Exact Inference for Treatment Effect in Blocked Experiments with Binary Outcomes

Jiaxun Li and Jacob Spertus and Philip B. Stark

DRAFT: November 16, 2023
Rough and Incomplete

Contents

1 Introduction

A collection of N subjects is randomized into two groups, one of size n and one of size $m = N - n$. The subjects in the first group are assigned to "treatment" and the others are assigned to "control." For each subject, a binary response is measured, e.g., survival versus death. Such *randomized, controlled trials* (RCTs) with binary treatments and binary outcomes have been studied at least since Fisher's seminal work Fisher [1935].

A convenient and conceptually clear way to mathematize binary experiments with binary outcomes is to represent each subject by a pair of *potential outcomes*: $y_j(0)$ is the response that subject j would have if assigned to control, and $y_j(1)$ is the response subject j would have if assigned to treatment. the response the subject would have if assigned to control and the response the subject would have if assigned to treatment [**?**]. If subject j is assigned to control, we observe $y_j(0)$; if subject j is assigned to treatment, we observe $y_j(1)$. We do not observe both $y_j(0)$ and $y_j(1)$ for any subject. Implicit in this representation of the experiment is *non-interference*: the response of subject j depends only on whether subject j is assigned to treatment or to control, and not on the assignment of any other subjects.[1]

Let $\boldsymbol{y} := ((y_j(1), y_j(0)))_{j=1}^N$ be the *potential outcomes* of the N subjects. The *average treatment effect* (ATE) is the mean of the responses that would have resulted if every subject had been assigned to the active treatment, minus the mean of the responses that would have resulted result if every subject had been assigned to control. That is, the ATE is

$$\tau = \tau(\boldsymbol{y}) := \frac{1}{N}\sum_{j=1}^{N} y_j(1) - \frac{1}{N}\sum_{j=1}^{N} y_j(0).$$

The ATE is an interesting measure of the effectiveness of treatment, and a common target of inference in RCTs.[2]

[1] This would not be a good assumption in some circumstances, for instance, studying the effect of vaccination on the propagation of a communicable disease.

[2] There was a longstanding dispute between Fisher and Neyman about the "correct" null hypothesis to test [**?**Wu and Ding, 2021]. Fisher advocated testing the "strong" null that

Many applications involve tests and confidence sets for the ATE in binary experiments with binary outcomes, from medical experiments **?** to online marketing **?** to FIX ME. Charles et al. [2009] found that around half of clinical trials calculated their sample size based on a binary outcome.

Analyses of such trials often rely on asymptotic theory, resulting in tests and confidence sets that can be anti-conservative in practice [**?**]. Even Fisher noted that the normal approximation could be inadequate for this problem—and that it could be avoided [Fisher, 1935, Section 21, p.50].

A number of methods have been proposed for making exact or conservative inferences about the ATE in experiments with binary treatments and binary outcomes when the assignment to treatment is by *simple random sampling*, i.e., when every subset of n of the N subjects is equally likely to be given the active treatment **??**Chiba [2015], Rigdon and Hudgens [2015], Li and Ding [2016], Aronow et al. [2023].

A simple, computationally efficient, but statistically over-conservative approach uses the fact that under simple random sampling, the number of 1s in the treatment group has a hypergeometric distribution with parameters $N = N$, $n = n$, and $G = \sum_{j=1}^{N} y_j(1)$; and the number of 1s in the control group has a hypergeometric distribution with parameters $N = N$, $n = N - n$, and $G = \sum_{j=1}^{N} y_j(0)$. These two random variables are not independent, but confidence sets for $\frac{1}{N}\sum_{j=1}^{N} y_j(0)$ and $\frac{1}{N}\sum_{j=1}^{N} y_j(1)$ can be combined using the union bound: a lower $1-\alpha$ confidence bound for the ATE can be found by subtracting an upper $1-\alpha/2$ confidence bound for the mean response in the control group from a lower $1-\alpha/2$ confidence bound for the mean response in the treatment group. An upper $1-\alpha$ confidence bound for the ATE can be found by subtracting a lower $1-\alpha/2$ confidence bound for the mean response in the control group from an upper $1-\alpha/2$ confidence bound for the mean response in the treatment group [Rigdon and Hudgens, 2015, Li and Ding, 2016].

