test, the power curve displays feature similar to that observed in the F test (Figure 4.1): the type I error rate is inflated under the TBD, deflated under the block designs, and preserved under the RAR. Note also that the test of average effect of B gives the highest power, and the test of treatment interaction gives the lowest power for all $\Delta > 0$. Given that the expectation of treatment effect under the alternative hypotheses are $\bar{A} = 0.625\Delta$, $\bar{B} = 1.125\Delta$, $\bar{AB} = 0.125\Delta$, respectively, the observation is expected.

4.3.2 Outliers

The presence of outliers is modeled by the Cauchy distribution, $Cauchy(x_0, \gamma)$, where x_0, γ are the location and scale parameters. Patient responses to treatment j are sampled from $Cauchy(\beta_j, 1)$,

where $\beta_1 = 0, \beta_2 = \Delta, \beta_3 = 1.5\Delta, \beta_4 = 1.75\Delta, \Delta \in \{0, 0.1, \ldots, 1\}$. The null and the alternative hypothesis correspond to $\Delta = 0$ and $\Delta > 0$, respectively. We first compare the randomization test using the ratio of mean squares statistic with the randomization test using Kruskal Wallis H statistic. Figure 4.4 shows that the power of the test using the H statistic is considerably higher than that using the ratio of mean squares statistic. In the mean time, the influence of randomization procedures on power is not evident, indicated by overlapping curves. Furthermore, we obtain the power curve of the formal Kruskal-Wallis H test (Figure 4.4). Comparing the two plots, we see that the power of the formal H test approximates the power of the randomization test in this situation.

4.3.3 Conclusions

The statistical validity of the randomization test is demonstrated by the above simulations in analyzing data from multi-armed randomized clinical trials, in terms of overall treatment comparison, pairwise comparison, and comparison in a factorial design. The type I error rate is preserved by construction, as the randomization distribution of any test statistic is always correct for the set of data under repeated experiments. We conclude that periodic balance of treatment allocation in randomization procedure improves the sensitivity in detecting existing treatment effects in the presence of time trend, which is observed alike in the simulation results of all three categories of treatment comparison. Second, in the presence of outliers, the variability in patient responses is not related to the sequential order of treatment assignment, and the power of detecting existing treatment effects in an overall test appears less affected by the randomization procedures compared, but is increased when changing the test statistic from the ratio of mean squares to the H statistic. Third, the population-based test, either parametric or nonparametric, is not always valid for analyzing data from (multi-armed) randomized clinical trials, except when the RAR or CR is employed. This observation agrees with Kempthorne's conclusion from the mathematical proof on early occasions (Kempthorne, 1952a, 1955) that the normal theory test serves as an approximation to the randomization test.

4.4 Case Study

4.4.1 Multiple Tumor Recurrence Data for Patients with Bladder Cancer

As an illustration of the method, we consider the data from a randomized trial conducted by Byar, Blackard, and the Veteran's Administration Co-operate Urology Research Group (1980). The data are found in Andrews and Herzberg (1985). A total of 121 patients recruited at ten hospitals were randomly assigned to one of the three treatments with equal probability: placebo, pyridoxine, or thiotepa instillation. The primary outcome is the rate of tumor recurrence per 100 patient months follow-up (Byar, 1980). The treatment effect model assumes that the survival time in month between two recurrences follows the exponential distribution defined by a constant recurrence rate. To evaluate the difference in recurrence rate among treatment groups, the test statistic was chosen to be the ratio of two group means (i.e., the sum of survival times over the number of uncensored recurrences), and was compared to an F-distribution (Byar, 1980, Byar et al., 1977, Gehan, 1975). The five patients who were lost to follow up at the beginning were excluded in the analysis (Byar, 1980). One-sided tests showed that thiotepa differs significantly from placebo ($p = 0.012$) and from pyridoxine ($p = 0.019$), and the difference between placebo and pyridoxine is not significant (Byar, 1980). To repeat the analysis, we compute the pairwise p-value by the parametric test mentioned above and the conditional randomization test, where treatment assignments are re-randomized via complete randomization and $L = 2,000,000$. We deliberately choose a large L to guarantee the precision of the p-value estimate. The results from the parametric test agree with the original analysis, but the p-value estimates from the conditional randomization test do not show similar significance for the first two comparisons (Table 4.1). We may therefore infer that, for this experiment, the parametric test underestimates the p-value when the significance level is small in comparison to the conditional randomization test.

4.4.2 Gallstones Data from The National Cooperative Gallstone Study

Another trial we considered is the National Cooperative Gallstone Study that was conducted to determine the efficacy and safety of using chenodiol for dissolution of gallstones (Schoenfield and

Table 4.1: *One-sided p-value from the parametric test and the conditional randomization test*

	Thiotepa vs placebo	Thiotepa vs pyridoxine	Pyridoxine vs placebo
Conditional randomization test	0.07	0.15	0.51
Parametric test	0.019	0.012	0.51
Original results	0.019	0.012	not significant

Lachin, 1981). In the major study, 916 patients were randomly assigned to one of the three treatments: high dose of chenodiol, low dose of chenodiol, or placebo. The primary outcome is the proportion of patients whose gallstones were completely dissolved during the 24 months of follow up. The comparison of the probabilities of events across time was evaluated by the Mantel chi-squared test ($p < 0.0001$) based on the total cohort of 916 patients (Schoenfield and Lachin, 1981). The randomization followed a generalized big stick design[1] with imbalance parameter $b = \theta(j)\sqrt{2j}/3$, where $\theta(j) = 2$ for $j < 8, \theta(j) = 12/j + 1/2$ for $j \geq 8, j = 1, \ldots, n$, and treatment assignment sequences were separately generated, inspected, and modified to satisfy the study requirements before being assigned to the ten participating clinics (see Lachin, Marks, and Schoenfield (1981) for details). Two clinics withdrew and were replaced during the study and an additional sequence was generated for this reason, which makes an exact re-randomization difficult. In computing the randomization test, we combine the patients in a withdrawn clinic with a newly recruited clinic, and, at each re-randomization, generate one sequence for each clinic with regard to the big stick design without additional considerations. The overall comparison is evaluated by the ratio of mean squares statistic in Table 1, the randomization-based analog of the population-based F-test. The test ($L = 2,000,000$) gives a significant result ($\hat{p} < 0.0001$), which agrees with the original analysis.

[1]This method of generalizing the big stick design is not applicable to the situation where more than three treatments are involved.

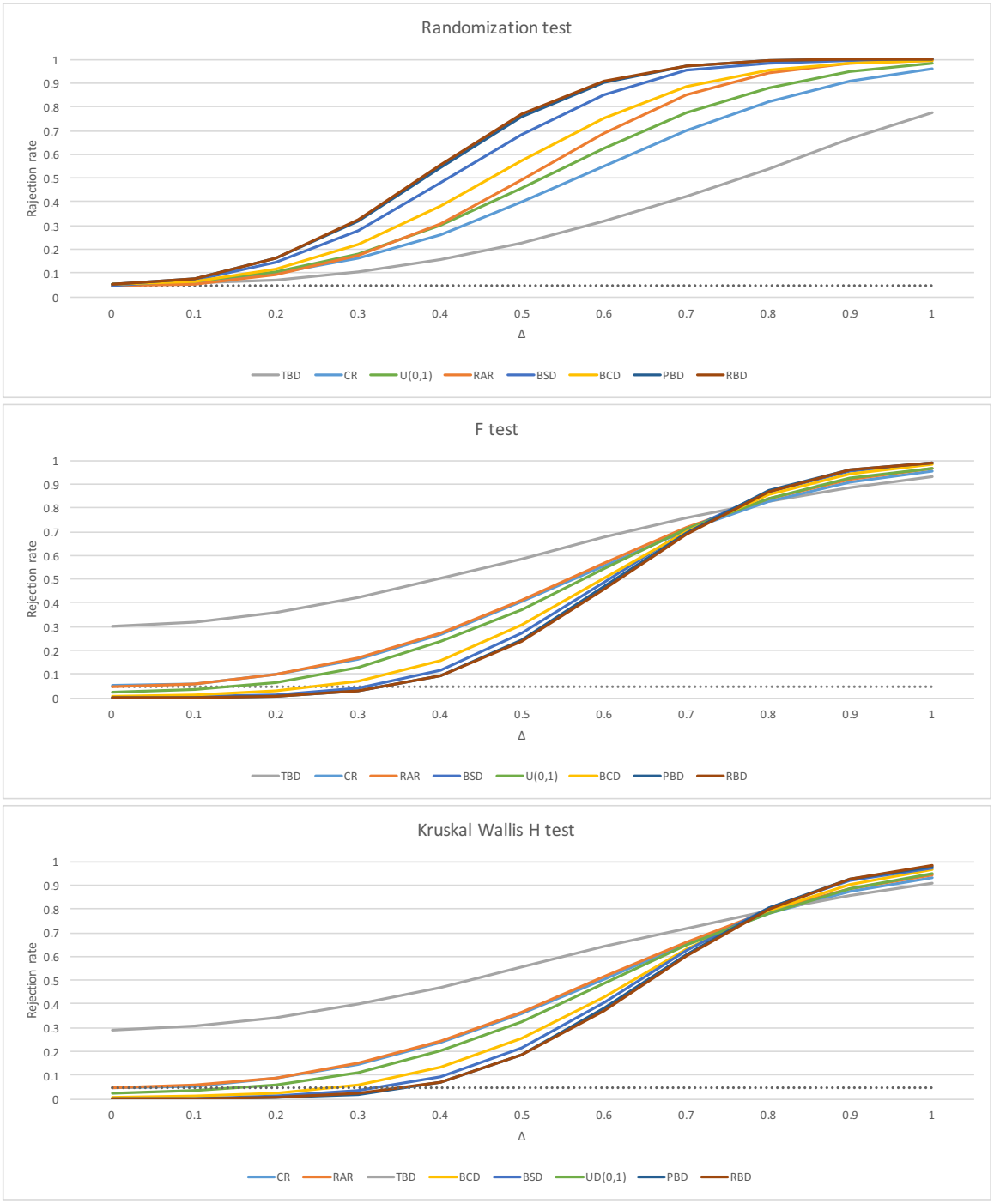

Figure 4.1: *Power curves of the randomization test with the ratio of mean square statistic, F-test, and Kruskal-Wallis H test under a linear drift and eight randomization procedures*

