Statistica Sinica **24** (2014), 773-797
doi:http://dx.doi.org/10.5705/ss.2011.227

RANDOMIZATION INFERENCE FOR THE TRIMMED MEAN OF EFFECTS ATTRIBUTABLE TO TREATMENT

Xingdong Feng, Yang Feng, Yuguo Chen and Dylan S. Small

Shanghai University of Finance and Economics, Google Inc., University of Illinois at Urbana-Champaign and University of Pennsylvania

Abstract: Randomization is described by Fisher (1935) as the reasoned basis for inference about the effectiveness of treatments. Fisher advocated both using randomization in designing experiments and using "randomization inference" to analyze experiments that have been randomized. Randomization inference is inference that assumes only the physical act of randomization for its validity. It provides exact, distribution free inferences in randomized experiments. In this paper, we expand the scope of randomization inference by developing randomization inference for the trimmed mean of effects attributable to treatment. Trimmed means of the effects attributable to treatment are interpretable summaries of the treatment effect that are robust to outliers. We connect the inference problem for trimmed means of effects attributable to treatment to a multiple choice knapsack problem, and use an efficient combinatorial optimization algorithm.

Key words and phrases: Knapsack problem, observational study, randomization inference, trimmed mean.

1. Introduction

Randomization is described by Fisher (1935) as the reasoned basis for inference about the effectiveness of treatments. Fisher advocated both using randomization in designing experiments and using "randomization inference" to analyze experiments that have been randomized. Randomization inference is inference that assumes only the physical act of randomization for its validity. It provides exact, distribution free inferences in randomized experiments.

Traditionally, randomization inference has focused on constant additive and multiplicative treatment effects (e.g., Moses (1965); Lehmann (1963, 1975)). However, constant additive or multiplicative treatment effects are often unrealistic (Hill (2002)). Rosenbaum (2001, 2002a) developed an approach for making randomization inferences in the presence of nonconstant treatment effects. For ordered outcomes, Rosenbaum shows how to make inferences for (a) displacement effects, the number of treated subjects whose outcome if taking the control would be below a certain quantile of the distribution of outcomes if all subjects took the control and whose outcome would be above the quantile if taking the treatment, and (b) the number of comparisons between outcomes of a treated subject and a control subject which would be reversed if the treated subject were instead assigned to the control. Lee (2000) presented conditions under which the sign of the median treatment effect is identified.

In this paper, we expand the scope of randomization inference by developing randomization inference for the trimmed mean of effects attributable to treatment. The effects attributable to treatment are the effects of treatment for the subjects in the experiment who were assigned to the treatment. We find confidence intervals for the trimmed means of the effects attributable to treatment. We focus on the trimmed means because these are concise summaries of the treatment effect that are robust to outliers. We show that randomization inference for the trimmed mean of the effects attributable to treatment can be formulated as a multiple choice knapsack problem. Knapsack problems are a well researched class of combinatorial optimization problems (Martello and Toth (1990)).

In Section 2, we establish the setup and notation, analogous to Rosenbaum (2001). In Section 3, we state the statistical problem we are trying to solve – we would like to perform hypothesis tests and construct confidence intervals for the trimmed mean of effects attributable to treatment. In Section 4, we show how to solve this problem by formulating our problem as a multiple choice knapsack problem. In Section 5, we apply our method to analyze a study of the effect of intrinsic vs. extrinsic motivation on creative writing and a study of remedying education in India. Finally, some conclusions are given in Section 6.

2. Setup and Notations

Consider N subjects. Suppose that m out of N subjects are randomly chosen to be exposed to treatment, and the remaining $N-m$ subjects are exposed to a control condition. Define the random variables Z_i, $i = 1, \ldots, N$, as

$$Z_i = \begin{cases} 1, & \text{The } i\text{th subject was exposed to treatment,} \\ 0, & \text{otherwise.} \end{cases}$$

It follows that $\sum_{i=1}^{N} Z_i = m$. Use $\mathbb{B}$ to denote the set of all the possible values of the vector $\mathbf{Z} = (Z_1, \ldots, Z_N)^T$, so $|\mathbb{B}| = N!/[m!(N-m)!]$, where $|\cdot|$ is used to indicate the cardinality of the set $\mathbb{B}$. We assume that the experiment is completely randomized so that Z is selected at random from $\mathbb{B}$, so $P(\mathbf{Z} = z) = 1/|\mathbb{B}|$ for each $z \in \mathbb{B}$.

Let $Y_i^{(t)}$ denote the response subject i would have if she were assigned the treatment and $Y_i^{(c)}$ the response she would have if assigned the control. In randomization inference, the two potential responses under treatment and control are treated as fixed features, and only the vector $\mathbf{Z}$ is random (Rosenbaum (2001); Fisher (1935)). $Y_i^{(t)}$ is observed if $Z_i = 1$, and $Y_i^{(c)}$ is observed if $Z_i = 0$. Hence, the observed response from subject i is $Y_i = Z_i Y_i^{(t)} + (1 - Z_i) Y_i^{(c)}$, a random variable depending on the random assignment of the treatment Z_i (Fisher (1935)). Obviously, we can observe both the vector $\mathbf{Z}$ and the vector $\mathbf{Y} = (Y_1, \ldots, Y_N)$, which include the information of the treatment assignment and the corresponding responses. In this article, we assume that the treatment has a nonnegative effect in that $Y_i^{(t)} \geq Y_i^{(c)}$, for $i = 1, \ldots, N$, as assumed by Rosenbaum (2001). As discussed by Rosenbaum (2001), the model of a nonnegative treatment effect cannot be verified or refuted by inspecting the responses of individuals, because $Y_i^{(t)}$ and $Y_i^{(c)}$ are never jointly observed on the same person. However, Hamilton (1979) noted that the standard epidemiological measures of effect on a binary response can be reinterpreted in terms of the model of a nonnegative treatment effect and these measures are all consistent with the assumptions whenever treated subjects are at increased risk. The nonnegative treatment effect may not hold marginally, but hold for a subgroup of subjects defined by covariates, in which case the analysis can be done for this subgroup.

The treatment effects for the subjects are denoted as $A_1 = Y_1^{(t)} - Y_1^{(c)}, \ldots, A_N = Y_N^{(t)} - Y_N^{(c)}$. Suppose that the treated subjects are indexed by $i_1, \ldots, i_m$. The effects attributable to treatment are the treatment effects for the treated subjects, $A_{i_1} = Y_{i_1}^{(t)} - Y_{i_1}^{(c)}, \ldots, A_{i_m} = Y_{i_m}^{(t)} - Y_{i_m}^{(c)}$. Let $A_{(1)}^{(t)}, \ldots, A_{(m)}^{(t)}$ denote the increasingly ordered effects $\{A_{i_j}, j = 1, \ldots, m\}$. Our quantities of interest are the τ trimmed mean of treatment effects on subjects of the treatment group,

$$\bar{A}_\tau = \Big(m - 2\Big[\frac{m\tau}{2}\Big]\Big)^{-1} \sum_{k=[m\tau/2]+1}^{m-[m\tau/2]} A_{(k)}^{(t)}, \tag{2.1}$$

where $[m\tau/2]$ is the largest integer less than or equal to $m\tau/2$ (Dasgupta (2008, p.271)). Note that $\bar{A}_\tau$ is a random variable (not a fixed parameter) that indicates how the observed treatment group would have responded differently under the control; $\bar{A}_\tau$ is random because the treatment group is random. The regular mean is the case of $\tau = 0$. We focus on the $\tau > 0$ trimmed means because they are robust estimates of location (Huber (2009)) so that our inference will not be seriously affected by outliers.

