Nonparametric tests for umbrella alternatives in stratified datasets

Josh Larock

Thesis submitted to the Faculty of Science in partial fulfillment of the requirements
for the degree of
Master of Science Mathematics and Statistics[1]

Department of Mathematics and Statistics
Faculty of Science
University of Ottawa

[1]The M.Sc. program is a joint program with Carleton University, administered by the Ottawa-Carleton Institute of Mathematics and Statistics

Abstract

This thesis considers the problem of hypothesis testing for umbrella alternatives when there are two groups, or strata, of observations. The proposed methods extend a previously established general framework of hypothesis testing based on rankings to stratified datasets by first aligning the strata. The tests based on the Spearman and Kendall distances between ranking vectors lead to the traditional aligned-rank tests and new methods which account for "misalignment" under the alternative hypothesis. Asymptotic null distributions and simulation studies are given for the Spearman distance. Diagnostic tools for the misalignment issue are illustrated alongside the proposed tests on a dataset of IQ scores of coma patients. Extensions to three or more strata and "adaptive" tests are provided as future research directions.

KEY WORDS: nonparametric tests; ranks; aligned-ranks; umbrella alternatives; stratified data; non-null tests; Spearman and Kendall distance.

Acknowledgement

I would first like to express my appreciation to my supervisor Professor Mayer Alvo for his guidance during the thesis process and for the many opportunities he has provided me over the past two years. Working with him in the summer of 2021 as an undergraduate student was pivotal in my decision to pursue graduate studies and a career involving methodology research. He has taught me countless lessons on statistics, research and life that I am forever grateful for.

I would also like to thank the department for providing a great academic environment with many inspirational and welcoming professors. I would like to personally thank the Director of the graduate program Professor Gilles Lamothe for his kindness towards me as a curious student in my first probability course and later for his patience and guidance throughout the graduate program.

Lastly, I want to express my gratitude to my friends and family, who have been a constant source of support throughout this journey. I am especially thankful for the many enriching conversations with my friends Liam and Mourad over the years. Above all, I am eternally indebted to my girlfriend Sophie, my father Jeff, and my grandfather Larry for their unwavering support and encouragement, without which this thesis would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis considers the problem of testing for umbrella alternatives when two groups, or strata, are involved. Umbrella patterns are present in some applications where there is a tendency for the response variable to increase in value up to a point and then to subsequently decrease. As examples, such orderings may arise when measuring the effectiveness of a drug over time, when judging the reaction to increasing dosage levels of a drug, or when studying some type of mental or physical performance across age groups.

In the unstratified case where there is only one group or stratum, the interest is in testing the hypothesis that the data exhibit an umbrella ordering. This problem was first considered by Mack and Wolfe (1981) and later by Simpson and Margolin (1986), Hettmansperger and Norton (1987) and Shi (1988). Chen and Wolfe (1990) adapted the Mack-Wolfe statistic to the case of unknown peak location. See also Chen (1991) and more recently Millen and Wolfe (2005), who introduced modifications and presented a simulation study. Kössler (2006) compared the test statistics due to Chen-Wolfe, Hettmansperger-Norton and Shi. Alvo (2008) considered the unstratified problem using a general hypothesis testing framework proposed by Alvo and Pan (1997).

Less research has been done on methods for stratified data. Pan and Wolfe (1995, 1996) considered the hypothesis that the location of the peaks is the same for the two strata, though there may be differences in magnitude between them. Both a parametric test which assumes the underlying populations are normally distributed (1996) and a nonparametric test based on ranks (1995) were proposed.

In this thesis, we consider two strata of subjects, each of which are assumed to be parallel under both the null hypothesis of no treatment effects and the alternative hypothesis of umbrella patterns. We extend the results provided in Alvo (2008) to develop nonparametric tests for umbrella alternatives with two strata.

For the first stratum, let $X^{(1)}_{i(1)}, ..., X^{(1)}_{i(m_{1i})}, i = 1, ..., k$, be independent random samples from the k treatment groups, with $X^{(1)}_{i(l)}, l = 1, ..., m_{1i}$ having an absolutely continuous distribution function $F_{1i}(x)$. Similarly, for the second stratum, let $X^{(2)}_{i(1)}, ..., X^{(2)}_{i(m_{2i})}, i = 1, ..., k$, be k independent random samples with $X^{(2)}_{i(l)}, l = 1, ..., m_{2i}$ having an absolutely continuous distribution function $F_{2i}(x)$. We denote the sample where the peak of the umbrella pattern is located by p, with $1 \leq p \leq k$.Additionally, let $n_{1i} = \sum_{h=1}^{i} m_{1h}$ and $n_{2i} = \sum_{h=1}^{i} m_{2h}$ for $i = 1, ..., k$, with $n_1 = n_{1k}, n_2 = n_{2k}$, $\tilde{n}_1 = n_1 - m_{1p}, \tilde{n}_2 = n_2 - m_{2p}$, and $n_{10} = n_{20} = 0$. We further assume that the samples for the two strata are independent of one another. Under the null hypothesis,

$$H_0 : F_{1i}(x) = F(x) \text{ for all } x, \tag{1.0.1}$$

$$F_{2i}(x) = F(x + \delta) \text{ for all } x, \tag{1.0.2}$$

where δ is the difference in magnitudes between the two strata. Under the alternative, each of the two strata is assumed to follow an umbrella ordering. Hence, we suppose

$$H_1 : F_{1i}(x) = F(x + \gamma_i) \text{ for all } x, \tag{1.0.3}$$

$$F_{2i}(x) = F(x + \gamma_i + \delta) \text{ for all } x, \tag{1.0.4}$$

where

$$\gamma_1 \leq ... \leq \gamma_{p-1} \leq \gamma_p \geq \gamma_{p+1} \geq ... \geq \gamma_k, \tag{1.0.5}$$

or alternatively

$$\gamma_1 \geq ... \geq \gamma_{p-1} \geq \gamma_p \leq \gamma_{p+1} \leq ... \leq \gamma_k, \tag{1.0.6}$$

with at least one strict inequality. We remark that $F_{1i}(x) = F_{2i}(x - \delta)$ for all $i = 1, ..., k$, an assumption of parallelism similar to what is often considered in the aligned-rank test discussed next and also multivariate profile analysis.

Our tests first "align" the two strata following the approach of Hodges and Lehmann (1962). The general approach for $s > 1$ strata is to compute for each stratum a measure of location for the pooled sample of all k treatment groups and subtract the measure from the same pooled sample. For stratum $i = 1, ..., s$, let $X^{(i)} = \{X_1^{(i)}, X_2^{(i)}, ..., X_k^{(i)}\}$ be the pooled sample across all treatment groups and θ_i be a scalar location estimate of $X^{(i)}$. Additionally, for treatment group $q = 1, 2, ..., k$ define the aligned sample for each treatment group from stratum $i = 1, ..., s$ as

$$\{X_q^{(i)} - \theta_i\} = \{X_{q(l)}^{(i)} - \theta_i,\ l = 1, ..., m_{iq}\}, \tag{1.0.7}$$

and define

$$X_q^* = \{X_q^{(i)} - \theta_i,\ i = 1, ..., s\} = \{X_{q(l)}^{(i)} - \theta_i,\ i = 1, ..., s,\ l = 1, ..., m_{iq}\}. \tag{1.0.8}$$

Rank tests for stratified experiments are typically constructed using either the within-stratum ranks or the "aligned-ranks" of the entire dataset (Mehrotra, Lu, and Li, 2010). We now illustrate these two contrasting approaches in their simplest forms for the stratified two-treatment problem with $s > 1$ strata. For $i = 1, ..., s$, let $X_1^{(i)} = \{X_{1(l)}^{(i)}, l = 1, ..., m_{i1}\}$ be the values from the first treatment group belonging to stratum i following the distribution F_{i1} and let $X_2^{(i)} = \{X_{2(l)}^{(i)}, l = 1, ..., m_{i2}\}$ be the values from the second treatment group belonging to stratum i, following the distribution F_{i2}. We wish to test that for all $i = 1, ..., s$,

$$H_0 : F_{i1}(x) \quad = \quad F_{i2}(x) \tag{1.0.9}$$

for all x under the null hypothesis against the alternative that for all $i = 1, ..., s$,

$$H_1 : F_{i1}(x) \quad = \quad F_{i2}(x + \gamma) \tag{1.0.10}$$

for all x with $\gamma \neq 0$ for the two-sided alternative hypothesis and either $\gamma > 0$ or $\gamma < 0$ for the one-sided alternative hypothesis.

The first approach using the within-stratum ranks was introduced by Wilcoxon (1946). The test statistic was given by

$$\sum_{i=1}^{s} U(X_1^{(i)}, X_2^{(i)}), \tag{1.0.11}$$

where

$$U(X_1^{(i)}, X_2^{(i)}) = \sum_{j=1}^{m_{i1}} \sum_{l=1}^{m_{i2}} sign(X_{1(j)}^{(i)} - X_{2(l)}^{(i)}) \tag{1.0.12}$$

is the Mann-Whitney-Wilcoxon statistic (Wilcoxon,1945, Mann and Whitney, 1947). A generalized class of tests

$$T_{vE} = \sum_{i=1}^{s} c_i U(X_1^{(i)}, X_2^{(i)}) \tag{1.0.13}$$

were studied by van Elteren (1959), which includes the test of Wilcoxon (1946) as a special case when the regression coefficients or weights $c_1, ..., c_s$ are equal. It was proved that under the assumption of equal treatment effects across strata, the weights resulting in the locally most powerful test are

$$c_i = (m_{i1} + m_{i2} + 1)^{-1}. \tag{1.0.14}$$

Furthermore, van Elteren (1959) showed that under the null hypothesis of no treatment effect, the test statistics given by (1.0.13) are asymptotically normally distributed with mean 0 and variance

$$Var(T_{vE}) = \frac{1}{3} \sum_{i=1}^{s} c_i^2 m_{i1} m_{i2} (m_{i1} + m_{i2} + 1). \tag{1.0.15}$$

Alternatively, Hodges and Lehmann (1962) proposed an aligned-rank test which is performed in two steps. The first step consists of the alignment procedure described above with $k = 2$ to obtain samples X_1^* and X_2^*. The second step consists of computing the Mann-Whitney-Wilcoxon test statistic using the entire dataset. The resulting test statistic is then

$$T_{align} = U(X_1^*, X_2^*), \tag{1.0.16}$$

with U defined as in (1.0.12).

The choice of stratum alignment estimators $\theta_i,\ i = 1, ..., s$, has been discussed in the literature. The one-sample Hodges-Lehmann estimator (Hodges and Lehmann, 1963) has been found to be robust for different underlying distributions (Mehrotra, Lu, and Li, 2010). This estimator for stratum $i = 1, ..., s$ can be computed by the median pairwise average of the pooled data within the stratum,

$$Median\{\frac{X_{(j)}^{(i)} + X_{(l)}^{(i)}}{2},\ 1 \leq j \leq l \leq n_i\}, \tag{1.0.17}$$

with $X_{(j)}^{(i)}$ and $X_{(l)}^{(i)}$ being the j^{th} and l^{th} elements of the pooled sample $X^{(i)}$. The origin of this estimator will be discussed in Chapter 5.

The structure of the thesis is as follows: In Chapter 2, we review the general theory of hypothesis testing based on rankings proposed by Alvo and Pan (1997) and extend it to aligned stratified samples with a specific focus on umbrella alternatives. We also provide adjustments to our methods which address situations where the alignment procedure fails to align the strata properly. Asymptotic null distributions are derived and simulation results are given in Chapters 3 and 4, respectively. In Chapter 5, we discuss nonparametric point estimators for a common treatment effect in the stratified two-treatment problem, which we show can also be used to assess the results of the alignment procedure. In Chapter 6, we illustrate the use of the methods proposed on a real dataset consisting of measurements of mental performance in coma patients. We conclude the thesis in Chapter 7 and provide a discussion on extending the methods to three or more strata and developing "adaptive" tests that remain robust to all possible misalignments.

Chapter 2

Aligned-rank tests through a general theory of hypothesis testing

2.1 A general theory of hypothesis testing based on ranks

Alvo and Pan (1997) proposed a general approach to hypothesis testing based on the ranks of the observations. The method consists of defining a set of permutations induced by the observations and an additional set of extremal permutations that are "most in agreement" with the alternative hypothesis. The test statistic is then constructed based on a measure of the distance between these two sets. This approach has led to both plausible new forms of tests as well as to existing tests that have previously been established in the literature. Notably, the use of both the Spearman and Kendall distances lead to the Mann-Whitney-Wilcoxon test when applied to the two-sample location problem. The theory can be used to construct tests based on a specified alternative hypothesis for a wide variety of problems. The resulting tests can then be analyzed using well known theoretical properties of rank-based tests. The general procedure, when there are a total of n observations, is as follows:

Let $\mathcal{P} = \{\mu : [\mu(1), ..., \mu(n)]\}$ be the set of all permutations of the integers $1, 2, ..., n$, and let $d(\mu, \nu)$ be a distance function between permutations μ and ν.

Step 1: Rank all the observations together so that the smallest gets rank 1, the next smallest rank 2 etc.. Let the $n-$dimensional vector π represent the ranks of the data. In view of the continuity assumption on the distributions, ties among the observations occur with probability zero.

Step 2: Define $\{\pi\}$ to be the subclass of permutations "equivalent" to the observable permutation π in the sense that ranks occupied by identically distributed random variables are exchangeable within each distribution separately.

Step 3: Define E to be an extremal subclass of $\mathcal{P}$ consisting of all permutations which are "most in agreement with H_1". The extremal set E does not necessarily correspond to the entire critical region but rather consists of those permutations which provide the strongest evidence in favour of the alternative.

Step 4: Let $d(\mu, \nu)$ be a distance function between two permutations μ, ν. Define the distance between $\{\pi\}$ and an extremal set E by computing the sum of all pairwise distances between them:

$$d(\{\pi\}, E) = \sum_{\mu \epsilon \{\pi\}} \sum_{\nu \epsilon E} d(\mu, \nu) \tag{2.1.1}$$

Small values of $d(\{\pi\}, E)$ are inconsistent with the null hypothesis and consequently lead to rejection of H_0. The construction described above is similar to that of Critchlow (1992), but uses the sum of the pairwise distances instead of the minimum distance between permutation sets in Step 4.