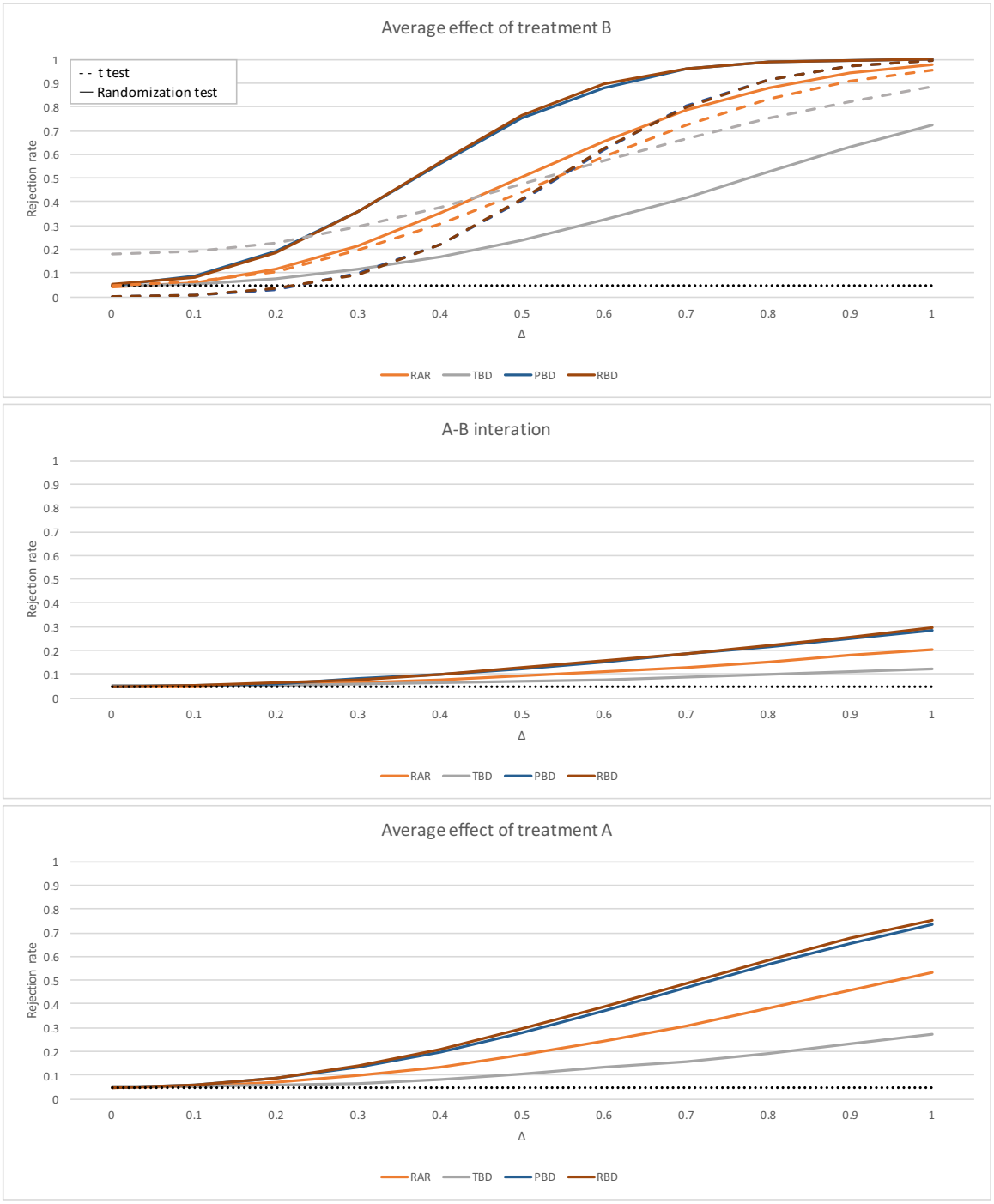

Figure 4.2: *Power curves of the randomization test of average treatment effects and treatment interaction in a factorial design under a linear drift and four randomization procedures*

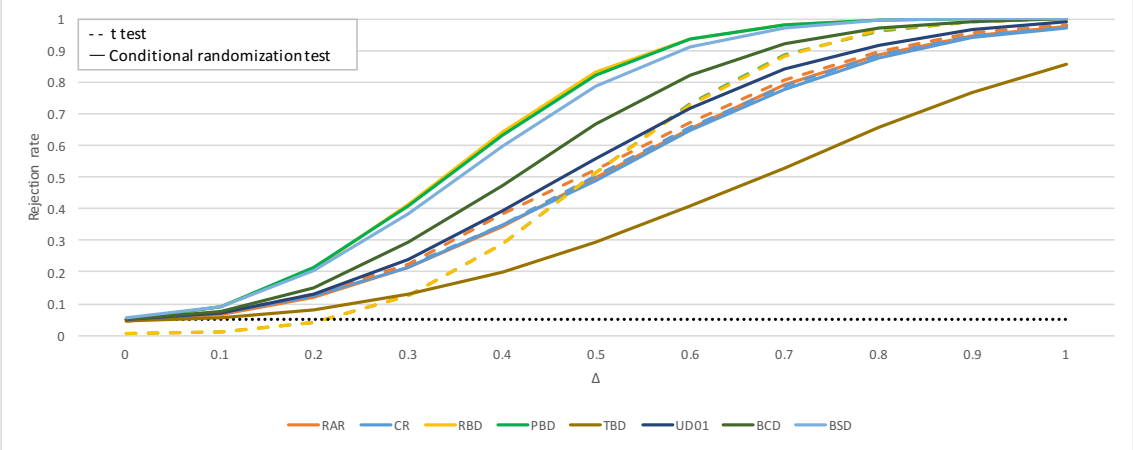

Figure 4.3: *Power curves of the conditional randomization test and the t-test in pairwise comparison under a linear drift*

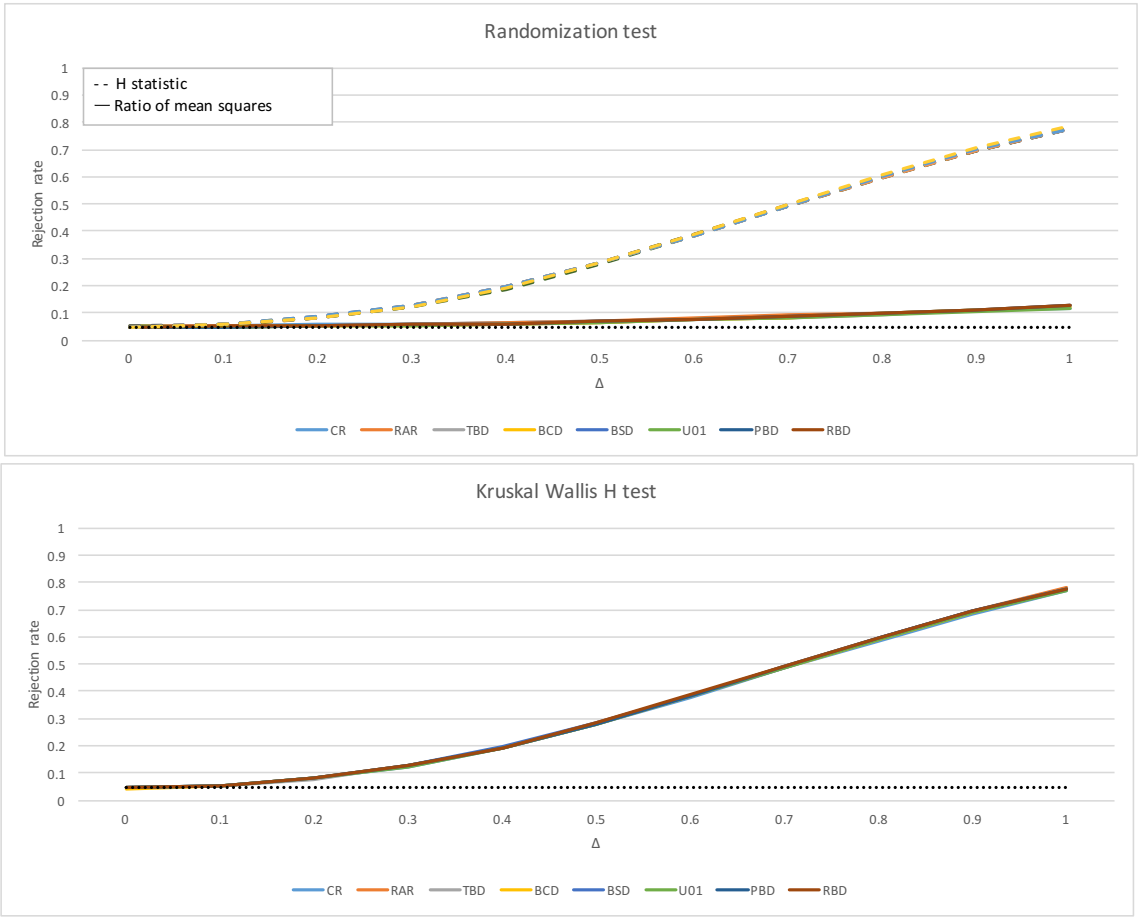

Figure 4.4: *Power curves of the randomization test and the Kruskal-Wallis H test under a outliers model and eight randomization procedures*

Chapter 5: Confidence Interval Procedures

5.1 Introduction

In this chapter, we explore another important aspect of randomization-based inference, namely estimating a confidence interval for a treatment effect. Statistical estimation is considered to be the primary method of reporting evidence from a sample of data (Bhattacharyya and Johnson, 1977). The estimation has predictive value for circumstances under which it was developed. The predictive value is the prerequisite for, rather than the result of, statistical analysis. In the analysis of clinical trials, a point estimator calculated from the sample data provides an estimate of the treatment difference in terms of a single number, even though the standard error, as a statement of accuracy, can be attached to it. An alternative approach to estimation, that is, estimation by interval, was introduced by Neyman (1937). The approach extends the notion of error bound and produces an interval of numerical values of the treatment difference, to be calculated from the data and upon the hypothesis of the mathematical model of the treatment difference, that is acceptable under the prescribed type I error rate.