To see this, consider the following example: the responses of the control group are $1, 3, 5, \ldots, 39$ and the responses of the treatment group are $2, 4, 6, \ldots, 40$. Now consider testing the hypothesis that the untrimmed mean of effects attributable to treatment is equal to $1,000,000$, which appears implausible. One member of this hypothesis is that the treatment effect is $20,000,000$ for the treated subject with response equal to 2 and the treatment effect is 0 for the rest of the subjects;

the corresponding potential responses under control for the treated subjects are $-19999998, 4, 6, \ldots, 40$. Then, the ranks of the potential responses under control of the control subjects $(2, 3, 5, 7, \ldots, 39)$ are very similar to those of the treated subjects $(1, 4, 6, 8, \ldots, 40)$, so that we cannot reject this hypothesis using a usual rank test like the Wilcoxon rank sum test that compares the ranks of the potential responses under control of the control subjects to those of the treated subjects (the two-sided p-value is 0.81) or the variant of Conover and Salsburg's (1988) rank test described in Section 4.4 (the two-sided p-value is 0.77). Furthermore, this hypothesis cannot be rejected using a permutation t-test that compares the potential response under control of the control subjects to those of the treated subjects (the p-value is 0.71). Now consider instead testing the null hypothesis that the median of the effects attributable to treatment is $1,000,000$. Every null in this composite null must have at least ten subjects with treatment effects greater than or equal to $1,000,000$. Thus, for the observed data and every null hypothesis in the composite null, there will be at least ten treated subjects who have potential responses under control less than $-999,000$, which would make the ranks of the potential responses under control of the treatment group quite different from those of the control group. Consequently, using the variant of Conover and Salsburg's rank test described in Section 4.4, we reject the hypothesis that the median of the effects attributable to treatment is $1,000,000$ (p-value < 0.01).

For technical reasons, we will assume that there are no ties among the potential responses under control:

Assumption 1. The potential responses under control $Y_1^{(c)}, \ldots, Y_N^{(c)}$ are distinct.

For continuous responses, Assumption 1 is not a strong assumption because the set of potential responses under control that have ties has probability zero.

3. Inference for the Trimmed Mean of Effects Attributable to Treatment

Our goal is to construct a confidence interval for the trimmed mean of effects attributable to treatment, $\bar{A}_\tau$. We will do this by inverting one-sided hypothesis tests. Specifically, consider the hypothesis test, $H_0 : \bar{A}_\tau \leq C$ vs. $H_a : \bar{A}_\tau > C$. To test this null, we will use rank tests. Rank tests have been found to have good power properties for testing treatment effects while being robust to outliers (Lehmann (1975)). We will illustrate our method using Wilcoxon's rank sum test but we will show that the method can also be used with generalized rank tests in Section 4.4.

Denote the treatment effects as a vector $\mathbf{A} = (A_1 = Y_1^{(t)} - Y_1^{(c)}, \ldots, A_N = Y_N^{(t)} - Y_N^{(c)})^T$. We could test $H_0 : \mathbf{A} = \mathbf{A}^*$ by creating adjusted responses that subtract the treatment effects from the responses of subjects who received treatment and leave alone the control subjects' responses, and then applying Wilcoxon's rank sum test to test for a difference between the control and treated adjusted responses. The null hypothesis $H_0 : \bar{A}_\tau \leq C$ is composite, consisting of many $\mathbf{A}$. Consequently we will reject the composite null if and only if every member of the null hypothesis is rejected. Let $\bar{A}_\tau(\mathbf{A}, \mathbf{Z})$ denote the trimmed mean of effects attributable to treatment for effects $\bar{A}$. To carry out the test, we need to find out if there is any $\mathbf{A}$ with $\bar{A}_\tau(\mathbf{A}, \mathbf{Z}) \leq \mathbf{C}$ such that the Wilcoxon's rank sum test applied to the adjusted responses is not rejected. In other words, we need to find out, does there exist an $\mathbf{A}$ such that $\bar{A}_\tau(\mathbf{A}, \mathbf{Z}) \leq C$ and Wilcoxon's rank sum test does not reject the null hypothesis of no effect when applied to $\mathbf{Y} - \mathbf{Z} \odot \mathbf{A}$ where $\odot$ is element by element multiplication?

To describe how the Wilcoxon rank sum test is applied in more detail, consider a specific group of effects attributable to treatment, $A_1 = a_1, \ldots, A_N = a_N$, which satisfies the null hypothesis,

$$\Big(m - 2\Big[\frac{m\tau}{2}\Big]\Big)^{-1} \sum_{k=[m\tau/2]+1}^{m-[m\tau/2]} a_{(k)}^{(t)} \leq C,$$

where $a_{(1)}^{(t)}, \ldots, a_{(m)}^{(t)}$ are the increasingly ordered values $\{a_{i_j}, j = 1, \ldots, m\}$ (the treated subjects are indexed by $i_1, \ldots, i_m$, so we have $Z_{i_j} = 1$). Let Y_i^*'s be the potential response under control if this null hypothesis is true:

$$Y_i^* = \begin{cases} Y_i - a_i & Z_i = 1, \\ Y_i & Z_i = 0. \end{cases} \tag{3.1}$$

Let r_i^* denote the rank of Y_i^* among $Y_1^*, \ldots, Y_N^*$, ($r_i^* = 1, \ldots, N$, and $r_1^* = 1$ for the smallest observation and $r_N^* = N$ for the largest one). The Wilcoxon rank sum test statistic is

$$W = \sum_{j=1}^{m} r_{i_j}^*. \tag{3.2}$$

Under Assumption 1 that there are no ties among the potential responses under control, the distribution of W does not depend on $\mathbf{A}$.

For the one-sided test of $H_0 : \bar{A}_\tau \leq C$ vs. $H_a : \bar{A}_\tau > C$, we reject for large values of W. Thus, to find the p-value for this composite null hypothesis, we carry out the following steps:

(P1) find a specific vector $\mathbf{A}$ with $\bar{A}_\tau \leq C$ that minimizes W;

(P2) move the treatment observations accordingly and calculate potential outcomes under control Y^* based on (3.1) for the $\mathbf{A}$ from Step (P1);

(P3) compute the value of W for this $\mathbf{A}$, which we denote by $W_{\mathbf{A}}$ and calculate the p-value as $P_{H_0}(W \geq W_{\mathbf{A}})$ where P_{H_0} is the usual null distribution of the Wilcoxon rank sum statistic for testing that two distributions are the same when there are m observations from one group and $N-m$ from the other, and there are no ties in the ranks (Lehmann (1975)).

Actually, we need only a specific vector $(a_{i_1}, \ldots, a_{i_m})^T$ for the treatment group in Step (P1) because no movement is applied for the control group in Step (P2).

For the one-sided test of $H_0 : \bar{A}_\tau \geq C$ vs. $H_a : \bar{A}_\tau < C$, we carry out the same steps as (P1)-(P3) except that we reject for small values of W, so we seek to maximize W under H_0.

For the two-sided test of $H_0 : \bar{A}_\tau = C$ vs. $H_a : \bar{A}_\tau \neq C$, we combine the two one-sided tests of $H_0 : \bar{A}_\tau \leq C$ vs. $H_a : \bar{A}_\tau > C$ and $H_0 : \bar{A} \geq C$ vs. $H_a : \bar{A}_\tau < C$. Specifically, for a level α test of $H_0 : \bar{A}_\tau = C$ vs. $H_a : \bar{A}_\tau \neq C$, we reject if the p-value for either the one-sided test of $H_0 : \bar{A}_\tau \leq C$ vs. $H_a : \bar{A}_\tau > C$ or the one-sided test of $H_0 : \bar{A}_\tau \geq C$ vs. $H_a : \bar{A}_\tau < C$ is less than $\alpha/2$. To find a $1-\alpha$ confidence interval for $\bar{A}_\tau$, we invert the level α test of $H_0 : \bar{A}_\tau = C$ vs. $H_a : \bar{A}_\tau \neq C$.

The challenging step in (P1)$-$(P3) is (P1). In Sections 4.1$-$4.2, we discuss how to carry out (P1) and then in Section 4.3, we discuss the corresponding step for testing $H_0 : \bar{A}_\tau \geq C$ vs. $H_a : \bar{A}_\tau < C$ of maximizing W.

4. Optimization

In this section, we show how step (P1) in Section 3 can be formulated as a multiple choice knapsack problem and solved. Although we focus on the trimmed mean $\bar{A}_\tau$, it is helpful to start by considering the regular mean which is a special case of the trimmed mean ($\tau = 0$) because certain simplifications arise.

