A RANDOM-LINEAR-EXTENSION TEST

BASED ON CLASSIC NONPARAMETRIC PROCEDURES

A Dissertation
Submitted to
the Temple University Graduate Board

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

---

by

Jun Cao

August, 2009

Copyright © 2009

by

Jun Cao

ABSTRACT

A RANDOM-LINEAR-EXTENSION TEST BASED ON CLASSIC NONPARAMETRIC PROCEDURE

Jun Cao

DOCTOR OF PHILOSOPHY

Temple University, 2009

Professor Woollcott Smith, Chair

Most distribution free nonparametric methods depend on the ranks or orderings of the individual observations. This dissertation develops methods for the situation when there is only partial information about the ranks available.

A random-linear-extension exact test and an empirical version of the random-linear-extension test are proposed as a new way to compare groups of data with partial orders. The basic computation procedure is to generate all possible permutations constrained by the known partial order using a randomization method similar in nature to multiple imputation. This random-linear-extension test can be

simply implemented using a Gibbs Sampler to generate a random sample of complete orderings. Given a complete ordering, standard nonparametric methods, such as the Wilcoxon rank-sum test, can be applied, and the corresponding test statistics and rejection regions can be calculated. As a direct result of our new method, a single p-value is replaced by a distribution of p-values. This is related to some recent work on Fuzzy P-values, which was introduced by Geyer and Meeden in Statistical Science in 2005.

A special case is to compare two groups when only two objects can be compared at a time. Three matching schemes, random matching, ordered matching and reverse matching are introduced and compared between each other. The results described in this dissertation provide some surprising insights into the statistical information in partial orderings.

ACKNOWLEDGEMENTS

During the long course of my graduate study, I have always imagined the moment that I would write my dissertation acknowledgement. Gladly, this moment finally comes.

Writing a thesis could be a long and lonely journey, but thanks to my advisor, Woollcott Smith, for his encouragement when I almost gave up and the guidance and counseling throughout the process, I am able to make it till the end. His patience in reading draft after draft of my thesis, proposal and idea write ups continues to amaze me. No one should be subjected to the torture of reading my early attempts at technical writing, and thanks to Dr. Smith, no one will. Also, I would like to thank him for always being willing to meet or talk whenever I barged in, even when he was on trips or relaxing at his home.

I am deeply grateful to my committee members for their comments and suggestions. I have benefitted greatly from their advice. I thank Damaraju Raghavarao and Milton Parnes for the help with tightening up the notation and presentation of the proofs; Pallavi Chitturi for lending a sympathetic ear, investing time and energy, discussing ideas with me. Also heartfelt thanks to Edward Rosenthal for his willingness to read about and listen to these “strange” statistical ideas.

Lastly, I would like to thank my parents, my mother Shulan Wei and my father Yanming Cao who have given me more love than anybody could ever expect in a lifetime. They have always supported me unconditionally, have guided me at every step of the way, and have always been there for me. Also thanks to my fiancée Xiaoming Duan and many friends at Temple and Bristol-Myers Squibb who cherish me despite my eccentricities. Without all their help, I would not even have the opportunity to write this acknowledgement.

DEDICATION

To Many People,

JUN CAO

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This dissertation deals with nonparametric comparisons between groups when only partial orders are available. A special case is to compare two groups when only two objects can be compared at a time. The research in this area is summarized in the paper by Liu and David (1993); new results from my research work are contained in Chapters 4 and 5. Let us start with the following example:

Consider a situation in a tennis tournament comparing two tennis teams. A key element inherent in this situation is that the relative strengths of the players within each team are usually known. Even if the teams have not finished playing, we may be able to predict the probability that team A will win, given the results from the matches that already been played. There are different ways of matching players, such as random matching, or ordered matching based on the known ranks within teams. Usually in tournament play between teams, the players are matched based on their ranks within teams. This is what we define as the ordered matching. This example and three possible matching schemes are evaluated in Chapter 5.

Nonparametric comparison between groups based on partial orderings can also be used in nonrandomized experiments or observational studies, where the order between individuals is based on a set of ordered characteristics. An individual A has

a lower rank then the individual B if all the comparisons of characteristics have the same relationships. An epidemiological example was described by Morton et al. (1982). They sampled the distribution of lead in the blood of children whose parents worked in a battery factory that used lead in production. Only children of employees were enrolled in the study. The aim of this investigation was to analyze a potential relationship between the level of lead in the children's blood and the intensity of lead exposure at their parents' workplace. An important confounder of this relationship, which was measured in the study, was the parents' individual hygiene practices that can reduce the lead contamination in the children's home environment. Thus, the partial order of the categories is given by $(y_{1i},\ z_{1i}) < (y_{2j},\ z_{2j})$, if and only if ($y_{1i} < y_{2j}$ and $z_{1i} \le z_{2j}$) or ($y_{1i} \le y_{2j}$ and $z_{1i} < z_{2j}$), where y_{lk} and z_{lk} are the level of lead in children's blood and the level of parents' individual hygiene practices for the k^{th} individual in the l^{st} group.

In Section 1.1, I will briefly explain why exact test is preferable. The concept of Fuzzy decisions and Fuzzy p-values by Geyer and Meeden (2005) are introduced in Section 1.2.

1.1 Why Exact Test?

Nonparametric tests are popular mainly because they make minimal assumptions about the data distribution. But most nonparametric tests do require the assumption that the data set is large enough for the test statistics obtained by

evaluating the tail area of the limiting distribution, instead of actually deriving the true distribution of the test statistic and evaluating its tail. P-values based on the large-sample assumption are *asymptotic p-values*, while the p-values based on the true distribution of test statistics are *exact p-values*.

For large and well-balanced data sets, using the asymptotic p-values or the exact p-values makes little difference. But for small, unbalanced and heavily tied data, the exact and asymptotic p-values can be quite different, which may lead to different conclusions. This concern was mentioned by R. A. Fisher in the preface to the first edition of Statistical Methods for Research Workers (1925):

"The traditional machinery of statistical processes is wholly unsuited to the needs of practical research. Not only does it take a cannon to shoot a sparrow; but it misses the sparrow! The elaborate mechanism built on the theory of infinitely large samples is not accurate enough for simple laboratory data. Only by systematically tackling small problems on their merits does it seem possible to apply accurate tests to practical data."

Although an exact p-value is preferred for scientific inference, it is harder to compute practically. In Chapter 4, I will propose a randomized exact test based on partial ordering using the Gibbs Sampler algorithm. This can be considered an extension of the common exact test based on complete orderings such as the Wilcoxon rank-sum test.

1.2 Fuzzy Decisions and Fuzzy P-values

In Chapter 4, a p-value is generated for each random linear extension. Thus, as a direct result of our new method, a single p-value is replaced by a distribution of p-values. This is related to some recent work by Geyer and Meeden (2005) on Fuzzy P-values.

In most nonparametric tests based on rank statistics, the test statistics have a discrete distribution, and thus an exact α level test can only be constructed by randomizing at the critical points (Lehmann, 1998). Tests based on the standard p-values, (e.g., reject the null hypothesis when $P \le \alpha$), are conservative when the test statistics are discrete.

A randomized test provides the value of the randomized decision function $\phi(x)$, which is the probability of rejecting the null hypothesis given the statistics X. Usually, $\phi(x)$ equals to 0 (1) for X inside (outside) the rejection region, and $0 \le \phi(x) \le 1$ on the boundary. The above classical randomized test contrasts with the "*fuzzy P-value*" recently introduced by Geyer and Meeden (2005), which they describe as: "*If $\mathbb{P}$ is a random variable that has a uniform distribution between 0 and 1 unconditionally under the null hypothesis, then a random variable having the conditional distribution of $\mathbb{P}$ given observed data Y is an exact fuzzy P-value*" (Geyer and Meeden, 2005). The test will reject H_0 when $\mathbb{P} \le \alpha$ with probability α under H_0, because

$$E\left\{\Pr\left[\mathbb{P} \le \alpha \mid Y\right]\right\} = \Pr\left(\mathbb{P} \le \alpha\right) = \alpha .$$

A fuzzy P-value is not a single number but a random variable that has a probability distribution. A *fuzzy decision* is the reported value of $\phi(x)$, which is a random variable that has a probability distribution, while the classical randomized decision is a realization of the random variable. Below is the construction of a fuzzy P-value for the sign test given in Geyer and Meeden (2005).

In cases where no tie appears, the fuzzy P-value for the upper-tailed one-sided sign test is constructed in the following way that it is uniformly distributed between $\Pr(T > t)$ and $\Pr(T \geq t)$, where the test statistics here is the sum of the signs of the paired comparisons. *T* is a random variable having the null distribution of the test statistics and *t* is the observed value of the test statistics.

In cases where ties appear, which is a special case of partially ordered data, fuzzy method adds infinitesimal jitter to the data to break the ties. Suppose the observed data have *l*, *t*, and *u* points below, tied and above zero. After jittering, we have $l + K$ points below, and $u + t - K$ points above, where *K* is a binomial distribution with $p = \frac{1}{2}$, and $n = t$. For example, assume we observed 2, 4, 14 data points below, tied, and above zero. If a one-sided sign test is applied, the PDF and CDF of the fuzzy P-value are shown in Figure 1 and Figure 2. In Figure 1, the areas under the five steps are in the ratios 1: 4: 6: 4: 1, which are the ratios of the densities of a binomial distribution with $p = \frac{1}{2}$, and $n = 4$. The step boundaries were

calculated by $P(L \leq 2 + B)$, where L follows a binomial distribution with $p = \frac{1}{2}, n = 20$ and $B = -1, 0, ..., 4$.

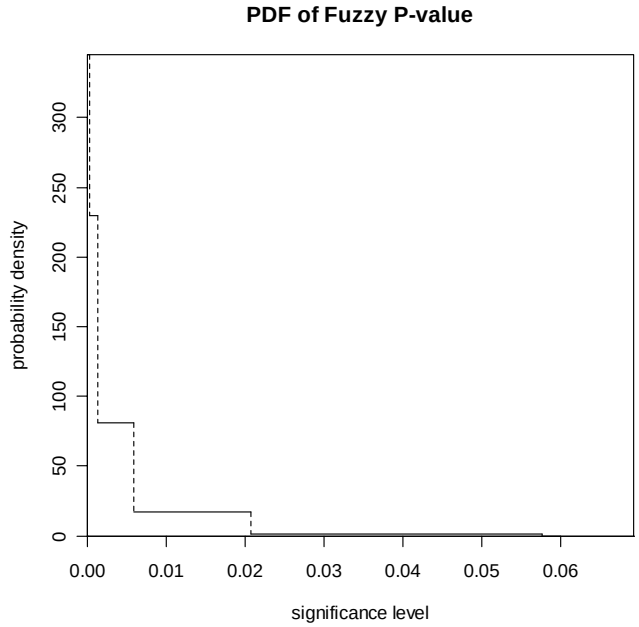

Figure 1: PDF of Fuzzy P-values for Two-sided Sign Test

A way of interpreting fuzzy P-values is explained by Geyer (2005) as if possible fuzzy P-values are concentrated below 0.01, which is then strong evidence against the null hypothesis. If the whole distribution of a fuzzy P-value is concentrated above 0.2, then the evidence against the null hypothesis is so weak that it is practically nonexistent. In the middle, it is equivocal.

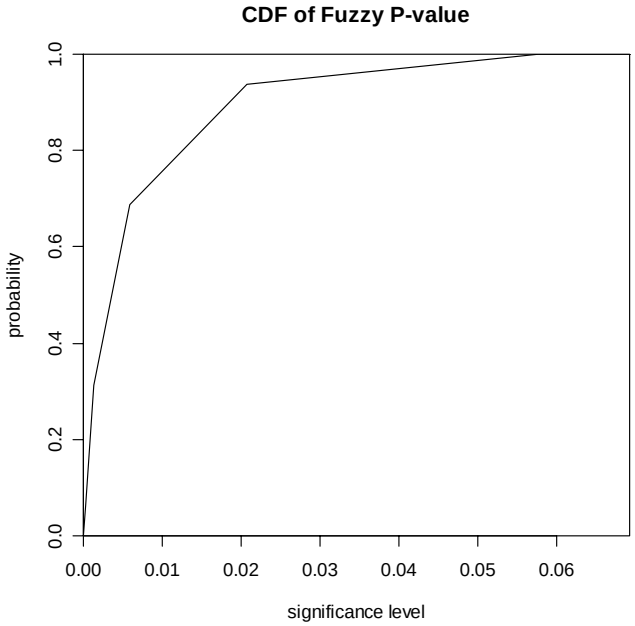

Figure 2: CDF of Fuzzy P-values for Two-sided Sign Test

For the sign test example mentioned above, the evidence against the null hypothesis is not overwhelming, since a part of the distribution of the fuzzy P-value is above 0.01; however, it is fairly strong. Let us compare the result with three popular methods of dealing with ties mentioned by Hollander and Wolfe (1999). The first method is the "standard method" (Dixon and Massey, 1951), which simply ignores the ties; another method is the "randomized method", which counts ties as favoring the null or alternative at random with equal probabilities; the last method is the "conservative method", which counts all ties as favoring the null hypothesis. A comparison of these three popular methods of dealing with ties is the subject of a paper by Pratt (1955). In his paper, the "standard method" which is omitting the ties has the highest power among a large class of tests. This class includes, in particular,

the “conservative method” and the “randomized method” which breaks each tie at random.