Alvo (2008) applied this general theory to umbrella alternatives, where there is an increase (decrease) until a certain treatment group, followed by a subsequent decrease (increase) across the remaining treatment groups. For the rest of this section, let $X_{i(1)}, ..., X_{i(m_i)}, i = 1, ..., k$, be k independent random samples with $X_{i(l)}, l = 1, ..., m_i$ having an absolutely continuous distribution function $F_i(x)$. Under the null hypothesis

$$H_0 : F_i(x) = F(x) \tag{2.1.2}$$

for all x. Under the alternative, the data is assumed to follow an umbrella pattern across treatment groups. Hence, it is supposed that

$$H_1 : F_i(x) = F(x + \gamma_i), \tag{2.1.3}$$

where depending on the context, either

$$\gamma_1 \leq ... \leq \gamma_{p-1} \leq \gamma_p \geq \gamma_{p+1} \geq ... \geq \gamma_k, \tag{2.1.4}$$

or

$$\gamma_1 \geq ... \geq \gamma_{p-1} \geq \gamma_p \leq \gamma_{p+1} \leq ... \leq \gamma_k, \tag{2.1.5}$$

with at least one strict inequality. We note that for a set of k treatment groups, an umbrella alternative contains the ordered alternatives as special cases when peak $p = k$ or $p = 1$. These special cases have been studied in their own right for the unstratified case by Terpestra (1952), Jonckheere (1954), Page (1963) and Alvo and Cabilio (1995). For the rest of this section, let $n_i = m_1 + ... + m_i, i = 1, ..., k, n = n_k, and \tilde{n} = n - m_p$.

Let $\mathcal{P} = \{\mu : [\mu(1), ..., \mu(n)]\}$ be the set of all permutations of the integers $1, 2, ..., n$, and let $d(\mu, \nu)$ be a distance function between permutations μ and ν.

Step 1: Rank all the observations together so that the smallest gets rank 1, the next smallest rank 2 etc.. Let the $n-$dimensional vector

$$\pi = [\pi(1), ..., \pi(m_1) | \pi(m_1 + 1), ..., \pi(m_1 + m_2) | ... | ..., \pi(n)] \tag{2.1.6}$$

represent the ranks of the $\left\{\left(X_{i(l)}\right), i = 1, ..., k, l = 1, ..., m_i\right\}$ and grouped by populations $F_i, i = 1, ..., k$. In view of the continuity assumption on the distributions, ties among the observations occur with probability zero.

Step 2: Define $\{\pi\}$ to be the subclass of permutations "equivalent" to the observable permutation π in the sense that ranks occupied by identically distributed random variables are exchangeable within each treatment group separately. This subclass consists of all the permutations π where the rankings within each treatment group are permuted among themselves only. The cardinality of $\{\pi\}$ is given by the product $\Pi_i m_i!$.

Step 3: Define E to be an extremal subclass of $\mathcal{P}$ consisting of all permutations

which are "most in agreement" with H_1. The extremal set E does not necessarily correspond to the entire critical region but rather consists of those permutations which provide the strongest evidence in favour of the alternative.

The enumeration of the extremal set E is a two-stage procedure. First, choose the relative order of the $(p-1)$ pairs $(F_i), ..., (F_{p-1})$ among $(F_1), .., (F_{p-1}), (F_{p+1}), ..., (F_k)$. This can be done in $c = \binom{k-1}{p-1}$ ways. Then partition the integers $1, ..., n$ in accordance with the prescribed ordering of the treatment groups while taking into account corresponding sample sizes. The extremal set E is finally obtained by permuting the integers within each treatment group. Population F_p where the peaks occur are always allocated respectively the last m_p integers, namely positions $\tilde{n}+1, ..., n$ The cardinality of E is therefore equal to $c(\Pi_i m_i!)$.

Step 4: Let $d(\mu, \nu)$ be a distance function between two permutations μ, ν and define the distance between the two sets $\{\pi\}$ and an extremal set E by computing the sum of all pairwise distances between them*:*

$$d(\{\pi\}, E) = \sum_{\mu \epsilon \{\pi\}} \sum_{\nu \epsilon E} d(\mu, \nu) \tag{2.1.7}$$

Small values of $d(\{\pi\}, E)$ are inconsistent with the null hypothesis and consequently lead to rejection of H_0.

The commonly used Spearman and Kendall distances in this context are given respectively as

$$\begin{aligned} d_S(\mu, \nu) =& \frac{1}{2} \sum_{i=1}^{k} \sum_{l=1}^{m_i} \{\mu(i(l)) - \nu(i(l))\}^2 \\ =& \frac{n(n^2-1)}{12} - \sum_{i=1}^{k} \sum_{l=1}^{m_i} [\nu(i(l))]\{u(i(l)) - \frac{n+1}{2}\} \end{aligned} \tag{2.1.8}$$

and

$$d_K(u, \nu) = \frac{n(n-1)}{2} - \sum_{i_1(l) < i_2(l')} sign\{u(i_1(l)) - u(i_2(l'))\} sign\{\nu(i_1(l)) - \nu(i_2(l'))\}. \tag{2.1.9}$$

For the rest of this section, only end results are given for conciseness, as the

techniques used in the derivations of the test statistics are repeated more in depth in the next chapter.

The test based on the Spearman distance is found to be

$$S_p = \sum_{i=1}^{k} m_i \nu_i (\bar{\pi}_i - \frac{n+1}{2}), \tag{2.1.10}$$

where

$$\bar{\pi}_i = \frac{1}{m_i} \sum_{l=1}^{m_i} \pi(i(l)) \tag{2.1.11}$$

for $i = 1, ..., k$ and

$$\nu_i = \begin{cases} c^{-1} \sum_{j=i}^{k+i-p} a_{ij} \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} + \frac{1+m_i}{2} & \text{if } i < p, \\ \tilde{n} + \frac{1+m_p}{2} & \text{if } i = p, \\ c^{-1} \sum_{j=k-i+1}^{k-i+p} b_{ij} \binom{j-1}{k-i} \binom{k-1-j}{i-1-p} + \frac{1+m_i}{2} & \text{if } i > p, \end{cases} \tag{2.1.12}$$

with

$$a_{ij} = n_{i-1} + n - n_{k+i-j}, \tag{2.1.13}$$

and

$$b_{ij} = n - n_i + n_{j+i-k-1}. \tag{2.1.14}$$

As $min(m_i) \to \infty$, where $\frac{m_i}{n} \to \lambda_i > 0$, the test based on the Spearman distance can be shown to normally distributed under the null hypothesis with mean 0 and variance

$$Var_{H_0}(S_p) = (n+1)^2 \frac{\sum_{i=1}^{k} m_i (\nu_i - \bar{\nu})}{12}, \tag{2.1.15}$$

where $\bar{\nu} = \frac{1}{k} \sum_{i=1}^{k} \nu_i$. Additionally, the covariance between the tests for peaks p and p' was given by

$$Cov_{H_0}(S_p, S_{p'}) = \sigma^2 (n+1)^2 \sum_{i=1}^{k} m_i (\nu_{i,p} - \bar{\nu})(\nu_{i,p'} - \bar{\nu}), \tag{2.1.16}$$

where

$$\sigma^2 = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{m_i}(\frac{j}{n+1} - \frac{1}{2})^2. \tag{2.1.17}$$

On the other hand, the test based on the Kendall distance was given by

$$D_p = \sum_{1\le i_1<i_2\le k} W(i_1, i_2)U(i_1, i_2), \tag{2.1.18}$$

where

$$U(i_1, i_2) = \sum_{l=1}^{m_{i_1}}\sum_{l'=1}^{m_{i_2}} sign\{\mu(i_1(l)) - \mu(i_2(l))\} \tag{2.1.19}$$

and

$$W(i_1, i_2) = \begin{cases} -c & \text{if } 1 \le i_1 \le i_2 \le p, \\ c - 2H(i_1, i_2) & \text{if } 1 \le i_1 < p < i_2 \le k, \\ c & \text{if } p \le i_1 < i_2 \le k, \end{cases} \tag{2.1.20}$$

with

$$H(i_1, i_2) = \sum_{j=i_1}^{k+i_1-i_2} \binom{j-1}{i_1-1}\binom{k-j-1}{p-1-i_1} = \sum_{j=k+i_1-i_2}^{k+p-i_2} \binom{j-1}{k-i_2}\binom{k-j-1}{i_2-p-1}. \tag{2.1.21}$$

Additionally, it was shown that for the statistics with known peaks p and p' as $n \to \infty$,

$$|Corr(S_p, S_{p'}) - Corr(D_p, D_{p'})| \to 0 \tag{2.1.22}$$

The above result is useful for constructing a test for an unknown peak location through the maximum-type procedure by Hettmansperger and Norton (1987) using the k standardized test statistics for all possible peaks. This approach will be described in Section 2.3 for our proposed methods. Such tests for an unknown peak location in the unstratified case were shown to have good power in a simulation study

(Alvo, 2008). In the next section, we develop similar tests for known peaks in the case of stratified data.

2.2 Extensions to stratified data

In what follows, we extend the approach presented in the previous section to the case of two strata. Let $\mathcal{P} = \{\mu : [\mu(1), ..., \mu(n)]\}$ be the set of all permutations of the integers $1, 2, ..., n = n_1 + n_2$, and let $d(\mu, \nu)$ be a distance function between permutations μ and ν. The procedure is as follows:

Step 1: For stratum $i = 1, 2$ calculate the one-sample Hodges-Lehmann estimator θ_i from the pooled sample $X^{(i)} = \{X_1^{(i)}, ..., X_k^{(i)}\}$ given in (1.0.17). For each treatment group $q = 1, ..., k$ within stratum $i = 1, 2$, replace the raw observations with the aligned data

$$X_q^{(i)} \leftarrow \{X_q^{(i)} - \theta_i\} = \{X_{q(l)}^{(i)} - \theta_i,\ l = 1, ..., m_{iq}\}. \tag{2.2.1}$$

Step 2: Following the alignment in Step 1, the magnitudes of the strata are now comparable. We therefore rank all the observations together so that the smallest gets rank 1, the next smallest rank 2 etc.. Let the $n-$dimensional vector

$$\begin{aligned} \pi \ = \ & [\left(\pi(X_{1(1)}^{(1)}), ..., \pi(X_{1(m_{11})}^{(1)}) | \pi(X_{1(1)}^{(2)}), ..., \pi(X_{1(m_{21})}^{(2)})\right) | ... | ... \\ & ... | ... | \left(\pi(X_{k(1)}^{(1)}, ..., \pi(X_{k(m_{1k})}^{(1)}) | \pi(X_{k(1)}^{(2)}, ..., \pi(X_{k(m_{2k})}^{(2)})\right)] \end{aligned} \tag{2.2.2}$$

represent the ranks of the $\left\{\left(X_{i(l)}^{(1)} | X_{i(l^*)}^{(2)}\right), i = 1, ..., k, l = 1, ..., m_{1i}, l^* = 1, ..., m_{2i}\right\}$ and grouped by populations $(F_{1i}|F_{2i})$. In view of the continuity assumption on the distributions, ties among the observations occur with probability zero. Furthermore, let $\bar{\pi}_{1i} = \frac{1}{m_{1i}} \sum_{j=1}^{m_{1i}} \pi(X_{ij}^{(1)})$ and $\bar{\pi}_{2i} = \frac{1}{m_{2i}} \sum_{j=1}^{m_{2i}} \pi(X_{ij}^{(2)})$.

Step 3: Define $\{\pi\}$ to be the subclass of permutations "equivalent" to the observable permutation π in the sense that ranks occupied by identically distributed random variables are exchangeable within each distribution separately. This subclass consists of all the permutations π where the rankings within each population are permuted among themselves only. The cardinality of $\{\pi\}$ is given by the product

$\left(\Pi_i \left(m_{1i}!m_{2i}!\right)\right)$.

Step 4: Define $E_{2>1}$ to be an extremal subclass of $\mathcal{P}$ consisting of all permutations which are "most in agreement with H_1" with observations from stratum 2 ordered above those from stratum 1 within each treatment group. The extremal set $E_{2>1}$ does not necessarily correspond to the entire critical region but rather consists of those permutations which provide the strongest evidence in favour of the alternative. In the present context, permutations in $E_{2>1}$ are such that ranks occupied by observations from F_{1i} are always less than those from $F_{1i'}$, if $i < i' \leq p$, whereas the reverse is true if $p \leq i < i'$. Similarly, ranks occupied by observations from F_{2i} are always less than those from $F_{2i'}$, if $i < i' \leq p$, whereas the reverse is true if $p \leq i < i'$. Moreover, ranks attributed to a distribution consist of consecutive integers.

The enumeration of the extremal set $E_{2>1}$ is a two-stage procedure. First, choose the relative order of the $(p-1)$ pairs $\left(F_{1,i}|F_{2,i}\right), ..., \left(F_{1,p-1}|F_{2,p-1}\right)$ among $\left(F_{1,1}|F_{2,1}\right), .., \left(F_{1,p-1}|F_{2,p-1}\right), \left(F_{1,p+1}|F_{2,p+1}\right), ..., \left(F_{1,k}|F_{2,k}\right)$. This can be done in $c = \binom{k-1}{p-1}$ ways. Then partition the integers $1, ..., n$ in accordance with the prescribed ordering of the treatment groups while taking into account corresponding sample sizes. The extremal set $E_{2>1}$ is finally obtained by permuting the integers within each population. Populations $F_{1,p}, F_{2,p}$ where the peaks occur are always allocated respectively the last $m_{1,p} + m_{2,p}$ integers, namely positions $\tilde{n}_1 + \tilde{n}_2 + 1, ..., n$ The cardinality of $E_{2>1}$ is therefore equal to $c\left(\Pi_i \left(m_{1i}!m_{2i}!\right)\right)$.

Step 5: Let $d\left(\mu, \nu\right)$ be a distance function between two permutations μ, ν and define the distance between $\{\pi\}$ and extremal set $E_{2>1}$ by computing the sum of all pairwise distances between them*:*

$$d\left(\{\pi\}, E_{2>1}\right) = \sum_{\mu \epsilon \{\pi\}} \sum_{\nu \epsilon E_{2>1}} d\left(\mu, \nu\right), \tag{2.2.3}$$

where again we specifically use the distances of Spearman and Kendall, defined respectively as

$$d_S\left(\mu, \nu\right) = \frac{1}{2} \sum_{i=1}^{k} \left[\sum_{l=1}^{m_{1i}+m_{2i}} \left[\mu\left(i\left(l\right)\right) - \nu\left(i\left(l\right)\right)\right]^2 \right]$$

$$= \frac{n\left(n^2-1\right)}{12} - \sum_{i=1}^{k} \sum_{l=1}^{m_{1i}+m_{2i}} \left[\nu\left(i\left(l\right)\right)\right] \left[\mu\left(i\left(l\right)\right) - \frac{n+1}{2}\right] \quad (2.2.4)$$

and

$$d_K(\mu,\nu) = \frac{n(n-1)}{2} - \sum_{i_1(l)<i_2(l')} sign\{\mu(i_1(l)) - \mu(i_2(l'))\} sign\{\nu(i_1(l)) - \nu(i_2(l'))\}. \quad (2.2.5)$$

Small values of $d\left(\{\pi\}, E_{2>1}\right)$ are inconsistent with the null hypothesis and consequently lead to rejection of H_0.