4.1. Testing the hypothesis that the untrimmed mean of attributable effects is less than or equal to C

We seek to minimize W over all $\mathbf{A}$ such that the mean of attributable effects on subjects of the treatment group, $\bar{A}_{\tau=0}$, is less than or equal to a fixed value of C, where W is defined in (3.2). We can think of this as the problem of "moving" the treated observations in such a way that the cost of the moves is less than or equal to C and the sum of the ranks of the moved treated observations among the pool of moved treated observations and original control observations is minimized. The moved treated observations correspond to the potential responses under control of the treated subjects and the costs of the moves correspond to the treatment effects. The sum of the ranks of the moved treated observations among the pool of moved treated observations and control observations is W. Note that the value of W changes only when at least one response value, denoted as Y, in the treatment group is changed to be a smaller one than those samples of the control group that were originally smaller than Y (a small positive number is used as needed to make sure that the changed response value is slightly smaller than a sample of the control group, so no ties will be present after we change some response values in the treatment group).

To clarify the relationship between the movement and the rank changes, we introduce the following matrices. The first matrix is

$$\mathbf{D} = \begin{pmatrix} d_{10} & d_{11} & \cdots & d_{1,N-m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m0} & d_{m1} & \cdots & d_{m,N-m} \end{pmatrix}, \tag{4.1}$$

where

$$d_{ij} = \begin{cases} 0 & i = 1, \ldots, m, \; j = 0, \\ \max(0, Y^{(t)}_{(m+1-i)} - Y^{(c)}_{(N-m-j+1)}) & i = 1, \ldots, m, \; j = 1, \ldots, N-m, \end{cases}$$

and $Y^{(t)}_{(i)}$ and $Y^{(c)}_{(j)}$ are the ith and the jth order statistics within the treatment group and the control group, respectively. Here d_{ij} is the cost of moving the ith largest observation in the treatment group before the jth largest observation in the control group. It is clear that $d_{ij} \leq d_{i,j+1}$ and $d_{ij} \geq d_{i+1,j}$. The second matrix is used to count the reduced ranks for the treatment observations associated with the movements given by the matrix $\mathbf{D}$:

$$\mathbf{V} = \begin{pmatrix} v_{10} & v_{11} & \cdots & v_{1,N-m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{m0} & v_{m1} & \cdots & v_{m,N-m} \end{pmatrix}, \tag{4.2}$$

where

$$v_{ij} = \begin{cases} 0 & d_{ij} = 0 \\ v_{i,j-1} + 1 & d_{ij} > 0. \end{cases}$$

Here v_{ij} is the amount of rank reduction we can obtain if we move the ith largest observation in the treatment group before the jth largest observation in the control group. Furthermore, for $i = 1, \ldots, m$ and $j = 1, \ldots, N-m$, we define a sequence of binary variables as

$$\delta_{ij} = \begin{cases} 1 & Y^{(t)}_{(m+1-i)} \text{ is moved before } Y^{(c)}_{(N-m+1-j)}, \\ 0 & \text{otherwise,} \end{cases}$$

and $\delta_{i0} = 1$ if $Y^{(t)}_{(m+1-i)}$ is kept unchanged, and $\delta_{i0} = 0$ otherwise. These binary variables are used to indicate how the treatment observations are moved.

When the regular mean $\bar{A}_{\tau=0} \leq C$, then we must have $\sum_{j=1}^{m} A_{i_j} \leq mC$. Hence, to find a vector of $(A_{i_1}, \ldots, A_{i_m})^T$ that minimizes W, we can consider the following integer programming problem:

$$\begin{aligned} \max_{\delta} & \sum_{i=1}^{m} \sum_{j=0}^{N-m} \delta_{ij} v_{ij}, \\ s.t. & \sum_{i=1}^{m} \sum_{j=0}^{N-m} \delta_{ij} d_{ij} \leq mC, \\ & \sum_{j=0}^{N-m} \delta_{ij} = 1,\ i = 1, \ldots, m. \end{aligned}$$

If we have $\delta_{ij} = 1$ $(j \geq 1)$ in the above optimization problem, then the ith largest subject $Y^{(t)}_{(m+1-i)}$ in the treatment group is moved before $Y^{(c)}_{(N-m+1-j)}$ by changing d_{ij}. If $\delta_{i0} = 1$, then $Y^{(t)}_{(m+1-i)}$ is kept unchanged. This integer programming is a well-known problem in optimization: the multiple choice knapsack problem (MCKP). Although this is a non-deterministic polynomial-time hard (NP-hard) problem (Martello and Toth (1990)), more efficient algorithms than the classical branch-and-bound algorithm have been developed and these algorithms often yield an exact solution in a reasonable amount of time. We used the hybrid dynamic programming/branch-and-bound algorithm of Dyer, Riha, and Walker (1995) to give the solution of MCKP. A brief review on MCKP is given in Appendix A.3.

4.2. Testing the hypothesis that the trimmed mean of attributable effects is less than or equal to C

In this section, we consider the null hypothesis that the trimmed mean is no greater than a constant C. We first consider the upper trimmed mean, which is useful to lead to the method for the regular trimmed mean case. The upper trimmed mean is the smallest $m - [m\tau/2]$ effects attributable to treatment.

4.2.1. The general method

When the $\tau/2$ upper trimmed mean of attributable effects is given as a constant no greater than C, we have to consider the orders of the effects for the subjects in the treatment group. Since the $[m\tau/2]$ largest attributable effects in the treatment group can be infinity, the $[m\tau/2]$ largest attributable effects in the treatment group are given by moving the $[m\tau/2]$ largest treatment observations to negative infinity (we can move them to somewhere before the smallest observation of the control group in practice) because this procedure can give the largest treatment effects among all the possible schemes for moving $[m\tau/2]$ out of m treatment observations, so the movement of the largest $[m\tau/2]$ treatment observations are determined ($\delta_{i,N-m} = 1$, for $i = 1, \ldots, [m\tau/2]$). Hence, to determine the moving scheme of other treatment observations, we only need to consider the lowest $(m - [m\tau/2])$ rows of the matrix $\mathbf{D}$, denoted by $\mathbf{D}_1$, as

$$\mathbf{D}_1 = \begin{pmatrix} d_{[m\tau/2]+1,0} & d_{[m\tau/2]+1,1} & \cdots & d_{[m\tau/2]+1,N-m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{m0} & d_{m1} & \cdots & d_{m,N-m} \end{pmatrix}.$$

It is clear that the upper trimmed mean is

$$\Big(m - \Big[\frac{m\tau}{2}\Big]\Big)^{-1} \sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} d_{ij}$$

for a moving scheme $\{\delta_{ij}\}$. With the restriction that $\bar{A}_\tau \leq C$, we need to consider the MCKP

$$\begin{aligned} &\max_{\delta} \sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} v_{ij} \\ s.t. \quad &\sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} d_{ij} \leq \Big(m - \Big[\frac{m\tau}{2}\Big]\Big) C, \\ &\sum_{j=0}^{N-m} \delta_{ij} = 1, \ i = \Big[\frac{m\tau}{2}\Big] + 1, \ldots, m. \end{aligned}$$

Thus, with the upper trimmed mean controlled, we need to move the $[m\tau/2]$ largest treatment observations to somewhere before the smallest observation of the control group ($\delta_{i,N-m} = 1$, for $i = 1, \ldots, [m\tau/2]$), and move the other treatment observations based on the scheme given by the above optimization solution $\{\delta_{ij}, \ i = [m\tau/2] + 1, \ldots, m\}$.

Unfortunately, it is a much more complicated problem to determine the $[m\tau/2]$ smallest treatment effects with the restriction on the trimmed mean, while the $[m\tau/2]$ largest treatment effects can always be thought as infinity. We cannot consider just a single optimization problem to determine the moving scheme of the treatment group for minimizing the rank W, and we use a sequence of optimization problems to obtain it.

We sort the distinctive entries of the matrix $\mathbf{D}_1$ in an increasing order, and denote the sorted entries as

$$d_{(1)}, \ldots, d_{(l)}, \tag{4.3}$$

where $d_{(1)} = 0$, which implies no movements. Then, we determine the index m_1 such that $d_{(m_1)} \le C < d_{(m_1+1)}$. We use $d_{(m_1)}$ to consider an initial optimization problem with feasible solutions in (4.6). Clearly, for the treatment group, the value $d_{(m_1)}$ is the upper bound that the smallest $[m\tau/2]$ effects attributable to treatment can reach based on matrix $\mathbf{D}_1$ because we need $\bar{A}_\tau \le C$.