If we apply these three methods to the above sign test example, we will get $p = 0.0042$ by the “standard method”, and $p = 0.0577$ by the “conservative method”. Geyer (2005) claimed that the fuzzy P-value is not too liberal or too conservative, and provides more information than the conventional P-values.

Geyer and Meeden (2005) believe that this distribution of $\mathbb{P}$ given Y, instead of a number generated by random realization of it, is what a statistician or scientist should report in a situation where randomized tests are appropriate. Fuzzy P-value has proven useful in latent variable problems by Thompson and Geyer (2005), where fuzzy P-value provides not only an expression of the strength of the evidence against the null hypothesis, but also an expression of the uncertainty due to the lack of knowledge about the latent variables. Those features are illustrated with simulated data mimicking a real genetic linkage detection problem by Thompson and Geyer (2005). It will be shown that the methods in Chapter 3 and 4 can be used to obtain both a classic exact randomized α level test or a distribution of “fuzzy P-values”.

CHAPTER 2

TOOLS PART I: LINEAR EXTENSION AND PARTIAL ORDERING

In this chapter, we will introduce some basic ideas to be used in Chapter 4 when constructing a randomized exact test based on incomplete information about the observed ranks. Definitions of partial order, linear extension and statistical poset are introduced in Section 2.1. For completeness, the classic randomized Wilcoxon test for comparing complete orders is introduced in Section 2.2. The Markov Chain Monte Carlo simulation method for generating linear extension for a partial order is described in Section 2.3. The connection between random linear extension and the well known multiple imputation method (Rubin, 1987) is briefly explained in Section 2.4.

2.1 Partial Order, Linear Extension and Statistical Poset

A *partial order* on a set $x = (x_1, x_2, \ldots, x_N)$ is a set of consistent pairwise restrictions $x_{i_k} < x_{j_k}$ for some collection of pairs $(i_1, j_1), ..., (i_K, j_K)$. It satisfies the transitivity condition:

if $x_i < x_k$ and $x_k < x_j$ then $x_i < x_j$.

A partially ordered set is also called a *partial ordering* or *poset.*

A *linear extension I* of a partially ordered set *P* is a permutation of the elements $x_1, x_2, \ldots, x_N$, such that $x_i < x_j$ in *P* implies $x_i < x_j$ in *I*.

A *complete order* is a set of pairwise restrictions satisfies the conditions for a partial order plus an additional condition known as the comparability condition:

- for any $i, j \in (1, ..., N)$, either $x_i < x_j$ or $x_j < x_i$.

For example, we have a partial order on a set $x = (x_1, x_2, x_3, x_4)$ as $\{(x_1 < x_2), (x_3 < x_4)\}$. The directed graph of this four elements poset is shown in Figure 3.

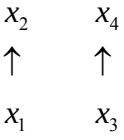

Figure 3: A Directed Graph of a Four Elements Poset

According to the definition, its linear extensions consist of those permutations of (x_1, x_2, x_3, x_4) that have x_1 smaller than x_2, and x_3 smaller than x_4. These permutations are (x_1, x_2, x_3, x_4), (x_1, x_3, x_2, x_4), (x_1, x_3, x_4, x_2), (x_3, x_1, x_2, x_4), (x_3, x_1, x_4, x_2), (x_3, x_4, x_1, x_2).

Linear extensions of a partial order could be used in many cases. Consider an artificial situation in a chess tournament, stronger players in the game always beat weaker players. Instead of having the tournament continued until all the players are ranked, we may be able to predict the probability of player A ending up ranked 1st,

given the results from the matches that were already played. If we can sample uniformly from the set of possible rankings under constraints imposed by the observed partial ordering from the matches already played, we will be able to estimate the probability of a particular final ranking. This is equivalent to generating all random linear extensions of a partial order, which is on a set $(x_1, x_2, \ldots, x_N)$ of N players, and has a set of pairwise restrictions $x_{i_k} < x_{j_k}$ for selected K pairs $(i_1, j_i), \ldots, (i_K, j_K)$, those are, the K chess matches already played. This random sampling was discussed by Matthews (1991). To draw a random order constrained by a fixed poset, he proposed simulating a reflecting random walk in an n-dimension polynomial polytope that is a subset of N-dimensional space, since the distribution of the random walk is uniform in the polytope.

Linear extensions and partial orders are also widely used to solve environmental and ecological problems. An example is the human-environment interface which was conducted by United Nations Environment Programme (UNEP) to compare and rank the countries according to environmental quality. Three indicators (air, water and land) were evaluated and given values between 0 and 1 for each of 106 countries. In this case, Patil and Taillie (2004) suggested using the cumulative rank-frequency (CRF) ordering, which can be estimated using discrete Markov Chain Monte Carlo (MCMC) methods (Aldous, 1987).

2.2 A Randomized Wilcoxon Test for Complete Orderings

The methodology proposed in Chapter 4 and 5 will apply a randomized Wilcoxon test to the complete orderings generated by a simulation algorithm. In this section, we will review this randomized Wilcoxon test.

Let $X = (x_1, ..., x_n)$, with cumulative distribution function $F_X(X)$, and $Y = (y_1, ..., y_m)$, with cumulative distribution function $F_Y(Y) = F_Y(X + \theta)$. Assume we have the complete order of data, we can construct a randomized Wilcoxon Rank-sum test for $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$. The randomized decision function of the test is

$$\phi_0(T) = \begin{cases} 1 & T < C_1 \text{ or } T > C_2 \\ \gamma & T = C_1,\ T = C_2 \\ 0 & C_1 < T < C_2 \end{cases}$$

where $T = S_1 + \cdots + S_n$, and $S = (S_1, \ldots, S_n)$ are the ranks of $X = (x_1, ..., x_n)$ based on the complete orderings of all $n+m$ data points from the two groups.

In this way, we will have a random variable *D* that takes the value 1, "reject H_0", with probability $\phi_0(T)$, and takes value 0, "accept H_0", with probability $1 - \phi_0(T)$. If a classical randomized test is preferred, we can generate a Uniform(0.1) random variable *U* and "reject H_0" if $U < \phi_0(T)$ and "accept H_0" otherwise. However, if a fuzzy P-value (Geyer and Meeden, 2005) is preferred, we can provide the distribution of *D* instead of a random realization of it, that is $P(D = 1) = \phi_0(T)$ and $P(D = 0) = 1 - \phi_0(T)$.

2.3 Markov Chain Monte Carlo Method for Uniform Discrete Sampling

For a Markov Chain $\left\{x^{(i)}\right\}$ on a finite state space *V* with stationary distribution π, we know that the sample mean of an arbitrary *h* function of $\left\{x^{(i)}\right\}$,

$$H_t = t^{-1}\sum_{l=1}^{t} h\left(x^{(i)}\right),$$

approaches the expectation of *h* under the stationary distribution π,

$$\bar{h} \equiv \sum_{v \in V} \pi(v) h(v),$$

as $t \to \infty$. The MCMC simulation method simulates s process with a stationary distribution π on a space *V*. When simulating the distribution π directly is not easy, MCMC provides a procedure to construct a Markov chain with stationary distribution as π and then simulate the Markov chain.

An application of this method to uniform combinatorial distributions is described by Aldous (1987), and is summarized below.

Let V_N be a finite "combinational" set such as: $\{0,1\}^N$ or all permutations of $\{1,\ldots,N\}$. If we can simulate a uniform random element U of V_N, then the empirical average of repeated simulations gives an estimator of $\bar{h}$ with error $O\left(t^{-1/2}\right)$. For the simple sets V_N, such as in the examples above, it is possible to sample uniformly from the sampling frame of all possible outcomes. But for other

cases, direct simulation might be difficult, since the sampling frame is too large to list and enumerate. For example, consider the complex set V_N as below.

(i). All permutations X_N with a specified upper bound on $\sum iX_N(i) \le C$, where X_N denotes a random permutation of $\{1, \ldots, N\}$.

(ii). All $N \times N$ matrices with positive integer entries and specified row and column sums.

The Markov Chain Monte Carlo (MCMC) method can be applied to generate uniform random samples from the complex sets V_N. Aldous (1987) suggested updating a point v of V_N to another element v' by a "minimal change". For the examples above, the minimal changes were defined as follows.

(i). Given $X_N^{(t)}$, pick i_1, i_2 and let $Y_N^{(t)}$ be a "trial permutation", where $Y_N^{(t)}(i_1) = X_N^{(t)}(i_2)$, $Y_N^{(t)}(i_2) = X_N^{(t)}(i_1)$, $Y_N^{(t)}(i) = X_N^{(t)}(i)$ for $i \ne i_1, i_2$. If $\sum iY_N^{(t)}(i) \le C$, then $X_N^{(t+1)} = Y_N^{(t)}$, else $X_N^{(t+1)} = X_N^{(t)}$.

(ii). Given $A^{(t)} = \left(a_{i,j}\right)$, pick i_1, i_2, j_1, j_2 and let $\tilde{A}^{(t)}$ be the matrix $\tilde{a}_{i_1 j_1} = a_{i_1 j_1} + 1$, $\tilde{a}_{i_1 j_2} = a_{i_1 j_2} - 1$, $\tilde{a}_{i_2 j_1} = a_{i_2 j_1} - 1$, $\tilde{a}_{i_2 j_2} = a_{i_2 j_2} + 1$, $\tilde{a}_{ij} = a_{ij}$ for $i \ne i_1, i_2$ and $j \ne j_1, j_2$. If $\tilde{A}^{(t)}$ is nonnegative, then $A^{(t+1)} = \tilde{A}^{(t)}$, else $A^{(t+1)} = A^{(t)}$.

2.4 Multiple Imputation

Generating a complete ordering from a partial order has a connection to the now classic method, multiple imputation, first introduced by Rubin (1987). Multiple imputation (Rubin 1987) is a simulation based approach to deal with incomplete data. Although there are many methods to deal with incomplete data, including ad-hoc methods such as case deletion, mean substitution, and maximum likelihood methods, multiple imputation has become one of the most popular methods.

One of the basic objectives in multiple imputation is to enable the user to use complete-data procedures. There are three steps involved in multiple imputation:

Step 1: impute the missing data

Step 2: analyze each complete data set using a complete data procedure

Step 3: combine the results from each complete data to provide a final result

First, let's talk about imputing the missing data. Rubin's (1987) imputation method consists of simulating *D* independent versions of the missing data from the posterior predictive conditional probability mass function

$$p\left(X_{mis} = x_{mis} \mid X_{obs} = x_{mis},\ M = m\right),$$

where X_{obs} is the observed part of data, X_{mis} is the missing part of data and to create $M = \left[m_{i,j}\right]$ is a set of missing indicators as $m_{i,j} = \begin{cases} 1 & \text{if } I_{ij} \text{ is missing} \\ 0 & \text{if } I_{ij} \text{ is observed} \end{cases}$. Here, $I_{ij} = \begin{cases} 1 & \text{if } x_i < x_j \\ 0 & \text{if } x_i \geq x_j \end{cases}$. In this way, it generates draws from a posterior predictive distribution for the missing data given the observed data. We also assume that the

missing-ordering mechanism does not depend on the missing values X_{mis}, which is defined as missing at random when:

$$p(M = m \mid X = x) = p(M = m \mid X_{obs} = x_{obs}) \text{ for all } X_{mis}.$$

After the complete data sets are generated, common complete data procedures can be applied. We will then have the estimates of $E\left(h(\theta) \mid X_{obs}, X_{mis}\right)^j$ and $V\left(h(\theta) \mid X_{obs,} X_{mis}\right)^j$ from each complete data set, where h is a function of θ, which are the parameters of complete data distribution, and $j = 1, \ldots, D$.

The last step of multiple imputation involves combining the statistics from each complete data set to provide one final result. Rubin (1987) developed the rules for combining the estimates of $E\left(h(\theta) \mid X_{obs}, X_{mis}\right)^j$ and $V\left(h(\theta) \mid X_{obs,} X_{mis}\right)^j$ based on normal approximation. Additional combination rules were suggested by Meng and Rubin (1992) for combining likelihood ratio test (LRT) statistics based on the asymptotic distribution of LRT statistics.

2.5 Some Poset Statistics and Tests for Coherence

Rosenbaum (1991) proposed some poset statistics to measure the agreement of two partially ordered samples. The statistics is a function $h\left(X_1, X_2\right)$ of two matrix arguments, where each has N rows and M_1, M_2 columns. Let $\rho^*\left(X_i\right)$ be a simple rank-score function obtained by separately ranking each column of X_i, with

average ranks in case of ties and using the sum of the ranks in a row as the score for that row.

For $M_1 \geq 1$ or $M_2 \geq 1$, one statistics suggested by Rosenbaum (1991) is $\rho_1 * (X_1)^T \rho_2 * (X_2)$. Write $\rho_{X_1} = \rho_1 * (X_1)$, $\rho_{X_2} = \rho_2 * (X_2)$, $\overline{\rho}_{X_1} = (1/N)\rho_1 * (X_1)$ and $\overline{\rho}_{X_2} = (1/N)\rho_2 * (X_2)$. Under the null hypothesis, the moments of the statistics are $E\left(\rho_{X_1}{}^T \rho_{X_2}\right) = N\overline{\rho}_{X_1}\overline{\rho}_{X_2}$ and

$$Var\left(\rho_{X_1}{}^T \rho_{X_2}\right) = \left[1/(N-1)\right]\left[\sum(\rho_1 * (X_1)_i - \overline{\rho}_{X_1})^2\right]\left[\sum(\rho_2 * (X_2)_j - \overline{\rho}_{X_2})^2\right].$$

For $M_1 \geq 1$ and there are two treatment groups indicated by binary coordinates of the $M_2 = 1$ column of X_2, the statistics suggested by Rosenbaum is $\rho_1 * (X_1)^T X_2$, which is also the sum of M_1 Wilcoxon ranks sum statistics.