Step 6: Repeat steps 2 through 5 now with extremal set $E_{1>2}$ chosen such that the values of stratum 1 are above stratum 2 within each treatment group. A test based on both extremal sets $E = E_{2>1} \cup E_{1>2}$ can be obtained by defining the total distance

$$d\left(\{\pi\}, E\right) = d\left(\{\pi\}, E_{2>1}\right) + d\left(\{\pi\}, E_{1>2}\right) \quad (2.2.6)$$

In what follows, we derive tests following the above procedure using the Spearman and Kendall distances.

2.2.1 The test statistic for a known peak corresponding to Spearman distance

In this section, we derive the test statistic corresponding to Spearman distance under the extremal set E when the location of the peak is known. Throughout, we shall assume that permutations defined by the extremal sets are arranged in columns indexed by $1 \leq i\left(l\right) \leq n$ in such a way that ranks are in increasing order for populations $\left(F_{1i}, F_{2i}\right), i \leq p$ and in decreasing order when $i \geq p$.

We first begin with $d\left(\{\pi\}, E_{2>1}\right)$. Under the hypothesis of parallelism, we suppose the values of the second stratum are above those of the first stratum for each

treatment group $i = 1, ..., k$. Suppose that (F_{1i}, F_{2i}) , $i < p$ as a vector is in relative position j and hence populations (F_{11}, F_{21}) ...,$(F_{1,i-1}, F_{2,i-1})$ are in relative positions chosen from among the first $(j-1)$ positions. Populations $(F_{1,i+1}, F_{2,i+1})$ ..., $(F_{1,p-1}, F_{2,p-1})$ are then assigned positions chosen from $(j+1), ..., (k-1)$. This can happen with frequency $\binom{j-1}{i-1}\binom{k-1-j}{p-1-i}$. The positions of the remaining populations are then automatically determined. Together populations (F_{11}, F_{21}) , ..., $(F_{1,i-1}, F_{2,i-1})$ and $(F_{1,k+i-j+1}$, $F_{2,k+i-j+1})$, ..., $(F_{1,k}, F_{2,k})$ are assigned the first a_{ij} integers where

$$\begin{aligned} a_{ij} &= \left(\sum_{h=1}^{i-1} m_{1h} + \sum_{h=1}^{i-1} m_{2h}\right) + \left(\sum_{h=k+i-j+1}^{k} m_{1h} + \sum_{h=k+i-j+1}^{k} m_{2h}\right) \\ &= (n_{1i-1} + n_1 - n_{1k+i-j}) + (n_{2i-1} + n_2 - n_{2k+i-j}) \end{aligned} \tag{2.2.7}$$

Populations (F_{1i}, F_{2i}) are assigned integers $(a_{ij}+1), ..., (a_{ij} + m_{1i} + m_{2i})$ whose sum is equal to $\left(a_{ij} + \frac{m_{1i}+m_{i2}+1}{2}\right)(m_{1i} + m_{i2})$. Consequently, F_{1i} is assigned integers $a_{ij} + 1, ..., a_{ij} + m_{1i}$, whereas F_{2i} is assigned integers $a_{ij} + m_{1i} + 1, ..., a_{ij} + m_{1i} + m_{2i}$. The process of permuting ranks within individual population F_{1i} implies that each entry will contribute to the sum $(m_{1i} - 1)!$ times. Similarly for F_{2i}, each entry will contribute to the sum $(m_{2i} - 1)!$ times. Hence, for (F_{1i}, F_{2i}) for each entry taking into account the permutations, we have

$$(\Pi\,(m_{1i}!m_{2i}!)) \left[\left(a_{ij} + \frac{m_{1i}+1}{2}\right), a_{ij} + m_{1i} + \frac{m_{2i}+1}{2}\right], i < p \tag{2.2.8}$$

Finally, summing separately over each j

$$\begin{aligned} (\Pi\,(m_{1i}!m_{2i}!))\,[&\sum_{j=i}^{k+i-p} \left(a_{ij} + \frac{m_{1i}+1}{2}\right)\binom{j-1}{i-1}\binom{k-1-j}{p-1-i}, \\ &\sum_{j=i}^{k+i-p} \left(a_{ij} + m_{1i} + \frac{m_{2i}+1}{2}\right)\binom{j-1}{i-1}\binom{k-1-j}{p-1-i}] \end{aligned} \tag{2.2.9}$$

On the other hand, for the data vector we have for each entry in (F_{1i}, F_{2i})

$$(\Pi\,(m_{1i}!m_{2i}!)) \left[\bar{\pi}_{1i} - \frac{n+1}{2}, \bar{\pi}_{2i} - \frac{n+1}{2}\right], \tag{2.2.10}$$

where $\bar{\pi}_{1i}$ and $\bar{\pi}_{2i}$ represent the averages of the ranks for the populations F_{1i}, F_{2i} respectively.

Similarly for $i > p$, we may define

$$\begin{aligned} b_{ij} &= \left[\left(\sum_{h=i+1}^{k} m_{1h}\right) + \left(\sum_{h=1}^{j+i-k-1} m_{1h}\right)\right] \\ &+ \left[\left(\sum_{h=i+1}^{k} m_{2h}\right) + \left(\sum_{h=1}^{j+i-k-1} m_{2h}\right)\right] \\ &= (n_1 - n_{1i} + n_{1j+i-k-1}) \\ &+ (n_2 - n_{2i} + n_{2j+i-k-1}) \end{aligned} \tag{2.2.11}$$

Hence the Spearman distance is

$$d_S\left(\{\pi\}, E_{2>1}\right) = \frac{n\left(n^2-1\right)}{12} c\left(\Pi\left(m_{1i}! m_{2i}!\right)\right)^2 - c\left(\Pi\left(m_{1i}! m_{2i}!\right)\right)^2 \left(S_{1p} + S_{2p}\right), \tag{2.2.12}$$

where

$$S_{1p} = \sum_{i=1}^{k} m_{1i}\nu_{1i}\left[\bar{\pi}_{1i} - \frac{n+1}{2}\right] \tag{2.2.13}$$

$$S_{2p} = \sum_{i=1}^{k} m_{2i}\nu_{2i}\left[\bar{\pi}_{2i} - \frac{n+1}{2}\right], \tag{2.2.14}$$

and

$$v_{1i} = \begin{cases} c^{-1} \sum_{j=i}^{k+i-p} \left(a_{ij} + \frac{m_{1i}+1}{2}\right) \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} & i < p \\ \left(\tilde{n}_1 + \tilde{n}_2 + \frac{m_{1p}+1}{2}\right) & i = p \\ c^{-1} \sum_{j=k-i+1}^{k-i+p} \left(b_{ij} + \frac{m_{1i}+1}{2}\right) \binom{j-1}{k-i} \binom{k-1-j}{i-1-p} & i > p \end{cases} \tag{2.2.15}$$

$$v_{2i} = \begin{cases} c^{-1} \sum_{j=i}^{k+i-p} \left(a_{ij} + m_{1i} + \frac{m_{2i}+1}{2}\right) \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} & i < p \\ \left(\tilde{n}_1 + \tilde{n}_2 + m_{1p} + \frac{m_{2p}+1}{2}\right) & i = p \\ c^{-1} \sum_{j=k-i+1}^{k-i+p} \left(b_{ij} + m_{1i} + \frac{m_{2i}+1}{2}\right) \binom{j-1}{k-i} \binom{k-1-j}{i-1-p} & i > p \end{cases} \tag{2.2.16}$$

In like manner, we can also obtain the statistic for the partial extremal set $E_{1>2}$ where F_{1i} are assumed to be above F_{2i} for $1 \leq i \leq k$ by swapping the labels, leading to

$$d_S\left(\{\pi\}, E_{1>2}\right) = \frac{n\left(n^2-1\right)}{12} c\left(\Pi\left(m_{1i}! m_{2i}!\right)\right)^2 - c\left(\Pi\left(m_{1i}! m_{2i}!\right)\right)^2 \left(S_{1p}^{'} + S_{2p}^{'}\right), \tag{2.2.17}$$

where

$$S_{1p}^{'} = \sum_{i=1}^{k} m_{1i} \nu_{1i}^{\prime} \left[\bar{\pi}_{1i} - \frac{n+1}{2}\right] \tag{2.2.18}$$

$$S_{2p}^{'} = \sum_{i=1}^{k} m_{2i} \nu_{2i}^{'} \left[\bar{\pi}_{2i} - \frac{n+1}{2}\right], \tag{2.2.19}$$

and

$$v_{1i}^{'} = \begin{cases} c^{-1} \sum_{j=i}^{k+i-p} \left(a_{ij} + m_{2i} + \frac{m_{1i}+1}{2}\right) \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} & i < p \\ \left(\tilde{n}_1 + \tilde{n}_2 + m_{2p} + \frac{m_{1p}+1}{2}\right) & i = p \\ c^{-1} \sum_{j=k-i+1}^{k-i+p} \left(b_{ij} + m_{2i} + \frac{m_{1i}+1}{2}\right) \binom{j-1}{k-i} \binom{k-1-j}{i-1-p} & i > p \end{cases} \tag{2.2.20}$$

$$v_{2i}^{'} = \begin{cases} c^{-1} \sum_{j=i}^{k+i-p} \left(a_{ij} + \frac{m_{2i}+1}{2}\right) \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} & i < p \\ \left(\tilde{n}_1 + \tilde{n}_2 + \frac{m_{2p}+1}{2}\right) & i = p \\ c^{-1} \sum_{j=k-i+1}^{k-i+p} \left(b_{ij} + \frac{m_{2i}+1}{2}\right) \binom{j-1}{k-i} \binom{k-1-j}{i-1-p} & i > p. \end{cases} \tag{2.2.21}$$

The test $d_S(\{\pi\}, E)$ based on the entire extremal set will have weights which are simply the sum of both the original and the swapped weights, given as

$$v_{1i} + v_{1i}^{'} = v_{2i} + v_{2i}^{'} = \begin{cases} 2c^{-1} \sum_{j=i}^{k+i-p} \left(a_{ij} + \frac{m_{1i}+m_{2i}+1}{2}\right) \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} & i < p \\ 2\left(\tilde{n}_1 + \tilde{n}_2 + \frac{m_{1p}+m_{2p}+1}{2}\right) & i = p \\ 2c^{-1} \sum_{j=k-i+1}^{k-i+p} \left(b_{ij} + \frac{m_{1i}+m_{2i}+1}{2}\right) \binom{j-1}{k-i} \binom{k-1-j}{i-1-p} & i > p \end{cases} \tag{2.2.22}$$

which is equivalent to the unstratified test of Alvo (2008) with $F_i := (F_{1i}, F_{2i})$ for $1 \leq i \leq k$. This implies that using the entire defined extremal set leads to the traditional aligned-rank approach in the case of the Spearman distance. It is therefore instructive to recall that for the unstratified case when $m_i = m$ observations are taken from each population, it was shown that $a_{ij} = b_{ij} = (j-1)\, m$ and

$$v_i \;\; = \begin{cases} c^{-1} \sum_{j=i}^{k+i-p} \left(jm + \frac{1-m}{2}\right) \binom{j-1}{i-1} \binom{k-1-j}{p-1-i} & i < p \\ km + \left(\frac{1-m}{2}\right) & i = p \\ c^{-1} \sum_{j=k-i+1}^{k-i+p} \left(jm + \frac{1-m}{2}\right) \binom{j-1}{k-i} \binom{k-1-j}{i-p-1} & i > p, \end{cases} \tag{2.2.23}$$

using the identity (12.16) in Feller (1968, p.65)

$$v_i \;\; = \begin{cases} n\frac{i}{p} + \left(\frac{1-m}{2}\right) & i \leq p \\ n\frac{(k+1-i)}{k+1-p} + \left(\frac{1-m}{2}\right) & i > p \end{cases} \tag{2.2.24}$$

It follows that

$$S_p = mn \left\{ \sum_{i \leq p} \frac{i}{p} \left[\bar{\pi}_i - \frac{n+1}{2} \right] + \sum_{i > p} \frac{(k+1-i)}{(k+1-p)} \left[\bar{\pi}_i - \frac{n+1}{2} \right] \right\}. \tag{2.2.25}$$

2.2.2 The test statistic for a known peak corresponding to Kendall distance

In this section, we first review the derivation of the unstratified test statistic based on the Kendall distance before extending to the stratified case. Consider for now there is only one observation per sample. Fix $1 \leq i_1 < p < i_2 \leq k$. Suppose that the integer j_1 is assigned to F_{i_1} and integer j_2 is assigned to F_{i_2}, with $j_2 > j_1$. The frequency with which this can occur is given by

$$\binom{j_1 - 1}{i_1 - 1} \binom{j_2 - j_1 - 1}{j_2 + i_2 - i_1 - k - 1} \binom{k - j_2}{i_2 - p}, \; j_2 > j_1. \tag{2.2.26}$$

If q is the number of samples among $F_{(i_1+1)}, ..., F_{(p-1)}$ which are assigned ranks greater than j_1 but less than j_2, then we must have

$$q + i_1 + k - i_2 = j_2 - 1. \tag{2.2.27}$$

Their ranks are chosen from $j_1 + 1, ..., j_2 - 1$. Summing over j_2 we obtain the total number of negatives,

$$H(i_1, i_2) = \sum_{j_1 = i_1}^{k + i_1 - i_2} \binom{j_1 - 1}{i_1 - 1} \binom{k - 1 - j_1}{p - 1 - i_1}. \tag{2.2.28}$$

Alternatively, by instead summing over j_1 we obtain

$$H(i_1, i_2) = \sum_{j_1 = k + i_1 - i_2}^{k + p - i_2} \binom{j_2 - 1}{k - i_2} \binom{k - 1 - j_2}{i_2 - p - 1}. \tag{2.2.29}$$

It follows that the sum over the signs is given by the positives less the negatives $c - 2H(i_1, i_2)$. It follows that the Kendall test statistic becomes

$$\sum_{i_1<i_2} W(i_1,i_2)sign\{\pi(i_1)-\pi(i_2)\}, \tag{2.2.30}$$

where

$$W(i_1,i_2)=\begin{cases}-c, & \text{if } 1\le i_1\le i_2\le p\\ c-2H(i_1,i_2) & \text{if } 1\le i_1\le p\le i_2\le k\\ c & \text{if } p\le i_1\le i_2\le k.\end{cases} \tag{2.2.31}$$