It is difficult to formulate an optimization problem when we need to consider the orders of the treatment effects. To simplify this problem, we consider the matrix,

$$\mathbf{D}_1^* = \begin{pmatrix} d^*_{[m\tau/2]+1,0} & d^*_{[m\tau/2]+1,1} & \cdots & d^*_{[m\tau/2]+1,N-m} \\ \vdots & \vdots & \ddots & \vdots \\ d^*_{m0} & d^*_{m1} & \cdots & d^*_{m,N-m} \end{pmatrix}, \tag{4.4}$$

and the associated indication matrix,

$$\mathbf{U} = \begin{pmatrix} u_{[m\tau/2]+1,0} & u_{[m\tau/2]+1,1} & \cdots & u_{[m\tau/2]+1,N-m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{m0} & u_{m1} & \cdots & u_{m,N-m} \end{pmatrix}, \tag{4.5}$$

where $d^*_{ij} = \max\{d_{ij}, d_{(m_1)}\}$, $u_{ij} = 0$ if $d^*_{ij} \le d_{(m_1)}$, and $u_{ij} = 1$ if $d^*_{ij} > d_{(m_1)}$. We use the matrix $\mathbf{D}_1^*$ so that we impose an upper bound $d_{(m_1)}$ for the smallest $[m\tau/2]$ trimmed treatment effects. Thus, we can consider a multidimensional multiple choice knapsack problem (MMKP) by replacing those trimmed treatment effects that are less than $d_{(m_1)}$ with $d_{(m_1)}$. The trouble caused by ordering the smallest treatment effects can be avoided by this modification. In this paper, we use the numerical algorithm of Sbihi (2007) to give the exact solution for the MMKP. A brief review on MMKP is given in Appendix A.4.

This MMKP can be expressed mathematically as

$$\begin{aligned} &\max_{\delta} \sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} v_{ij} \qquad (4.6)\\ &s.t. \quad \sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} u_{ij} d^*_{ij} + \Big(m - 2\Big[\frac{m\tau}{2}\Big] - \sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} u_{ij}\Big) d_{(m_1)} \\ &\qquad\quad \le \Big(m - \Big[\frac{m\tau}{2}\Big]\Big) C, \\ &\sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} u_{ij} \le m - 2\Big[\frac{m\tau}{2}\Big], \\ &\sum_{j=0}^{N-m} \delta_{ij} = 1, \ i = \Big[\frac{m\tau}{2}\Big] + 1, \ldots, m. \end{aligned}$$

If $\sum_{j=0}^{N-m} \delta_{ij} u_{ij} = 1$, then the ith subject is moved by $\sum_{j=0}^{N-m} \delta_{ij} u_{ij} d^*_{ij}$, which is strictly greater than $d_{(m_1)}$. If $\sum_{j=0}^{N-m} \delta_{ij} u_{ij} = 0$, the corresponding treatment effect based on $\mathbf{D}_1$ is no greater than $d_{(m_1)}$, composed of both the trimmed and the non-trimmed effects when we calculate the trimmed mean. The first constraint is imposed such that the trimmed mean $\bar{A}_\tau \leq C$, and by rearranging the items, the first constraint can be expressed as

$$\sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} u_{ij} (d^*_{ij} - d_{(m_1)}) \leq \left(m - 2\left[\frac{m\tau}{2}\right]\right)(C - d_{(m_1)}).$$

The second constraint requires that the total number of those subjects moved by more than $d_{(m_1)}$ is limited by $m - 2[m\tau/2]$.

The optimization problem (4.6) gives the optimal solution based on $\mathbf{D}^*_1$ that leads to the maximum of total reduced ranks for the treatment group, and we use $\delta_{\max}$ to record the moving scheme $\{\delta_{ij}\}$, and use $R_{\max}$ to record the total reduced ranks, which is $\sum_{i=[m\tau/2]+1}^{m} \sum_{j=1}^{N-m+1} \delta_{ij} v_{ij}$. Then, we sequentially consider $d_{(m_1-1)}, d_{(m_1-2)}, \ldots, d_{(1)}$, which is in decreasing order, update the matrices $\mathbf{D}^*_1$ and $\mathbf{U}$ accordingly, and use the optimization problem (4.6) to give the moving schemes of the treatment group. We update $\delta_{\max}$ and $R_{\max}$ to record the moving scheme that gives the maximum of the total reduced ranks and the corresponding total reduced ranks, respectively. A summary of the algorithm is given in Appendix A.1.

The constraints of (4.6) and the fact that $0 \leq d_{ij} \leq d^*_{ij}$ imply that any feasible solution of (4.6) leads to a moving scheme that satisfies the restriction of $\bar{A}_\tau \leq C$. Also, the objective function of (4.6) aims to minimizing W under given constraints by maximizing total reduced ranks of the treatment group, where W is defined in (3.2), so it suffices to show that any moving scheme $\{\tilde{\delta}_{ij}\}$, that minimizes W given the restriction that $\bar{A}_\tau \leq C$, is a feasible solution of (4.6) for some $d_{(k)}$. The result is summarized in a theorem with proof given in Appendix A.2.

Theorem 1. *By sequentially considering $d_{(m_1)}, \ldots, d_{(1)}$, where $d_{(l)}$ is defined in* (4.3) *and m_1 is the index such that $d_{(m_1)} \leq C < d_{(m_1+1)}$, and solving a sequence of MMKP problems* (4.6), *we obtain an optimal moving scheme $\{\tilde{\delta}_{ij}\}$ that minimizes W given the restriction that $\bar{A}_\tau \leq C$, where W is defined in* (3.2).

Theorem 1 ensures that an optimal solution $\{\tilde{\delta}_{ij}\}$ is obtained by sequentially considering (4.6) because there exists $d_{(k)} \leq d_{(m_1)}$ such that the moving scheme $\{\tilde{\delta}_{ij}\}$ is a feasible solution of (4.6) with $d_{(m_1)}$ replaced by $d_{(k)}$.

By inverting the hypothesis testing given the trimmed fraction τ, we can give the confidence interval (CI) for the trimmed mean of attributable effects at the confidence level α. The details about how to invert the hypothesis testing to CI are given in Section 9.2.1 of Casella and Berger (2002). We use different "capacity" C, and give the lower bound of the CI by considering the maximum of the real "capacities" used by the solutions with the p-values no greater than $\alpha/2$ from the above optimization procedure for testing the hypothesis $H_0 : \bar{A}_\tau \leq C$ versus $H_a : \bar{A}_\tau > C$. Similarly, we can determine the upper bound of the CI by using the minimum of the real "capacities" that give the p-values no greater than $\alpha/2$ for testing the hypothesis $H_0 : \bar{A}_\tau \geq C$ versus $H_a : \bar{A}_\tau < C$ discussed in the following section. This inverting procedure will be used in Section 5 to give the CIs.

4.2.2. An explanatory example

In this section, we use an example to illustrate how the treatment observations are moved for testing the hypotheses $H_0 : \bar{A}_\tau \leq C$ versus $H_a : \bar{A}_\tau > C$, as discussed in the previous section. Suppose that the treatment group is $\{1, 3, 5\}$ and the control group is $\{0, 2, 4\}$, and we consider $C = 1.5$ and $\tau = 2/3$, where $\tau = 2/3$ implies the median of attributable effects.

First, we move the largest treatment observation to somewhere before the smallest observation of the control group, and then we have the matrices $\mathbf{D}_1$ and $\mathbf{V}_1$ as follow,

$$\mathbf{D}_1 = \begin{pmatrix} 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{V}_1 = \begin{pmatrix} 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Second, we sort the entries of $\mathbf{D}_1$ as $d_{(1)} = 0, d_{(2)} = 1, d_{(3)} = 3$ in increasing order, and determine the index m_1 such that $d_{(m_1)} \leq C < d_{(m_1+1)}$. Since $C = 1.5$, we have $m_1 = 2$. With $d_{(m_1)} = 1$, we have the modified matrices,

$$\mathbf{D}_1^* = \begin{pmatrix} 1 & 1 & 1 & 3 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

For this MMKP problem, the optimal movement scheme of the treatment group is that $1 \to 0^-, 3 \to 2^-, 5 \to 0^-$, where a^- denote $a - \epsilon$, where ϵ is a small positive number, and the total of the reduced ranks is 5 and the "capacity" used by the median of attributable effects is 1.