In chapter 4, Rosenbaum's linear rank test will be compared with the proposed randomized linear-extension test using the lead evaluation data (Morton, 1982).

CHAPTER 3

TOOLS PART II: THE GIBBS SAMPLER

The random-linear-extension test based on classic nonparametric procedures proposed in Chapter 4 needs a way of uniformly generating all complete orderings based on a partial order. Here, we show that the Gibbs Sampler is a natural way of doing this. The Gibbs Sampler is a Markov Chain Monte Carlo (MCMC) simulation method that eventually generates a draw from the joint distribution. It generates random variables from marginal distribution without calculating the density. The purpose of the Gibbs Sampler is to approximate the joint distribution and the estimates for the parameters of interest.

The Gibbs Sampler was introduced by Geman and Geman (1984), who studied image-processing models. Additionally, Gelfand and Smith (1990) pointed out that the algorithm can also be used to simulate from posterior distributions.

The advantages of Gibbs Sampler and the other sampling-based methods are their conceptual simplicity and ease of implementation. All that is required are the relevant conditional probability structures and the techniques for efficient generation of appropriate random variables.

In this chapter, we will introduce the basic statistical concepts related to Gibbs Sampler. The Gibbs Sampler algorithm and updating strategies are introduced in Section 3.1. The ways of evaluating convergence of the algorithm are explained in Section 3.2.

3.1 The Algorithm and Updating Strategies

The Gibbs Sampler eventually generates a draw from the distribution $P\left(x_1,\ldots,x_p\right)$ of a set of p random variables $X_1,\ldots,X_p$, utilizing the available conditional distributions $P\left(x_j \mid x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p\right)$, where $j=1,\ldots,p$. The iterative sampling scheme is implemented as shown in the following steps:

I. An initial starting value of parameters is required. Initial values $x_1^{(0)},\ldots,x_p^{(0)}$ are chosen to satisfy the constraint.

II. Then, one at a time, a new value for each parameter of interest is sampled by drawing from the conditional distributions, given the current values of the other parameters at iteration i:

$$
\begin{aligned}
x_1^{(i+1)} &\sim p\left(x_1 \mid x_2^{(i)},x_3^{(i)},\ldots,x_p^{(i)}\right)\\
x_2^{(i+1)} &\sim p\left(x_2 \mid x_1^{(i)},x_3^{(i)},\ldots,x_p^{(i)}\right)\\
x_3^{(i+1)} &\sim p\left(x_3 \mid x_1^{(i)},x_2^{(i)},x_4^{(i)},\ldots,x_p^{(i)}\right)\\
&\vdots\\
x_p^{(i+1)} &\sim p\left(x_p \mid x_1^{(i)},x_2^{(i)},\ldots,x_{p-1}^{(i)}\right)
\end{aligned}
$$

III. Once all parameters of interest have been sampled, return to step II. After t iterations, we have $\left(x_1^{(t)},\ldots,x_p^{(t)}\right)$.

This is the most widely adopted version of the Gibbs Sampler, which uses deterministic sweep updating strategy by Amit and Grenander (1991) updating the components in the natural order. There are several other updating strategies following orders other than the natural order. Some are described below.

One updating strategy is the random sweep strategy, which states that at the i^{th} update, we generate a uniformly distributed random variable k over $\left(1,\ldots,p\right)$ and update $x_k^{(i+1)}$. We repeat this p times and treat the entire updates as one iteration to keep it consistent with the other strategies.

Another updating strategy is the random permutation strategy, which is that at the i^{th} update we generate a random permutation of $\left(1,\ldots,p\right)$ and update the components in that order. The resulting Gibbs Sampler is called Random Permutation Gibbs Sampler (RPGS). This random permutation Gibbs Sampler will be used in Chapter 4 for generating complete orderings from the known partial orders.

3.2 Convergence

The solutions for the problem of determining the MCMC algorithm (including Gibbs Sampler) convergence have been concentrated in two areas. The first is to theoretically determine the number of iterations that will ensure

convergence; the second is to apply diagnostics tools to the output produced by the algorithm.

Suppose that $\left\{x^{(t)}, t = 0,1,\ldots\right\}$ is a Markov Chain with stationary density function $\pi = P\left(x_1, \cdots x_p\right)$, where $x^{(t)} = \left(x_1^{(t)}, \ldots, x_p^{(t)}\right)$. Let $h\left(x\right)$ be a measurable function of $x = \left(x_1, \ldots, x_p\right)$. The convergence of the expectation of $h\left(x^{(t)}\right)$ at time t to the expectation of $h\left(x\right)$ under the stationary distribution π ensures the proper behavior of the chain $\left\{x^{(t)}\right\}$ regardless of the initial value $x^{(0)}$.

A theoretical description of convergence properties involves the speed of convergence of $P\left(x^{(t)} \mid x^{(0)}\right)$ to π. The rate of convergence of the Markov chain can be described as how quickly the expectations of arbitrary measurable functions approach their stationary values.

With a reasonable number of simulations, Geman and Geman (1984) showed that the following results hold using the deterministic sweep updating strategy.

Result 3.1. (Geman and Geman 1984) For the Gibbs sampling algorithm described as in Section 3.1,

a) $\left(x_1^{(t)}, \ldots, x_p^{(t)}\right)$ converges to a draw $\left(x_1, \ldots, x_p\right)$ from the joint distribution $P\left(x_1, \ldots, x_p\right)$, as $t \to \infty$. In fact, assuming that each variable is visited infinitely often, convergence follows under any visiting scheme, rather than requiring that each variable be visited in repetitions of the natural order.

b) The convergence in part a) is exponential in t using the L_1 norm.

c) (Ergodic Theorem) For any measurable function $h(x)$, $x = (x_1, \ldots, x_p)$ whose expectation exists,

$$\lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{t} h\left(x_1^{(i)}, \ldots, x_p^{(i)}\right) \xrightarrow{a.s.} E\left(h\left(x_1, \ldots, x_p\right)\right).$$

Theoretical bounds on rates of convergence for the appropriate Markov Chain were also discussed by Amit and Grenander (1991), Roberts and Polson (1994), Fishman (1996), Roberts and Sahu (1997). Theoretical results provide convergence rates which can assist in selecting between competing algorithms, but the rates are available only up to an arbitrary constant, which provide little use for a stopping rule to guarantee that the number of iterations is sufficient. Lower bounds on convergence rates have been calculated in special cases and are often too conservative to be of any practical value. As a result, convergence diagnostics based on the output of single or multiple chains are usually used to detect convergence.

Cowles and Carlin (1996) presented a review of 13 diagnostics. For example, Gelfand and Smith (1990) suggest monitoring density estimates from D independent Gibbs sequences, and choosing the number of iterations to be the first point at which these densities appear to be the same under a "felt-tip pen test". Raftery and Lewis' (1992) method provides the total number of iterations that should be run by running a single-chain Gibbs Sampler with minimum number of iterations that would be needed to obtain the desired precision of estimation. This method is based on the two-state Markov Chain theory, as well as standard sample size formulas involving

binomial variance. Geweke (1992) recommends using methods from spectral analysis to assess the convergence of the Gibbs Sampler when the analysis is targeted to estimate the mean of some function *h* of the parameters. Ritter and Tanner (1992) suggest monitoring a sequence of weights that measure the discrepancy between the sampled and the desired distribution. Heidelberger and Welch (1983) recommend using the Cramer-von-Mises statistics to test the null hypothesis that the sampled values are from a stationary distribution. The methods that will be used in Chapter 4 and 5 are the Raftery and Lewis' method (1992) and the Heidelberger and Welch's (1983) method.

CHAPTER 4

A RANDOM-LINEAR-EXTENSION TEST BASED ON CLASSIC NONPARAMETRIC PROCEDURES

A randomized exact test is introduced here as a new way to compare two groups of data with partial orders. A randomization method similar in nature to multiple imputation (Rubin 1987) is proposed. The basic computation procedure is to generate all possible permutations constrained by the known partial order. This can be simply implemented using a Gibbs Sampler described in Chapter 3. Given a complete ordering generated by the Gibbs Sampler, standard nonparametric methods, such as the Wilcoxon rank-sum test, can be applied, and the corresponding test statistics and rejection regions can be calculated. By combining the statistical decisions generated from multiple imputed complete orderings, we are able to obtain a randomized exact test to compare two groups of partially ordered data. To illustrate this procedure, the Gibbs sampler introduced in Chapter 3 is used to generate a complete ordering based on the partial ordering.

In Section 4.1, we will introduce a random-linear-extension test based on the Wilcoxon test. In Section 4.2, the algorithm for conducting the proposed random-linear-extension test is explained in detail and the convergence of the algorithm is

evaluated in Section 4.3. In section 4.4, this method is compared with the method proposed by Rosenbaum (1991).

4.1 A Random-Linear-Extension Test based on Wilcoxon Test

Consider a situation where only partial order is observed for various reasons. Denote the observed binary orderings by

$$\mathcal{I}_{obs} = \left\{ \left(i, j, I_{ij} \right) \mid m_{i,j} = 0 \right\},$$

where $\mathcal{I} = \left\{ \left(i, j, I_{ij} \right) \right\}$ is the complete order,

$$I_{ij} = \begin{cases} 1 & \text{if } x_i < x_j \\ 0 & \text{if } x_i \geq x_j \end{cases} \quad \text{and} \quad m_{i,j} = \begin{cases} 1 & \text{if } I_{ij} \text{ is missing} \\ 0 & \text{if } I_{ij} \text{ is observed} \end{cases}.$$

We also assume that the missing-order mechanism does not depend on the missing values $\mathcal{I}_{mis}$. Using Rubin's (1987) terminology, it is defined as missing at random when the conditional mass function satisfies:

$$p(M = m \mid \mathcal{I}) = p(M = m \mid \mathcal{I}_{obs}) \text{ for all } \mathcal{I}_{mis}.$$

In fact, for the case considered here, M is determined exactly by $\mathcal{I}_{obs}$. Therefore, the probability mass function is identical for each set of missing indicators M. With the complete order $\mathcal{I}$, the computation of test statistics T is usually straightforward, for example, if a Wilcoxon rank-sum test is used,

$$T = S_1 + \cdots + S_n,$$

where $S = (S_1, \ldots, S_n)$ are the ranks of the observations from the first group. However, with the partially ordered data, the computation of *T* is impossible, since the ranks are unknown.

In the case of partially ordered data, a Gibbs Sampler, described in Chapter 3, can be used to generate a uniform draw from all possible complete orderings satisfying its known constraints. Once its complete ordering is generated, techniques for data with complete order can be applied. The procedure outlined above can be thought of as a special case of multiple imputation by Rubin (1987). The unusual characteristic of the proposed methods is that each imputation is generated uniformly at random using a Gibbs Sampler. To provide a single α level decision, a randomized exact test is proposed here by combining the repeated imputation inferences.

For example, if *D iid* imputations for missing values $\mathcal{I}_{mis}$ are simulated as $\mathcal{I}_{*mis}^{(l)},\ l = 1, ..., D,$ the corresponding *D* complete orders can be constructed as

$$\left\{\mathcal{I}_{*}^{(l)}; l = 1, ..., D\right\} = \left\{\left(\mathcal{I}_{obs} \cup \mathcal{I}_{*mis}^{(l)}\right); l = 1, ..., D\right\}.$$

Then data analysis based on *D* complete orderings can be applied to compute $T_{*l},\ l = 1, ..., D,$ where T_{*l} is calculated based on the ranks of the imputed complete orderings.

For each imputated complete ordering, the randomized decision function of an unbiased randomized test for $H_0 : \theta = 0$ versus $H_1 : \theta > 0$, where $X \sim F_X(X)$ and $Y \sim F_Y(Y) = F_X(X + \theta)$, is

$$\phi_{*l}\left(T_{*l}\right)=\begin{cases}1 & T_{*l}>C\\ \gamma & T_{*l}=C\\ 0 & T_{*l}<C\end{cases},$$

where T_{*l} is the sufficient statistics for θ and C is calculated from Mann-Whitney-Wilcoxon distribution to satisfy $E_{H_0}\left(\phi_{*l}\left(T_{*l}\right)\mid \mathcal{I}_{obs}\cup \mathcal{I}_{*mis}^{(l)}\right)=\alpha$.

To construct an exact randomized test for $H_0:\theta=0$ versus $H_1:\theta>0$ based on the observed partial ordering, we propose a random-linear-extension test with the randomized decision function as

$$\phi_p\left(T_D\right)=\frac{1}{D}\sum_{l=1}^{D}\phi_{*l}\left(T_{*l}\right). \qquad (4.1)$$

where $\phi_{*l}\left(T_{*l}\right),\ l=1,\ldots,D,$ are the randomized decision functions for each imputed complete ordering, and $T_D=\left(T_{*1},\ldots,T_{*D}\right)$. We should have $T_D\in T_P$, where T_P represent the test statistics generated from all possible complete orderings which satisfy the observed partial order *P*. By averaging the randomized decision functions generated from multiple imputed complete orderings, we can obtain one randomized decision for each observed partial order *P*.