The formation and interpretation of confidence interval estimation procedure is very distinct between population-based and randomization-based inference. In population-based inference, the confidence interval is usually determined by the inversion of a hypothesis test. In randomization-based inference, the one-to-one correspondence does not exist, and the interval is constructed without introducing the notion of repeated sampling of patients responses from some distribution. Instead, the confidence interval of a constant additive effect Δ, for instance, is understood to be a set of Δ values for which the hypothesis H_Δ that the treatment difference is Δ for each and every patient is not rejected at the prescribed significant level based on the given set experimental data (Edgington and Onghena, 2007, Kempthorne, 1977, 1979). The mathematical model of the treatment difference is specified independently. To search for the confidence limits, the method is to consider

a sequence of values, $\Delta_1, \ldots, \Delta_n$, and perform a corresponding one-sided randomization test of H_{Δ_i} for each $\Delta_i, i = 1, \ldots, n$, until the desired significant level is approached. The problem in its practical aspect is not merely a computational one, but also involves probability theory, specifically, convergence theory.

On the basis of the Robbins-Monro process, a search algorithm is developed by Garthwaite (1996). The coverage probability of the confidence limit estimates is unbiased and have small variance as the number of search steps goes large, and the process is computationally efficient (Garthwaite, 1996). However, the re-randomization procedure in Garthwaite's search process permutes patient outcomes while holding the treatment assignments unchanged, which is not equivalent to more general circumstances in randomized clinical trials where the number of assignments for each treatment is not fixed and the randomization sequences are not equally likely. Moreover, in testing a hypothesis $H_\Delta, \Delta \neq 0$, the computational procedure developed by Garthwaite modifies patient outcomes only initially, and then all the permutations (i.e., the analogue of re-randomizations) are sampled from the modified data set without further change. An equitable procedure would require modifying patient outcome data according to the hypothesis at each time a re-randomization of treatment assignments is generated (Edgington and Onghena, 2007, Kempthorne, 1977, 1979). A comparison of the computational procedure is presented in Table 5.1. Suppose the interest is in estimating a constant additive effect Δ between two treatment A and B, and the hypothesis is that $\Delta = \Delta_0$. Although it is not designed with an appropriate randomization test procedure, the algorithm can be applied and extended to determining the confidence intervals of additive effects in the randomization context.

5.2 Garthwaite's Robbins-Monro Search Process

5.2.1 The Robbins Monro Process

Let us review the Robbins Monro process (Robbins and Monro, 1951) in the context of randomized clinical trials. The process defines an estimation of the confidence limit by a convergent sequence using binary random variables. Let $\boldsymbol{X} = (\boldsymbol{T}, \boldsymbol{Y})$ be a random vector, where $\boldsymbol{T} = (T_1, \ldots, T_n)$ is

Table 5.1: *Process of computing the permutation test and the randomization test for testing hypothesis* $\Delta = \Delta_0$

	Permutation Test		Randomization Test
(1)	Add Δ_0 to the responses of patients in group A.	(1)	Generate a new treatment assignment sequence.
(2)	Do nothing to the responses in group B.	(2)	If a patient in group A was re-assigned to B, add $-\Delta_0$ to the response.
(3)	Re-randomize the responses into two groups.	(3)	If a patient in group B was re-assigned to A, add Δ_0 to the response.
(4)	Calculate a new estimate for Δ.	(4)	Calculate a new estimate for Δ.
(5)	Repeat (3)-(4) for L times, estimate a p-value.	(5)	Repeat (1)-(4) for L times, estimate a p-value.

the treatment assignment sequence and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is the patient responses. The distribution of $\boldsymbol{T}$, P_ϕ, is derived from a randomization procedure ϕ. Let Ω be the reference set (of treatment assignment sequences). Let s_{obs} be the observed test statistic value. Note that s_{obs} is unchanged under H_Δ for any Δ value. Define a sequence of binary randomization variables

$$z_i = \begin{cases} 0 & \text{if } S_{L_i}(\boldsymbol{x}) < s_{obs} \\ 1 & \text{if } S_{L_i}(\boldsymbol{x}) \geq s_{obs}, \end{cases}$$

where $S_{L_i}(\boldsymbol{x})$ is the test statistic calculated from $\boldsymbol{x}$ under the hypothesis of $\Delta = L_i$. Select L_1 be the best guess of the (lower) confidence limit at level $\alpha, 0 < \alpha < 1$, and $\{a_i\}$ be a sequence of positive constants of type $1/i$, or, more generally, with $\sum_i a_i^2 < \infty$. Then define the values $L_2, L_3, \ldots$ according to the rule

$$L_{i+1} - Li = a_i(\alpha - z_i). \tag{5.1}$$

Denote the p-value of s_{obs} under the hypothesis $\Delta = L_i$ by α_i, where α_i is computed as

$$\alpha_i = P(S_{L_i}(\boldsymbol{x}) \geq s_{obs}) = \sum_{t \in \Omega} I(S_{L_i}(\boldsymbol{x}) \geq s_{obs}) P_\phi(\boldsymbol{T} = \boldsymbol{t}).$$

Observe that $P_\phi(z_i = 1|\Delta = L_i) = \alpha_i, P_\phi(z_i = 0|\Delta = L_i) = 1-\alpha_i$. We thereby have the equation $E_\phi(z_i|L_i) = \alpha_i$, which can be viewed as a function of L_i. It has been proved that $\lim_i L_i = \Delta_L$ in quadratic mean and thus in probability, where Δ_L is the root of the equation $E_\phi(z|\Delta_L) = \alpha$, that is, the level α (lower) confidence limit (Garthwaite and Buckland, 1992, Robbins and Monro, 1951). Robbins and Monro (1951) remarked that the efficiency of $\{L_i\}$ is decided by the choice of L_1 and $\{a_i\}$, and the convergent sequence defined as above has the advantage of being distribution-free over other more efficient estimator of Δ_L.

5.2.2 Algorithm Overview

We summarize the algorithm with reference to the paper (Garthwaite, 1996). The process may be considered as stepping from one estimate of Δ to the next, either forward or backward depending on the value of test statistic calculated from a single re-randomization taken at that estimate. In other words, the re-randomization is the source of randomness of the search process. The length of each step is decided by a constant, whose value determines the asymptotic properties of the process (Garthwaite, 1996).

Let (Δ_L, Δ_U) be the $100(1 - 2\alpha)\%$ equal-tailed confidence interval for Δ estimated from the data. Under some natural monotonicity conditions (Cox and Hinkley, 1974, Garthwaite and Buckland, 1992), the lower and upper confidence limits would be sought such that H_{Δ_L} is rejected in favor of $\Delta > \Delta_L$ at level α, and H_{Δ_U} is rejected in favor of $\Delta < \Delta_U$ at level α. Let $\hat{\Delta}$ be the point estimate of Δ from the data. Let L_i, U_i be the estimates of Δ_L, Δ_U after i steps of the search. A permutation of patient outcomes, denoted by $\boldsymbol{y}$, is taken from the data set modified according to

H_{U_i}. The estimate of upper limit is updated by

$$U_{i+1} = \begin{cases} U_i - c_i\alpha/i & \text{if } S^{\boldsymbol{y}}_{U_i} > S^{obs}_{U_i} \\ U_i + c_i(1-\alpha)/i & \text{if } S^{\boldsymbol{y}}_{U_i} \leq S^{obs}_{U_i}, \end{cases} \tag{5.2}$$

where $S^{\boldsymbol{y}}_{U_i}$ is the test statistic for this permutation under H_{U_i} and $S^{obs}_{U_i}$ is the test statistic for the original patient outcome sequence under H_{U_i}. In this way, the confident limit estimate takes random walk at each step, either goes upward by $c_i(1-\alpha)/i$ or goes downward by $c_i\alpha/i$, depending on $S^{\boldsymbol{y}}_{U_i}$. With similar notations, the lower confidence limit estimate is updated by

$$L_{i+1} = \begin{cases} L_i + c_i\alpha/i & \text{if } S^{\boldsymbol{y}}_{L_i} < S^{obs}_{L_i} \\ L_i - c_i(1-\alpha)/i & \text{if } S^{\boldsymbol{y}}_{L_i} \geq S^{obs}_{L_i}. \end{cases} \tag{5.3}$$

We can see that Equation (5.2) and (5.3) are derived from Equation (5.1) by taking $a_i = c_i/i$.

Let the p-value of the observed patient outcomes under H_{U_i} be α_i (i.e., $P(S_{U_i} \leq S^{obs}_{U_i}) = \alpha_i$,). Then the expected distant from step i to step $i+1$ is

$$E\left(U_{i+1} - U_i\right) = \alpha_i c_i(1-\alpha)/i - (1-\alpha_i)c_i\alpha/i = c_i(\alpha_i - \alpha)/i.$$

The value tends to zero as i goes large. Under weak regularity conditions (Blum, 1954, Garthwaite, 1996), the sequence of the confidence limit estimates attains the true upper confidence limit Δ_U with probability one, which is the root of the equation $P(S_{\Delta_U} \leq S^{obs}_{\Delta_U}) = \alpha$.