In the general case of unequal sample sizes, the extremal set is determined by additionally permuting within each sample. The weight function will remain unchanged, as it is still only a function of the terms $\pi(i_1(l))-\pi(i_2(l'))$, which are determined entirely by the ordering of the treatment groups. The data set on the other hand consists of permuting the ranks within each sample, leading to a double sum

$$U(i_1,i_2)=\sum_{l=1}^{m_{i_1}}\sum_{l'=1}^{m_{i_2}} sign\{\pi(i_1(l))-\pi(i_2(l'))\}. \tag{2.2.32}$$

Hence the test based on the Kendall distance when the sample sizes are unequal for an extremal set E becomes

$$D_p=\sum_{i_1<i_2} W(i_1,i_2)U(i_1,i_2), \tag{2.2.33}$$

where $U(\cdot,\cdot)$ is the Mann-Whitney-Wilcoxon statistic given by (1.0.12). The tests of Mack and Wolfe (1981) and Millen and Wolfe (2005) also employ the sums of Mann-Whitney statistics across samples. Notably they do not include comparisons across the peak, which can result in a loss of efficiency (Hettmansperger and Norton 1987). The extension to data with two strata is easily found from the construction of the extremal set, where for each $1\le i\le k$, (F_{1i},F_{2i}) are grouped together. It follows that the comparison between treatment groups remained unchanged, but there are now also within-treatment comparisons which depend on the specific ordering of (F_{1i},F_{2i}) chosen for the extremal set. The tests for each of the two components of the full extremal set are given by

$$D_{(1>2)p} = D_p + \sum_{j=1}^{k} U(i_{1j}, i_{2j}) \tag{2.2.34}$$

and

$$D_{(2>1)p} = D_p + \sum_{j=1}^{k} U(i_{2j}, i_{1j}), \tag{2.2.35}$$

where D_p is the conventional aligned-rank test and i_{1j} and i_{2j} represent the j^{th} population of the first and second stratum, respectively. We note that the additional terms are equal to the stratified test statistic of Wilcoxon (1946) given by (1.0.11) with the strata and treatment information of each observation being swapped after the alignment step. Similar to the Spearman test, the sum over both orderings of the strata within each treatment group removes the stratum information, as

$$U(i_{1j}, i_{2j}) = -U(i_{2j}, i_{1j}). \tag{2.2.36}$$

Based on the above relationship, the test using the entire extremal set E is the traditional aligned-rank test corresponding to D_p in Alvo (2008).

2.3 The use of partial extremal sets for unequal sample sizes

In this chapter, we have shown that the test statistics based on the Spearman and Kendall distances constructed using the entire extremal set in the stratified case lead to the established aligned-rank test statistic corresponding to the unstratified test based on the same distances.

We note that in some situations, a test based on only one of the two partial extremal sets as discussed in Chapter 2 can be preferable. This suggestion is explained by the fact that the alignment step does not properly align the data under the alternative hypothesis when there are unequal sample sizes.

We demonstrate this issue with the alignment by considering the difference in Hodges-Lehmann estimators of the pooled samples of each stratum in the two-

treatment problem. We simulate 100000 sets of samples under the alternative hypothesis with a treatment effect of 2 for a given set of sample sizes and densities. For each of the 100000 iterations, we calculate the difference between the two alignment estimators $\theta_1 - \theta_2$, where θ_1 and θ_2 are the one-sample Hodges-Lehmann estimators of the pooled samples $X_1^{(1)}, ..., X_k^{(1)}$ and $X_1^{(2)}, ..., X_k^{(2)}$, respectively. The average across all iterations is given for each combination of sample size and underlying density in Table 2.1.

The distributions are standardized so that the variance is always equal to 1, with density functions of the normal, double-exponential, logistic and exponential densities, respectively given by

$$f_N(x, \mu) = \frac{1}{\sqrt{2\pi}} \exp(\frac{1}{2}(x - \mu)^2), \ -\infty < x < \infty, \tag{2.3.1}$$

$$f_D(x, \mu) = \frac{1}{\sqrt{2}} \exp(-\sqrt{2}|x - \mu|), \ -\infty < x < \infty, \tag{2.3.2}$$

$$f_L(x, \mu) = \frac{\exp((x - \mu)/\sqrt{\frac{3}{\pi^2}})}{\sqrt{\frac{3}{\pi^2}}(1 + \exp((x - \mu)/\sqrt{\frac{3}{\pi^2}}))}, \ -\infty < x < \infty, \tag{2.3.3}$$

and

$$f_E(x, \mu) = e^{-x}, \ x > 0. \tag{2.3.4}$$

Table 2.1 shows that when there is no difference in sample sizes, the average difference in alignment estimators is approximately 0. This result indicates an unbiased alignment procedure, implying the stratum information is of no value after alignment. The traditional aligned-rank approach obtained through the use of the full extremal set E is recommended in this case, as the resulting test statistic ignores stratum information after alignment.

On the other hand, the difference in alignment estimators is not equal to 0 when the sample sizes are unequal, indicating a bias in the alignment procedure. When present under the assumption of parallelism, this "misalignment issue" results in a stochastic ordering of the samples after alignment matching one of the partial

Sample sizes				Mean difference between stratum alignment estimates			
m_{11}	m_{12}	m_{21}	m_{22}	Normal	D-exp	Logistic	Exp
15	15	15	15	0.00	0.00	0.00	0.00
14	16	16	14	0.15	0.14	0.14	0.16
13	17	17	13	0.29	0.28	0.29	0.32
12	18	18	12	0.44	0.42	0.44	0.48
11	19	19	11	0.59	0.58	0.59	0.63
10	20	20	10	0.73	0.73	0.74	0.77

Table 2.1: Mean differences between stratum alignment estimates when $k = 2, \delta = 0$ and $\gamma_2 - \gamma_1 = 2$

extremal sets $E_{2>1}$ and $E_{1>2}$. The test statistics based on these partial extremal sets are therefore of interest, as they capture additional evidence against the null hypothesis H_0.

Table 2.1 also shows that for the chosen treatment effect of 2, there is an approximately linear relationship between the average difference between the two alignment estimators $\theta_1 - \theta_2$ and the difference in sample sizes $m_{21} - m_{22}$. The underlying density used in the simulation does not effect the results to the same extent as the sample sizes. The average difference between the two alignment estimators $\theta_1 - \theta_2$ is larger when the underlying density is exponential when compared to the symmetric densities used.

We proceed by deriving the asymptotic properties of the tests based on the Spearman distance with only the partial extremal sets and compare them in simulations to the corresponding traditional aligned-rank approach obtained using the full extremal set.

Chapter 3

Asymptotic distributions of the test statistics

3.1 Asymptotic distributions of the test statistics based on the Spearman distance when the peak location is known

In this section, we find the asymptotic distribution of the two-strata test statistics under the null hypothesis when the location of the peak location is known. To prove the asymptotic distribution of the two-strata test statistic when the peak location is known based on the partial extremal set $E_{2>1}$, we require the following alternative notation: let for $i = 1, 2, ..., 2k$,

$$\tilde{m}_i = \begin{cases} m_{1i} & i \leq k \\ m_{2(i-k)} & i > k \end{cases} \tag{3.1.1}$$

$$\tilde{\nu}_i = \begin{cases} \nu_{1i} & i \leq k \\ \nu_{2(i-k)} & i > k \end{cases} \tag{3.1.2}$$

$$\tilde{\bar{\pi}}_i = \begin{cases} \bar{\pi}_{1i} & i \leq k \\ \bar{\pi}_{2(i-k)} & i > k \end{cases} \tag{3.1.3}$$

$$\tilde{\pi}(X_{i(l)}) = \begin{cases} \pi(X^{(1)}_{i(l)}) & i \leq k \\ \pi(X^{(2)}_{(i-k)(l)}) & i > k. \end{cases} \tag{3.1.4}$$

Using this notation,

$$S_{(2>1)p} \equiv S_{1p} + S_{2p} = \sum_{i=1}^{2k} \tilde{m}_i \tilde{\nu}_{i,p} \left[\tilde{\bar{\pi}} - \frac{n+1}{2} \right]. \tag{3.1.5}$$

Alternatively, the test statistic is equal to the normalized linear rank statistic

$$S_{(2>1)p} = (n+1) \sum_{i=1}^{2k} \tilde{\nu}_{i,p} \sum_{l=1}^{\tilde{m}_i} \left[\frac{\tilde{R}_{i(l)}}{n+1} - \frac{1}{2} \right]. \tag{3.1.6}$$

As with the asymptotic results of the aligned-rank test of Hodges and Lehmann (1962), the following theorem assumes the alignment step has been applied to the dataset, resulting in each of the $2k$ samples being equal under the null hypothesis H_0.

Theorem 1. *Assume that* $\min(\tilde{m}_i) \to \infty$ *with*

$$\frac{\tilde{m}_i}{n} \to \lambda_i > 0. \tag{3.1.7}$$

Define

$$\tilde{\sigma}_p^2 = (n+1)^2 \frac{\sum_{i=1}^{2k} \tilde{m}_i \left(\tilde{\nu}_{i,p} - \bar{\tilde{\nu}}_p \right)^2}{12}, \tag{3.1.8}$$

where $\bar{\tilde{\nu}}_p = \frac{1}{2k} \sum_{i=1}^{2k} \tilde{\nu}_{i,p}$.

Then the test statistic $S_{(2>1)p}$ *corresponding to the Spearman distance and the partial extremal set* $E_{2>1}$ *is asymptotically normal with mean* 0 *and variance equal to* σ_p^2.

Proof: Representing $S_{(2>1)p}$ as the normalized linear rank statistic (3.1.6), since $\frac{\tilde{m}_i}{n}$ converges to a constant, it follows that as $\min(\tilde{m}_i) \to \infty$,

$$\begin{aligned} \max_i \frac{(\tilde{\nu}_{i,p} - \bar{\tilde{\nu}})^2}{\sum_{i=1}^{2k} \tilde{m}_i (\tilde{\nu}_{i,p} - \bar{\tilde{\nu}})^2} &\leq \frac{1}{\min \tilde{m}_i} \max_{i \leq 2k} \frac{(\tilde{\nu}_{i,p} - \bar{\tilde{\nu}})^2}{\sum_{i=1}^{2k} (\tilde{\nu}_{i,p} - \bar{\tilde{\nu}})^2} \\ &\to 0. \end{aligned} \tag{3.1.9}$$

Following Hájek, Šidák, and Sen (1999, pg. 183-184), the above result proves the asymptotically normality of $S_{(2>1)p}$. The mean and variance follow from Hájek, Šidák, and Sen (1999, pg. 61-62). ∎

Similar results follow for the test statistic $d_S(\{\pi\}, E_{1>2}) \equiv S_{(1>2)p} \equiv S'_{1p} + S'_{2p}$.

3.2 Asymptotic distributions of the test statistics based on the Spearman distance when the peak location is unknown

The test based on the Spearman distance using the partial extremal set $E_{2>1}$, as in the unstratified case, may be constructed using the approach of Hettmansperger and Norton (1987). We first define the standardized statistics for all possible peaks as

$$S^*_{(2>1)p} = \frac{S_{(2>1)p}}{\tilde{\sigma}^2_p}, \quad p = 1, ..., k. \tag{3.2.1}$$

The null hypothesis H_0 is rejected whenever the test statistic

$$S_{(2>1)\max} = \max_p S^*_{(2>1)p.} \tag{3.2.2}$$

is large. As with the case of a known peak location, the following theorem assumes the alignment step has been applied to the dataset, resulting in each of the $2k$ samples being equal under the null hypothesis H_0.

Theorem 2. *Let the vector* $S = (S_{(2>1)1,...,}S_{(2>1)k})^T$ *and let* $Cov(S) = BB^T$. *Under* H_0, *if* $\min \tilde{m}_i \to \infty$ *with* $\tilde{m}_i/n \to \lambda_i > 0$, $i = 1,,,k$, *then* S *has asymptotically*

the distribution of BZ, where Z is multivariate normal with mean 0 *and covariance matrix I. Consequently, S_{max} has asymptotically the distribution of* $\max BZ$.

Proof: Using the same notation as in the previous section, we note that the components of the vector S are linear combinations of the average ranks over each population. Let $d_1, ..., d_k$ be a set of arbitrary coefficients and consider the linear combination

$$\sum_{p=1}^{k} d_p S^*_{(2>1)p} = \sum_{i=1}^{2k} d'_i(\tilde{\pi}_i - \frac{n+1}{2}), \tag{3.2.3}$$

where

$$d'_i = (\sum_{p=1}^{k} \frac{d_p \tilde{\nu}_{i,p}}{\tilde{\sigma}_p} \tilde{m}_i). \tag{3.2.4}$$

The asymptotic normality of (3.2.3) follows as in the proof of Theorem 3.
From Hájek, Šidák, and Sen (1999, p.62), we calculate

$$Cov(S^*_{(2>1)p}, S^*_{(2>1)p'}) = \sigma^2 (n+1)^2 \sum_{i=1}^{2k} \tilde{m}_i (\frac{\tilde{\nu}_{i,p} - \bar{\tilde{\nu}}_p}{\tilde{\sigma}_p})(\frac{\tilde{\nu}_{i,p'} - \bar{\tilde{\nu}}_{p'}}{\tilde{\sigma}_{p'}}) \tag{3.2.5}$$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{2k} \sum_{l=1}^{\tilde{m}_i} (\frac{l}{n+1} - \frac{1}{2})^2. \tag{3.2.6}$$

∎

As noted by Hettmansperger and Norton (1987), while the distribution of $\max BZ$ does not have a closed-form solution, it can easily be simulated. The approximate $p-$value of the test is given by

$$P(\max BZ \geq s_{(2>1)\max}), \tag{3.2.7}$$

where $s_{(2>1)\max}$ is the observed value of $S_{(2>1)\max}$.

Chapter 4

Simulation results

In this section, we report results on a limited simulation study for the distribution of the Spearman statistics for known and unknown peak locations based on the full extremal set and each of the two partial extremal sets. All simulations were performed using the R programming language. Four families of underlying distributions were considered: the normal, the double exponential, the logistic, and the exponential. The normal has short to medium tails, the logistic and double exponential have longer tails, whereas the exponential is skewed. The variance for each distribution was set equal to 1. Let $k = 5, p = 3$ and let the distributions $F_1, ..., F_5$ have location parameters equal to $0, w, 2w, w, 0$ respectively under the alternative, with $w = 1/4$ when the peak location is known and $w = 1/3$ when the peak location is unknown. We considered the case where the sample sizes are equal (Table 4.1) as well as when they are unequal (Table 4.2 and Table 4.3). In the unequal sample size case described in Table 4.3, the difference between the sample sizes remains the same as the value of m increases and therefore the impact of the unequal sample sizes diminishes for larger n. On the other hand, the difference between two sample sizes grows linearly as n gets larger in the case described in Table 4.2.