Lemma 4.1. *The randomized test based on (4.1) is an unbiased test if all* $\phi_{*l}\left(T_{*l}\right),\ l=1,\ldots,D,$ *are unbiased exact tests based on the imputed complete orderings.*

Proof: Recall the law of conditional expectations:

$$E\left[E\left[X\mid Y\right]\right]=E\left[X\right]$$

Thus,

$$
\begin{aligned}
& E\left(\phi_{*l}\left(T_{*l}\right) \mid \mathcal{I}_{obs}\right) \\
&= E\left(E\left(\phi_{*l}\left(T_{*l}\right) \mid \mathcal{I}_{mis}^{(l)} \cup \mathcal{I}_{obs}\right) \mid \mathcal{I}_{obs}\right) \\
&= \alpha.
\end{aligned}
$$

Referring to Lemma 4.1, we have

$$
\begin{aligned}
& E\left(\phi_{D}\left(T_{D}\right) \mid \mathcal{I}_{obs}\right) \\
&= E\left(\frac{1}{D}\sum_{l=1}^{D}\phi_{*l}\left(T_{*l}\right) \mid \mathcal{I}_{obs}\right) \\
&= \frac{1}{D}\sum_{l=1}^{D}E\left(\phi_{*l}\left(T_{*l}\right) \mid \mathcal{I}_{obs}\right) \\
&= \alpha.
\end{aligned}
$$

In this way, we conclude that the randomized test constructed with randomized decision function $\phi_D\left(T_D\right)$ is an unbiased test for the observed partially ordered data.

4.2 An Empirical Version of the Random-Linear-Extension Test

After the complete orderings are imputated from the known partial orderings, another way of constructing a randomized exact test is to build a randomized test based on the empirical distribution of the test statistics.

First, let's simulate *m* random data sets under null hypothesis satisfying the known partial order. Then we can propose an empirical random-linear-extension test with the randomized decision function as

$$
\phi_e\left(\overline{T_k}\right) = \begin{cases} 1 & \overline{T_k} > C_e \\ \gamma_e & \overline{T_k} = C_e \\ 0 & \overline{T_k} < C_e \end{cases} \qquad (4.2).
$$

For the data set k, $k = 1, 2, \ldots, m,$ simulated data set, $\overline{T_k} = \frac{1}{D}\sum_{l=1}^{D} T_{kl}$ is the within sample mean value of a test statistic for D random linear extensions generated and C_e is calculated from the empirical distribution of $\overline{T}_k$ to satisfy $E_{H_0}\left(\phi_e\left(\overline{T}_k\right)\right) = \alpha.$ When m is large enough, the randomized test constructed with $\phi_e\left(\overline{T}\right)$ is an unbiased randomized test for the observed partially ordered data.

4.3 Algorithms for Conducting the Proposed Random-Linear-Extension Tests

We propose the following randomized exact test algorithms which can be used in most partial ordering situations and will produce consistent results. There are two parts to the algorithm. In the first part we impute ranks based on partial order, and in the second part we combine multiple imputed results to construct a randomized exact test. Combining the arguments in Sections 4.1, 4.2 and the Random Permutation Gibbs Sampler (RPGS) described in Chapter 3, the steps are as follows:

(1) Let $Z = (x_1, \ldots, x_m, y_1, \ldots, y_n)$, $A_i = \left\{z_k : z_k < z_i, k \in \left(1, ..., m+n\right)\right\}$, where A_i denotes the set of all points that are lower than $z_{i,}$, and $B_i = \left\{z_k : z_k > z_i, k \in \left(1, ..., m+n\right)\right\}$, where B_i denotes the set of all points that are greater than $z_{i,}$.

(2) Start with an initial vector, which satisfies the partial order.

(3) Within a block of $m + n$ steps, choose the updating order following a random permutation of coordinates $(1, 2, \ldots, m + n)$.

(4) Using the coordinates i chosen in step (3), generate a random uniform number u_i on the interval $(\max_{z_k \in A_i}(z_k), \min_{z_k \in B_i}(z_k))$, and let $z_i = u_i$ in the vector Z.

(5) Repeats (3) and (4) for t times, then take the ranks across groups.

For the random-linear-extension procedure proposed in Section 4.2, the following steps are:

(6) Compare the ranks from two groups by a classic nonparametric test, such as Wilcoxon rank-sum test.

(7) Repeat (3) to (6) independently for *D* times, and then construct the random-linear-extension test by averaging the randomized decision functions generated from the multiple imputed complete orderings.

For the empirical random-linear-extension test proposed in Section 4.2, the following steps are:

(6) Calculate the test statistics, such as the sum of ranks within one team.

(7) Repeat (3) to (6) independently for *D* times, and then generate the randomized exact test based on empirical distribution of the test statistics.

Constructing the Initial Vector

To conduct the above steps, we need a way of generating a vector of initial values, which satisfies the partial order. A way to do this is to create a

$(m+n)\times(m+n)$ rank matrix $M = \left[m_{i,j} \right]$. Now consider the general case where a partial order on a set $x = (x_1, x_2, \ldots, x_N)$ is defined by a set of consistent pairwise restrictions $x_i < x_j$. Let $m_{i,j}$ equals to 1 if and only if $x_i < x_j$. By adding the rows of rank matrix ***M*** and ranking the sums, we will get a vector of initial values, which satisfies the partial order. The proof is as following:

Lemma 4.2. *The rank vector of the row sums of rank matrix **M** satisfies the partial order.*

Proof: Denote the row sums of matrix ***M*** as $(s_1, s_2, \ldots, s_{m+n})$ and its corresponding rank vector as $(r_1, r_2, \ldots, r_{m+n})$. To show that the rank vector of the row sums satisfies the partial order, we need to prove that $r_i > r_j$ for any paired relationship in the partial order, such as $z_i > z_j$.

Let $A_j = \left\{ z_k : z_k < z_j,\ k \in (1, ..., m+n) \right\}$, where A_j denotes the set of all points that are lower than z_j. Then $A_i = \left\{ z_k : z_k < z_i,\ k \in (1, ..., m+n) \right\}$ includes z_j and all the points that are lower than z_j. Thus, $A_j \in A_i$. Since the row sums s_i and s_j are the number of observations in A_i and A_j, respectively. We can conclude that $s_i > s_j$. Therefore, the corresponding ranks $r_i > r_j$.

4.4 Convergence of the Iteration Steps

The convergence of the iteration steps is essential for the consistency of the proposed random-linear-extension test based on classic nonparametric procedures. If the Gibb sampler iterations have not proceeded long enough, the simulations may be seriously biased. We describe the convergence tests applied in the following paragraph. The results for special designs are shown in the Appendix B.

The convergence diagnostic applied here is based on the Heidelberger and Welch's test (1983) and the Raftery and Lewis' method (1992). Heidelberger and Welch's test (1983) uses the Creamer-von-Mises statistic to test the null hypothesis that the sampled values come from a stationary distribution. If the overall stationarity of the whole chain fails this test, the first 10% of the chain is discarded and the remainder is reanalyzed. The sequence proceeds until the last 50% of the chain remains, and if this portion of the chain fails, the chain is said to fail the stationarity test. Additionally, a half-width test is applied on the stationary portion of the chain, which assures the sufficiency of the chain length to provide a reliable estimate to the mean of the parameter. Raftery and Lewis' (1992) method reports the total number of iterations that should be run by running a single-chain Gibbs Sampler with minimum number of iterations that would be needed to obtain the desired precision of estimation. This method is based on the two-state Markov chain theory, as well as standard sample size formulas involving binomial variance.

For the random linear extension Wilcoxon test, given the number of iterations as 10000, the sum of the simulated ranks from the first group was

calculated and outputted at the end of each iteration. Heidelberger and Welch's convergence diagnostic was performed on the chain of the rank sums. Raftery and Lewis' (1992) method was also applied to report the total number of iterations that should be run. For different partial ordering schemes or number of observations in each group, the number of iterations required might be different. Those numbers are listed in Table 12 of Appendix B.

For example, two groups with 20 observations per group are compared using a random matching scheme as described in section 5.1. The iteration passed the stationarity test with p-value 0.193, and also passed the half-width test with estimated half width as 1.62. The number of 'burn in' iterations is estimated to be 16, and the number of iterations required after 'burn in' is 4836, in order to estimate the 0.025 quantile with a 90% CI within an accuracy of 0.01. The autocorrelation plot, trace plot and density plot of the simulated rank sums are shown in Figures 4 to 6.

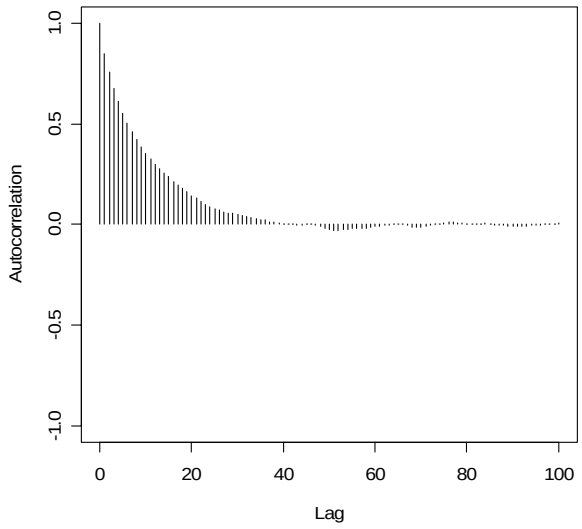

Figure 4: Autocorrelation Plot of the Simulated Rank Sums by Comparing Two Groups in Randomly Selected Pairs (n per group= 20, number of iterations=10000).

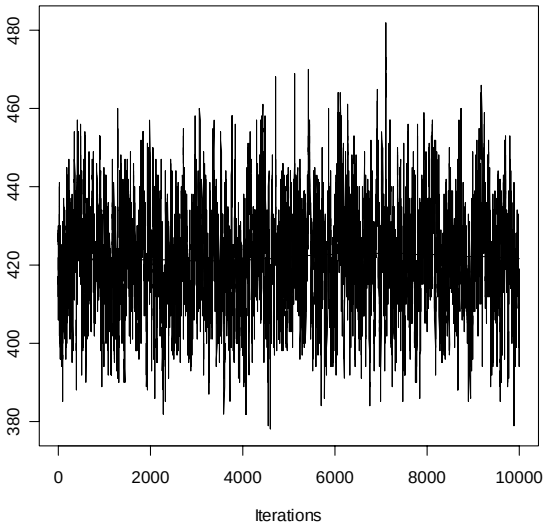

Figure 5: The Trace Plot of the Simulated Rank Sums by Comparing Two Groups in Randomly Selected Pairs (n per group= 20, number of iterations=10000)

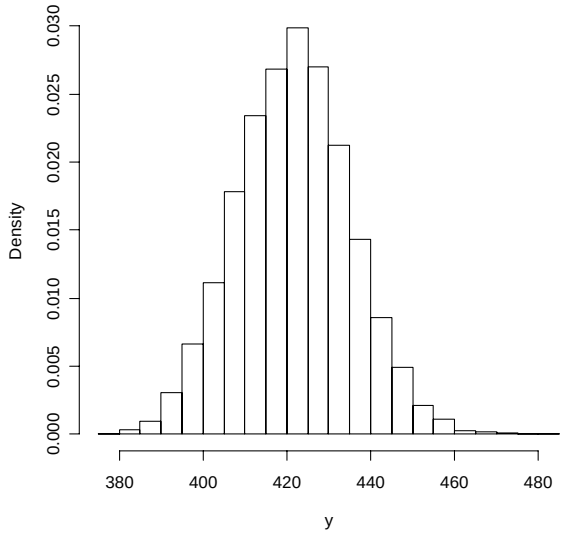

Figure 6: The Density Plot of the Simulated Rank Sums by Comparing Two Groups in Randomly Selected Pairs (n per group= 20, number of iterations=10000)

4.5 Application

For the epidemiological example described in Chapter 1, where the blood lead levels of children whose parents worked in a battery factory was measured, the proposed random-linear-extension test can be applied.

Since the parents' individual hygiene practices can reduce the lead contamination in the children's home environment, the relationship between the level of lead in the children's blood and the intensity of lead exposure at their parents' workplace are confounded with the parents' individual hygiene. Thus, we would anticipate agreement between children's lead levels and the partial order of the categories by $\left(z_{1i},\ z_{2i}\right) < \left(z_{1j},\ z_{2j}\right)$, if and only if $(z_{1i} < z_{1j}$ and $z_{2i} \le z_{2j})$ or

($z_{1i} \leq z_{1j}$ and $z_{2i} < z_{2j}$), where z_{1k} and z_{2k} are the parent's exposures to lead (0 = low, 1 = medium, 2 = high) and the level of parents' individual hygiene practices (0 = good, 1 = median, 2 = poor) for the k^{th} individual. The measured blood levels together with the partial order of exposure levels and hygiene levels are presented in Figure 7. Numeric values in boxes are blood lead concentrations in individual children.