The positive step length constant c_i is formulated as $c_i = k(U_i - \hat{\Delta})$ in a search for the upper limit, or $c_i = k(\hat{\Delta} - L_i)$ in a search for the lower limit. The constant k is chosen to be

$$k = 2/\{z_\alpha(2\pi)^{-1/2}\exp(-z^2_\alpha/2)\}.$$

By this choice of k, c_i is twice its optimal value with regard to minimizing the variance of U_{i+1} (and thus the variance of the coverage probability of interval) when Δ is normally distributed. The value c_i is overestimated to guarantee that the estimators would converge to the true confidence limit and with probability one, given that the efficiency of the variance is not dramatically affected (see Garthwaite and Buckland (1992) for details). Moreover, although the optimal value of k depends on the distribution of the point estimate of Δ, it does not vary greatly across many distributions (not including the Cauchy distribution and the two parameterizations of the exponential distribution); the above choice of k is recommended in the absence of better information (Garthwaite and Buckland, 1992).

The starting values for searches are prepared as follows. Modify the data according to $H_{\hat{\Delta}}$. Generate $(2-\alpha)/\alpha$ data permutations of the modified data and estimate Δ at each time. Let t_1, t_2 be the second smallest and second largest estimates. The starting values are given by $\hat{\Delta} \pm (t_2 - t_1)/2$. The search would start with letting i equal to m, $m = \min\{50, 0.3(2-\alpha)/\alpha\}$, to mitigates the rapid change in the early part of a search. The search is continued for a predetermined number of steps and the last value is regarded as the confidence limit. The number of steps is suggested to be larger than the number of data permutations used in estimating the randomization test. If $5,000$ sequences were re-generated in the estimation, then an efficient number of steps is recommended to be $6,000$. The choice of number may also be guided by the asymptotic properties of the Robbins-Monro process.

5.2.3 Extension to Randomized Clinical Trials

The Robbins-Monro search process provides estimates that converge to the confidence limits at the designated coverage probability with asymptotic property and unbiasedness. A j-step searching process is equivalent to performing j Monte Carlo re-randomization tests each with the number of re-randomizations $L = 1$. Nonetheless, further considerations are expected when it comes to the choice of step length constant and the number of searching steps. We have seen that k, a key element in the formation of step length constant c_i that guards the asymptotic property of the searching process, is formulated with reference to the distribution of parameter Δ (Garthwaite, 1996). In population-based inference, it is apparent to derive the distribution of Δ from the assumed patient

population. However, in randomization-based inference, patient responses are no longer assumed to be random samples. To extend the theoretical foundation to this context, it is imperative to provide another basis from which a reasonable distribution of Δ can be derived. We may consider using the randomization distribution of a treatment difference under the null hypothesis.

Garthwaite and Buckland (1992) recommend monitoring the progress of the search and restarting the search to accelerated the convergence. Since the publication of the papers (Garthwaite, 1996, Garthwaite and Buckland, 1992), computational facility has been greatly improved. The consideration for computation time may not be as important. Once the step length constant has been chosen, the standard error of the coverage probability reduces as the number of searching steps increases (Garthwaite and Buckland, 1992). For this reason, we may deliberately choose a relatively large number of steps in searching for the confidence limit. A more critical aspect in deciding the stopping rule is the evaluation of whether the search process is still moving towards the limit, or having converged to the limit and is oscillating around the value. Other important components in the search process are the choice of test statistic and the formulation of the treatment effect model by which an alternative hypothesis is tested and the patient responses are modified. Both statistical and clinical considerations are needed for a sensible choice.

5.3 Preliminary Study: Effect of Malarial Infection on Lizards

We first re-examine the lizard data discussed in the paper (Garthwaite, 1996) with a proper randomization test procedure and the randomization test applied in the paper, which is referred to as "data permutation test" in below in order to distinguish from the randomization test. The data are the distances in meters each of the thirty lizards could ran in two minutes. It is worth noting that the research is not a randomized experiment but a field study (Schall et al., 1982):

> infected lizards (group A):
> 16.4, 29.4, 37.1, 23.0, 24.1, 24.5, 16.4, 29.1, 36.7, 28.7, 30.2, 21.8, 37.1, 20.3, 28.3;
> uninfected lizards (group B):
> 22.2, 34.8, 42.1, 32.9, 26.4, 30.6, 32.9, 37.5, 18.4, 27.5, 45.5, 34.0, 45.5, 24.5, 28.7.

The difference of group means are $\hat{\Delta} = \bar{y}_B - \bar{y}_A = 5.36$. For a randomized test of $H_{\hat{\Delta}}$, we re-generate the treatment assignment sequence L times by the *random allocation rule*. At each time, the outcomes are modified accordingly and a new Δ is estimated. Specifically, if a lizard from group A is re-randomized to group B, we add 5.36 to the datum. If a lizard from group B is re-randomized to group A, we subtract 5.36 from the datum. To determine the starting points for searches at confidence level 95%, we perform a Monte-Carlo re-randomization test of $H_{\hat{\Delta}}$ with $L = 79$ $(= (2-\alpha)/\alpha, \alpha = 0.025)$ and estimate Δ for each re-randomization. The second smallest and second largest estimates of Δ are obtained, and the starting points are computed from them. See Table 5.2 for a summary. The other required starting value m is 24 $(= \min\{50, 0.3(2-\alpha)/\alpha\}, \alpha = 0.025)$.

The test statistic is the mean of group B. For this statistic, a search process updated with the data permutation test and a search process updated with the randomization test are equivalent except for the starting values. In other words, the difference between the two results is caused mainly by the randomness in re-randomization. In searching for the limits, 6,000 steps were taken. The estimated 95% confidence intervals for Δ are $(-0.22, 10.96)$ from the randomization test and $(-0.10, 10.82)$ from the data permutation test. We continue the search process up to 30,000 steps (see Figure 5.1). The 95% confidence interval for Δ turns out to be $(-0.23, 10.90)$ from the randomization test and $(-0.25, 10.98)$ from the data permutation test.

Next, we examine the influence of the starting values. We obtain another set of starting values by increasing the number of re-randomizations L from 79 to $20,000$. The lower and upper $100\alpha\%$ estimate of Δ are taken to be the 500^{th} and the 19501^{th} values from the $20,000$ re-randomization estimates, and the starting points for searches are computed from the two values. See Table 5.2 for a summary. The confidence limits estimated from a search with the randomization test are $(-0.20, 10.96)$ at step $6,000$ and $(-0.23, 10.90)$ at step $30,000$. The confidence limits estimated from the data permutation test are $(-0.11, 10.82)$ at step $6,000$ and $(-0.25, 10.98)$ at step $30,000$. The change of starting values does not change the confidence limit estimates significantly.

To assess the accuracy of the estimates, the one-sided p-value is computed by the Monte Carlo re-randomization test ($L = 1,000,000$) at various values of Δ. It shows that a 95% confidence

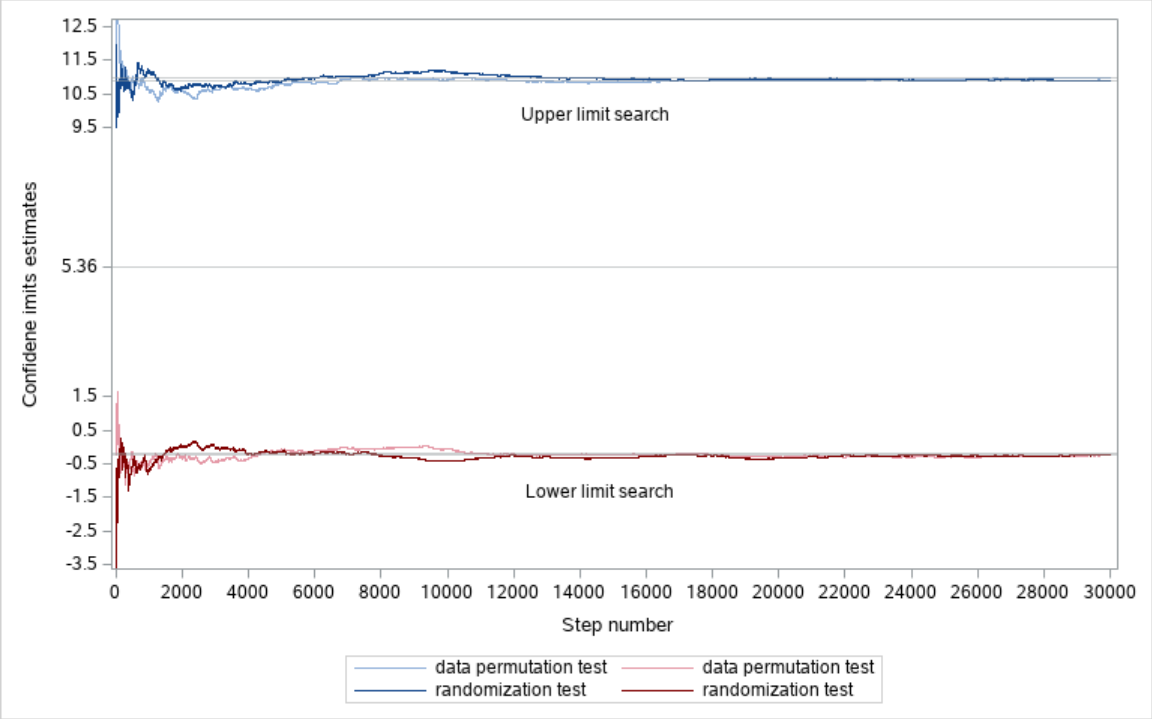

Figure 5.1: 95% *confidence limits estimates for* Δ *from randomization tests and data permutation tests*

interval for Δ is $(-0.27, 10.97)$, which is very close to the estimates given by the data permutation test at step $30,000$, but is somewhat difference from $(-0.30, 10.69)$, the confidence interval estimates given by the paper.