For each value of m, we used 10000 repetitions. The simulations were done for the Spearman test statistics under the extremal sets E, $E_{1>2}$, and $E_{2>1}$.

Stratum \ Treatment	1	2	3	4	5
1	m	m	m	m	m
2	m	m	m	m	m

Table 4.1: Equal sample size case, $m = 5, 6, 7, 8, 9, 10, 15$

Stratum \ Treatment	1	2	3	4	5
1	$m-2$	m	$m+4$	m	$m-2$
2	m	m	m	m	m

Table 4.2: Nearly equal sample size case, $m = 5, 6, 7, 8, 9, 10, 15$

Stratum \ Treatment	1	2	3	4	5
1	m	$2m$	$3m$	$2m$	m
2	$2m$	$2m$	$2m$	$2m$	$2m$

Table 4.3: Unequal sample size case, $m = 3, 4, 5, 6, 7$

4.1 Simulation results for known peak location

When the location of the peak is known and the sample sizes are equal in Table 4.4, we find that the test based on the full extremal set E has an empirical level closer to the nominal level than either of the tests based on the partial extremal sets $E_{2>1}$ and $E_{1>2}$. We also see that the test based on E has approximately $1-2\%$ higher empirical power than the tests based on the partial extremal sets in most cases. The tests based on $E_{2>1}$and $E_{1>2}$ in turn have equal empirical levels and powers.

It is seen in Tables 4.5 and 4.6 that the test based on the partial extremal set $E_{2>1}$ tends to have an empirical level closer to the nominal value compared to the test based on E when the sample sizes are unequal. The test based on $E_{2>1}$ also tends to outperform the test based on E, with empirical power values between $1-2\%$ higher in the more unbalanced scenarios. The test based on the partial extremal set $E_{1>2}$ has a much worse performance than the other two tests and also does not maintain an empirical level near the nominal level as compared to the other two tests.

The difference in performance between the tests based on $E_{2>1}$ and E can be explained by the misalignment issue describes in Section 2.3. When the sample sizes are equal, there is no misalignment under either the null or alternative hypotheses and therefore the tests based on partial extremal sets unnecessarily account for information independent of the problem. On the other hand, the misalignment issue is present under the alternative hypothesis when the sample sizes are unequal. Our choices of unequal sample size designs have a larger proportion of the stratum 1 observations near the peak compared to the proportion of stratum 2 observations. As a result, stratum 1 would be below stratum 2 after the alignment and that additional information would be captured by the test based on the extremal set $E_{2>1}$. The test based on $E_{1>2}$ takes the same information as evidence instead in favour of the null hypothesis, explaining the worse performance in the cases considered. The two tests based on extremal sets would have the opposite results in cases where the alignment step leads to the observations of stratum 1 being above those of stratum 2.

In most cases considered, the performances of the tests based on $E_{2>1}$ and E differ to a greater extent when the underlying distribution is normal or logistic compared to when the underlying distribution is double exponential or exponential. In the case of equal sample sizes, the relative advantage the test based on E has compared to the test based on $E_{2>1}$ appears to diminish by a small margin as the value of m grows. We suspect this is a result of a higher proportion of the test statistic based on $E_{2>1}$ being made up of between-treatment comparisons when m is larger. In the first case of unequal sample sizes, the relative disadvantage of the test based on E compared to the test based on $E_{2>1}$ diminishes by a similar margin as m gets larger due to the shrinking difference in proportions of the observations near the peak for the two strata. For the second case of unequal sample sizes, the results do not appear to be influenced by the value of m, as the misalignment does not lessen for large values.

		Level				Power			
Test	m	Normal	D-exp	Logistic	Exp	Normal	D-exp	Logistic	Exp
E	5	0.0516	0.0480	0.0526	0.0497	0.3480	0.4399	0.3660	0.5740
	6	0.0509	0.0485	0.0508	0.0517	0.3802	0.5059	0.4193	0.6419
	7	0.0523	0.0517	0.0502	0.0464	0.4354	0.5650	0.4660	0.7229
	8	0.0493	0.0515	0.0501	0.0518	0.4727	0.6017	0.5093	0.7625
	9	0.0476	0.0450	0.0523	0.0503	0.5174	0.6536	0.5556	0.8092
	10	0.0507	0.0503	0.0461	0.0501	0.5558	0.6940	0.5959	0.8461
	15	0.0511	0.0476	0.0482	0.0492	0.6965	0.8378	0.7476	0.9503
$E_{1>2}$	5	0.0464	0.0430	0.0465	0.0465	0.3246	0.4152	0.3449	0.5516
	6	0.0444	0.0433	0.0458	0.0478	0.3616	0.4829	0.3976	0.6225
	7	0.0460	0.0457	0.0454	0.0430	0.4130	0.5445	0.4456	0.6991
	8	0.0442	0.0466	0.0436	0.0472	0.4512	0.5803	0.4870	0.7385
	9	0.0427	0.0387	0.0462	0.0472	0.4950	0.6323	0.5333	0.7917
	10	0.0455	0.0450	0.0419	0.0483	0.5345	0.6753	0.5752	0.8296
	15	0.0457	0.0419	0.0418	0.0469	0.6780	0.8260	0.7318	0.9417
$E_{2>1}$	5	0.0465	0.0431	0.0469	0.0446	0.3264	0.4186	0.3456	0.5485
	6	0.0446	0.0433	0.0460	0.0457	0.3597	0.4841	0.3967	0.6193
	7	0.0466	0.0456	0.0453	0.0443	0.4150	0.5432	0.4464	0.7006
	8	0.0436	0.0465	0.0437	0.0492	0.4509	0.5816	0.4864	0.7424
	9	0.0417	0.0389	0.0466	0.0473	0.4934	0.6313	0.5334	0.7897
	10	0.0453	0.0443	0.0431	0.0468	0.5327	0.6745	0.5751	0.8312
	15	0.0459	0.0416	0.0414	0.0427	0.6783	0.8257	0.7308	0.9423

Table 4.4: Level and power for the Spearman based tests: $k = 5, p = 3$, equal sample sizes and known peak location

		Level				Power			
Test	m	Normal	D-exp	Logistic	Exp	Normal	D-exp	Logistic	Exp
E	5	0.0437	0.0409	0.0391	0.0460	0.3187	0.4229	0.3540	0.5486
	6	0.0482	0.0465	0.0459	0.0434	0.3635	0.4853	0.4061	0.6310
	7	0.0410	0.0449	0.0449	0.0446	0.4155	0.5518	0.4837	0.7046
	8	0.0479	0.0430	0.0469	0.0484	0.4772	0.6016	0.5128	0.7632
	9	0.0445	0.0488	0.0489	0.0484	0.5138	0.6643	0.5683	0.8099
	10	0.0451	0.0489	0.0506	0.0481	0.5502	0.6947	0.6019	0.8553
	15	0.0484	0.0520	0.0494	0.0519	0.7075	0.8494	0.7566	0.9568
$E_{1>2}$	5	0.0476	0.0452	0.0434	0.0488	0.3327	0.4386	0.3685	0.5637
	6	0.0495	0.0497	0.0471	0.0450	0.3747	0.4958	0.4166	0.6420
	7	0.0433	0.0464	0.0459	0.0433	0.4198	0.5575	0.4875	0.7093
	8	0.0484	0.0435	0.0478	0.0492	0.4799	0.6060	0.5159	0.7630
	9	0.0441	0.0487	0.0489	0.0464	0.5148	0.6656	0.5656	0.8070
	10	0.0443	0.0487	0.0497	0.0464	0.5489	0.6918	0.6005	0.8544
	15	0.0466	0.0495	0.0468	0.0518	0.7009	0.8457	0.7503	0.9535
$E_{2>1}$	5	0.0315	0.0304	0.0297	0.0369	0.2661	0.3655	0.3016	0.4888
	6	0.0373	0.0345	0.0347	0.0344	0.3146	0.4362	0.3616	0.5821
	7	0.0317	0.0352	0.0342	0.0395	0.3673	0.5056	0.4377	0.6594
	8	0.0382	0.0351	0.0361	0.0411	0.4350	0.5590	0.4694	0.7236
	9	0.0366	0.0403	0.0391	0.0419	0.4777	0.6235	0.5252	0.7801
	10	0.0383	0.0403	0.0412	0.0433	0.5102	0.6591	0.5619	0.8279
	15	0.0400	0.0450	0.0414	0.0458	0.6771	0.8284	0.7303	0.9458

Table 4.5: Level and power for the Spearman based tests: $k = 5, p = 3$, nearly equal sample sizes and known peak location

		Level				Power			
Test	m	Normal	D-exp	Logistic	Exp	Normal	D-exp	Logistic	Exp
E	3	0.0417	0.0425	0.0438	0.0440	0.3519	0.2704	0.3799	0.6019
	4	0.0438	0.0426	0.0425	0.0454	0.4343	0.3320	0.4691	0.7106
	5	0.0423	0.0418	0.0389	0.0420	0.5070	0.3959	0.5571	0.8061
	6	0.0428	0.0412	0.0421	0.0484	0.5679	0.4433	0.6127	0.8708
	7	0.0460	0.0460	0.0412	0.0466	0.6183	0.4928	0.6843	0.9173
$E_{2>1}$	3	0.0434	0.0460	0.0464	0.0462	0.3647	0.2834	0.3927	0.6139
	4	0.0466	0.0450	0.0447	0.0455	0.4469	0.3425	0.4815	0.7246
	5	0.0466	0.0439	0.04411	0.0446	0.5210	0.4084	0.5698	0.8171
	6	0.0462	0.0434	0.0449	0.0490	0.5812	0.4572	0.6236	0.8782
	7	0.0483	0.0484	0.0439	0.0466	0.6320	0.5039	0.6969	0.9211
$E_{1>2}$	3	0.0311	0.0319	0.0323	0.0371	0.2995	0.2220	0.3235	0.5391
	4	0.0325	0.0308	0.0310	0.0355	0.3753	0.2791	0.4080	0.6593
	5	0.0307	0.0311	0.0276	0.0332	0.4461	0.3420	0.4951	0.7554
	6	0.0316	0.0302	0.0320	0.0388	0.5047	0.3887	0.5572	0.8291
	7	0.0352	0.0346	0.0320	0.0400	0.5624	0.4325	0.6227	0.8836

Table 4.6: Level and power for the Spearman based tests: $k = 5, p = 3$, unequal sample sizes and known peak location

4.2 Simulation results for unknown peak location

When the location of the peak is unknown and the sample sizes are equal in Table 4.7, we again find that the test based on the full extremal set E has an empirical level closer to the nominal level than either of the tests based on the partial extremal sets $E_{2>1}$ and $E_{1>2}$. We also see that similar to the known peak case, the test based on E has higher empirical power than the tests based on the partial extremal sets in most cases and the tests based on $E_{2>1}$and $E_{1>2}$ in turn have equal empirical levels and powers.

Unlike for the known peak case, when the sample sizes are unequal and the peak location is unknown in Tables 4.8 and 4.9, the test based on the partial extremal set $E_{2>1}$ did not maintain an empirical level as close to the nominal level as that of the test based on E. We instead find that the test based on $E_{2>1}$ tends to have an empirical level slightly lower than the test based on E. On the other hand, the test based on $E_{2>1}$ tends to outperform the test based on E in terms of empirical power, similar to when the peak location is known. The test based on the partial extremal set $E_{1>2}$ again has a much worse performance than the other two tests and also does not maintain an empirical level near the nominal level as well as the other two tests.

The difference in performance between the tests based on $E_{2>1}$ and E can once again be explained by the misalignment issue describes in Section 2.3. As when the peak location is known, the performances of the tests based on $E_{2>1}$ and E for unknown peak location differ to a greater extent when the underlying distribution is normal or logistic compared to when the underlying distribution is double-exponential or exponential. The difference in power across the different underlying densities considered follows the same pattern as the unstratified case (Alvo, 2008). We find that m influences the differences in empirical powers similar to the previous section. The similarity between the results for here and the last section is expected, as the tests for unknown peak location is constructed using the k possible tests for known peak location.