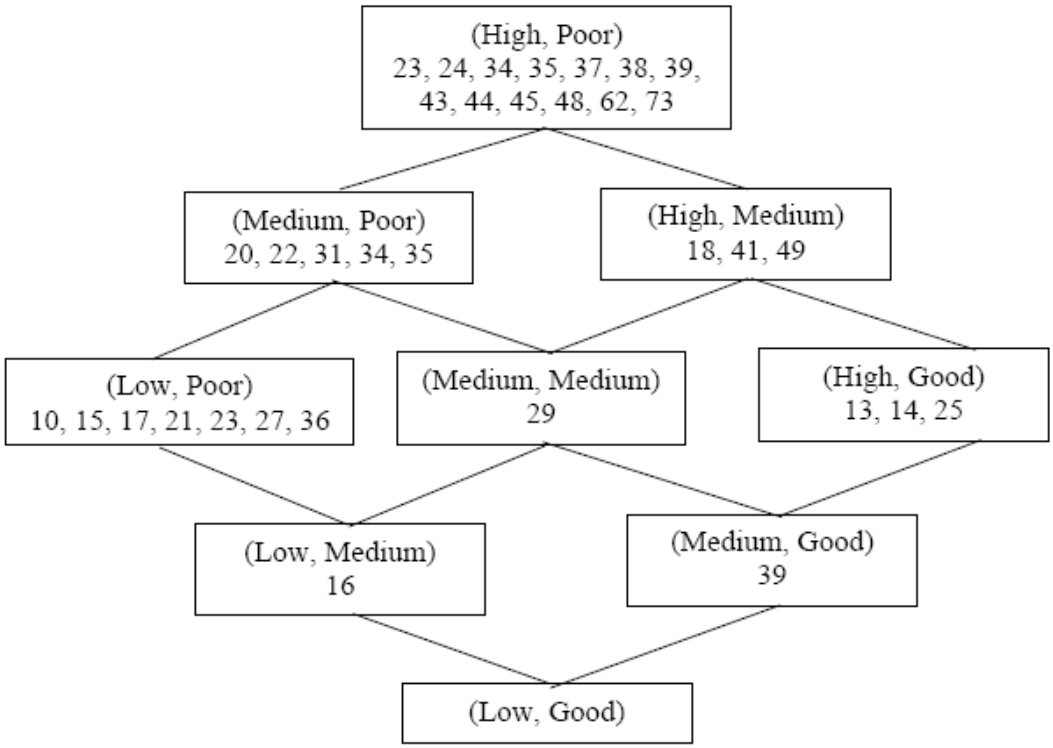

Figure 7: The Partial Order of Exposure Levels and Hygiene Levels of the Lead Data (Morton, 1982)

Rosenbaum (1991) proposed a linear rank statistics to measure agreement between the children's blood lead levels and the partial ordering of the categories

given by the parent's exposure to lead and their hygiene levels. This statistics is explained in Section 2.5. Using the central limit theorem, we conclude an approximate one-sided significance level as 0.0007.

Using the proposed random-linear-extension procedure, we can randomly generate 1000 complete orderings which satisfy the observed partial order. For each of the generated complete orderings, the Spearman's rank correlation, between the children's blood lead levels and the complete ordering of the categories given by the parent's exposure to lead and the their hygiene levels, was calculated and tested. The resulting distribution of the Fuzzy p-values is shown in Figure 8. Comparing with the linear rank test proposed by Rosenbaum, 86% of the Fuzzy p-values are smaller than 0.0007, which is the resulted p-value from the linear rank test.

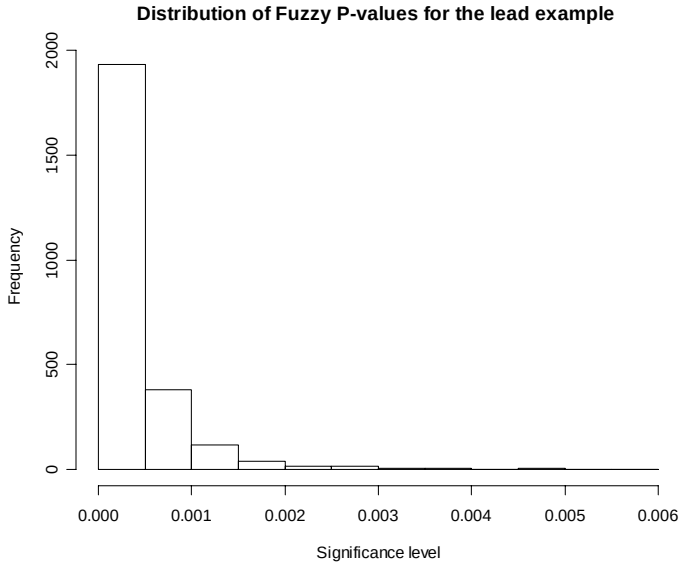

Figure 8: The PDF of the Fuzzy P-values Generated by the Random-linear-extension Test Based on Spearman's Rank Correlation for the Lead Data

CHAPTER 5

SPECIAL DESIGNS

For the tennis tournament example described in Chapter 1, where two groups are matched by pairs, the random-linear-extension test proposed in Chapter 4 can be applied. The ranks of all the players from both groups can be considered as a partial ordering, where only the ordering within each group and the ordering between pairs are available. The underline assumption of this problem would be that the complete ordering of all the players is known. When the relative relationship within each group is available, a matching by rank is appealing. Three of the many available matching schemes, including random matching, positive ordered matching and reverse ordered matching, are evaluated and compared in the following sections.

Random matching, positive ordered matching and reverse ordered matching are described in Sections 5.1, 5.2 and 5.3 respectively. The expected numbers of comparisons that prefer one group over the other for these three matching schemes are compared in Section 5.4. In Section 5.5, the power from the proposed random-linear-extension test based on the Wilcoxon test described in Chapter 4, under those three matching schemes is estimated by simulation. The results are compared with the power from the randomized sign test. A different conclusion about the power of various matching schemes is drawn at the end of this chapter and compared to the

results by Liu and David (1993). The results described in this chapter provide some surprising insights into the statistical information in partial orderings.

To simplify the problem, I assume equal number of observations within groups in the following sections. The hypothesis to test is,

$$H_0 : F_x(x) = F_y(x) \text{ versus } H_1 : F_x(x) = F_y(x+\theta),\ \theta > 0.$$

5.1 Random Case

Consider two known complete orders $x_{(1)} < x_{(2)} \cdots x_{(n-1)} < x_{(n)}$ and $y_{(1)} < y_{(2)} \cdots y_{(n-1)} < y_{(n)}$. The random matching scheme compares $x_{(i)}$ with $y_{(r_i)},\ i = 1, 2, \cdots, n,$ where $r = (r_1, r_2, \ldots, r_n)$ is a random permutation of $(1, 2, \ldots, n)$. This matching scheme can be shown as a directed graph in Figure 9.

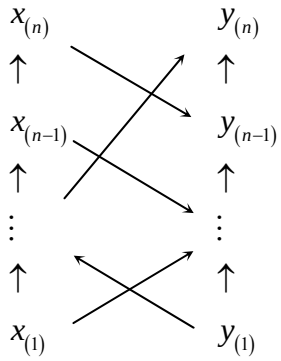

Figure 9: Directed Graph for the Random Case

As n comparisons between $x_{(i)}$ and $y_{(r_i)},\ i = 1, 2, \cdots, n,$ are independent from each other, this test is a standard binomial test. Therefore, under the null hypothesis, the probability that k different $x_{(i)}$ less than $y_{(r_i)}$ equals to $\binom{n}{k}\frac{1}{2}^n$.

5.2 Positive Ordered Case

Consider two known complete orders $x_{(1)} < x_{(2)} \cdots x_{(n-1)} < x_{(n)}$ and $y_{(1)} < y_{(2)} \cdots y_{(n-1)} < y_{(n)}$. The positive matching scheme compares the observations with the same ranks from each group, which is to compare $x_{(i)}$ with $y_{(i)},\ i = 1, 2, \ldots, n$. An example of this matching scheme can be shown as a directed graph in Figure 10.

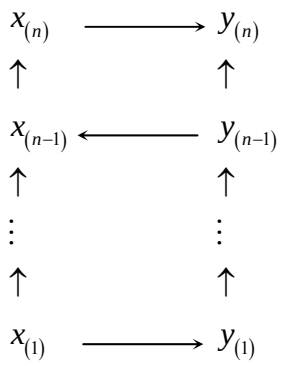

Figure 10: Directed Graph for the Positive Ordered Case

This distribution of k different $x_{(i)}$ lower than $y_{(i)}$ can be derived using the Result 5.1 (Feller, 1950).

Result 5.1 *(W. Feller)* Let $L_{2k,2n}$ be the number of paths from the origin to the point $2n$ of the x-axis such that $2k$ of its sides lie above the x-axis and $2n-2k$ below, and L_{2n} be the number of paths from the origin to the point $2n$ of the x-axis such that all $2n$ of its sides lie above the x-axis, $(k = 1, 2, \cdots, n)$. Then $L_{2k,2n} = L_{2n}$, independently of *k*.

The above result is an extension of the classic ballot theorem result of Bertrand, 1887. Feller's proof depends on a direct application of the reflection principle. Theorem 5.1 below is proved by setting up a straightforward correspondence between Feller's result and the distribution of the random variable that counts the number of times $x_{(i)}$ is less than $y_{(i)}$.

Theorem 5.1 Let $x_{(i)}$ and $y_{(i)}$ denote the *i-th* ordered observations within each group. Under the null hypothesis that $F_x(x) = F_y(x)$ the probability that there are exactly *k* pairs where $x_{(i)} < y_{(i)}$ is $1/(n+1)$, where $k = 0, 1, \cdots n$.

Proof. Let $w_{(j)}$, $j = 1, 2, \cdots 2n$ denote the order statistics for the combined values from group X and group Y. Let $\varepsilon_j = +1$ if the observation in *j-th* position is from group X and $\varepsilon_j = -1$ if the observation in the *j-th* position is from group Y. Set $S_0 = 0$ and $S_j = \sum_{i=1}^{j} \varepsilon_1$, for $j = 1, 2, \cdots, 2n$, then the sequence S_j, for $j = 0, 1, \cdots, 2n$ corresponds to the random paths generated by randomly drawing, without replacement, "ballots" from a ballot box containing exactly n

ballots labeled X and n ballots labeled Y. Each possible outcome of comparing $x_{(i)}'s$ with $y_{(i)}'s$, is represented by a path which connects the points $\left(j, S_j \right)$, for $j = 0, 1, \cdots 2n$. The edges of the path start at the origin (0, 0) and end at the point $\left(2n, 0\right)$.

We need to show that exactly $2k$ edges above the x-axis corresponds to $x_{(i)} < y_{(i)}$ holding for exactly k subscripts. Let m_i, $i = 1, 2, \ldots n$ denote the position of the order statistics $x_{(i)}$ in the combined sequence $w_{(j)}$, $j = 1, 2 \cdots 2n$; that is $w_{(m_i)} = x_{(i)}$ for $i = 1, 2, \cdots n$. Now if $S_{m_i} > 0$ then $x_{(i)} < y_{(i)}$; this is because there are more observations from group X than from group Y in the combined ranked sequence, w_j, up to and including the position m_i, where $w_{m_i} = x_{(i)}$. A similar argument shows that if $S_{m_i} \leq 0$ then $x_{(i)} > y_{(i)}$; this is because there must have been at least i draws from group Y before the m_i-th draw. We have just shown that $S_{m_i} > 0$ if and only if $x_{(i)} < y_{(i)}$. Every point $(m_i,\ S_{m_i})$, where $S_{m_i} > 0$ creates two edges above the x-axis, one edge going up from $(m_i - 1,\ S_{m_i} - 1)$ to $(m_i,\ S_{m_i})$ which must be balance by a succeeding edge going down, since $S_{2n} = 0$.

From Result 5.1, the probability of $2k$ sides above the x-axis is $1 / (n + 1)$, which is equivalent to the probability of k different $x_{(i)}$ lower than $y_{(i)}$ equals to $1 / \left(n + 1\right)$. This completes the proof.

5.3 Reverse Ordered Case

Consider two known complete orders $x_{(1)} < x_{(2)} \cdots x_{(n-1)} < x_{(n)}$ and $y_{(1)} < y_{(2)} \cdots y_{(n-1)} < y_{(n)}$. The reverse matching scheme is to compare $x_{(i)}$ with $y_{(n-i)},\ i = 1, 2, \cdots n$. This matching scheme can be shown as a directed graph in Figure 11.

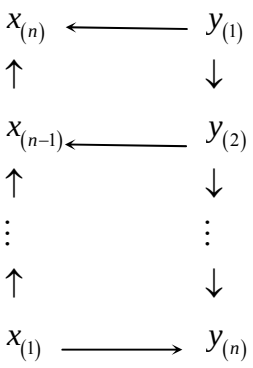

Figure 11: Directed Graph for Reverse Ordered Case

Let

$$Z_k = \begin{cases} 1 & \text{if } x_{(k)} > y_{(n+1-k)} \\ 0 & \text{if } x_{(k)} < y_{(n+1-k)} \end{cases}$$

It is clear that the sequence $z = z_1,\ z_2, \ldots, z_n$ has at most one transition from 0 to 1. Let K denote the position of the last zero in the sequence, then we have

$z = (1, 1, \ldots, 1) \rightarrow K = 0$
$z = (0, 1, \ldots, 1) \rightarrow K = 1$
$z = (0, 0, 1, \ldots, 1) \rightarrow K = 2$
...
$z = (0, 0, \ldots, 0) \rightarrow K = n.$

Theorem 5.2 The distribution of *K* follows the a hypergeometric distribution with

$$P\left(K = k\right) = \frac{\binom{n}{k}\binom{n}{n-k}}{\binom{2n}{n}}. \qquad (5.1)$$

where *K* denotes the largest *k* satisfies $x_{(k)} < y_{(n+1-k)}$.