Fourth, we consider $d_{(m_2)} = 0$, so the matrices need to be updated as follows,

$$\mathbf{D}_1^* = \begin{pmatrix} 0 & 0 & 1 & 3 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and the associated rank-change matrix

$$\mathbf{U} = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

For this MMKP problem, the optimal movement scheme of the treatment group is $1 \to 1, 3 \to 2^-, 5 \to 0^-$ or $1 \to 0^-, 3 \to 3, 5 \to 0^-$, and the total of the reduced ranks is 4. Thus, the first scheme is better because it gives a larger reduction of the total ranks of the treatment group. Hence, we adjust the treatment group to be $\{0^-, 2^-, 0^-\}$, so the p-value is 0.95 based on the Wilcoxon rank sum test as described in (P3) of Section 3.

4.3. Testing the hypothesis that the trimmed mean of attributable effects is greater than or equal to C

As stated in Section 3, for the hypothesis of $H_0 : \bar{A}_\tau \geq C$ versus $H_a : \bar{A}_\tau < C$, we need to find a moving scheme for the treatment group such that the total reduced ranks is minimized with the regular mean or the trimmed mean of attributable effects no less than C. Hence, we consider moving the treatment observations to somewhere right after the appropriate control observations (some small numbers may be needed to adjust the relative positions of the observations in cases of the presence of ties), respectively. Thus, we can define similar matrices $\mathbf{D}'$ and $\mathbf{V}'$ as the matrices $\mathbf{D}$ and $\mathbf{V}$ used in Section 4.1, respectively, with the entries d_{ij} and v_{ij} replaced by

$$d'_{ij} = \begin{cases} \max(0, Y^{(t)}_{(m+1-i)} - Y^{(c)}_{(N-m+1-j)}) & i = 1, \ldots, m, \ j = 1, \ldots, N-m, \\ +\infty & i = 1, \ldots, m, \ j = N-m+1, \end{cases}$$

and

$$v'_{ij} = \begin{cases} 0 & d'_{ij} = d'_{i1} \\ v_{i,j-1} + 1 & d'_{ij} > d'_{i1}. \end{cases}$$

Thus, given the constraint for the regular mean of attributable effects, we can consider the following variant of the MCKP to maximize W, where W is defined in (3.2),

$$\begin{aligned} &\min_{\delta'} \sum_{i=1}^{m} \sum_{j=1}^{N-m+1} \delta'_{ij} v'_{ij} \\ s.t. &\sum_{i=1}^{m} \sum_{j=1}^{N-m+1} \delta'_{ij} d'_{ij} \geq mC, \\ &\sum_{j=1}^{N-m+1} \delta'_{ij} = 1, \ i = 1, \ldots, m, \end{aligned}$$

where δ' indicates how the treatment observations are moved. Let $K = \max\big\{\max_{i,j}\{d_{ij}\}, C\big\}$. Then we can consider an equivalent problem with $d'_{i,N-m+1} = +\infty$ replaced by $d'_{i,N-m+1} = K$. Let $v'_i = \max_j\{v'_{ij}\}$. Since $\sum_{j=1}^{N-m+1}\delta'_{ij} = 1$, it is clear that this problem is equivalent to the following MCKP by changing the sign of its objective function and constraint coefficients and adding some constants to them, respectively,

$$\begin{aligned}
&\max_{\delta}\sum_{i=1}^{m}\sum_{j=1}^{N-m+1}\delta'_{ij}(v'_i - v'_{ij})\\
s.t.\ &\sum_{i=1}^{m}\sum_{j=1}^{N-m+1}\delta'_{ij}(K - d'_{ij}) \le m(K-C),\\
&\sum_{j=1}^{N-m+1}\delta'_{ij} = 1,\ i = 1,\ldots,m.
\end{aligned}$$

More generally, for the trimmed mean of attributable effects considered in the hypothesis testing, we can replace those d_{ij}, v_{ij} and C in Section 4.2 by $K - d'_{ij}$, $v'_i - v'_{ij}$ and $K - C$, respectively, and directly apply the method developed in Section 4.2 to give the solution that maximizes W.

4.4. Generalized rank test

The Wilcoxon rank sum test compares the center of the treated and control groups and does not compare the groups' dispersions. For obtaining powerful tests of the trimmed mean of attributable treatment effects, it is valuable to consider alternative tests that compare both the groups' centers and dispersions. To see the need for comparing both the centers and dispersions, consider an example. The responses of the control group are $1, 3, 5, \ldots, 39$ and the responses of the treatment group are $2, 4, 6, \ldots, 40$. Consider the seemingly implausible hypothesis that the median of the effects attributable to treatment is $1,000,000$ or greater. Any member of this composite null hypothesis has at least 10 treated subjects with potential responses under control less than $-999,000$ and consequently, the group that received treatment dominates the left tail of the potential responses under control distribution for any member of the null hypothesis. To be able to reject this seemingly implausible null hypothesis, we would like a test that takes on an extreme value when the left tail of the distribution is dominated by one group.

Conover and Salsburg (1988) proposed tests that take on extreme values when the right tail of the distribution of the combined groups' responses is dominated by one group. By using the reverse ranks of the responses, defined as

$N - r_i + 1$, where r_i is the rank of the ith subject, and applying Conover and Salsburg's tests to these reverse ranks, we obtain a test that takes on extreme values when the left tail of the distribution is dominated by one group. Conover and Salsburg's tests look at the rank sum of transformed responses. Suppose that f is the transformation function, and consider the $m \times (N - m + 1)$ rank matrix

$$\mathbf{R} = \begin{pmatrix} r_1 & r_1 & \cdots & r_1 \\ \vdots & \vdots & \ddots & \vdots \\ r_m & r_m & \cdots & r_m \end{pmatrix},$$

where r_j is the rank of the jth treatment observation among all the observations including both the treatment and the control groups.

The procedure given in the previous section can be extended to account for the generalized ranks by replacing the matrix $\mathbf{V}$ given in (4.2) with the matrix $\widetilde{\mathbf{V}} = f(\mathbf{R}) - f(\mathbf{R} - \mathbf{V})$. Furthermore, we use the matrix $\mathbf{D}$ given in (4.1) and the procedure as described in Section 4 to give the movement scheme of the treatment group. Then, we can carry out the rank test similar to the Wilcoxon rank sum test by replacing the original ranks with the transformed ranks.

Conover and Salsburg (1988) considered in particular the transformation $f(R) = R^{q-1}$. The Wilcoxon rank sum test is of the form $f(R) = R^{q-1}$ with $q = 2$. Conover and Salsburg (1988) suggested $q = 5$, and Podgor and Gastwirth (1994) showed that the rank test with this transformation has good power for location-scale alternatives.

Markowski and Hettmansperger (1982) generalized the Wilcoxon's rank test by considering ranks of groups. As another example, we can also find a function f to map the original rank matrix $\mathbf{R}$ to the matrix of group ranks. Then, we can invert this test by using fewer ranks. However, the matrix $\mathbf{D}$ stays unchanged, so we need to solve an MMKP with the different objective function, and few ranks cannot save us from the "curse" of large datasets.

4.5. Large sample size

Since both MCKP and MMKP are NP-hard assignment problems, the time to obtain the exact solutions increases exponentially with sample sizes. Thus, for some large datasets, the proposed procedure may not be able to give results in reasonable time, which limits its application. To extend the application of the proposed method, we can consider some approximation methods that lead to near optimal solutions. Chapter 11 of Kellerer, Pferschy, and Psinger (2004) and Han, Leblet, and Simon (2010) give good reviews of approximation algorithms for the MCKP and the MMKP, respectively.