Table 5.2: *Estimates of Δ and starting points for searches.*

		Estimates of Δ		Starting points for search	
		Lower $100\alpha\%$	Upper $100\alpha\%$	Lower limit	Upper limit
$L = 79$	Randomization test	1.299	10.917	0.551	10.169
	Data permutation test	-5.731	5.904	-0.457	11.177
$L = 20,000$	Randomization test	1.573	10.611	0.841	9.879
	Data permutation test	-5.203	5.251	-0.236	10.384

Table 5.3: *One-sided p-value from randomization test.*

Δ	$P(S_\Delta \geq S_\Delta^{obs})$	Δ	$P(S_\Delta \leq S_\Delta^{obs})$
-0.10	0.028	10.90	0.026
-0.20	0.026	10.96	0.025
-0.27	0.025	10.97	0.025
-0.28	0.024	10.96	0.024

5.4 Application to Randomized Clinical Trials: Multiple Tumor Recurrence Data for Patients with Bladder Cancer

5.4.1 Confidence Limits for Difference

We now extend the approach to analyzing data from randomized clinical trial. The data set, given by Andrews and Herzberg (1985), was collected from a three-treatment randomized trial evaluating the effects of placebo, pyridoxine, and thiotepa instillation on patients with bladder cancer (Byar et al., 1977). A total of 121 patients recruited at ten hospitals were randomly assigned to one of the three treatments with equal probability. The primary outcome is the rate of tumor recurrence per 100 patient months follow-up, and the treatment effect model in the analysis assumes that the

recurrence takes place according to the exponential distribution with a constant rate (Byar et al., 1977). A one-sided parametric test of the ratio of recurrence rates shows that difference between thiotepa and placebo is significant ($p = 0.012$); that is, the thiotepa group has lower recurrence rate than the placebo group. Therefore, we estimate the one-sided 95% confidence limit of the treatment effect measured by the difference between the two recurrence rates.

Let treatment A be placebo and treatment B be thiotepa. Denote the recurrence rate of group j by $\lambda_j, j = A, B$. Let $\Delta = \lambda_A - \lambda_B$. A one-sided 95% confidence interval of Δ, (Δ_L, ∞), is obtained by using the conditional randomization test re-randomized according to complete randomization. The lower confidence limit should be sought that H_{Δ_L} is rejected in favor of $\Delta > \Delta_L$ at level 0.05. We choose the test statistic to be Δ. To test an alternative hypothesis H_{Δ^*}, we propose an additive treatment effect model and modify the patient outcomes as follows. If all the patients assigned to treatment A were re-randomized to treatment B, then the recurrence rate of the group would be decreased by Δ^*. Therefore, if a patient from group A is re-randomized to group B, we multiple the patient's observed number of recurrences by $(\lambda_A - \Delta^*)/\lambda_A$. Similarly, if a patient from group B is re-randomized to group A, we multiple the observed number of recurrences by $(\lambda_B + \Delta^*)/\lambda_B$. Note that some patients do not have tumor recurrence during the follow up period.

Next we determine the starting values for search. The point estimator from the data is $\hat{\Delta} = \hat{\lambda}_A - \hat{\lambda}_B = 0.0189$, which also is the observed test statistic value under an alternative hypothesis. To compute the starting value for search at the desired level, we generate $20,000$ re-randomization sequences and obtain the $1001th$ and $19000th$ values of the re-randomization estimates of Δ, which are -0.00298 and 0.0406 respectively. The starting value for the lower limit search is therefore -0.00289. Another starting value m is 12 according to the formula Section 5.2. The step length constant is set to be $c_i = k(\hat{\Delta} - L_i)$, where $k = 2/g$ and $g = z_\alpha(2\pi)^{-1/2}\exp(-z_\alpha^2/2)$. After $60,000$ steps, the estimated one-sided confidence interval is $(-0.00319, \infty)$ (Figure 5.2). The one-sided p-value for $H_{\Delta=-0.00319}$ is 0.05 from a conditional randomization test ($L = 1,000,000$), which verifies the confidence level of the lower limit estimate.

The asymptotic properties of the searching process is influenced by the choice of k. It is seen from Figure 5.2 that the variance of the confidence limit estimates increases when k increases from

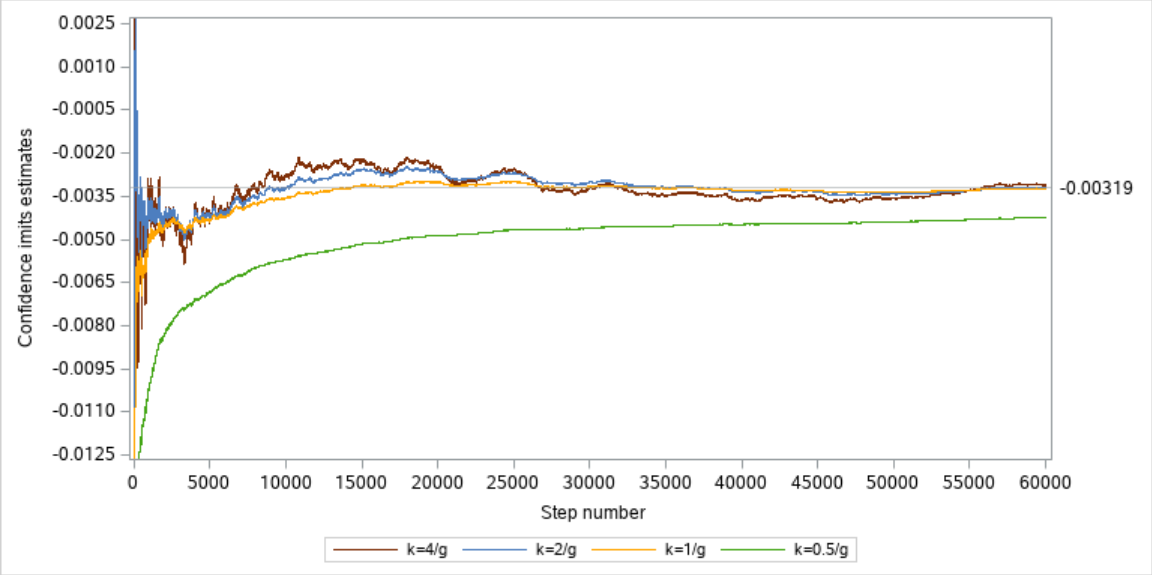

Figure 5.2: *Lower confidence limit estimates for Δ from the conditional randomization test.*

$2/g$ to $4/g$, and decrease when k decreases from $2/g$ to $1/g$. However, the influence of k becomes less evident as the number of steps goes large. But when k is decreased to $0.5/g$, the estimate is far lower than the true 95% confidence limit even at step $60,000$ and the search process moves towards the limit very slowly.

5.4.2 Confidence Limits for Ratio

In the previous section, we obtained the 95% lower confidence limit for the difference between the two recurrence rates, from which the lower confidence limit for the ratio of the two recurrence rates can also be derived. Now we explore the way to directly calculate a confidence interval for the ratio of the two recurrence rates. Let $\Delta = \lambda_A/\lambda_B$. To test a hypothesis H_{Δ^*}, we propose an additive treatment effect model in the following manner. If a patient from group A is re-randomized

to group B, we multiple the patient's observed number of recurrences by $1/\Delta^*$ so that the patient's observed number of recurrences would decrease to λ_A/Δ^*. Likewise, if a patient from group B is re-randomized to group A, we expect that the patient's observed number of recurrences would increase to $\Delta^*\lambda_B$. Thus we multiple the patient's observed number of recurrences by Δ^*.

Let Δ be the test statistic. The observed test statistic is $\hat{\Delta} = \hat{\lambda}_A/\hat{\lambda}_B = 1.497$. To determine the starting values for searching the confidence limits, we re-randomize $20,000$ times. At each time, we modify the patient's outcomes according to the treatment effect model as if to test the hypothesis $H_{\hat{\Delta}}$ and compute the estimate for Δ. The starting values are calculated from the $1001th$ and $19000th$ values among the $20,000$ estimates, which are $t_1 = 0.776$ for searching the lower confidence limit and $t_2 = 2.218$ for searching the upper confidence limit.

Noticing that the Robbins-Monroe process is an algorithm based on addition and subtraction, so, to calculate a confidence interval for ratio, a monotonic transformation of Δ needs to be applied so that the confidence limits for Δ can be approached linearly. Let $\Delta' = \log \Delta = \log \lambda_A - \log \lambda_B$. Then the 95% confidence limits for Δ' can be identified using the Robbins-Monroe process. To find the lower confidence limit, for example, let the starting value L_1 be $\log t_1$. Update the lower limit estimate by

$$L_{i+1} = \begin{cases} L_i + c_i\alpha/(m-i+1) & \text{if } S^{\boldsymbol{y}}_{\exp(L_i)} < \hat{\Delta} \\ L_i - c_i(1-\alpha)/(m-i+1) & \text{if } S^{\boldsymbol{y}}_{\exp(L_i)} \geq \hat{\Delta}. \end{cases} \tag{5.4}$$

Note that the estimate of ratio Δ at step i is given by $\exp(L_i)$ due to the $\log$ transformation. The step length constant c_i is defined as $c_i = k(\log \hat{\Delta} - L_i)$. The choice of m and k follows from the previous section (i.e., $m = 12, k = 2/g$).