Proof. The relationship between the two groups can be simplified as,

$$\begin{array}{ccc} x_{(n)} & & y_{(1)} \\ \vdots & > & \vdots \\ x_{(k+1)} & & y_{(n-k)} \\ \vee & & \wedge \\ x_{(k)} & & y_{(n-k+1)} \\ \vdots & < & \vdots \\ x_{(1)} & & y_{(n)} \end{array}.$$

For $p = k+1, \ldots, n$ and $q = 1, \ldots, k$, we have $x_{(p)} > y_{(q)}$ and $x_{(p)} > x_{(q)}$, also $y_{(p)} > x_{(q)}$ and $y_{(p)} > y_{(q)}$. Thus, $x_{(p)}$, $y_{(q)}$ are greater then $x_{(q)}$, $y_{(p)}$ for $\forall p \in \left(k+1, \ldots, n\right)$ and $\forall q \in \left(1, \ldots, k\right)$. In this way, we are able to identify that *k* of the largest n observations are from group Y, and *n-k* of the largest n observations are from group X. Under the null hypothesis, we can conclude that *K* follows a special hypergeometric distribution with

$$P\left(K = k\right) = \frac{\binom{n}{k}\binom{n}{n-k}}{\binom{2n}{n}}.$$

Lemma 5.1 Comparing two known complete orders $x_{(1)} < x_{(2)} \cdots x_{(n-1)} < x_{(n)}$ and $y_{(1)} < y_{(2)} \cdots y_{(n-1)} < y_{(n)}$ with reverse matching scheme is a special case of Brown-Mood median test (Brown and Mood, 1951).

Proof. In the Brown-Mood median test, let M denote the median of the pooled samples, which is any number between *n* and *n+1*. Let K_M be the number of observations in the group X, which are smaller than *M*. Since there are n observations greater than M combining X and Y groups, $n - K_M$ of those observations are from the group Y. Then the data can be presented as follows:

	X	*Y*	*Totals*
$< M$	K_M	$n- K_M$	*n*
$\geq M$	$n- K_M$	K_M	*n*
Totals	*N*	*n*	*2n*

In the reverse matching scheme, if *K* denotes the largest *k* satisfies $x_{(k)} < y_{(n+1-k)}$, the median of the pooled samples is any number between $x_{(k)}$ and $y_{(n+1-k)}$. Thus K is equivalent to K_M, and the distribution also follows the hypergeometric distribution in (5.1).

5.4 Random Case Compared with Positive Ordered and Reverse Ordered Cases

A similar situation was described by Liu and David (1993), which discuss the superiority of positive ordered matching compared to random matching and symmetric matching.

Definition 5.1 *(Liu and David)* $\pi = (\pi_1, \ldots, \pi_n)$ is said to be a symmetric permutation if $\pi_{n+1-i} = n + 1 - \pi_i$ for $i = 1, \ldots, n$.

The total numbers of comparisons prefer Y can be written as

$S(\pi) = \sum_{i=1}^{n} I\left(Y_{(i)} > X_{(\pi_i)}\right)$ under random matching;

$S(\pi^0) = \sum_{i=1}^{n} I\left(Y_{(i)} > X_{(i)}\right)$ under positive ordered matching;

$S(\pi^s) = \sum_{i=1}^{n} I\left(Y_{(i)} > X_{(\pi_i^s)}\right)$ under symmetric ordered matching,

where $\pi = (\pi_1, \ldots, \pi_n)$ is a permutation of $(1, \ldots, n)$, $\pi^0 = (1, \ldots, n)$, and π^s is a symmetric permutation.

Let $p_{i,j} = P(Y_{(i)} > X_{(j)})$. Then we have the expected value of $S(\pi)$, when π is randomly given, as

$$E\left(S(\pi)\right) = \frac{1}{n!}\sum_{\pi}\sum_{i=1}^{n} P\left(Y_{(i)} > X_{(\pi_i)}\right)$$
$$= \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} P\left(Y_{(i)} > X_{(j)}\right) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} p_{i,j}.$$

For the ordered matching, $E\left(S\left(\pi^0\right)\right) = \sum_{i=1}^{n} p_{i,i}$.

Result 5.2 *(Liu and David)* Let $\left(X_{(1)},\ldots,X_{(n)}\right)$ and $\left(Y_{(1)},\ldots,Y_{(n)}\right)$ be ordered statistics with cdf's $F\left(x\right)$ and $F\left(x+\theta\right)$, where $\theta \le 0$. Then

$$p_{i,i} + p_{j,j} \ge p_{i,j} + p_{j,i} \tag{5.2}$$

for any $1 \le i,\ j \le n$.

The proof of the above results is given by Liu and David (1993).

By Result 5.2, for any $\pi = \left(\pi_1,\ldots,\pi_n\right)$,

$$E\left(S\left(\pi^0\right)\right) = \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}(p_{i,i} + p_{j,j})$$
$$\ge \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}(p_{i,j} + p_{j,i}) = \frac{1}{n}\sum_{j=1}^{n}\sum_{i=1}^{n} p_{i,j} = E\left(S\left(\pi\right)\right).$$

Thus, the expected total number of comparisons that prefer Y under random matching is lower than the expected total number of comparisons that prefer Y under positive ordered matching.

Now we would like to compare the expected total number of comparisons that prefer Y, $E\left(S\left(\pi^s\right)\right) = \sum_{i=1}^{n} p_{i,\pi_i^s}$, for any symmetric permutations π^s, to the expected total number of comparisons that prefer Y, $E\left(S\left(\pi^0\right)\right) = \sum_{i=1}^{n} p_{i,i}$, under positive ordered matching.

Result 5.3 *(Liu and H. A. David)* Let $\left(X_{(1)},\ldots,X_{(n)}\right)$ and $\left(Y_{(1)},\ldots,Y_{(n)}\right)$ be the ordered statistics with cdf's $F\left(x\right)$ and $F\left(x+\theta\right)$, where $\theta \le 0$. Then

$$\sum_{i=1}^{n} p_{i,i} \ge \sum_{i=1}^{n} p_{i,\pi_i^s}$$

for any symmetric permutation π^s.

The proof of the above results is explained by Liu and David (1993).

Hence, the expected total number of comparisons that prefer Y under any symmetric permutation matching is lower than the expected total number of comparisons that prefer Y under ordered matching. A special case of symmetric permutation matching is reverse ordered matching, which is described in Section 5.3.

Theorem 5.3 *L*et $\left(X_{(1)},\ldots,X_{(n)}\right)$ and $\left(Y_{(1)},\ldots,Y_{(n)}\right)$ be the ordered statistics with cdf's $F\left(x\right)$ and $F\left(x+\theta\right)$, where $\theta \le 0$. Then

$$p_{i,(n+1-i)} + p_{j,(n+1-j)} \le p_{i,(n+1-j)} + p_{j,(n+1-i)} \tag{5.2}$$

for any $1 \le i,\ j \le n$.

Proof. Consider a four-dimensional function

$$D\left(x_{(n+1-i)},x_{(n+1-j)},y_{(i)},y_{(j)}\right) = I\left(y_{(i)} > x_{(n+1-i)}\right) + I\left(y_{(j)} > x_{(n+1-j)}\right) - I\left(y_{(i)} > x_{(n+1-j)}\right) - I\left(y_{(j)} > x_{(n+1-i)}\right)$$

where $x_{(n+1-i)} > x_{(n+1-j)}$ and $y_{(i)} < y_{(j)}$. Then all the possible orders of $x_{(n+1-i)}$, $x_{(n+1-j)}$, $y_{(i)}$, $y_{(j)}$ are:

(a) $x_{(n+1-j)} < x_{(n+1-i)} < y_{(i)} < y_{(j)}$ (b) $x_{(n+1-j)} < y_{(i)} < x_{(n+1-i)} < y_{(j)}$

(c) $x_{(n+1-j)} < y_{(i)} < y_{(j)} < x_{(n+1-i)}$ (d) $y_{(i)} < y_{(j)} < x_{(n+1-j)} < x_{(n+1-i)}$

(e) $y_{(i)} < x_{(n+1-j)} < y_{(j)} < x_{(n+1-i)}$ (f) $y_{(i)} < x_{(n+1-j)} < x_{(n+1-i)} < y_{(j)}$

By implementing the possible orders to $D\left(x_{(n+1-i),}x_{(n+1-j),}y_{(i),}y_{(j)}\right)$, we have:

$$D\left(x_{(n+1-i),}x_{(n+1-j),}y_{(i),}y_{(j)}\right) = \begin{cases} 1 & (e) \\ 0 & (a),\ (c),\ (d),\ (f). \\ -1 & (b) \end{cases}$$

Therefore, since $Y = X - \theta$ and $\theta \le 0$,

$$\begin{aligned} & p_{i,(n+1-i)} + p_{j,(n+1-j)} - p_{i,(n+1-j)} - p_{j,(n+1-i)} \\ &= E\left(D\left(x_{(n+1-i)}, x_{(n+1-j)}, y_{(i)}, y_{(j)}\right)\right) \\ &= P\left(Y_{(i)} < X_{(n+1-j)} < Y_{(j)} < X_{(n+1-i)}\right) - P\left(X_{(n+1-j)} < Y_{(i)} < X_{(n+1-i)} < Y_{(j)}\right) \\ &\le 0. \end{aligned}$$

By Theorem 5.3, for any random order $\pi = \left(\pi_1, \ldots, \pi_n\right)$, and the reverse order where $\pi^r = \left(n, n-1, ..., 1\right)$,

$$\begin{aligned} E\left(S\left(\pi^r\right)\right) &= \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(p_{i,(n+1-i)} + p_{j,(n+1-j)}\right) \\ &\le \frac{1}{2n}\sum_{j=1}^{n}\sum_{i=1}^{n}\left(p_{i,(n+1-j)} + p_{j,(n+1-i)}\right) = \frac{1}{n}\sum_{s=1}^{n}\sum_{t=1}^{n} p_{s,t} = E\left(S\left(\pi\right)\right). \end{aligned}$$

Thus, the expected total number of comparisons that prefer Y under random matching is larger than the expected total number of comparisons that prefer Y under reverse ordered matching.

To verify the above results and Theorem 5.1, the expected total numbers of comparisons that prefer Y under random, ordered and reverse matching schemes are

simulated and listed in Table 1. The simulated results are consistent with Theorem 5.1, Results 5.2 and 5.3.

Table 1: The Expected Total Numbers of Comparisons that Prefer Y under Random Ordered and Reverse Matching Schemes

Matching	*N=20,* $p = \Pr(X < Y)$				*N=40,* $p = \Pr(X < Y)$			
	0.50	***0.60***	***0.70***	***0.80***	***0.50***	***0.60***	***0.70***	***0.80***
Reverse	10.0	11.5	12.9	14.5	20.0	23.0	25.8	29.0
Random	10.1	12.1	13.9	15.9	20.1	24.1	27.9	32.0
Ordered	10.0	15.9	18.9	19.8	19.9	34.7	39.3	39.9

Although Liu and David (1993) use Results 5.1 and 5.2 to claim that ordered matching scheme was more powerful than the random matching scheme. The distributions of the test statistics explained in Sections 5.1, 5.2 and 5.3 are not the same under different matching schemes. Thus, the power of a test can not be evaluated simply by comparing the expected values of test statistics. Thus, the exact and simulated power of sign test under the random matching scheme, positive ordered matching scheme and reverse ordered matching scheme are calculated in the next section.

5.5 Power Comparisons with Randomized Sign Test

5.5.1 Random Case

The exact power of the one-sided randomized sign test for level of significance α, is given by:

$$\lambda(p) = \sum_{j=0}^{i} \binom{n}{j} p^j (1-p)^{n-j} + \gamma \binom{n}{i+1} p^{i+1} (1-p)^{n-i-1}$$

where i is the largest integer such that $\sum_{j=0}^{i} \binom{n}{j} (1/2)^n \le \alpha$. Here, p is the alternative population proportion.

The exact power of a one-sided randomized sign test at $\alpha = 0.05$ is shown in Table 2. Additionally, the power of the one-sided randomized sign test is simulated using the rejection region generated under the null hypothesis. The simulated power and 95% simulation interval are listed in Table 3. The values in those two tables are identical, which shows that the simulation algorithm is working correctly.

Table 2: Exact Power Table for One-sided Randomized Sign Test With $\alpha = 0.05$

	$p = \Pr(X < Y)$					
N	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.050	0.091	0.154	0.358	0.646	0.909
20	0.050	0.114	0.224	0.568	0.891	0.996
40	0.050	0.154	0.350	0.828	0.993	>0.999

Table 3: Simulated Power Table for Random Case Using One-sided Randomized Sign Test

	$p = \Pr(X < Y)$					
N	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.052 (0.038, 0.066)	0.100 (0.081, 0.119)	0.165 (0.142, 0.188)	0.374 (0.344, 0.404)	0.657 (0.628, 0.686)	0.921 (0.904, 0.938)
20	0.050 (0.036, 0.064)	0.128 (0.107, 0.149)	0.251 (0.224, 0.278)	0.562 (0.531, 0.593)	0.900 (0.881, 0.919)	0.996 (0.992, 1.000)
40	0.049 (0.036, 0.062)	0.159 (0.136, 0.182)	0.365 (0.335, 0.395)	0.848 (0.826, 0.870)	0.992 (0.986, 0.998)	>0.999

The simulated power of the proposed random-linear-extension test based on the Wilcoxon test and the corresponding 95% simulation interval are shown in Table 4. Additionally, the simulated power of the proposed empirical random-linear-extension test and the corresponding 95% simulation interval are shown in Table 5.