		Level				Power			
Test	m	Normal	D-exp	Logistic	Exp	Normal	D-exp	Logistic	Exp
E	5	0.0482	0.0517	0.0486	0.0501	0.3300	0.4546	0.3537	0.5675
	6	0.0487	0.0506	0.0455	0.0483	0.3788	0.5168	0.4307	0.6551
	7	0.0483	0.0515	0.0489	0.0455	0.4482	0.5754	0.4951	0.7324
	8	0.0483	0.0505	0.0494	0.0450	0.4890	0.6488	0.5443	0.8036
	9	0.0508	0.0508	0.0480	0.0468	0.5375	0.7017	0.5867	0.8461
	10	0.0476	0.0472	0.0470	0.0486	0.5862	0.7448	0.6369	0.8873
	15	0.0489	0.0471	0.0486	0.0511	0.7691	0.9045	0.8251	0.9792
$E_{1>2}$	5	0.0413	0.0435	0.0409	0.0481	0.3053	0.4165	0.3222	0.5405
	6	0.0416	0.0439	0.0386	0.0445	0.3464	0.4875	0.3990	0.6208
	7	0.0424	0.0441	0.0425	0.0429	0.4166	0.5527	0.4641	0.7052
	8	0.0440	0.0435	0.0415	0.0405	0.4553	0.6172	0.5159	0.7715
	9	0.0453	0.0441	0.0427	0.0436	0.5086	0.6698	0.5602	0.8271
	10	0.0411	0.0426	0.0390	0.0433	0.5592	0.7236	0.6004	0.8656
	15	0.0443	0.0407	0.0440	0.0473	0.7472	0.8920	0.8041	0.9730
$E_{2>1}$	5	0.0407	0.0423	0.0426	0.0445	0.3006	0.4172	0.3234	0.5353
	6	0.0416	0.0432	0.0390	0.0434	0.3475	0.4875	0.3927	0.6175
	7	0.0446	0.0432	0.0409	0.0413	0.4108	0.5460	0.4644	0.7019
	8	0.0435	0.0436	0.0420	0.0413	0.4622	0.6225	0.5151	0.7781
	9	0.0452	0.0435	0.0425	0.0427	0.5120	0.6757	0.5523	0.8235
	10	0.0405	0.0417	0.0407	0.0445	0.5552	0.7236	0.6066	0.8617
	15	0.0423	0.04396	0.0448	0.0460	0.7431	0.8899	0.8047	0.9731

Table 4.7: Level and power for the Spearman based tests: $k = 5, p = 3$, equal sample sizes and unknown peak location

		Level				Power			
Test	m	Normal	D-exp	Logistic	Exp	Normal	D-exp	Logistic	Exp
E	5	0.0436	0.0440	0.0433	0.0439	0.2966	0.4131	0.3326	0.5459
	6	0.0441	0.0469	0.0443	0.0461	0.3617	0.5045	0.4099	0.6467
	7	0.0454	0.0449	0.0477	0.0466	0.4309	0.5791	0.4798	0.7282
	8	0.0425	0.0453	0.0403	0.0480	0.4923	0.6482	0.5549	0.8002
	9	0.0462	0.0445	0.0439	0.0467	0.5538	0.7169	0.6047	0.8444
	10	0.0479	0.0466	0.0524	0.0486	0.5917	0.7627	0.6545	0.8975
	15	0.0469	0.0509	0.0506	0.0524	0.7818	0.9091	0.8350	0.9802
$E_{1>2}$	5	0.0409	0.0413	0.0423	0.0422	0.3189	0.4321	0.3513	0.5714
	6	0.0404	0.0435	0.0417	0.0451	0.3789	0.5145	0.4258	0.6592
	7	0.0425	0.0422	0.0454	0.0434	0.4306	0.5938	0.4829	0.7378
	8	0.0415	0.0417	0.0373	0.0452	0.4956	0.6531	0.5546	0.8006
	9	0.0426	0.0429	0.0424	0.0459	0.5525	0.7182	0.6069	0.8447
	10	0.0426	0.0449	0.0476	0.0463	0.5863	0.7555	0.6545	0.8924
	15	0.0424	0.0462	0.0469	0.0506	0.7710	0.9053	0.8313	0.9807
$E_{2>1}$	5	0.0329	0.0348	0.0345	0.0367	0.2308	0.3265	0.2572	0.4627
	6	0.0349	0.0376	0.0346	0.0376	0.2993	0.4203	0.3404	0.5678
	7	0.0361	0.0371	0.0385	0.0413	0.3587	0.5114	0.4068	0.6632
	8	0.0358	0.0377	0.0323	0.0432	0.4248	0.5824	0.4847	0.7464
	9	0.0388	0.0384	0.0381	0.0417	0.4916	0.6594	0.5467	0.8084
	10	0.0384	0.0413	0.0423	0.0455	0.5330	0.7184	0.5954	0.8684
	15	0.0406	0.0449	0.0448	0.0458	0.7393	0.8876	0.7991	0.9725

Table 4.8: Level and power for the Spearman based tests: $k = 5, p = 3$, nearly equal sample sizes and unknown peak location

		Level				Power			
Test	m	Normal	D-exp	Logistic	Exp	Normal	D-exp	Logistic	Exp
E	3	0.0556	0.0542	0.0518	0.0493	0.3534	0.4923	0.4164	0.6340
	4	0.0531	0.0525	0.0493	0.0546	0.4608	0.6129	0.5073	0.7557
	5	0.0431	0.0395	0.0452	0.0435	0.5469	0.7099	0.5984	0.8585
	6	0.0502	0.0500	0.0473	0.0484	0.6270	0.7847	0.6925	0.9182
	7	0.0459	0.0467	0.0488	0.0479	0.6991	0.8533	0.7455	0.9492
$E_{1>2}$	3	0.0558	0.0527	0.0499	0.0487	0.3794	0.5130	0.4357	0.6528
	4	0.0524	0.0504	0.0478	0.0537	0.4788	0.6322	0.5313	0.7768
	5	0.0524	0.0418	0.0433	0.0437	0.5670	0.7246	0.6221	0.8707
	6	0.0486	0.0493	0.0454	0.0452	0.6474	0.8023	0.7080	0.9280
	7	0.0440	0.0456	0.0457	0.0461	0.7187	0.8593	0.7655	0.9534
$E_{2>1}$	3	0.0455	0.0442	0.0417	0.0409	0.2835	0.4063	0.3258	0.5529
	4	0.0402	0.0407	0.0397	0.0432	0.3840	0.5322	0.4255	0.6892
	5	0.0356	0.0324	0.0334	0.0364	0.4617	0.6275	0.5118	0.8051
	6	0.0421	0.0375	0.0382	0.0400	0.5464	0.7189	0.6087	0.8826
	7	0.0374	0.0380	0.0382	0.0404	0.6165	0.7950	0.6695	0.9217

Table 4.9: Level and power for the Spearman based tests: $k = 5, p = 3$, unequal sample sizes and unknown peak location

Chapter 5

Estimators for common stratum treatment effect in stratified two-sample location problem

5.1 Introduction

This chapter discusses nonparametric estimators for the common stratum treatment effect in the two-treatment, $s-$strata problem with the assumption of parallelism. We propose the novel application of these estimators to estimate the misalignment issue central to the thesis. These estimators are typically associated with the tests discussed in Chapter 1 both in construction and application. Alternatively, there have been nonparametric hypothesis tests for the null hypothesis of equal treatment effects against the alternative of unequal treatment effects (i.e. Dai and Stern, 2022). The estimators discussed in this chapter would be of interest in the tests for equal treatment effects if the test does not reject the null hypothesis.

In the unstratified case, a common approach is to use estimators proposed by Hodges and Lehmann (1963) based on a rank test $h(X_1, X_2)$, where X_1 and X_2 refer to the first and second sample is given as follows:

Define

$$\Delta^*(X_1, X_2) = sup\{\Delta : h(X_1, X_2 - \Delta) > \mathbb{E}[h(X_1, X_2)|H_0]\} \tag{5.1.1}$$

and

$$\Delta^{**}(X_1, X_2) = inf\{\Delta : h(X_1, X_2 - \Delta) < \mathbb{E}[h(X_1, X_2)|H_0]\}. \tag{5.1.2}$$

The estimator is defined as

$$\hat{\Delta}(X_1, X_2) = [\Delta^*(X_1, X_2) + \Delta^{**}(X_1, X_2)]/2. \tag{5.1.3}$$

The definition above shows the similar alignment idea that the same authors proposed for the aligned-rank test around the same time. A more convenient form of the estimator when h is the Mann-Whitney-Wilcoxon test $U(\cdot, \cdot)$ given in Chapter 1 was shown to be

$$\hat{\Delta}(X_1, X_2) = med(V), \tag{5.1.4}$$

where V is the set of all pairwise differences $X_{1(i)} - X_{2(j)}$ for $i = 1, ..., m_1$, $j = 1, ..., m_2$. We will refer to this statistic as the two-sample Hodges-Lehmann estimator. The one-sample Hodges-Lehmann estimator used to align the strata in Chapter 2 is derived in a similar way based on the signed-rank test of Wilcoxon (1945).

In the remainder of this chapter, we present three estimators which are available in closed form and provide a small set of simulation results to study their performances. We then review bootstrap confidence intervals in the context of analyzing the misalignment issue central to this thesis.

5.2 Estimators for stratified samples

From the fact that both the one and two sample unstratified Hodges-Lehmann estimators have closed-form expressions, it follows that we can easily obtain a closed form estimator related the aligned-rank test described in Chapter 1. The estimator is given as

$$\hat{\Delta}_{Aligned}(X_1, X_2) = \hat{\Delta}(X_1^*, X_2^*), \tag{5.2.1}$$

where for treatment group $q = 1, 2$,

$$X_q^* = \{X_q^{(i)} - \theta_i,\ i = 1, ..., s\} = \{X_{ql}^{(i)} - \theta_i,\ i = 1, ..., s,\ l = 1, ..., m_{iq},\}. \quad (5.2.2)$$

are the aligned observations in treatment group q and θ_i is the one-sample Hodges-Lehmann estimator calculated from all observations of the i^{th} stratum for $i = 1, ..., s$ given by (1.0.17).

We note that this is not the Hodges-Lehmann type estimator for the aligned-rank test, but rather the Hodges-Lehmann type estimator of the unstratified data post-alignment. The true Hodges-Lehmann type estimator could be approximated through numerical methods, but this would be computationally expensive when combined with the bootstrap approach discussed in the next section and is ignored for our purposes.

Rashid (2003) proposed a general method which involves minimizing the sum of within-stratum dispersion functions. In the case of two treatment groups, the estimator simplifies to the weighted average of within-stratum two-sample Hodges-Lehmann estimators. For large sample sizes, the weights simplify and the estimator is given by

$$\hat{\Delta}_{Rashid}(X_1, X_2) = [\sum_{j=1}^{s} \frac{m_{j1}m_{j2}}{m_{j1} + m_{j2}} \hat{\Delta}(X_1^{(j)}, X_2^{(j)})]/[\sum_{j=1}^{s} \frac{m_{j1}m_{j2}}{m_{j1} + m_{j2}}]. \quad (5.2.3)$$

The author noted a motivation for this method being the lack of literature on a Hodges-Lehmann type estimator based on the van Elteren test. More recently, the Hodges-Lehmann type estimator based on the van Elteren test was discussed by O'Kelly (2003) as well as by Mehrotra, Lu, and Li (2010). However, in both papers it was explicitly stated that the estimator is obtained using numerical methods. We derive a closed-form version of the estimator in the Appendix, which to our knowledge is novel. The estimator is calculated as follows:

Step 1. Generate sets $V_i, i = 1, ..., s$ of all the pairwise differences $X_{1(j)}^{(i)} - X_{2(l)}^{(i)},\ j = 1, ..., m_{i1},\ k = 1, ..., m_{i2}$, similar to the two-sample Hodges-Lehmann estimator.

Step 2. Produce $r_i = \prod_{j \neq i}(m_{j1} + m_{j2} + 1)$ replicates of V_i for $i = 1, ..., s$.

Step 3. Define V as the set of length $\sum_{i=1}^{k} m_{i1}m_{i2}r_i$ comprising of all replicates for all the strata.

Step 4. The estimator $\hat{\Delta}_{vE}(X_1, X_2)$ is defined to be the median of V.

Instead of producing replicates, the estimator can equivalently be expressed as the weighted median of the pooled within-stratum pairwise differences with weights

$$w_i = \frac{r_i}{\sum_{j=1}^{s} m_{j1} m_{j2} r_j}, \ i = 1, ..., s \tag{5.2.4}$$

for each of the within-stratum pairwise differences from stratum i. Weighted medians have been used as robust alternatives to weighted averages in multiple statistical contexts in recent literature, including genetics (Bowden et al., 2016) and meta-analysis (Hartwig et al. , 2020).

The following tables exhibit the mean squared errors of the three estimators under the same two-strata, two-treatment scenarios seen in Table 2.1.

Sample sizes				Mean squared error			
m_{11}	m_{12}	m_{21}	m_{22}	Normal	D-exp	Logistic	Exp
15	15	15	15	0.072	0.048	0.063	0.030
14	16	16	14	0.072	0.048	0.063	0.030
13	17	17	13	0.073	0.048	0.064	0.030
12	18	18	12	0.075	0.050	0.066	0.031
11	19	19	11	0.077	0.052	0.067	0.033
10	20	20	10	0.080	0.054	0.071	0.034

Table 5.1: Mean squared error for the van Elteren Hodges-Lehmann estimator

Sample sizes				Mean squared error			
m_{11}	m_{12}	m_{21}	m_{22}	Normal	D-exp	Logistic	Exp
15	15	15	15	0.142	0.098	0.126	0.067
14	16	16	14	0.144	0.099	0.126	0.067
13	17	17	13	0.142	0.098	0.125	0.067
12	18	18	12	0.143	0.098	0.126	0.066
11	19	19	11	0.143	0.098	0.125	0.067
10	20	20	10	0.142	0.098	0.126	0.067

Table 5.2: Mean squared error for Rashid's estimator

Sample sizes				Mean squared error			
m_{11}	m_{12}	m_{21}	m_{22}	Normal	D-exp	Logistic	Exp
15	15	15	15	0.071	0.048	0.062	0.032
14	16	16	14	0.071	0.048	0.063	0.033
13	17	17	13	0.071	0.049	0.063	0.037
12	18	18	12	0.076	0.056	0.069	0.046
11	19	19	11	0.091	0.074	0.084	0.067
10	20	20	10	0.123	0.112	0.119	0.109

Table 5.3: Mean squared error for aligned-rank Hodges-Lehmann estimator

While no detailed comparison of these estimators exists in the literature more generally, we are specifically interested in their performance as a measure of misalignment between two strata. From Table 5.2, we see that Rashid's estimator had a higher mean squared error than the other two estimators in practically all of the situations, despite being the most robust to highly unbalanced sample sizes. Comparing Tables 5.1 and 5.3 show the aligned-rank based estimator had a lower mean squared error than the van Elteren based estimator for equal sample sizes, but a higher mean squared error for unequal sample sizes. This speaks to the misalignment issue discussed in Section 2.3 from a different perspective. For this reason, the van Elteren based estimator is used for our purposes, as the misalignment issue only occurs for unequal sample sizes.

5.3 Bootstrap confidence intervals to diagnose misalignment

Besides the typical applications, the estimators reviewed in the previous section can be used to provide a useful diagnostic tool for measuring the misalignment of two strata in aligned-rank tests as described in Chapter 2. When using the estimators for this purpose, the treatment and stratum labels for each observation are interchanged after the alignment step. Additionally, we can estimate the sampling distribution of the misalignment estimator by applying resampling methods before the alignment step. We describe the bootstrap approach that was first introduced by Efron (1979) adjusted to our context.

Step 1. For each bootstrap replicate, indexed by $b = 1, ..., B$:

(a) Generate bootstrap sample of size n by sampling with replacement m_{ij} observations from the original m_{ij} observations within each combination of strata $i = 1, 2$ and treatment group $j = 1, ..., k$.

(b) Perform the alignment step.

(c) Interchange the stratum and treatment group information for the full dataset.

(d) Compute the b^{th} replicate $\hat{\Delta}^{(b)}$ from the b^{th} bootstrap sample.

Step 2. The bootstrap estimate of $F_{\hat{\Delta}}(\cdot)$ is the empirical distribution of the replicates $\hat{\Delta}^{(1)}, ..., \hat{\Delta}^{(B)}$.

Confidence intervals can be obtained from this sample of bootstrapped estimates. While there are many ways of constructing confidence intervals from the sample of bootstrapped estimates, here we review the well known percentile confidence interval as well as the "bias corrected" and "adjusted for acceleration" (BCa) confidence interval, both developed by Efron in a series of articles (1981,1982,1985,1987).