After $30,000$ steps, the 95% lower and upper confidence limits for $\Delta = \lambda_A/\lambda_B$ are 0.949 and 2.524 respectively. The searching process is presented in Figure 5.3. We know from the previous section that the 95% lower confidence limit for $\Delta = \lambda_A - \lambda_B$ is -0.00319. Given that $\lambda_A = 87/1528, \lambda_B = 45/1183$, it can be easily verified that the two 95% lower confidence limits agree with each other.

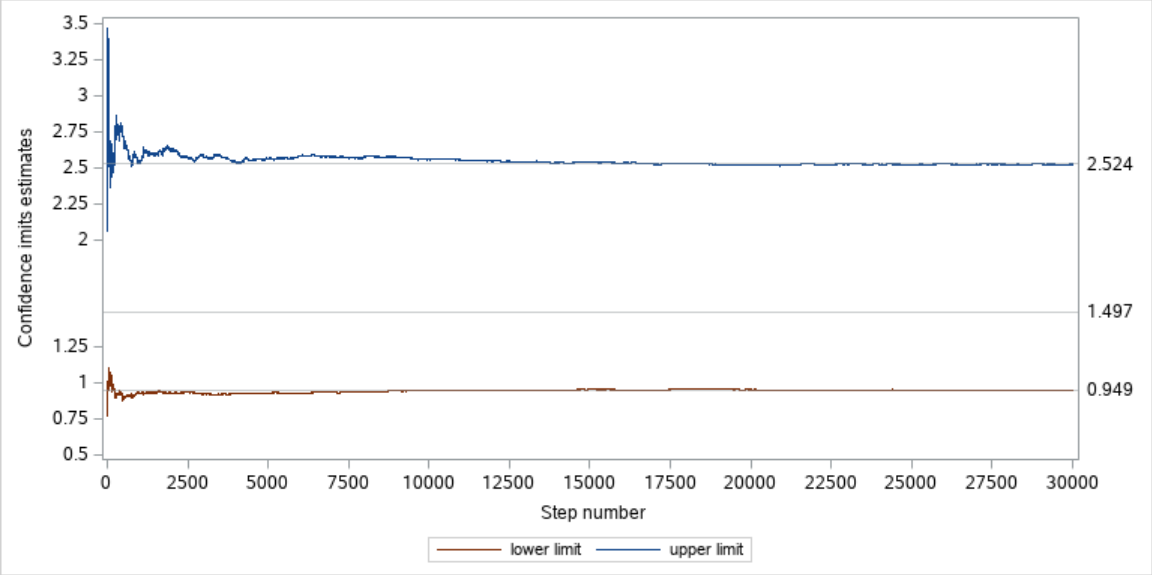

Figure 5.3: *Confidence limit estimates for the ratio of recurrence rates.*

5.5 Population-based and Randomization-based Confidence Intervals

In this section we compare the confidence intervals given by the population-based and randomization-based inference. We calculate from the lizard data (Section 5.3) a 95% confidence interval for the difference in group means given by the pooled two-sample t-test, and calculate from the bladder cancer data (Section 5.4) a 90% confidence interval for the ratio of the recurrence rates by a test based on F distribution (Gehan, 1975). The interval for ratio is determined as

$$F_\alpha/\hat{\Delta} < \Delta < F_{1-\alpha}/\hat{\Delta},$$

where F_α is the upper α percent point of the F distribution with degrees of freedom $df_1 = 2n_A, df_2 = 2n_B$, where $n_A = 87, n_B = 45$ are the number of recurrences in group A and B, respectively. The

results are summarized in Table 5.4. In both cases, population-based inference results in a shorter interval than that obtained from the randomization test.

Table 5.4: *Confidence interval estimate from population-based and randomization-based inference.*

	Lizard data		Bladder cancer data	
	Lower 2.5% limit	Upper 2.5%	Lower 5% limit	Upper 5% limit
Population-based	-0.23	10.95	1.115	2.045
Randomization-based	-0.27	10.97	0.949	2.524

Nevertheless, this is not always the case. We examine the population-based and the randomization-based 95% confidence intervals under eight randomization procedures when there are some heterogeneity in data. First we assume a time trend model. Patient responses to treatment j are sampled from normal distribution $N(\beta_j, 1)$ plus a linear drift ranging on the interval $(-1, 1]$, where $\beta_A = 0, \beta_B = 1$. Sample sized $n = 50$. The test statistic is chosen to be the difference in means for randomization-based estimation, or the t-statistic with the pooled standard deviation for population-based estimation.

Next we study a model of outliers in comparing recurrence rates. We assume that each patient is followed up for at least 36 months; the last event would be observed no sooner than the $36th$ month of follow-up. The number of recurrences and the total number of follow-up months is recorded for each patient. Let $Exp(\Delta)$ denote the exponential distribution with mean Δ. For patients assigned to treatment A, the event times are sampled from $Exp(10)$ with 10% random contamination sampled from $Exp(36)$. For patients assigned to treatment B, the event times are sampled from $Exp(15)$ with 10% random contamination sampled from $Exp(36)$. Therefore, the "true" ratio of recurrence rates Δ_A/Δ_B should be 1.5 (i.e., $15/10$). Test statistic is chosen to be the ratio of recurrence rates. The estimation of the population-based confidence interval follows from the discussion at the beginning of this section.

The performance of the confidence interval estimate is evaluated by the coverage probability and the length of the interval averaged over $1,000$ simulated data sets, each with sample sized $n = 50$. The values of the upper and lower confidence limits are determined after searching for $30,000$ steps. The results are presented in Table 5.5.

Table 5.5: *Comparison of confidence interval estimate from population-based and randomization-based inference*.

		Randomization-based		Population-based	
		Coverage probability	Interval length	Coverage probability	Interval length
Difference in means	CR	0.95	1.32	0.95	1.33
	RAR	0.96	1.31	0.96	1.33
	TBD	0.96	1.53	0.91	1.33
	UD(0,1)	0.96	1.24	0.96	1.33
	BCD	0.96	1.19	0.97	1.33
	BSD	0.96	1.18	0.97	1.34
	PBD	0.95	1.15	0.98	1.33
	RBD	0.94	1.14	0.97	1.34
Ratio of recurrence rates	CR	0.95	1.10	0.91	0.87
	RAR	0.95	1.08	0.90	0.87
	TBD	0.95	1.08	0.90	0.87
	UD(0,1)	0.96	1.07	0.90	0.87
	BCD	0.94	1.09	0.89	0.88
	BSD	0.95	1.07	0.89	0.87
	PBD	0.95	1.07	0.88	0.87
	RBD	0.94	1.07	0.90	0.88

When estimating the difference in group means in the presences of a time trend, the length of the population-based confidence interval is relatively preservative, whereas the length of the randomization-base interval varies from 1.14 to 1.53. Specifically, the block designs give the shortest interval length, and TBD gives the longeest interval length. This change of interval length resembles the change of power with regard to randomization procedure in the simulation study of the power of the test (Chapter 2, Chapter 5). While the coverage probability of the randomization-based interval is close to 0.95, the coverage probability of the population-based interval ranges from

0.91 to 0.98. This phenomena is related to the unpreserved type-I error rate of the population-based test in the previous simulation study (Chapter 2, Chapter 5).

When estimating the ratio of recurrence rates in the presence of outliers, the lengths of the population-based confidence interval are, on average, shorter than that of the randomization-base intervals (1.08 versus 0.87). But the population-based method basically gives a 90% interval rather than a 95% interval as it claims to be. Only the randomization-based interval preserves the coverage probability. In short, when heterogeneity in data renders the population model of the treatment effect insufficient, the randomization-based method maintains the confidence level.

The difference between the two methods is more profound than the apparent difference in numerical values. In randomization-based inference, the logic of interval estimation is simple: (i) a mathematical model of the treatment effect on an individual level is proposed and (ii) acceptable estimates of the parameter in terms of statistical significance are calculated from the experimental data, assuming the treatment effect model. In population-based inference, the mathematical model is imposed on the distribution of the data rather than directly on the treatment effect. Moreover, the goal of estimation consists in achieving a value or a set of values that should not differ very much from the "true" value of the parameter. The 95% confidence interval of a parameter has the interpretation that 95% of the intervals would cover the "true" parameter value under repeated sampling. A randomization-based confidence interval, on the other hand, does not has the connotation of covering a "true" parameter value with claimed probability.

5.6 Alternative Computational Method: The Bisection Method

The Robbins-Monro algorithm is a stochastic approximation algorithm. Now we introduce a numerical approximation algorithm–the bisection, or binary search method (Mauchly, 1949, as referred in Knuth 1998). This is a method based on the Intermediate Value Theorem. It finds a root of a given (continuous) function by narrowing down an interval that contains a root of the function. The method will split the interval into two equal halves and check which half interval contains a root of the function, and continue splitting the interval in halves until the resulting interval is sufficiently

small. Then the root is approximated by any value in the final interval. Here, the function is

$$f(x) = P(\text{observing an equally or more extreme test statistic}|H_x),$$

and the method would find solution $x = \Delta$ to the equation $f(x) = \alpha$. The starting interval $[x_1, x_2]$ would be arbitrarily chosen such that $f(x_1) < \alpha < f(x_2)$ and $f(x)$ is monotonic on the interval. According to the accuracy of the Monte Carlo p-value estimate (Plamadeala and Rosenberger, 2012), we propose that the iteration would be stopped when the interval is small enough that the difference between the p-value estimates at the two endpoints is smaller than $10\%\alpha$. The lower and upper confidence limits are determined separately.

We find that the numerical approximation method is more effective than the stochastic method when data are extreme. Consider again the simulation in Section 5.5 where we estimate the confidence limits for the ratio of event recurrence rates, denoted by Δ, in the presence of outliers. When the outliers are sampled from the Cauchy distribution, the Robbins-Monro algorithm does not give a convergent sequence for approximating the confident limits. Theoretically, it is possible to construct a convergent sequence by choosing a suitable step length constant c_i, but it is unclear that how this can be done efficiently in practice. On the other hand, the bisection method is able to produce a confidence limit estimate within a computable number of iterations. The details are as follows.