5. Applications

5.1. Motivation and creativity

Amabile (1985) studied the effect of extrinsic vs. intrinsic motivation on creative writing. Subjects with considerable experience in creative writing were randomly assigned to one of two treatment groups: 24 of the subjects were placed in an "instrinsic" group and 23 in the "extrinsic" group. The intrinsic group completed a questionnaire which involved ranking intrinsic reasons for writing; it was intended as a device to establish a thought pattern concerning intrinsic motivation – doing something because doing it brings satisfaction. The extrinsic group completed a questionnaire which involved ranking extrinsic reasons for writing; it was intended as a device to establish a thought pattern concerning extrinsic motivation – doing something because a reward is associated with its completion. After completing the questionnaire, all subjects were asked to write a poem in the Haiku style about "laughter." All poems were submitted to 12 poets, who evaluated them on a 40 point scale of creativity, based on their own subjective views. Judges were not told about the study's purpose. The outcome is the average response for each of the 12 subjects. Figure 1 shows a box plot of the data. The treatment effect does not appear to be additive as subjects assigned to the extrinsic group have more dispersed scores. Based on the discussion in Amabile (1985), we assume that the intrinsic motivation will not decrease creativity compared to the extrinsic motivation.

Table 1 provides inferences about the trimmed means of effects attributable to treatment; confidence intervals based on tests using two transformations of the ranks are considered, $f(R) = R^{q-1}$ with $q = 2$ (the Wilcoxon rank sum test) and $q = 5$ (the test suggested by Conover and Salsburg (1988)). We are confident that the untrimmed and $\tau = 0.2$ trimmed means of the attributable effects of intrinsic motivation on creativity are positive based on the confidence intervals derived from $q = 2$ and $q = 5$ tests; the lower confidence bounds are positive and the upper bounds are ∞ for both $q = 2$ and $q = 5$. For $\tau = 0.4$ and $\tau = 0.6$, the confidence intervals range from 0 to ∞. For the $\tau = 0.8$ trimmed mean of the attributable effects of intrinsic motivation on creativity, the lower confidence limit remains 0 but the upper confidence limit is finite, 9.88, for the $q = 2$ test. We also compared the two group means using the regular two-sided t-test (90% confidence interval: (1.7, 6.5)) and Wilcoxon test (90% confidence interval: (1.4, 5.9)). The greater width of our confidence intervals than the confidence intervals based on inverting the t-test and the Wilcoxon test comes from the fact that our confidence intervals are based on less assumptions, i.e., we do not use the normal distribution assumed by the t-test based confidence interval or the additive treatment effect model assumed by the Wilcoxon test based confidence interval.

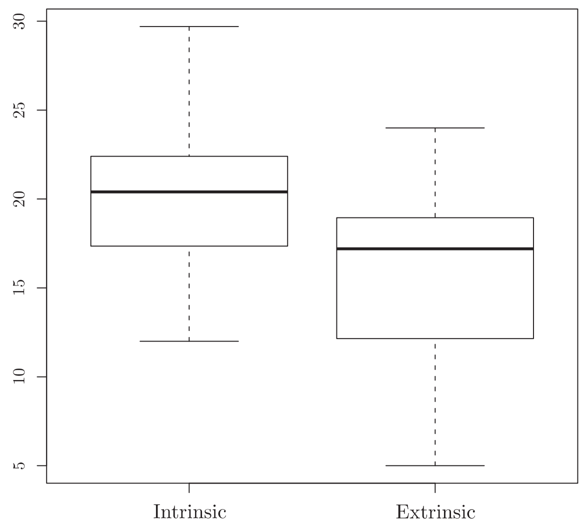

Figure 1. Box plot of the data from the effect of intrinsic vs. extrinsic motivation on creativitity experiment. The outcome is the average creativity score from a panel of 12 judges.

Table 1. For the effect of intrinsic vs. extrinsic motivation on creativity study, 90% confidence intervals for the trimmed mean of effects attributable to treatment given different trimmed fraction τ and different rank transformations with the transformation function $f(R) = R^{q-1}$.

Trimmed fraction	$q = 2$	$q = 5$
$\tau = 0$	$(0.54, \infty)$	$(0.64, \infty)$
$\tau = 0.2$	$(0.01, \infty)$	$(0.05, \infty)$
$\tau = 0.4$	$(0.00, \infty)$	$(0.00, \infty)$
$\tau = 0.6$	$(0.00, \infty)$	$(0.00, \infty)$
$\tau = 0.8$	$(0.00, 9.88)$	$(0.00, \infty)$

The algorithm is efficient for this dataset by using the exact solution of the MMKP (we can obtain the results in a few minutes with a Thinkpad T410 Notebook computer with Intel i5 CPU and 4G memory). However, the MMKP is an NP-hard problem, so it is time-consuming (and sometimes impractical) for large datasets. For this dataset, we used the method of Hifi, Michrafy, and Sbihi (2006) to assess the performance of the approximation in solving MMKPs. We used 100 iterations, and obtained very good results compared with the exact solutions. Actually in the confidence intervals, only one confidence bound from the approximation is different from the exact methods: the upper bound is 9.67 for the trimmed fraction $\tau = 0.8$ and $q = 2$ test.

5.2. Remedying education in India

Two randomized experiments were conducted by Banerjee et al. (2007), among which a computer-assisted learning program was offered to the students in grade 4 for two hours per week extra to the regular study. In this program, students played games by solving mathematical problems. The research focused on those weak students from poor families of Baroda, a city in Western India. In the sample for the study during 2003–2004, the computer-assisted learning program was applied in 55 schools serving as the treatment group, and the other 56 schools served as the control group. The average testing scores of each school were recorded as the raw data for analysis. Since the mathematical games were offered in the time extra to the regular study, it is reasonable to assume that the games help students to better understand mathematics taught in classes, which implies the attributable effects are nonnegative.

The MMKP is NP-hard, and the exact solution for this dataset takes too long to obtain (more than 5 hours). Instead we used the heuristic approximation algorithm of Hifi, Michrafy, and Sbihi (2006) to solve the MMKP in our method with 1000 iterations. It took around 50 minutes in a Thinkpad T410 Notebook computer with Intel i5 CPU and 4G memory.

Box plots of the data are shown in Figure 2 and confidence intervals for the effects of the computer-assisted learning program on mathematical ability are shown in Table 2. Table 2 suggests that it is plausible that the program aid does not improve mathematical ability for the involved students, which is consistent with the results given by the regular t test and Wilcoxon test used for comparing two group means at the significance level 0.05. When the trimmed fraction is above 0.4 for the $q = 2$ test and above 0.7 for the $q = 5$ test, the trimmed mean of attributable effects can exclude the extreme attributable effects (in this example, these extreme effects could be infinite when we inverting the second test of Section 4.3 to construct confidence intervals) such that the upper bounds of confidence intervals are bounded.

6. Conclusion

Randomization inference is well developed for additive treatment effects, but often treatment effects are not additive. Figures 1 and 2 are two examples. We developed new methods for randomization inference for trimmed means of effects attributable to treatment which provide useful information about the treatment effect regardless of whether the additive treatment effect model holds.

To compute randomization inferences for the regular mean or upper trimmed mean of attributable effects, we considered appropriate knapsack problems. For the regular mean or upper trimmed mean of attributable effects, we used the MCKP procedure, and for the regular trimmed mean case, we used the MMKP

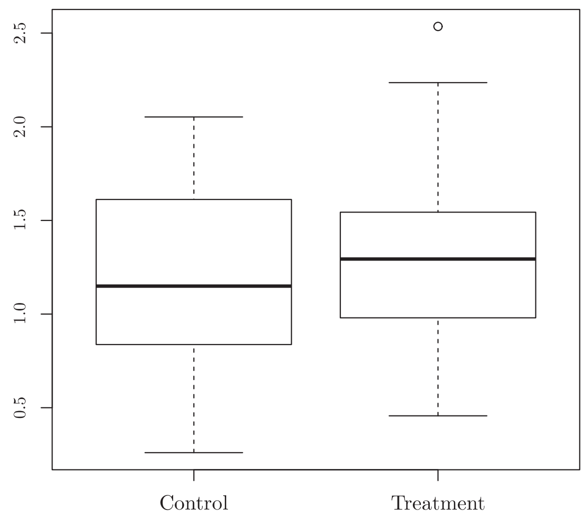

Figure 2. Box plot of the data from the computer-assisted learning program. The outcome is the average testing scores of participating students of each school.