A less conservative approach is to partition the null hypothesis $\tau(\boldsymbol{y}) = \tau_0$ into a union of potential outcome tables that have ATE equal to τ_0. Some of those tables can be ruled out algebraically—they are inconsistent with the observed counts—and some may be ruled out statistically Chiba [2015], Rigdon and Hudgens [2015], Li and Ding [2016], Aronow et al. [2023]. The hypothesis $\tau(\boldsymbol{y}) = \tau_0$ can be rejected if all of those tables can be ruled out.

Many randomized experiments with binary treatments and binary outcomes involve *blocking* or *stratification*, which we treat as synonymous. In a blocked experiment, the population of subjects is partitioned into K disjoint blocks within which subjects are randomized, independently across blocks. Blocking is common in clinical trials, where blocks may comprise subjects recruited at a particular center (often blocked further by gender and health covariates). Bruce et al. [2022] estimates that almost two-thirds of clinical trials use some form of stratification.

---

treatment has no effect whatsoever on any individual: $y_j(0) = y_j(1)$ for all j. Neyman advocated testing the "weak" null hypothesis that treatment has no effect *on average*: $\tau(\boldsymbol{y}) = 0$.

To the best of our knowledge, data from blocked experiments are generally analyzed using asymptotic methods. The only exact or conservative methods for making inferences about the ATE from blocked binary experments we know of are those of Rigdon and Hudgens [2015], Chiba [2017]. These methods become impractical when there are more than a few blocks of modest size.

This paper makes three contributions: it develops computationally tractable approaches to testing hypotheses about the ATE and forming confidence intervals for the ATE for blocked binary experiments with binary outcomes; it improves the statistical efficiency of some extant methods; and it compares the statistical and computational efficiency of a variety of methods using simulations. This paper also illustrates the methods using data from a clinical trial of vedolizumab versus placebo for chronic pouchitis, stratified by baseline antibiotic use.

2 Notation

A population of N subjects is partitioned into K strata. Stratum k contains N_k subjects of which n_k are assigned to active treatment by simple random sampling; the other $m_k = N_k - n_k$ are assigned to control. Assignments are independent across strata. The total number of subjects assigned to treatment is $n := \sum_{k=1}^{K} n_k$. Because the strata partition the population, $N = \sum_{k=1}^{K} N_k$. The potential outcomes for the jth subject in the kth stratum are $\boldsymbol{y}_{kj} = (y_{kj}(1), y_{kj}(0)) \in \{0,1\}^2$. Let $\boldsymbol{y}_k := (\boldsymbol{y}_{kj})_{j=1}^{N_k}$ be the potential outcomes for the subjects in the kth stratum, and let $\boldsymbol{y} := (\boldsymbol{y}_k)_{k=1}^{K}$ denote the entire collection of individual potential outcomes. The average treatment effect (ATE) is

$$\tau(\boldsymbol{y}) := \frac{1}{N} \sum_{k=1}^{K} \sum_{j=1}^{N_k} (y_{kj}(1) - y_{kj}(0)).$$

Let $Z_{kj} = 0$ if the jth subject in the kth stratum is assigned to control and $Z_{kj} = 1$ otherwise, and let $\boldsymbol{Z}_k := (Z_{kj})_{j=1}^{N_j}$. Define the *treatment vector* $\boldsymbol{Z} := (\boldsymbol{Z}_k)_{k=1}^{K}$. The observed outcome for the jth subject in the kth stratum is $Y_{kj} = Z_{kj} y_{kj}(1) + (1 - Z_{kj}) y_{kj}(0)$. The vector of observed outcomes for the kth stratum is $\boldsymbol{Y}_k := (Y_{kj})_{j=1}^{N_k}$ and the *observed outcome vector* is $\boldsymbol{Y} := (\boldsymbol{Y}_k)_{k=1}^{K}$. The usual unbiased estimator of $\tau(\boldsymbol{y})$ is

$$\hat{\tau}(\boldsymbol{Y}, \boldsymbol{Z}) := \frac{1}{N} \sum_{k=1}^{K} N_k \left[\frac{1}{n_k} \sum_{j=1}^{N_k} Z_{kj} Y_{kj} - \frac{1}{N_k - n_k} \sum_{j=1}^{N_k} (1 - Z_{kj}) Y_{kj} \right].$$

Define the *ATE in stratum k*:

$$\tau_k(\boldsymbol{y}) := \frac{1}{N_k} \sum_{j=1}^{N_k} (y_{kj}(1) - y_{kj}(0)),$$

and its unbiased estimator:

$$\hat{\tau}_k(\boldsymbol{Y}, \boldsymbol{Z}) = \frac{1}{n_k} \sum_{j=1}^{N_k} Z_{kj} Y_{kj} - \frac{1}{N_k - n_k} \sum_{j=1}^{N_k} (1 - Z_{kj}) Y_{kj}.$$

Then $\tau(\boldsymbol{y}) = \frac{1}{N} \sum_{k=1}^{K} N_k \tau_k(\boldsymbol{y})$ and $\hat{\tau}(\boldsymbol{Y}, \boldsymbol{Z}) = \frac{1}{N} \sum_{k=1}^{K} N_k \hat{\tau}_k(\boldsymbol{Y}, \boldsymbol{Z})$.