The event times are sampled from $Exp(\theta)$ with 10% random contamination sampled from $Cauchy(0, 1)/5$, where $\theta = 10$ for patients assigned to treatment A, and $\theta = 15$ for patients assigned to treatment B. We examine the example of estimating the lower confidence limit. The randomization procedure is complete randomization. The Robbins-Monro algorithm with step length constant parameter $k = 2/g$ (see the penultimate paragraph in Section 5.4.1) does not give a limit after $30,000$ steps of the search. For the bisection algorithm, we set the initial interval containing the limit to be $[0.5, 1.5]$. At each iteration, we estimate the p-value of the upper point, lower point, and midpoint of the interval by Monte Carlo re-randomization test with $L = 20,000$. After seven iterations, the estimate for the lower confidence limit is $\Delta = 0.83203$, and the confidence level is confirmed by the p-value of the estimate (Table 5.6).

Table 5.6: *Numerical approximation of the* 5% *lower confidence limit for the ratio of recurrence rates by the bisection method.*

Iteration	Interval	p-value at endpoints	Midpoint	p-value at midpoint
0	[0.5,1.5]	0.003, 0.496	1	0.128
1	[0.5, 1]	0.002, 0.127	0.75	0.028
2	[0.75, 1]	0.029, 0.123	0.875	0.065
3	[0.75, 0.875]	0.027, 0.063	0.8125	0.047
4	[0.8125, 0.875]	0.041, 0.063	0.84375	0.051
5	[0.8125, 0.84375]	0.044, 0.054	0.82813	0.049
6	[0.82813, 0.84375]	0.050, 0.056	0.83594	0.053
7	[0.82813, 0.83594]	0.050, 0.054	0.83203	0.052

Below is a brief comparison of the computational efficiency of the two algorithms for obtaining a 5% lower confidence limit under the outliers model in the previous section. The number of steps in Robbins-Monro algorithm is $30,000$ and the starting point is provided. The starting interval and the stopping rule of the bisection method follow from the above example. Computational time is 0.97 seconds for the bisection method and 0.13 seconds for the Robbins-Monro algorithm based on a regular laptop (with a 1.4 GHz Intel Core i5 processor and 4 GB 1600 MHz memory). The total number of modifications of patient responses according to the testing hypothesis is $2,100,000$ $(= 7$ iterations $\times 3$ points $\times 20,000$ Monte Carlo simulations $\times 50$ patients$)$ for the bisection method and $150,000 (= 30,000$ Monte Carlo simulations $\times$ 50 patients$)$. In this example, the difference in computational time does not have a noticeable impact to user experience. But if the object of estimation does not has a closed-form and requires applying an iterative method for each estimate, the difference may not be negligible, and may be reduced by decreasing the number of iterations.

5.7 Discussions

The discussion in this chapter has been focused on constant additive treatment effect in terms of a scalar parameter. Nonetheless, additivity of treatment effect does not exclude random errors. Let

x_i denote the response of patient i under some basic condition. Constant additivity implies that the response of patient i under treatment j is given by liner model

$$y_{ij} = x_i + t_j + e_{ij}, \tag{5.1}$$

where the e_{ij}'s are independent random errors not necessarily assuming a distribution (Kempthorne, 1955). In view that perfect patient-treatment additivity are rare and multiple covariates are common in clinical trials, we consider it is helpful to evaluate other model of treatment difference, such as a multiplicative model (Onghena, 2018) or, in more complex cases, regression models, and we are inviting future exploration on the topic.

We end this chapter with an excerpt (Kempthorne, 1992) that summarizes the randomization-based interval and its interpretation and remarks on different approaches to inference (randomization-based, population-based, Bayesian).

> Related to this process [the population-based method], but different from it, is the use of significance levels, often called P values. Inversion of the whole family of related significance tests of $\theta = \theta_0$ for a set of values of θ_0 gives a region of values of θ that *agree with the data* to a designated extent.
>
> My preference is to regard the regions so obtained as consonance regions that specify values of θ that are consonant with the data at chosen level.
>
> These procedures, however characterized by particular words, do not give probabilities of hypotheses such as probability that θ belongs to any chosen region of the parameter space.
>
> If, then, the aim of the whole exercise, design, performance and analysis of the experiment is the obtaining of such probabilities, the procedures are totally unsuccessful.
>
> The group of statisticians known as Bayesians take the position that the aim of all investigation must be the obtaining of such probabilities. Then it is obvious that one can reach the result with the introduction of a prior distribution. Unfortunately there is no logic that forces choice of a prior. It is the conclusion of this line of development

> that probability outcome is a belief probability that depends critically, obviously, on the prior belief probability.
>
> My opinion is that the processes of science and technology do not require belief probabilities. The processes of science and technology require obtaining of data under circumstances chosen by the investigators, and analysis of data, which consists of making judgment of whether the data are consonant with particular models suggested by previous investigators or of determining new models from the data that are obtained. . .
>
> The idea that analysis of astronomical data should use a parametric model determined by some θ with a prior belief distribution on θ seems to me to be an antithesis of scientific method.
>
> I therefore take the view that the Bayesian prescription, which is being heavily touted as the prescription by which all the uncertainty about this world in which we have to live can be handled, is not worth considering.

The application of Bayesian methods in the design and analysis of randomized clinical trials may yield further information that is helpful to design a scientific experiment, particularly if prior studies are available. If the researcher would like to adjust some measurements for the purpose of optimizing the design of the experiment, the researcher may run simulations based on a parametric model of data distribution to evaluate the performance at given values of parameters. An in-depth discussion on Bayesian methods is beyond the scope of this dissertation. In summary, it is unnecessary to always reduce the the processes of experimentation to a (hierarchy of) families of distribution functions in order to extrapolate, and it is unlikely that the level of confidence to predict in a population-based interpretation can eradicate the uncertainty in extrapolation with a claimed confidence level. Randomization-based estimation is more about checking a proposed model of how a treatment affects the response by real world data than checking the data by a stochastic distribution model.

Chapter 6: Conclusions and Future Work

A core of experimental inference is an objective, substantial basis for forming the probability distribution upon which the statistical significance of the test is shaped and calculated. In analyzing data from randomized clinical trials, such a basis is recognized as the randomization procedure. Essentially, randomized clinical trials are complex, designed experiments rather than sampling procedures with well-defined populations. It has been widely known for the last century that designed experiments must be analyzed differently from studies with random sampling. The experiment itself is a population, and replication, rather than repeated sampling, is the key factor in generalizability.

The contributions of the thesis are recapitulated as follows. First, we distinguish population-based inference from randomization-based inference, and examine the framework of randomization tests in terms of the hypothesis, the random mechanism, and the reference set. Perhaps the most obvious difference between the two inferential theories lies in the rationale for generalizing experimental conclusions as well as the meaning of repeating an experiment. If the observations are actually sampled from the proposed populations, generalizability is automatic, since the population of interest can be unambiguously defined by a distribution function. Valid generalization can be entrusted with a test that sufficiently represents the distributional characteristics. Repeating an experiment means repeated sampling from the population distribution. In randomization-based inference, it is understood that the general patient population can neither be well-defined nor characterized by a limited number of trials. The statistical significance of a hypothesis pertains to the specific trial being analyzed, and randomization allows us to replicate of a completed trial without actually replicating. The validity of the generalization relies on the design and proper conduct of the trial rather than on the accuracy of a statistical model of treatment effect.

Second, in exploring the statistical validity and the power of the test under heterogeneity in the patient responses, we discover that, while randomization tests preserve the type I error rate, the population-based test, either parametric or non-parametric, is not always valid. The study confirms

that the normal theory test can sometimes serve as an approximation to the randomization test, not the other way around, as Kempthorne discussed throughout his voluminous work on experimental design. We also find that, in the presence of time trends, periodic balance of treatment allocation in the randomization improves the sensitivity in detecting existing treatment effects. When the heterogeneity is not related to the sequential order of treatment assignment (e.g., outliers, heavy-tailed), the power of the test is less affected by the randomization procedure, but is increased when changing the test statistic from measurement-based to rank-based. The phenomena are observed alike in global hypothesis, pairwise comparisons, and factorial designs.

Third, the solution to the computational complexity in multiple comparisons culminates in an efficient approach for simulating the conditional reference set. Apart from the naive approach that samples a large number of treatment sequences and keep only those satisfying the condition, the convenient approach that samples by the corresponding two-armed randomization, we develop a approach that samples directly from the conditional reference set without loss of computation efficiency. Moreover, we make valid generalization of two randomization procedures, which heretofore were designed only for randomizing two treatments, so they can be applied to randomizing clinical trials with more than two treatments. Relevant mathematical proofs are provided.

Fourth, in developing the randomization-based estimation, we contextualize the definition of confidence interval and examined an efficient algorithm for computing a confidence limit based on Robbins-Monro process. We extend the algorithm from the context of permutation test to randomization test, and apply it to estimating confidence intervals for data from randomized clinical trials. We also compare the performance of randomization-based interval estimation with population-based interval estimation, and demonstrate that only the formal preserves the confidence level under heterogeneity in the patient outcomes.

Another important topic in popularizing the application of randomization tests in practice but is not covered in the thesis concerns the principle and method of handling missing outcomes data. Edgington (2007) discusses two ways of permuting the patient outcomes when there are missing outcomes. One method eliminates the missing records and permutes the remaining data. The other method permutes the entire data set without discrimination, as if there were no missing records.

A third approach, the worst-rank analysis, is described in Rosenberger and Lachin (2016). They suggest assigning the worst rank to the missing records when using a linear rank statistic.