Table 2. For the effect of the computer-assisted learning program on mathematical scores, 95% confidence intervals for the trimmed mean of effects attributable to treatment given different trimmed fraction τ and different rank transformations with the transformation function $f(R) = R^{q-1}$, where $\tau = 0.99$ refers to the median of effects attributable to treatment.

Trimmed fraction	$q = 2$	$q = 5$
$\tau = 0$	$(0.00, \infty)$	$(0.00, \infty)$
$\tau = 0.1$	$(0.00, \infty)$	$(0.00, \infty)$
$\tau = 0.2$	$(0.00, \infty)$	$(0.00, \infty)$
$\tau = 0.3$	$(0.00, \infty)$	$(0.00, \infty)$
$\tau = 0.4$	$(0.00, 0.33)$	$(0.00, \infty)$
$\tau = 0.5$	$(0.00, 0.36)$	$(0.00, \infty)$
$\tau = 0.6$	$(0.00, 0.39)$	$(0.00, \infty)$
$\tau = 0.7$	$(0.00, 0.42)$	$(0.00, 0.27)$
$\tau = 0.8$	$(0.00, 0.46)$	$(0.00, 0.25)$
$\tau = 0.9$	$(0.00, 0.49)$	$(0.00, 0.25)$
$\tau = 0.99$	$(0.00, 0.53)$	$(0.00, 0.25)$

procedure. Although the MMKP procedure can also be used for the regular mean case, we strongly suggest using the MCKP procedure for the regular mean because the exact MCKP algorithm is much more efficient than the existing exact MMKP algorithms. By inverting the rank tests, we constructed the confidence intervals for the trimmed means of effects attributable to treatment. Although the inference is based on solving NP-hard problems, the procedure is still computationally efficient for small or medium sample size. For example, in our first example, the proposed procedure can output results in a few minutes for the sample with the size no greater than 25. For relatively large datasets, the approximation methods can be used to solve MMKPs, which makes the proposed method practical in these cases. For example, in the second example analyzed in this paper, the proposed procedure used around 50 minutes to output results with the approximation method of Hifi, Michrafy, and Sbihi (2006) using 1,000 iterations to solve MMKPs. According to the discussion of Hifi, Michrafy, and Sbihi (2006), the approximation method used in this paper gives encouraging results in solving MMKPs, which can probably lead to good performance of the proposed method in this paper.

We have focused on inference for trimmed means of effects attributable to treatment, that has the desirable feature of being robust to outliers. Beyond the robustness properties, it is necessary to focus on trimmed means rather than untrimmed means to gain any power in randomization inference. For example, even if the treatment and control groups' responses are very similar, the width of the confidence interval for the untrimmed mean will be infinity whereas the confidence interval for the trimmed mean might be finite and informative.

Since the inferences for different trimming fractions τ provide different and useful information about the treatment effect, we recommend reporting a table of results for different trimming fractions τ as in Section 5. However, if one needs to choose a single τ to report conclusions for, then one must consider both the robustness properties of different trimming fractions and the considerations of power for randomization inference for trimmed means of attributable effects discussed above. For the robustness properties of different trimming fractions, Andrews et al. (1972) and Hettmansperger (1968) provide good discussion.

The trimmed means of effects attributable to treatment that we have developed inferences for are concise and robust measures of the effects of treatment. Other concise, robust measures include the trimean and Gastwirth's (1966) estimate ($0.3\times$ first tercile + $.4\times$ median + $.3\times$ second tercile). It is of interest to develop randomization inference for these measures in future work.

Acknowledgements

This research was partly supported by the National Science Foundation grants DMS-0806175, DMS-1106796 and SES-0961971, and the National Natural Science Foundation of China grant 11101254, Pujiang Project of Science and Technology Commission of Shanghai Municipality and Program for Changjiang Scholars and Innovative Research Team in University.

Appendix

A.1. Algorithm of sequential MMKPs

The following algorithm can be used to find a solution $\{\delta_{ij}\}$ that leads to the minimized Wilcoxon rank sum test statistic W as described in Section 4.2.

(S1) Let $\delta_{i,N-m} = 1$ and $\delta_{ij} = 0$ for $j \neq N-m$ and $i = 1, \ldots, [m\tau/2]$.

(S2) Sort the distinctive entries of the matrix $\mathbf{D}_1$ in an increasing order, and denote the sorted entries as $d_{(1)}, \ldots, d_{(l)}$, where $d_{(1)} = 0$.

(S3) Determine the index m_1 such that $d_{(m_1)} \leq C < d_{(m_1+1)}$, and solve the optimization problem (4.6); Denote the solution as $\{\delta_{ij}^{(1)}\}$; Let $R_{\max} = \sum_{i=[m\tau/2]+1}^{m} \sum_{j=1}^{N-m+1} \delta_{ij}^{(1)} v_{ij}$ and $\delta_{\max} = \{\delta_{ij}^{(1)}\}$; Let $k = 2$.

(S4) At the kth step for $k \geq 2$, use $d_{(m_1-k+1)}$ to update the matrices $\mathbf{D}_1^*$ and $\mathbf{U}$ accordingly, and solve the optimization problem (4.6); Denote the solution as $\{\delta_{ij}^{(k)}\}$; If $R_{\max} < \sum_{i=[m\tau/2]+1}^{m} \sum_{j=1}^{N-m+1} \delta_{ij}^{(k)} v_{ij}$, let $R_{\max} = \sum_{i=[m\tau/2]+1}^{m} \sum_{j=1}^{N-m+1} \delta_{ij}^{(k)} v_{ij}$ and $\delta_{\max} = \{\delta_{ij}^{(k)}\}$.

(S5) Let $k = k+1$ and continue Step (S4) if $k \leq m_1$; Output $R_{\max}$ and $\delta_{\max}$.

A.2. Proof of Theorem 1

Suppose that the moving scheme $\{\tilde{\delta}_{ij}\}$ maximizes the total reduced rank

$$\sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \delta_{ij} v_{ij},$$

under the restriction that $\bar{A}_\tau \leq C$. For $i = 1, \ldots, [m\tau/2]$, we must have $\tilde{\delta}_{i,N-m} = 1$ and $\tilde{\delta}_{ij} = 0$ for $j \neq N-m$ because the $[m\tau/2]$ largest treatment effects are always obtained by moving the largest observations in the treatment group to negative infinity.

Since $\sum_{j=0}^{N-m} \tilde{\delta}_{ij} = 1$, there exists k such that $d_{(k)} = \tilde{d}_{([m\tau/2])}$. Now we use $d_{(k)}$ to update the matrix $\mathbf{D}_1^*$, defined in (4.4). Let $\tilde{d}_{(i-[m\tau/2])}$ and $\tilde{d}^*_{(i-[m\tau/2])}$ denote the $(i-[m\tau/2])$th order statistic of $\sum_{j=0}^{N-m} \tilde{\delta}_{ij} d_{ij}$ and $\sum_{j=0}^{N-m} \tilde{\delta}_{ij} u_{ij} d^*_{ij}$, $i = [m\tau/2]+1, \ldots, m$, respectively, where u_{ij} is the entry of the matrix $\mathbf{U}$ (defined in (4.5)) at the ith row and the jth column, and d_{ij}'s and d^*_{ij}'s are the entries of the matrices $\mathbf{D}_1$ and $\mathbf{D}_1^*$, respectively.