The number of subjects in stratum k whose response if assigned to the active treatment would be a and whose response if assigned to control would be b is

$$N_{kab} := \sum_{j=1}^{N_k} \mathbf{1}\{y_{kj}(1) = a, y_{kj}(0) = b\}. \tag{1}$$

Let $\boldsymbol{N}_k := (N_{k11}, N_{k10}, N_{k01}, N_{k00})$. The *potential outcome table* $\boldsymbol{N}$ summarizes all the potential outcomes using $2 \times 2 \times K$ integers: $\boldsymbol{N} := (\boldsymbol{N}_k)_{k=1}^{K}$. The average treatment effect τ and the stratum-wise average treatment effect τ_k can be written as functions of $\boldsymbol{N}$:

$$\tau_k(\boldsymbol{N}) = \frac{1}{N_k} N_{k10} - N_{k01}, \quad k = 1, 2, \ldots, K \tag{2}$$

$$\tau(\boldsymbol{N}) = \frac{1}{N} \sum_{k=1}^{K} (N_{k10} - N_{k01}). \tag{3}$$

The data also can be summarized by a table of integers. The number of subjects in stratum k whose treatment assignment is a and whose response is b is

$$n_{kab} := \sum_{j=1}^{N_k} \mathbf{1}\{Z_{kj} = a, Y_{kj} = b\}. \tag{4}$$

Let $\boldsymbol{n}_k := (n_{k11}, n_{k10}, n_{k01}, n_{k00})$. The *observed table* is $\boldsymbol{n} := (\boldsymbol{n}_k)_{k=1}^{K}$. The estimated average treatment effect $\hat{\tau}$ and the stratum-wise average treatment effect estimator $\hat{\tau}_k$ can be written as functions of $\boldsymbol{n}$:

$$\hat{\tau}_k(\boldsymbol{n}) = \frac{1}{N_k} \left(\frac{n_{k11}}{n_k} - \frac{n_{k01}}{N_k - n_k} \right), \quad k = 1, 2, \ldots, K. \tag{5}$$

$$\hat{\tau}(\boldsymbol{n}) = \frac{1}{N} \sum_{k=1}^{K} N_k \left(\frac{n_{k11}}{n_k} - \frac{n_{k01}}{N_k - n_k} \right) \tag{6}$$

Let $\bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z})$ denote the observed table that would result from the treatment assignment $\boldsymbol{Z}$ applied to a canonical "unpacking" of the potential outcome table $\boldsymbol{N}$ into a full set of potential outcomes for each subject in each stratum.[3] The data $\boldsymbol{n}$ constrain the potential table $\boldsymbol{N}$ algebraically. A potential outcome table $\boldsymbol{N}$ is *compatible* with $\boldsymbol{n}$ if there is some treatment assignment $\boldsymbol{Z}$ for which $\boldsymbol{n} = \bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z})$.

[3]For example, one canonical unpacking sets $y_{kj}(1) = 1$ and $y_{kj}(0) = 1$ for the first N_{k11} subjects in stratum k; $y_{kj}(1) = 1$ and $y_{kj}(0) = 0$ for the next N_{k10} subjects in stratum k; $y_{kj}(1) = 0$ and $y_{kj}(0) = 1$ for the next N_{k01} subjects in stratum k; and $y_{kj}(1) = 0$ and $y_{kj}(0) = 0$ for the last N_{k00} subjects in stratum k.

3 Existing Solutions

[Rigdon & Hutchens; Chiba: explain why ours is an improvement (if it is) Limitations on number of strata.]

We begin by considering non-stratified cases. Recalling the methods proposed in Rigdon and Hudgens [2015] and Li and Ding [2016],there are mainly two ideas to deduce the exact confidence interval. One is based on the hypergeometric distribution and the other is based on inverting a series of permutation tests.