For randomization-based estimation, future investigations can be expected in the mathematical properties of the estimating process, for example, the accuracy of the confidence limit estimate, the method of determining the step length constant in order that the sequence of estimates would converge even if the data are extreme, so as to facilitate the application to a broader class of primary outcome variables and analyses.

In addition, the randomization procedures discussed in the thesis are procedures that preserve the allocation ratio at every treatment assignment (i.e., marginally, each treatment is equally likely to be assigned to each patient). Some discussion of randomization tests in nonstandard settings, such as covariate-adaptive and response-adaptive randomization, can be found in Proschan and Dodd (2019). Questions regarding how the p-value should be determined or applied arise when the randomization in a trial is almost a deterministic process (see Proschan and Dodd (2019) for some examples). We consider that the level of evidence in favor of a treatment effect in this situation may not be mechanically interpreted. Challenges of re-randomization are present when the clinical conditions are imperfect. Nonetheless, reliable conclusions in terms of scientific objectivity and statistical validity is attainable with the improvement in the design, documentation, and analysis of randomized clinical trials. We hope that the exposition of the purpose of statistical inference in the randomized clinical trials and techniques of how to perform the inference is coherent with the instructions and inspirations given by the great predecessors. We hope that the thesis will serve as one of the stepping stones for inquiring into these topics further.

Bibliography

Andrews, D. F. and Herzberg, A. M. (1985), *Data. A Collection of Problems from Many Fields for the Students and Research Worker*, New York: Springer, pp. 253–259.

Anscombe, F. J. (1948), "The validity of comparative experiments," *J Roy Statist Soc A*, 111, 181–211.

Armitage, P. (2003), "Fisher, Bradford Hill, and randomization," *Int J Epidemiol*, 32, 925–928.

Benjamini, Y. and Hochberg, Y. (1995), "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *J Roy Statist Soc B*, 57, 289–300.

Berger, V. (2018), *Randomization, Masking, and Allocation Concealment*, Boca Raton: CRC Press, 1st ed.

Bhattacharyya, G. K. and Johnson, R. A. (1977), *Statistical Concepts and Methods*, New York: Wiley.

Blum, J. R. (1954), "Approximation methods that converge with probability one," *Ann Math Statist*, 25, 390–394.

Byar, D. P. (1980), "The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa," in *Bladder Tumors and Other Topics in Urological Oncology*, eds. Pavone-Macaluso, M., Smith, P. H., and Edsmyr, F., New York: Plenum, pp. 363–370.

Byar, D. P., Blackard, C. P., and the VACURG (1977), "Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of stage I bladder cancer," *Urology*, 10, 556–561.

Cochran, W. G. (1980), "Fisher and the analysis of variance," in *R.A. Fisher: An Appreciation, Lecture Notes in Statistics*, eds. Fienberg, S. E. and Hinkley, D. V., New York: Springer, pp. 17–34.

Cox, D. R. (1982), "A remark on randomization in clinical trials," *Utilita Mathematica*, 21A, 245–252.

Cox, D. R. and Hinkley, D. V. (1974), *Theoretical Statistics*, London: Chapman and Hall, p. 213.

Dunn, O. J. (1961), "Multiple comparisons among means," *J Am Statist Assoc*, 56, 52–64.

Edgington, E. S. and Onghena, P. (2007), *Randomization Tests*, Boca Raton: CRC Press, 4th ed.

Efron, B. (1971), "Forcing a sequential experiment to be balanced," *Biometrika*, 58, 403–417.

Fisher, R. A. (1935a), *The Design of Experiments*, Edinburgh: Oliver and Boyd.

— (1935b), "Discussion to 'Statistics in agricultural research' by J. Wishart," *J Roy Statist Soc, Supplement*, 1, 26–61.

Galbete, A. and Rosenberger, W. F. (2015), "On the use of randomization tests following adaptive designs," *Statist Biopharm Res*, 26, 466–474.

Garthwaite, P. H. (1996), "Confidence interval from randomization tests," *Biometrics*, 52, 1387–1393.

Garthwaite, P. H. and Buckland, S. T. (1992), "Generating Monte Carlo confidence intervals by the Robbins-Monro process," *Appl Statist*, 41, 159–171.

Gehan, E. A. (1975), "Statistical methods for survival time studies," in *Cancer Therapy: Prognostic Factors and Criteria of Response*, ed. Staquet, M. J., New York: Raven Press, pp. 7–35.

Hinkelmann, K. and Kempthorne, O. (2008), *Design and Analysis of Experiments*, Hoboken, New Jersey: Wiley, 2nd ed.

Kempthorne, O. (1952a), *The Design and Analysis of Experiments*, New York: Wiley, pp. 135–160.

— (1952b), *The Design and Analysis of Experiments*, New York: Wiley, pp. 234–251.

— (1955), "The randomization theory of experimental inference," *J Am Statist Assoc*, 50, 946–967.

— (1966), "Some aspects of experimental inference," *J Am Statist Assoc*, 6, 11–34.

— (1969), "The behaviour of some significance tests under experimental randomization," *Biometrika*, 56, 231–248.

— (1977), "Why randomize?" *J Stat Plan Inference*, 1, 1–25.

— (1979), "Sampling inference, experimental inference and observation inference," *Sankhya B*, 40, 115–145.

— (1982), "Review of the book Randomization Tests by Edgington E S," *Biometrics*, 38, 864–867.

— (1987), "Discussion: What is an analysis of variance?" *Ann Statist*, 15, 925–929.

— (1992), "Intervention experiments, randomization and inference," *Lecture Notes-Monograph Series*, 13–31.

Knuth, D. E. (1998), *The Art of Computing Programming*, vol. 3, Boston: Addison-Wesley, 2nd ed.

Kruskal, W. H. and Wallis, A. W. (1952), "Use of ranks in one-criterion variance Analysis," *J Am Statist Assoc*, 47, 583–621.

Kuznetsova, O. M. and Tymofyeyev, Y. (2011), "Preserving the allocation ratio at every allocation with biased coin randomization and minimization in studies with unequal allocation," *Statist Med*, 31, 701–723.

Lachin, J. M., Marks, J. W., and Schoenfield, L. J. (1981), "Design and methodological considerations in the National Cooperative Gallstone Study: A multi-center clinical trial," *Controlled Clin Trials*, 2, 175–230.

Lehmann, E. L. and Romano, J. P. (2006), *Testing Statistical Hypotheses*, New York: Springer, 3rd ed.

Neyman, J. (1937), "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability," *Philos Trans Roy Soc A*, 236, 333–380.

Neyman, J. and Pearson, E. (1933), "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philos Trans R Soc Lond B Biol Sci A*, 231, 289–337.

Onghena, P. (2018), "Randomization and the Randomization Test: Two Sides of the Same Coin," in *Randomization, Masking, and Allocation Concealment*, ed. Berger, V., Boca Raton: CRC Press, chap. 13, pp. 185–203.

Parhat, P., Rosenberger, W. F., and Diao, G. (2014), "Conditional Monte Carlo Randomization Tests for Regression Models," *Statist Med*, 33, 3078–3088.

Pesarin, F. (2001), *Multivariate Permutation Tests: With Applications in Biostatistics*, New York: Wiley.

Plamadeala, V. and Rosenberger, W. F. (2012), "Sequential monitoring with conditional randomization tests," *Ann Statist*, 40, 30–44.

Proschan, M. A. and Dodd, L. E. (2019), "Re-randomization tests in clinical trials," *Statist Med*, dOI: 10.1002/sim.8093.

Robbins, H. and Monro, S. (1951), "A stochastic approximation method," *Ann Math Statist*, 33, 400–407.

Rosenberger, W. F. and Lachin, J. M. (2016), *Randomization in Clinical Trials: Theory and Practice*, Hoboken: Wiley, 2nd ed.

Rosenberger, W. F., Uschner, D., and Wang, Y. (2019), "Randomization: The forgotten component of the randomized clinical trial," *Statist Med*, 38, 1–12.

Schall, J. J., Bennett, A. F., and Putman, R. W. (1982), "Lizards infected with malaria: physiological and behavioral consequences," *Sciences*, 217, 1057–1059.

Schoenfield, L. J. and Lachin, J. M. (1981), "Chenodiol (chenodeoxycholic acid) for dissolution of gallstones: the National Cooperative Gallstaone Study," *Ann Intern Med*, 81, 257–282.

Soares, J. F. and Wu, C. F. J. (1982), "Some restricted randomization rules in sequential designs," *Commun Statist Theory Methods*, 12, 2017–2034.

Tamm, M. and Hilgers, R. D. (2014), "Chronological bias in randomized clinical trials arising from difference types of unobserved time trends," *Methods Inf Med*, 53, 501–510.

Tukey, J. W. (1949), "Comparing individual means in the analysis of variance," *Biometrics*, 5, 99–114.

Wei, L. J. (1977), "A class of designs for sequential clinical trials," *J Am Statist Assoc*, 78, 382–386.

— (1978), "An application of an urn model to the design of sequential controlled clinical trials," *J Am Statist Assoc*, 73, 559–563.

Zhang, L. and Rosenberger, W. F. (2011), "Adaptive randomization in clinical trials," *In Design and Analysis of Experiments*, Vol. III. (K. Hinkelmann, ed.). Wiley, Hoboken.

Biography

Yanying Wang was born in Wuhan, Hubei, China. She moved to Beijing, China with her parents while in high school. She received her Bachelor's degrees in Biological Science in Honors Program, and Mathematics and Applied Mathematics from China Agricultural University in July 2013. She came to Washington, DC in August 2014 and obtained her Master's degree in Statistics from The George Washington University in May 2016. In August 2016, she moved to Fairfax, Virginia to continue her study at George Mason University and was supported by a graduate teaching assistantship offered by the Department of Statistics during her doctorate years.