Table 4: Simulated Power Table for Random Case Using Wilcoxon-based Random-Linear-Extension Test

	$p = \Pr(X < Y)$					
N	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.051 (0.037, 0.065)	0.094 (0.076, 0.112)	0.152 (0.130, 0.174)	0.348 (0.318, 0.378)	0.627 (0.597, 0.657)	0.884 (0.864, 0.904)
20	0.053 (0.039, 0.067)	0.128 (0.107, 0.149)	0.257 (0.230, 0.284)	0.585 (0.554, 0.616)	0.902 (0.884, 0.920)	0.995 (0.991, 0.999)
40	0.044 (0.031, 0.057)	0.161 (0.138, 0.184)	0.394 (0.364, 0.424)	0.878 (0.858, 0.898)	0.997 (0.994, 0.100)	>0.999

Table 5: Simulated Power Table for Random Case Based on the Empirical Distribution of the Testing Statistics

	$p = \Pr(X < Y)$					
N	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.049 (0.036, 0.062)	0.105 (0.086, 0.124)	0.176 (0.152, 0.200)	0.408 (0.378, 0.438)	0.705 (0.677, 0.733)	0.941 (0.926, 0.956)
20	0.048 (0.035, 0.061)	0.131 (0.110, 0.152)	0.275 (0.247, 0.303)	0.635 (0.605, 0.665)	0.932 (0.916, 0.948)	0.999 (0.997, 1.000)
40	0.042 (0.030, 0.054)	0.173 (0.150, 0.196)	0.418 (0.387, 0.449)	0.919 (0.902, 0.936)	0.999 (0.997, 1.000)	>0.999

Based on the results from Tables 2, 4 and 5, the proposed random-linear-extension test based on the Wilcoxon test was shown to have about the same power compared to the randomized sign test, while the proposed empirical random-linear-extension test was shown to have about the same or more power compared to the randomized sign test,

5.5.2 Positive Ordered Case

For the positive ordered case, under the null hypothesis, the probability of k different $x_{(i)}$ lower then $y_{(i)}$ equals $1/(n+1)$. The power of the one-sided positive ordered sign test is simulated using the rejection region generated under the null hypothesis. The simulated power and corresponding 95% simulation interval are listed in Table 6.

Table 6: Simulated Power Table for Positive Ordered Case Based on One-sided Positive Ordered Sign Test

	$p = \Pr(X < Y)$					
n	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.052 (0.038, 0.066)	0.097 (0.079, 0.115)	0.150 (0.128, 0.172)	0.297 (0.269, 0.325)	0.446 (0.415 0.477)	0.535 (0.504, 0.566)
20	0.043 (0.030, 0.056)	0.130 (0.109, 0.151)	0.259 (0.232, 0.286)	0.603 (0.573, 0.633)	0.870 (0.849, 0.891)	0.985 (0.977, 0.993)
40	0.047 (0.034, 0.060)	0.181 (0.157, 0.205)	0.407 (0.377, 0.437)	0.864 (0.843, 0.885)	0.988 (0.981, 0.995)	>0.999

The simulated power of the proposed random-linear-extension test based on the Wilcoxon test and corresponding 95% simulation interval are shown in Table 7. Additionally, the simulated power of the proposed empirical random-linear-extension test and corresponding 95% simulation interval are shown in Table 8.

Table 7: Simulated Power Table for Positive Ordered Case Using Wilcoxon-based Random-linear-extension Test

	$p = \Pr(X < Y)$					
n	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.052 (0.038, 0.066)	0.089 (0.071, 0.107)	0.134 (0.113, 0.155)	0.243 (0.216, 0.270)	0.342 (0.313, 0.371)	0.388 (0.358, 0.418)
20	0.049 (0.036, 0.062)	0.115 (0.095, 0.135)	0.198 (0.173, 0.223)	0.376 (0.346, 0.406)	0.476 (0.445, 0.507)	0.499 (0.468, 0.530)
40	0.049 (0.036, 0.062)	0.155 (0.133, 0.177)	0.304 (0.275, 0.333)	0.537 (0.506, 0.568)	0.585 (0.554, 0.616)	0.593 (0.563, 0.623)

Table 8: Simulated Power Table for Positive Ordered Case Based on Empirical Distribution of the Testing Statistics

	$p = \Pr(X < Y)$					
n	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.061 (0.046, 0.076)	0.103 (0.084, 0.122)	0.160 (0.137, 0.183)	0.338 (0.309, 0.367)	0.460 (0.429, 0.491)	0.563 (0.532, 0.594)
20	0.043 (0.030, 0.056)	0.139 (0.118, 0.160)	0.273 (0.245, 0.301)	0.616 (0.586, 0.646)	0.891 (0.872, 0.910)	0.988 (0.981, 0.995)
40	0.057 (0.043, 0.071)	0.196 (0.171, 0.221)	0.435 (0.404, 0.466)	0.891 (0.872, 0.910)	0.995 (0.991, 0.999)	>0.999

Based on the results from Tables 6, 7 and 8, the proposed random-linear-extension test has significantly lower power compared to the positive ordered sign test. On the other hand, the proposed empirical random-linear-extension test has about the same power compared to the positive ordered sign test.

5.5.3 Reverse Ordered Case

For the reverse ordered case, under the null hypothesis, the probability of k different $x_{(i)}$ lower than $y_{(i)}$ follows Hypergeometric distribution. The power of the one-sided reverse ordered sign test are simulated using the rejection region generated under the null hypothesis. The simulated power and 95% simulation interval are listed in Table 9.

Table 9: Simulated Power Table for Reverse Ordered Case Based on One-sided Reverse Ordered Sign Test

	$p = \Pr(X < Y)$					
n	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.053 (0.039, 0.067)	0.088 (0.070, 0.106)	0.144 (0.122, 0.166)	0.347 (0.317, 0.377)	0.607 (0.577, 0.637)	0.891 (0.872, 0.910)
20	0.054 (0.040, 0.068)	0.117 (0.097, 0.137)	0.234 (0.208, 0.260)	0.586 (0.555, 0.617)	0.907 (0.889, 0.925)	0.996 (0.992, 1.00)
40	0.056 (0.042, 0.070)	0.144 (0.122, 0.166)	0.354 (0.324, 0.384)	0.830 (0.807, 0.853)	0.995 (0.991, 0.999)	>0.999

The simulated power of the proposed random-linear-extension test based on the Wilcoxon test and corresponding 95% simulation interval are shown in Table 10. Additionally, the simulated power of the empirical random-linear-extension test and corresponding 95% simulation interval are shown in Table 11.

Table 10: Simulated Power Table for Reverse Ordered Case Using Wilcoxon-based Random-linear-extension Test

	$p = \Pr(X < Y)$					
n	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.053 (0.039, 0.067)	0.085 (0.068, 0.102)	0.134 (0.113, 0.155)	0.313 (0.284, 0.342)	0.551 (0.520, 0.582)	0.846 (0.824, 0.868)
20	0.044 (0.031, 0.057)	0.108 (0.089, 0.127)	0.201 (0.176, .0.226)	0.488 (0.457, 0.519)	0.821 (0.797, 0.845)	0.985 (0.977, 0.993)
40	0.052 (0.038, 0.066)	0.131 (0.110, 0.152)	0.293 (0.265, 0.321)	0.741 (0.714, 0.768)	0.975 (0.965, 0.985)	>0.999

Table 11: Simulated Power Table for Reverse Ordered Case Based on Empirical Distribution of Testing Statistics

	$p = \Pr(X < Y)$					
n	***0.50***	***0.55***	***0.60***	***0.70***	***0.80***	***0.90***
10	0.057 (0.043, 0.071)	0.102 (0.083, 0.121)	0.159 (0.136, 0.182)	0.389 (0.359, 0.419)	0.661 (0.632, 0.690)	0.914 (0.897, 0.931)
20	0.044 (0.031, 0.057)	0.129 (0.108, 0.150)	0.255 (0.228, .0.282)	0.622 (0.592, 0.652)	0.928 (0.912, 0.944)	0.997 (0.994, 1.000)
40	0.047 (0.034, 0.060)	0.133 (0.112, 0.154)	0.321 (0.292, 0.350)	0.833 (0.810, 0.856)	0.995 (0.991, 0.999)	>0.999

Based on the results from Table 9, 10 and 11, the proposed random-linear-extension test has about the same or less power when compared to the reverse ordered sign test. On the other hand, the proposed empirical random-linear-extension test has about the same power when compared to the reverse ordered sign test.

5.5.4 Compare the Random, Ordered and Reverse Ordered Cases

By comparing Tables 2, 6, and 9, we see that the power achieved by using the randomized sign test is approximately the same between the random case and the reverse ordered case. Under the positive ordered case, when the sample size is large, the power reached is approximately the same or a little bit more comparing to the other two matching schemes. However, when the sample size is small, the power reached is significantly lower than the other two cases using the same method.

The above conclusions differ from the conclusions drawn by Liu and David (1993): “Ordered matching has more power to identify the stronger group than random matching”. This is due to the lack of consideration of the different test statistics distributions under various matching schemes. The power of testing could not be evaluated simply by the expected value of the number of comparisons that prefer Y. The distributions of the test statistics also need to be taken into consideration.

The goal of the random-linear-extension (RLE) test is to obtain an exact α level test, when one is not available. The RLE procedure achieves this goal. However, it is the case that this imputation procedure may reduce power when compared with other non-randomized procedures

CHAPTER 6

DISCUSSION

In previous chapters, I discussed the theory of randomized exact test based on partial ordering. Also, three special cases of partial orderings were analyzed using the proposed Wilcoxon random-linear-extension test and empirical random-linear extension test, and the power was compared with randomized sign tests by simulation. The simulation showed that the application of the proposed Wilcoxon random-linear-extension test on the random case and the reversed ordered case has approximately the same or a little less power compared to the randomized sign test, and significantly less power than the randomized positive ordered sign test on the positive ordered case. On the other hand, the empirical random-linear-extension test using the Wilcoxon test statistics has approximately the same or a little more power compared to the randomized sign test on the positive ordered case.

The general linear extension methodology can be applied in to other classic nonparametric methods, such as the Kruskal-Wallis test, which is a rank based ANOVA, and Spearman's rank correlation test (Hollander and Wolfe, 1999).

A test based on partial ordering has general appeal in medical studies where there are a number of numerical outcome measures for each patient. And we can

agree that Patient A has a better response than Patient B, if all numerical measures are greater on Patient A than on Patient B. A test based on this type of partial ordering would be a competitor to testing each outcome measure individually and then using an experimentwise multiple testing procedure, such as the Bonferroni procedure.

Similar application appears in environmental science and chemistry where a number of chemical indicators are used to assess overall environmental quality. A conventional approach is to assign a composite numeric score to each location or station by combining the indicator information, often in an *ad hoc* way. However, there are many cases where these indicators cannot logically be combinable into a single overall numerical index.

By introducing the partial ordering concept, we can evaluate the environmental samples by examining their relative positions in the indication space determined by their partial order. Using the proposed randomized-linear-extension procedure, we are able to study the rank of the objects that satisfies the partial order and derive the appropriate test based on their simulated complete orderings.

.

APPENDICES

APPENDIX A

Apply Available Methods to Combine Results from Independent Runs

In this paper, we applied two ways of combining results from independent runs. The first method is to obtain an exact α level test by "averaging across" statistical decisions (see Lemma 4.1); the second method is to apply the empirical distribution of the averaged test statistics.

In addition to these two methods, another way of combing results is to construct an asymptotic test on the averaged test statistics. Denote the observed binary orderings by $\mathcal{I}_{obs}$ and the missing binary orderings by $\mathcal{I}_{mis}$. Using the Gibbs Sampler described in Chapter 3, uniform draws can be generated from all possible complete orderings satisfying its known constraints. Once *D iid* imputations for missing values $\mathcal{I}_{mis}$ are simulated, the corresponding *D* complete orders can be constructed as

$$\left\{\mathcal{I}_*^{(l)};\ l = 1,...,D\right\} = \left\{\left(\mathcal{I}_{obs} \cup \mathcal{I}_{*mis}^{(l)}\right);\ l = 1,...,D\right\}.$$

Then the Mann-Whitney test statistics *U* for each generated complete ordering can be calculated as $U_{*l} = T_{*l} - n_1(n_1 + 1)/2$, where T_{*l} is the sum of ranks of the observations from the first group and $l = 1,\ldots,M$.

To combine the results from *D* independent runs, an asymptotic test of $\overline{U} = E\left(U \mid \mathcal{I}_{obs}\right)$ can be constructed as:

$$\frac{E\left(U \mid \mathcal{I}_{obs}\right) - E(U)}{\sqrt{Var\left(E\left(U \mid \mathcal{I}_{obs}\right)\right)}} \rightarrow N\left(0,1\right)$$

As $Var(U) = E\left(Var\left(U \mid \mathcal{I}_{obs}\right)\right) + Var\left(E\left(U \mid \mathcal{I}_{obs}\right)\right)$, the above formula can be rewritten as:

$$\frac{E\left(U \mid \mathcal{I}_{obs}\right) - E(U)}{\sqrt{Var(U) - E\left(Var\left(U \mid \mathcal{I}_{obs}\right)\right)}} \rightarrow N(0,1).$$

Here, $E(U)$ and $Var(U)$ are the known constant mean and variance of the Wilcoxon test statistics. Given a $\mathcal{I}_{obs}$, $\overline{U}_*$ is the simulated unbiased estimate of $E\left(U \mid \mathcal{I}_{obs}\right)$ and $Var\left(\overline{U}_*\right)$ is the best estimator of $Var\left(U \mid \mathcal{I}_{obs}\right)$. However, $Var\left(U \mid \mathcal{I}_{obs}\right)$ is correlated with $E\left(U \mid \mathcal{I}_{obs}\right)$. The relationships between $E\left(U \mid \mathcal{I}_{obs}\right)$ and $Var\left(U \mid \mathcal{I}_{obs}\right)$ for the random and ordered case are shown in Figure 12 and 13. Thus, an asymptotic test on $\overline{U}$ can not be constructed based on the observed partial orderings.