Suppose that $\hat{\Delta}_{(1)}, ..., \hat{\Delta}_{(B)}$ are the ordered bootstrap replicates of the statistic $\hat{\Delta}$. From the empirical distribution function of the replicates, compute the $\alpha/2$ quantile $\hat{\Delta}_{(\alpha/2)}$, and the $1-\alpha/2$ quantile $\hat{\Delta}_{(1-\alpha/2)}$. The $(1-\alpha)*100\%$ percentile bootstrapped confidence interval is simply given as

$$(\hat{\Delta}_{(\alpha/2)}, \hat{\Delta}_{(1-\alpha/2)}). \tag{5.3.1}$$

The BCa bootstrap confidence interval modifies the percentile bootstrap confi-

dence interval and has been shown to have better theoretical properties and performance in simulations. The bias correction factor is computed by

$$\hat{z}_0 = \Phi^{-1}(\frac{1}{B}\sum_{b=1}^{B} I(\hat{\Delta}_{(b)} < \hat{\Delta})), \tag{5.3.2}$$

where Φ is the distribution function of the standard normal distribution and $\hat{\Delta}$ is the estimate from the original sample. The acceleration (or skewness) factor is

$$\hat{a} = \frac{\sum_{i=1}^{n}(\bar{\Delta}^*_{(.)} - \Delta^*_{(i)})^3}{6[\sum_{i=1}^{n}(\bar{\Delta}^*_{(.)} - \Delta^*_{(i)})^2]^{3/2}} \tag{5.3.3}$$

where $\Delta^*_{(1)}, ..., \Delta^*_{(N)}$ are the jackknife or leave-one-out replicates and

$$\bar{\Delta}^*_{(.)} = \frac{1}{n}\sum_{i=1}^{n} \Delta^*_{(i)}.$$

It is important to note that as with the bootstrap resampling, the removal of the observation should be done prior to the alignment step when obtaining jackknife replicates. The $(1-\alpha) * 100\%$ BCa confidence interval is given by

$$(\hat{\Delta}_{(\alpha_1)}, \hat{\Delta}_{(\alpha_2)}), \tag{5.3.4}$$

where

$$\alpha_1 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{\alpha/2})}), \tag{5.3.5}$$

$$\alpha_2 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z_{1-\alpha/2}}{1 - \hat{a}(\hat{z}_0 + z_{1-\alpha/2})}), \tag{5.3.6}$$

and $z_\alpha = \Phi^{-1}(\alpha)$.

Chapter 6

Illustration on IQ data for coma patients

In this section, we illustrate the tests from Chapter 2 and the diagnostic methods from Chapter 4 on the real dataset of Wong, Monette, and Weiner (2001). The dataset contains information on 200 patients who sustained traumatic brain injuries resulting in comas of varying duration. After awakening from their comas, patients were periodically administered a standard I.Q. test. A score for mathematical I.Q. and another for verbal I.Q. were provided, with the average of the two scores being used in our analysis. Furthermore, we remove the data corresponding to follow-up tests some patients participated in.

We stratify by coma duration to demonstrate our methods. We include a short coma duration (< 7 days) stratum and a long coma duration (≥ 7 days) stratum. The total sample sizes of the strata are nearly equal, with 99 and 101 patients in the short and long coma duration strata, respectively. The sample sizes for each combination of coma duration group and age group are given in Table 6.1, while Table 6.2 includes the proportion of observations in each stratum that belong to different age groups.

Coma Duration \ Age Group	≤ 20	$(20, 30]$	$(30, 40]$	$(40, 55]$	> 55
Short	13	32	22	17	15
Long	22	47	12	13	7

Table 6.1: Sample sizes of the Wong Dataset

Coma Duration \ Age Group	≤ 20	$(20, 30]$	$(30, 40]$	$(40, 55]$	> 55
Short	0.13131	0.32323	0.22222	0.17172	0.151515
Long	0.21782	0.46535	0.11881	0.12871	0.069307

Table 6.2: Proportion of observations within each stratum across age groups

We see that the long coma duration stratum has a much higher proportion of patients in the ≤ 20 and $[20, 30)$ age groups and a much lower proportion in the other age groups compared to the short coma duration stratum. Under the alternative of a strict umbrella pattern, this imbalance in proportions in the younger age groups leads to a relatively lowered alignment estimate for the long coma duration stratum. The imbalance would result in the long coma duration stratum being systematically below the short coma duration stratum after the alignment step.

The one-sample Hodges-Lehmann alignment estimate for the short coma duration stratum was calculated to be 90.00004 while for the long coma duration stratum it was found to be 85.00005. Figure 6.1 shows the data after the alignment step.

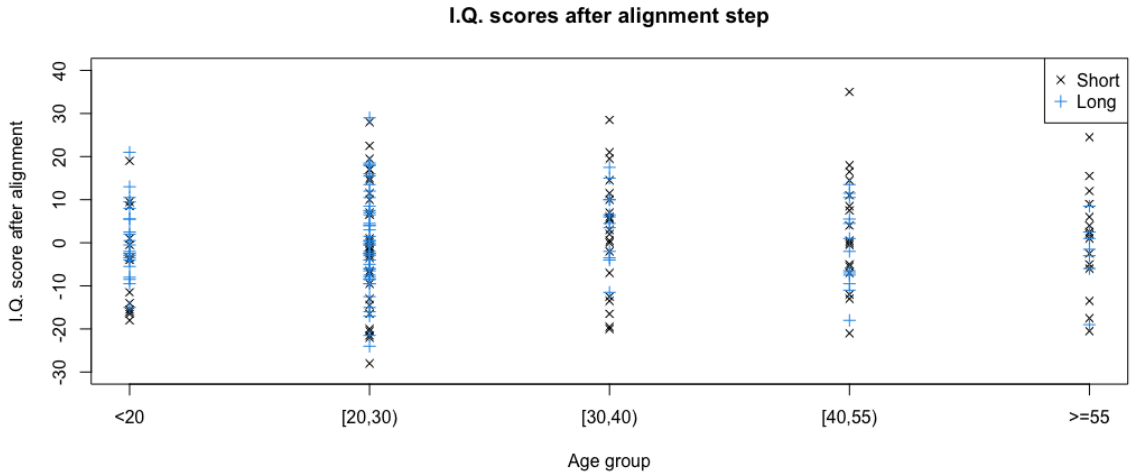

Figure 6.1: I.Q. scores of Wong dataset after alignment

The p-values for the unknown peak location tests described in Chapter 2 are 0.179 for the full extremal set, 0.177 for the partial extremal set with Short $>$ Long and 0.236 for the partial extremal set with Short $<$ Long. On the other hand, the p-values for the known peak tests with the peak at age group $[30, 40)$ are 0.047 for the full extremal set, 0.048 for the partial extremal set with Short $>$ Long and 0.067 for the partial extremal set with Short $<$ Long.

The van Elteren Hodges Lehmann estimate of misalignment calculated with the original data was found to be 0.9999918. To further understand the misalignment, we use the bootstrap approach discussed in Chapter 4. Figure 6.2 displays a histogram of the sample of bootstrap replicates. The variance of the sample of bootstrap replicates is approximately 0.513, while the skewness and excess kurtosis are approximately 0.279 and 0.215, respectively.

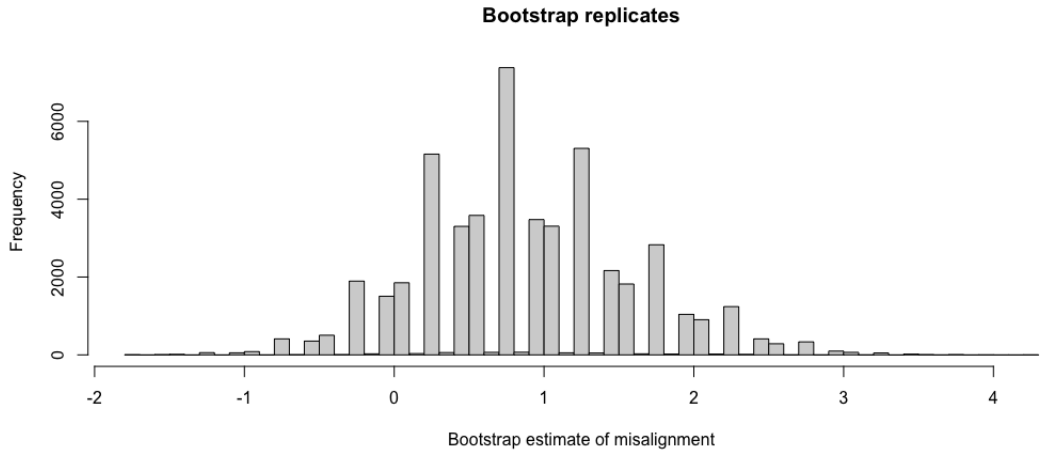

Figure 6.2: Histogram of bootstrap replicates

Based on 50000 bootstrap replicates, the 90% percentile confidence interval was calculated to be $[-0.249984, 2.249886]$. On the other hand, the 90% BCa bootstrap confidence interval was calculated to be $[0.000043, 2.499991]$. This speaks to the misalignment being nontrivial and further supports the use of the partial extremal set to account for the misalignment.

Chapter 7

Discussion

7.1 Conclusion

This thesis extended to stratified samples the general theory of hypothesis testing based on rankings by Alvo and Pan (1997). We pursue this extension by first aligning the data as is done in the aligned-rank test of Hodges and Lehmann (1962). Our extension leads to both the traditional aligned-rank tests while also providing new statistics which account for possible misalignments seen for unequal sample sizes. We focus on umbrella alternatives, which contain ordered alternatives as special cases. Our simulation study results show that when chosen correctly, the new statistics can offer an advantage over the traditional aligned-rank approach when the peak of the umbrella pattern is both known and unknown. We also provided estimators and confidence intervals which can be used to assess the misalignment issue and illustrated these diagnostic tools alongside the proposed tests on I.Q. data of coma patients.

Tests constructed here and more broadly through the general theory of Alvo and Pan (1997) are based on extremal sets corresponding to the alternative hypothesis. Non-null models like these allow for tests addressing a variety of alternative hypotheses to be constructed. This thesis presents a novel advantage of taking a non-null approach to hypothesis testing to account for issues related to pre-testing transformations on the data. In our context, the transformation is the alignment of the strata prior to testing, which we have shown only works correctly for equal sample sizes. We conclude the thesis by discussing future directions for research, which include extensions to more than two strata, adaptive tests to address the subjective choice

of extremal set, and the construction of tests robust to various umbrella shapes and underlying distributions.

7.2 Future research directions

7.2.1 Three or more strata

The methods from Chapter 2 to this point have been restricted to two strata, which is often not the case in practice. The way we have extended the general theory of hypothesis testing by Alvo and Pan (1997) easily accounts for the inclusion of more than two strata. For s strata, we would have $s!$ partial extremal sets corresponding to the $s!$ possible orderings of the strata inside each treatment group. In the cases of the Spearman and Kendall distances, it can easily be shown that the sum of the tests based on all partial extremal sets again results in the traditional aligned-rank test. However, with three or more strata, it may not be obvious which ordering is appropriate based solely on the sample sizes. This issue will be addressed in the next section.

We can also obtain $s!$ misalignment values from the s strata by considering each pairwise misalignment value. These pairwise misalignment values could be analyzed individually and bootstrap confidence intervals could be constructed. The multiple comparisons problem would exist if any type of inference is being made and the level of the confidence interval would need to be adjusted accordingly. Other quantities of interest could be the average and maximum misalignment values and their resulting bootstrap confidence intervals.

Another possible direction would be to obtain "coherent" misalignment estimates, satisfying

$$\hat{\Delta}_{12} + \hat{\Delta}_{13} = \hat{\Delta}_{23}. \tag{7.2.1}$$

This concept is used in network meta-analysis, with a goal of including both direct and indirect comparisons between treatments in an interpretable way (Dias et al., 2018). We could follow a similar approach as network meta-analysis, treating the strata as treatments and the misalignment estimates as the direct comparisons in the network. As in network meta-analysis, obtaining coherent estimates would allow for

an estimate of the strata ordering after alignment, which will be useful in the next section on adaptive tests. Additionally, the coherent estimates satisfy the triangle inequality and therefore the matrix

$$\begin{bmatrix} 0 & |\hat{\Delta}_{12}| & ... & |\hat{\Delta}_{1s}| \\ |\hat{\Delta}_{21}| & 0 & ... & ... \\ ... & ... & ... & ... \\ |\hat{\Delta}_{s1}| & ... & ... & 0 \end{bmatrix} \tag{7.2.2}$$

is a distance matrix. This fact would allow for further exploration through multivariate methods such as hierarchical clustering and heat maps.

7.2.2 Adaptive tests

To address the subjectivity of deciding between the partial and full extremal sets in Chapter 2, so called adaptive tests could be developed in the future. Adaptive tests make use of the data itself to adjust the test procedure. The idea when first proposed in papers by Randles and Hogg (1973), Hogg (1974), and Hogg, Fisher, and Randles (1975) involves a discrete set of possible procedures which are chosen based on a "selector" statistic calculated from the data. There are also continuously adaptive tests, which have a continuous mapping of the selector statistic to a continuum of possible procedures (Ruberg,1986). Another continuously adaptive set of tests was proposed more recently by Alvo (2021), who used the Pearson system of equations to estimate a locally most powerful score function based on pooled sample skewness and traditional kurtosis.

Two interesting observations were made by Neuhäuser and Hothorn (2005). Firstly, that maximum-type tests as seen in this thesis are a special case of adaptive permutation tests. Secondly, by performing adaptive tests as permutation tests and calculating a separate selector statistic for each permutation, both the independence between the selector statistic and the possible test statistics as well as the continuousness of the underlying distribution is no longer necessary. This is useful in our context, as both the misalignment estimator and the procedures depend on the alignment step.

For the case of two strata, one possible way of pursuing a continuously adaptive

approach would be to use the bootstrap sample of misalignment values discussed in Section 3.3. The adaptive statistic would be defined as

$$S_p^{adapt} = w_{(1>2)} S_{p(1>2)} + w_{(2>1)} S_{p(2>1)}, \tag{7.2.3}$$

with weights

$$w_{(1>2)} = \frac{1}{B} \sum_{i=1}^{B} [\mathbf{1}(\hat{\Delta}_{(i)} < 0) + \frac{1}{2} \mathbf{1}(\hat{\Delta}_{(i)} = 0)] \tag{7.2.4}$$

and

$$w_{(2>1)} = \frac{1}{B} \sum_{i=1}^{B} [\mathbf{1}(\hat{\Delta}_{(i)} > 0) + \frac{1}{2} \mathbf{1}(\hat{\Delta}_{(i)} = 0)] \tag{7.2.5}$$

which is the sample proportion of the bootstrap replicates indicating a specific ordering. This construction would result in the use of a test similar to the traditional aligned-rank test when there is little misalignment in the data, while putting most of the weight on one of the partial extremal sets in the presence of misalignment.