Suppose that $\tilde{d}^*_{([m\tau/2]+l)} = d_{(k)}$ for $l = 0, \ldots, p-1$, where p is an integer greater than 1. Thus, $\tilde{d}^*_{(l)} = \tilde{d}_{(l)}$ for $l \geq [m\tau/2]+p$, since $d^*_{ij} = d_{ij}$ if $d_{ij} \geq d_{(k)}$

in the definition of the matrix $\mathbf{U}$. It then follows that

$$\sum_{l=[m\tau/2]+1}^{m-[m\tau/2]} \tilde{d}_{(l)} = \sum_{l=[m\tau/2]+p}^{m-[m\tau/2]} \tilde{d}^*_{(l)} + (p-1)d_{(k)},$$
$$\sum_{l=[m\tau/2]+p}^{m-[m\tau/2]} \tilde{d}^*_{(l)} = \sum_{i=[m\tau/2]+1}^{m} \sum_{j=0}^{N-m} \tilde{\delta}_{ij} u_{ij} d^*_{ij}.$$

Note that the moving scheme $\{\tilde{\delta}_{ij}\}$ satisfies the restriction that

$$\Big(m - 2\Big[\frac{m\tau}{2}\Big]\Big)^{-1} \sum_{l=[m\tau/2]+1}^{m-[m\tau/2]} \tilde{d}_{(l)} \le C.$$

Hence, if we use $\tilde{d}_{([m\tau/2])}$ to replace $d_{(m_1)}$ in the optimization problem (4.6), then $\{\tilde{\delta}_{ij}\}$ is a feasible solution of (4.6). Therefore, after we enumerate $d_{(m_1)}, \ldots, d_{(1)}$, we can give the optimal moving scheme $\{\delta_{ij}\}$ that minimize W. $\diamond$

A.3. Multiple choice knapsack problem (MCKP)

The MCKP has been addressed in rich literatures, and various algorithms have been proposed to give the optimal solution. Assume that there are m classes and r_i items in the ith class. The MCKP is the problem to give the selection scheme if we need to select exactly one item from each class, subject to a constraint, and maximize some criteria. Mathematically, the MCKP can be defined to be of the form

$$\begin{aligned} &\max_{\delta} \sum_{i=1}^{m} \sum_{j=1}^{r_i} v_{ij}\delta_{ij} \\ s.t.\ &\sum_{i=1}^{m} \sum_{j=1}^{r_i} w_{ij}\delta_{ij} \le C_0, \\ &\sum_{j=1}^{r_i} \delta_{ij} = 1,\ i = 1, \ldots, m, \\ &\delta_{ij} \in \{0, 1\},\ i = 1, \ldots, m,\ j = 1, \ldots, r_i, \end{aligned}$$

where the nonnegative coefficients v_{ij} and w_{ij} are the values and the costs of the jth item in the ith class, respectively. Here δ_{ij} indicates if the item j of the ith class is picked or not.

We use the algorithm developed by Dyer, Riha, and Walker (1995). They considered a hybrid algorithm with the breadth-first search strategy when a searching tree is expanded. They partitioned the MCKP into stage subproblems, and used Lagrangian duality to give tight bounds for the bounding tests on the subproblems. By eliminating any partial solutions that cannot be optimal, this algorithm efficiently reduces the branch space. It is not our focus to analyze how the numerical algorithm works, so we do not describe detailed steps of the algorithm here.

A.4. Multidimensional multiple choice knapsack problem (MMKP)

The MMKP is a harder variant of the knapsack problem than the MCKP. This problem also is to give a selection scheme by picking one item from each class, but it allows multiple constraints. It can be formally stated as

$$
\begin{aligned}
& \max_{\delta} \sum_{i=1}^{m} \sum_{j=1}^{r_i} \delta_{ij} v_{ij} \\
s.t. \quad & \sum_{i=1}^{m} \sum_{j=1}^{r_i} \delta_{ij} w_{kij} \leq C_k, \quad k = 1, \ldots, n \\
& \sum_{j=1}^{r_i} \delta_{ij} = 1, \ i = 1, \ldots, m, \\
& \delta_{ij} \in \{0, 1\}, \ i = 1, \ldots, m, \ j = 1, \ldots, r_i,
\end{aligned}
$$

where w_{kij} is the nonnegative weight of the kth constraint equation and C_k is the kth constraint, which are nonnegative.

Sbihi (2007) expanded a tree by considering the classes one by one, and each node of the tree corresponding to one item is added in the order of the decreasing v values for each class. Then, the best-first search strategy is applied by searching the first feasible solution by checking all the constraints. We do not address the details about how to expand the searching tree because it is not our purpose here.

References

Amabile, T. (1985). Motivation and creativity: effects of motivational orientation on creative writers. *J. Personality and Social Psychology* **48**, 393-399.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972). *Robust Estimates of Location: Survey and Advances.* Princeton University Press, Princeton.

Banerjee, A., Cole, S., Duflo, E. and Linden, L. (2007). Remedying education: evidence from two randomized experiments in India. *The Quarterly J. Economics* **122**, 1235-1264.

Casella, G. and Berger, R. L. (2002). *Statistical Inference.* 2nd edition. Duxbury Press, CA.

Conover, W. J. and Salsburg, D. S. (1988). Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond" to treatment. *Biometrics* **44**, 189-196.

Dasgupta, A. (2008). *Asymptotic Theory of Statistics and Probability.* Springer, NY.

Dyer, M. E., Riha, W. O., and Walker, J. (1995). A hybrid dynamic programming/branch-and-bound algorithm for the multiple-choice knapsack problem. *J. Comput. Appl. Math.* **58**, 43-54.

Fisher, R. A. (1935). *Design of Experiments.* Olive and Boyd, Edinburgh.

Gastwirth, J. L. (1966). On robust procedures. *J. Amer. Statist. Assoc.* **61**, 929-948.

Hamilton, M. A. (1979). Choosing the parameter for 2×2 and $2 \times 2 \times 2$ table analysis. *Amer. J. Epidemiology* **109**, 362-375.

Han, B., Leblet, J., and Simon, G. (2010). Hard multidimensional multiple choice knapsack problems, an empirical study. *Comput. Oper. Res.* **37**, 172-181.

Hettmansperger, T.P. (1968). On the trimmed Mann-Whitney statistic. *Ann. Math. Statist.* **39**, 1610-1614.

Hifi, M., Michrafy, M., and Sbihi, A. (2006). A reactive local search-based algorithm for the multiple-choice multi-dimensional knapsack problem. *Comput. Optim. Appl.* **33**, 271-285.

Hill, J. (2002). Comment on "Covariance adjustment in randomized experiments and observational studies" by Paul R. Rosenbaum. *Statist. Sci.* **17**, 307-309.

Huber, P. J. (2009). *Robust Statistics.* 2nd edition. John Wiley, NJ.

Kellerer, H., Pferschy, U., and Psinger, D. (2004). *Knapsack Problems*, Springer-Verlag, Germany.

Lee, M.-J. (2000). Median treatment effect in randomized trials. *J. Roy. Statist. Soc. Ser. B* **62**, 595-604.

Lehmann, E. L. (1963). Nonparametric confidence intervals for a shift parameter. *Ann. Math. Statist.* **34**, 1507-1512.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods Based on Ranks.* Holden-Day, San Francisco, CA.

Markowski, E. and Hettmansperger, T. (1982). Inference based on simple rank step score statistics for the location model. *J. Amer. Statist. Assoc.* **77**, 901-909.

Martello, S. and Toth, P. (1990). *Knapsack Problems: Algorithms and Computer Implementations.* John Wiley, NJ.

Moses, L. (1965). Confidence limits from rank test. *Technometrics* **7**, 257-260.

Podgor, M. J. and Gastwirth, J. L. (1994). On non-parametrric and generalized tests for the two-sample problem with location and scale change alternatives. *Statist. Medicine* **13**, 747-758.

Rosenbaum, P. R. (2001). Effects attributable to treatments: inference in experiments and observational studies with a discrete pivot. *Biometrika* **88**, 219-231.

Rosenbaum, P. R. (2002). Attributing effects to treatment in matched observational studies. *J. Amer. Statist. Assoc.* **97**, 183-192.

Sbihi, A. (2007). A best first search exact algorithm for the Multiple-choice Multidimensional Knapsack Problem. *J. Comb. Optim.* **13**, 337-351.

School of Statistics and Management, Shanghai University of Finance and Economics, and Key Laboratory of Mathematical Economics(SUFE), Ministry of Education, Shanghai 200433, China.

E-mail: feng.xingdong@mail.shufe.edu.cn

Ads Metrics, Google Inc., Pittsburgh, PA 15206, USA.

E-mail: yfeng@google.com

Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA.

E-mail: yuguo@illinois.edu

Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: dsmall@wharton.upenn.edu

(Received September 2011; accepted March 2013)