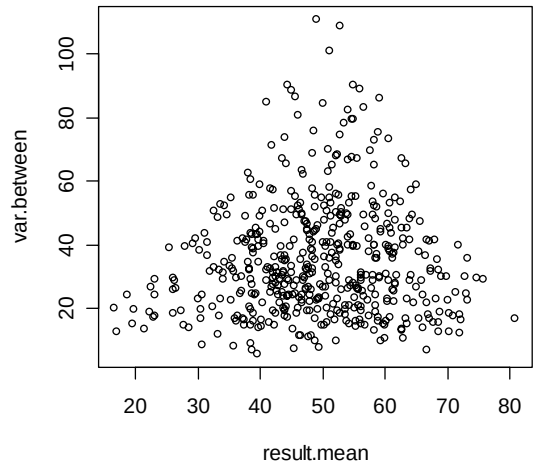

Figure 12: $E\left(U \mid \mathcal{I}_{obs}\right)$ **vs.** $Var\left(U \mid \mathcal{I}_{obs}\right)$ **for the Random Case**

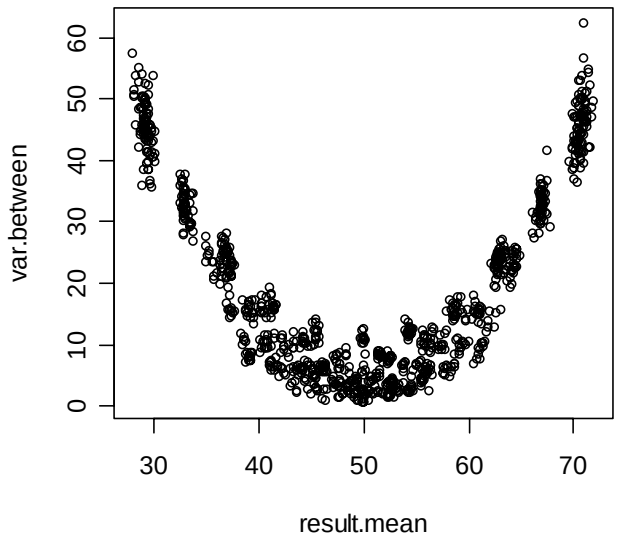

Figure 13: $E\left(U \mid \mathcal{I}_{obs}\right)$ **vs.** $Var\left(U \mid \mathcal{I}_{obs}\right)$ **for the Ordered Case**

APPENDIX B

Optimize the Program

Within the three special designs mentioned in Chapter 5, the Gibbs Sampler algorithm to generate the random-linear-extension test converges slowly under the ordered matching scheme.

Raftery and Lewis (1992) proposed a method to find the number of iterations needed to achieve a given precision in this case, based on an initial run. The method can be implemented by R CODA package. The method finds the number of iterations needed to estimate $P(U \leq u \mid data)$ within $\pm r$ with probability *s*, when the correct answer is *q* and *U* is the test statistics. It returns the number of burn-in iterations to be discarded, the total number of iterations required and the dependence factor, which measures the increase in the number of iterations due to dependence in the sequence.

Table 12 below includes the burn-in time, the dependence factors and the number of iteration required for three special cases, while $r = 0.01$, $q = 0.025$ and $s = 95\%$.

Table 12: Convergence Results Using the Raftery and Lewis's (1992) Method (r = 0.01, q = 0.025 and s = 95%)

	Random			*Ordered*			*Reverse*		
N per group	***Burn-in***	***Dep. factor***	***Total***	***Burn-in***	***Dep. factor***	***Total***	***Burn-in***	***Dep. factor***	***Total***
10	8	2.94	2754	8	4.52	4236	10	3.87	3236
20	16	5.16	4836	15	5.64	5280	20	6.05	5672
40	24	7.23	6771	27	8.98	8421	52	14.8	13880

To speed up the program, parts of the codes are written and compiled in Visual Basic C++, Then use R code to call the compiled dynamic link file. The C++ code and the flow of the R code are included in the following pages.

Visual C++ Codes for Creating the DLL file

```
// Test.cpp : Defines the exported functions for the DLL application.
//
#include "stdafx.h"
#include "stdlib.h"
#include "time.h"


void simumat(double *x, double rankmat[6][6], long *nsample)

{
double uinew;
double ulower;
double uuper;
srand(time(NULL));

for (int j = 0;j < (*nsample*2);j++)
{
double imin=0;
double imax=1;

for (int i = 0;i < (*nsample*2);i++)
{
uuper = x[i] * rankmat[i][j];

if ((uuper < imax) && (uuper!=0) )
imax=uuper;

ulower = x[i] * rankmat[j][i];

if (ulower > imin)
imin=ulower;

uinew = ((double)rand()/(double)RAND_MAX)* (imax-imin) + imin;

}

x[j] =uinew;
}
}

// Test.def
LIBRARY"CSimu"
EXPORTS
Simumat
```

Flow of R Codes for the Random-linear-extension Tests

1. Load the compiled Visual C++ code by the following R code:

```
dyn.load("Csimu.dll","simumat", "cdecl")
Test<-function(x,rankmat,u, nsample,niter,order) .C("simumat",
as.double(x),as.double(rankmat),as.double(u),as.integer(nsample,
as.integer(niter),as.integer(order))
```

2. Create a R function "General.case" with the following parameters:

```
n.sample is the sample size per group
n.sim.it is the number iterations
M.burn is the number if initial iterations to be discarded
k.thin is where every kth iterate is used
Qt.ran is the 95% quantile for the random case
Qt.ord is the 95% quantile for the ordered case
Qt.rev is the 95% quantile for the reverse case
```

3. Construct a matrix "rank.mat" to indicate the relationship between all pairs. If the relationship is known to be greater, place a "1" in the rank matrix, otherwise, place a "0".
4. Assign the initial ranks by rank the row sums of "rank.mat". If ties appear, assign the ranks randomly.
5. Simulate the ranks satisfying the rank matrix by calling the compiled Visual C++ codes.
6. Calculate the sum of simulated ranks for group1.
7. Discard the first M.burn simulated ranks sums and thin the simulated results by using every kth iterate.

8. Apply the Wilcoxon random-linear-extension test or the empirical randomized-linear-extension test on the simulated rank sums.

REFERENCES

1.Aldus, D. (1987). On the Markov Chain Simulation Method for Uniform Combinatorial Distributions and Simulated Annealing. *Probability in the Engineering and Informational Sciences,* 1, 33-46.

2.Amit, Y. and Grenander, U. (1991). Comparing Sweep Strategies for Stochastic Relaxation. *Journal of Multivariate Analysis*, 37, 197-222.

3.Andre, D. (1887). Solution directe du probl`eme resolu par M. Bertrand.*Comptes Rendus de l'Academie des Sciences*, Paris 105, 436–437.

4.Bertrand, J. (1887). Solution d'un probl`eme. *Comptes Rendus de l'Academie des Sciences*, Paris 105**,** 369.

5.Brown, G. W. and Mood, A. M. (1951). On Median Tests for Linear Hypotheses. *The Proceeding of the Second Berkeley Symposium on Mathematical Statistics and Probability,* 2, 159-166.

6.Casella, G. and George, E. I. (1992). Explaining the Gibbs Sampler. *The American Statistician*, 46, 167-174.

7.Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*, Chapman & Hall.

8.Cowles, M. K. and Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, 91, 883-904.

9.Deshpande, J. V. (1968). On Continuity of a Partial Order. *Proceedings of the American Mathematical Society*, 19, 383-386.

10.Dixon, W. J. and Massey, F. J. (1951). *An Introduction to Statistical Analysis.* McGraw-Hill Book Co.

11.Dixon, W. J. and Mood, A. M. (1946). The Statistical Sign Test. *Journal of the American Statistical Association*, 41, 557-566.

12.Feller, William (1950). *An Introduction to Probability Theory and Its Applications.* John Wiley & Sons, Inc.

13.Fisher, R. A. (1925). *Statistical Method for Research Workers*. Oliver and Boyed.

14.Fishman, G. S. (1996). Coordinate Selection Rules for Gibbs Sampling. *The Annals of Applied Probability*, 6, 444-465.

15.Gefeller, O. and Pralle, L. (1997). Nonrandomized Comparative Clinical Studies. *Proceedings of the International Conference on Nonrandomized Comparative Clinical Studies in Heidelberg*, April 10-11.

16.Gelfand, A.E. and Smith, A.F.M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. *Journal of the American Statistical Association*, 85, 398-409.

17.Geman, S. and Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6, 721-741.

18.Geweke, J. (1992). Evaluating the Accuracy of Sampling-based Approaches to Calculating Posterior Moments. *Bayesian Statistics* 4, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.). Clarendon Press, Oxford.

19.Geyer, C. J. (1992). Practical Markov Chain Monte Carlo (with discussion). *Statistical Science*, 7, 473–511.

20.Geyer, C. J., and Meeden, G.D. (2005). Fuzzy and Randomized Confidence Intervals and P-values. *Statistical Science*, 20, 358-366.

21.Geyer, C. J. (2005). Fuzzy P-values and Ties in Nonparametric Tests. (Submitted)

22.Geyer, C. J. and Thompson, E. A. (2007). Fuzzy P-values in Latent Variable Problems. *Biometrika*, 94, 49-60.

23.Heidelberger, P., and Welch, P. D. (1983). Simulation Run Length Control in the Presence of an Initial Transient. *Operations Research*, 31, 1109-1144.

24.Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods (2nd ed.)*. Wiley-Interscience.

25.Lehmann, E. L. (1998). *Nonparametric Statistical Methods Based on Ranks*. Prentice Hall.

26.Liu, Jingyu and David, H. A. (1993). Comparing Two Groups of Ranked Objects by Pairwise Matching. *Journal of Statistical Planning and Inference*, 35, 157-169.

27.Mathews, P. (1991). Generating a Random Linear Extension of a Partial Order. *The Annals of Probability,* 19, 1367-1392.
28.Meng, X.L. and Rubin, D.B. (1992). Performing Likelihood Ratio Tests with Multiple-imputed Data Sets. *Biometrika,* 79, 103-111.

29.Mood, A. M. (1954). On the Asymptotic Efficiency of Certain Nonparametric Two-Sample Tests. *The Annals of Mathematical Statistics*, 25, 514-522.

30.Morton, D., Saah, A., Silberg, S., Owens, W., Roberts, M., Saah, M. (1982). Lead Absorption in Children of Employees in a Lead Related Industry. *American Journal of Epidemiology,* 115, 549-555.

31.Patil, G. P. and Taillie, C. (2004). Multiple Indicators, Partially Ordered Sets, and Linear Extensions: Multi-criterion Ranking and Prioritization. *Environmental and Ecological Statistics*, 11, 199-228.

32.Pratt, Joseph (1955). The Treatment of Ties in Some Nonparametric Tests. *The Annals of Mathematical Statistics*, 26, 368-386.

33.Raftery, A. E., and Lewis, S. (1992). How Many Iterations in the Gibbs Sampler? *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (eds.), Oxford University Press.

34.Ritter, C., and Tanner, M. A. (1992). Facilitating the Gibbs Sampler: The Gibbs Stopper and the Griddy-Gibbs Sampler. *Journal of the American Statistical Association,* 87, 861-868.

35.Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer.

36.Roberts, G.O. and Polson, N. G. (1994). On the Geometric Convergence of the Gibbs Sampler. *Journal of the Royal Statistical Society, Series B (Methodological)*, 56, 377-384.

37.Roberts, G. O. and Sahu, S. K (1997). Updating Schemes, Correlation Structure, Blocking and Parameterization for the Gibbs Sampler. *Journal of the Royal Statistical Society, Series B (Methodological)*, 59, 291-317.

38.Rosenbaum, P. R. (1991). Some Poset Statistics. *The Annals of Statistics*, 19, 1091-1097.

39.Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons.

GLOSSARY

Complete order: A *complete order* is a set of pairwise restrictions satisfies the conditions for a partial order plus an addition condition known as the comparability condition:

- for any $i,\ j \in (1, ..., N)$, either $x_i < x_j$ or $x_j < x_i$.

Complete ordering: A completely ordered set is called a *complete ordering.*

Fuzzy Decision: A *fuzzy decision* is the reported value of $\phi(x)$, which is a random variable that has a probability distribution, while the classics randomized decision is a realization of the random variable.

Fuzzy P-value: If $\mathbb{P}$ is a random variable that has a uniform distribution between 0

and 1 unconditionally under the null hypothesis, then a random variable having the conditional distribution of $\mathbb{P}$ given observed data Y is an exact *fuzzy P-value*.

Linear Extension: A *linear extension I* of a partially ordered set *P* is a permutation of the elements $x_1,\ x_2, \ldots,\ x_N$, such that $x_i < x_j$ in *P* implies $x_i < x_j$ in *I*.

Partial Order: A *partial order* on a set $x = (x_1, x_2, \ldots, x_N)$ is a set of consistent pairwise restrictions $x_{i_k} < x_{j_k}$ for some collection of pairs $(i_1, j_1), \ldots, (i_K, j_K)$. It satisfies the transitivity condition:

- if $x_i < x_k$ and $x_k < x_j$ then $x_i < x_j$.

Partial Ordering / Poset: A partially ordered set is also called a *partial ordering* or *poset.*