The use of coherent estimates and their resulting ranking of the strata as discussed in the previous section would allow us to generalize this idea to more than two strata. A tie in the rankings from a bootstrapped set of coherent estimates could contribute equally to all statistics which do not contradict any of the remaining inequalities. As an example, a ranking

$$\text{Stratum } 3 > \text{Stratum } 1 = \text{Stratum } 4 > \text{Stratum } 2 \tag{7.2.6}$$

would contribute equal weight to the statistics $S_{p(3>1>4>2)}$ and $S_{p(3>4>1>2)}$.

Appendix: Derivation of Hodges-Lehmann estimator of van Elteren test

Suppose first that $\sum_{i=1}^{k} m_{i1}m_{i2}r_i$ is odd, say $\sum_{i=1}^{s} m_{i1}m_{i2}r_i = 2z+1$ for some integer z and let $V^{(q)}$ be the $q-$th smallest element of the set V. Here we express the van Elteren test using an alternative definition of the Mann-Whitney-Wilcoxon statistic

$$U_i(X_1^{(i)}, X_2^{(i)}) = \sum_{j=1}^{m_{i1}} \sum_{l=1}^{m_{i2}} \mathbf{1}(X_{1j}^{(i)} > X_{1l}^{(i)}), \tag{7.2.7}$$

where

$$\mathbf{1}(X_{1j}^{(i)} > X_{1l}^{(i)}) = \begin{cases} 1, & \text{if } X_{1j}^{(i)} > X_{1l}^{(i)} \\ 0, & \text{else.} \end{cases} \tag{7.2.8}$$

$$
\begin{aligned}
\Delta^{**} =& inf(\Delta : \sum_{i=1}^{s} \frac{U_i(X_1^{(i)}, X_2^{(i)} - \Delta)}{m_{i1} + m_{i2} + 1} < \sum_{i=1}^{s} \frac{m_{i1} m_{i2}}{2(m_{i1} + m_{i2} + 1)}) \\
=& inf(\Delta : \sum_{i=1}^{s} \frac{U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1))}{(\prod_{j=1}^{s}(m_{j1} + m_{j2} + 1))} < \sum_{i=1}^{s} \frac{m_{i1} m_{i2}}{2(m_{i1} + m_{i2} + 1)}) \\
=& inf(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1)) < \sum_{i=1}^{s} \frac{m_{i1} m_{i2}(\prod_{j=1}^{s}(m_{j1} + m_{j2} + 1))}{2(m_{i1} + m_{i2} + 1)}) \\
=& inf(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1)) < \frac{1}{2} \sum_{i=1}^{s} m_{i1} m_{i2}(\prod_{j \neq i} m_{j1} + m_{j2} + 1)) \\
=& inf(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1)) < z + \frac{1}{2}) \\
=& V^{(z+1)}.
\end{aligned}
\tag{7.2.9}
$$

Similarly,

$$
\begin{aligned}
\Delta^{*} =& sup(\Delta : \sum_{i=1}^{s} \frac{U_i(X_1^{(i)}, X_2^{(i)} - \Delta)}{m_{i1} + m_{i2} + 1} > \sum_{i=1}^{s} \frac{m_{i1} m_{i2}}{2(m_{i1} + m_{i2} + 1)}) \\
=& sup(\Delta : \sum_{i=1}^{s} \frac{U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1))}{(\prod_{j=1}^{s}(m_{j1} + m_{j2} + 1))} > \sum_{i=1}^{s} \frac{m_{i1} m_{i2}}{2(m_{i1} + m_{i2} + 1)}) \\
=& sup(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1)) > \sum_{i=1}^{s} \frac{m_{i1} m_{i2}(\prod_{j=1}^{s}(m_{j1} + m_{j2} + 1))}{2(m_{i1} + m_{i2} + 1)}) \\
=& sup(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1)) > \frac{1}{2} \sum_{i=1}^{s} m_{i1} m_{i2}(\prod_{j \neq i} m_{j1} + m_{j2} + 1)) \\
=& sup(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j \neq i}(m_{j1} + m_{j2} + 1)) > z + \frac{1}{2}) \\
=& V^{(z+1)}
\end{aligned}
\tag{7.2.10}
$$

.

Thus, $\hat{\Delta} = \frac{V^{(z+1)} + V^{(z+1)}}{2} = V^{(z+1)}$.

On the other hand, if $\sum_{i=1}^{k} m_{i1} m_{i2} r_i$ is even, say $\sum_{i=1}^{s} m_{i1} m_{i2} r_i = 2z$ for some

integer z, the last two lines of the above derivations are now

$$\begin{aligned}\Delta^{**} =& ... \\ =& inf(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j\neq i}(m_{j1} + m_{j2} + 1)) < \frac{1}{2}\sum_{i=1}^{s} m_{i1}m_{i2}(\prod_{j\neq i} m_{j1} + m_{j2} + 1)) \\ =& inf(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j\neq i}(m_{j1} + m_{j2} + 1)) < z) \\ =& V^{(z+1)}\end{aligned} \tag{7.2.11}$$

and

$$\begin{aligned}\Delta^{*} =& ... \\ =& sup(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j\neq i}(m_{j1} + m_{j2} + 1)) > \frac{1}{2}\sum_{i=1}^{s} m_{i1}m_{i2}(\prod_{j\neq i} m_{j1} + m_{j2} + 1)) \\ =& sup(\Delta : \sum_{i=1}^{s} U_i(X_1^{(i)}, X_2^{(i)} - \Delta)(\prod_{j\neq i}(m_{j1} + m_{j2} + 1)) > z) \\ =& V^{(z)}\end{aligned} \tag{7.2.12}$$

.

Consequently $\hat{\Delta} = \frac{V^{(z)}+V^{(z+1)}}{2}$.

In both cases,

$$\hat{\Delta} = med(V). \tag{7.2.13}$$

Bibliography

[1] Alvo, M. and Cabilio, P. (1995). Testing Ordered Alternatives in the Presence of Incomplete Data. *Journal of the American Statistical Association*, 90, 1015-1024.

[2] Alvo, M. and Pan, J. (1997). A General Theory of Hypothesis Testing Based on Rankings. *Journal of Statistical Planning and Inference*, 61, 219-248.

[3] Alvo, M. and Charbonneau, M. (1997). The use of Spearman's Footrule in testing for trend when data is incomplete. *Communications in Statistics-Simulation*, 26, 193-213.

[4] Alvo, M. and Cabilio, P. (2005). General Scores Statistics on Ranks in the Analysis of Unbalanced Designs. *The Canadian Journal of Statistics*, 33, 115-129.

[5] Alvo, M. (2008). Nonparametric tests of hypotheses for umbrella alternatives. *Canadian Journal of Statistics*, 36. 143-156.

[6] Alvo, M. (2023) Empirical Bayes on a shoestring and other applications, *Communications in Statistics - Theory and Methods*, 52 (7), 2228-2239,

[7] Barlow, R.E., Bartholomew, D.J., Bremner, J.M., and Brunk, H.D. (1972). *Statistical Inference under Order Restrictions*. Wiley and Sons, N.Y.

[8] Bowden, J., Davey Smith, G., Haycock, P.C., and Burgess, S. (2016). Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.*, 40, 304-314.

[9] Cabilio, P. and Tilley, J. (1999). Power Calculations for Tests of Trend with Missing Observations. *Environmetrics*, 10, 803-816.

[10] Chen, Y.I. and Wolfe, D.A. (1990). A Study of Distribution-free Tests for Umbrella Alternatives. *Biometrical Journal*, 32, 47-57.

[11] Chen, Y.I. (1991). Notes on the Mack-Wolfe and Chen-Wolfe Tests for Umbrella Alternatives. *Biometrical Journal*, 33, 281-290.

[12] Critchlow, D. (1992). On Rank Statistics: An Approach via Metrics on the Permutation Group. *Journal of Statistical Planning and Inference*, 32, 325-346.

[13] Dai, M. and Stern, H.S. (2022). A U-statistic-based test of treatment effect heterogeneity. *Journal of Nonparametric Statistics*, 34(1), 141-163.

[14] Daniels, H.E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society B*, 12, 171-181.

[15] Dias, S., Ades, A.E., Welton, N.J., Jansen, J.P., and Sutton, A.J. (2018). Network meta-analysis for decision-making. John Wiley & Sons.

[16] Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7, 1-26

[17] Efron, B. (1981). Non-parametric standard errors and confidence intervals. *Canadian Journal of Statistics,* 9, 139-172.

[18] Efron, B. (1982). The jackknife, the bootstrap, and other resampling plans. In Regional Conference Series in Applied Mathematics, No. 38 (pp. 92-127). Philadelphia: SIAM.

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

[20] Efron, B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association,* 82, 171-185.

[21] Feller, W. (1968). Probability Theory. New York: Wiley and Sons.

[22] Hájek, J., Šidák, Z., Sen, P.K. (1999). Theory of Rank Tests (2nd ed.). New York: Academic Press.

[23] Hartwig, F. P., Davey Smith, G., Schmidt, A. F., Sterne, J. A. C., Higgins, J. P. T., and Bowden, J. (2020). The median and the mode as robust meta-analysis estimators in the presence of small-study effects and outliers. *Research Synthesis Methods,* 11(3), 397-412.

[24] Hettmansperger, T. P. and Norton, R. M. (1987). Tests for Patterned Alternatives in k-Sample Problems. *Journal of the American Statistical Association,* 82, 292-299.

[25] Hodges, J.L., and Lehmann, E.C. (1962). Rank Methods for Combination of Independent Experiments in the Analysis of Variance. *Annals of Mathematical Statistics,* 33(2), 482-497.

[26] Hodges, J.L., and Lehmann, E.C. (1963). Estimates of Location Based on Rank Tests. *Annals of Mathematical Statistics,* 34(2), 598-611

[27] Hogg, R. V. (1974). Adaptive robust procedures: a partial review and some suggestions for future applications and theory. *Journal of the American Statistical Association,* 69, 909-927.

[28] Hogg, R. V., Fisher, D. M., and Randles, R. H. (1975). A two-sample adaptive distribution-free test. *Journal of the American Statistical Association,* 70, 656-661.

[29] Jonckheere, A. R. (1954). A test of significance for the relation between m rankings and k ranked categories. *British Journal of Statistical Psychology,* 7, 93-100.

[30] Kössler, W. and Buning, H. (2000). The asymptotic power and relative efficiency of some c-Sample rank tests of homogeneity against umbrella alternatives. *Statistics,* 34, 1-26.

[31] Kössler, W. (2006). Some c-Sample rank tests of homogeneity against umbrella alternatives with unknown peak. *Journal of Statistical Computation and Simulation,* 76, 57-74.

[32] Mack, G. A. and Wolfe, D. A. (1981). k-Sample rank tests for umbrella alternatives. *Journal of the American Statistical Association,* 76, 175-181.

[33] Mann, H.B. (1945). Nonparametric tests against trend. *Econometrica*, 13, 245-259.

[34] Mann, H.B. and Whitney, D.R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *Annals of Mathematical Statistics*, 18 (1), 50-60.

[35] Mehrotra, D. V., Lu, X., and Li, X. (2010). Rank-based analyses of stratified experiments: Alternatives to the van Elteren. *The American Statistician,* 64, 121-130.

[36] Millen, B.A. and Wolfe, D.A. (2005). A Class of Nonparametric Tests for Umbrella Alternatives. *Journal of Statistical Research*, 39, 7-24.

[37] Miller, G.I. (1990). Living in the Environment: An Introduction to Environmental Science (6th ed.). Wadsworth.

[38] O'Kelly, M. (2003). Calculating Estimates of an Effect in Stratified Nonparametric Analysis. In A. Rizzi, M. Vichi, and H. Locarek-Junge (Eds.), *18th International Workshop on Statistical Modelling* (pp. 349-352).

[39] Neuhäuser, M., Seidel, D., Hothorn, L.A., and Urfer, W. (2000). Robust trend tests with application to toxicology. *Environmental and Ecological Statistics*, 7, 43-56.

[40] Neuhäuser, M. and Hothorn, L.A. (2005). Maximum tests are adaptive permutation tests. *Journal of Modern Applied Statistical Methods*, 5(2), 4.

[41] Page, E.B. (1963). Ordered Hypothesis for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, 58, 216-230.

[42] Pan, G. (1996). Distribution-free Tests for Umbrella Alternatives. *Communication in Statistics, Theory and Methods*, 25, 3185-3194.

[43] Pan, G. and Wolfe, D.A. (1995). Distribution-Free Test Based on Ranks for Comparing Groups with Umbrella Orderings. *Journal of Nonparametric Statistics*, 5(4), 409-416.

[44] Pan, G. and Wolfe, D.A. (1996). Comparing Groups with Umbrella Orderings. *Journal of the American Statistical Association*, 91, 311-317.

[45] Randles, R. H. and Hogg, R. V. (1973). Adaptive distribution-free tests. *Communications in Statistics,* 2, 337-356.

[46] Robertson, T., Wright, F.T., and Dystra, R.L. (1988). *Order Restricted Statistical Inference*. Wiley and Sons.

[47] Ruberg, S.J. (1986). A Continuously Adaptive Nonparametric Two-Sample Test. *Communications in Statistics - Theory and Methods*, 15(10), 2899-2920.

[48] Shi, N.Z. (1988). Rank Test Statistics for Umbrella Alternatives. *Communication in Statistics, Theory and Methods*, 17, 2059-2073.

[49] Simpson, D.G. and Margolin, B.H. (1986). Recursive nonparametric testing for dose-response relationships subject to downturns at high doses. *Biometrika,* 73, 589-596.

[50] Terpestra, T.J. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present. *Indagationes Mathematicae*, 14, 327-333.

[51] van Elteren, P.H. (1960). On the combination of independent two sample tests of Wilcoxon. *Bulletin of the Institute of International Statistics*, 37, 351-361.

[52] Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, 1, 80-83.

[53] Wilcoxon, F. (1946). Individual comparisons of grouped data by ranking methods. *Journal of Entomology*, 39, 269-270.

[54] Wong, P., Monette, G., and Weiner, N.I. (2001). Mathematical models of cognitive recovery. *Brain Injury*, 15, 519-530.