In this section, we ignore the strata-script in the notation since there is only one stratum. For example, we use $\boldsymbol{N} = (N_{11}, N_{10}, N_{01}, N_{00})$ to denote the potential outcome table and $\boldsymbol{n} = (n_{11}, n_{10}, n_{01}, n_{00})$ to denote the observed table.

3.1 Hypergeometric Confidence Interval

Denote $N_{1+} = N_{10} + N_{11}$ as the number of subjects whose response would be 1 if assigned to the active treatment and $N_{+1} = N_{01} + N_{11}$ as the number of subjects whose response would be 1 if assigned to control. Then, the treatment effect can be expressed as $\tau(\boldsymbol{N}) = (N_{10} - N_{01})/N = (N_{1+} - N_{+1})/N$. To obtain an exact confidence interval, we can combine confidence intervals for N_{1+} and N_{+1} with the Bonferroni adjustment.

Since the treated and control subjects are simple random samples of the N subjects in a completely randomized experiment, we have the following distributions:

$$n_{11} \sim \text{HyperGeo}(N_{1+}, N, n), \quad n_{01} \sim \text{HyperGeo}(N_{+1}, N, N-n).$$

Inference for hypergeometric parameters is a classic and well-studied problem. Using either the classic methods or more recent optimal procedures (Wang [2015]), we can obtain $(1-\alpha/2)$ confidence intervals for N_{1+} and N_{+1}, denoted as $[N_{1+}^L, N_{+1}^U]$ and $[N_{+1}^L, N_{+1}^U]$, respectively. Then, by applying the Bonferroni adjustment, we can determine that $[(N_{1+}^L - N_{+1}^U)/N, (N_{1+}^U - N_{1+}^L)/N]$ is a exact $(1-\alpha)$ confidence interval for τ.

In fact, there are other methods available to derive confidence intervals based on the hypergeometric distribution, as discussed in Rigdon and Hudgens [2015] and Li and Ding [2016]. The approach we have introduced here is the most relevant to our work.

3.2 Inverted Permutation Test

Instead of directly considering confidence intervals, we begin by examining hypothesis testing. Let's consider the following hypothesis:

$$H_0(\delta) : y_i(1) - y_i(0) = \delta_i, \text{ for all } i = 1, \ldots, N$$

where $\delta = (\delta_1, \delta_2, \ldots, \delta_N)$ is a known vector. This hypothesis is known as the *sharp null* [Rubin1980], as it implies that all individual treatment effects

are determined under $H_0(\delta)$. This allows us to impute all missing potential outcomes based on the observed data. With known potential outcomes, the only random component of any test statistic $T(\boldsymbol{Y}, \boldsymbol{Z})$ is the treatment vector $\boldsymbol{Z}$. Consequently, the distribution of $\boldsymbol{Z}$ fully determines the distribution of $T(\boldsymbol{Y}, \boldsymbol{Z})$ under $H_0(\delta)$. In a completely randomized experiment, the distribution of $\boldsymbol{Z}$ is uniform over the set:

$$\mathcal{Z}(N, n) = \Big\{ \boldsymbol{Z} \in \{0,1\}^N, \sum_{i=1}^{N} Z_i = n \Big\}.$$

We can then measure how likely the observed data are under $H_0(\delta)$. For example, if larger values of T are more extreme, we can calculate the tail probability:

$$\mathbb{P}\Big(T(\boldsymbol{Y}, \boldsymbol{Z}) \geq T(\boldsymbol{Y}, \boldsymbol{Z}^{obs}) \Big), \tag{7}$$

which is commonly referred to as the *p-value*. Here, $\boldsymbol{Z}^{obs}$ represents the observed treatment vector, and the probability is calculated with respect to $\boldsymbol{Z}$. Following previous work [Rigdon and Hudgens, 2015, Li and Ding, 2016, Aronow et al., 2023], we select $|\hat{\tau}(\boldsymbol{Y}, \boldsymbol{Z}) - \tau(\boldsymbol{y})|$ as the test statistic. Recall that in the case of binary outcomes, we can summarize the observed data by $\boldsymbol{n}$. Under $H_0(\delta)$, we can also summarize the potential outcomes in a table $\boldsymbol{N}$. Thus, the probability (7) can be rewritten as:

$$p(\boldsymbol{N}, \boldsymbol{n}) := \mathbb{P}\left(|\tau(\boldsymbol{N}) - \hat{\tau}(\bar{\boldsymbol{n}}(\boldsymbol{N}, \boldsymbol{Z}))| \geq |\tau(\boldsymbol{N}) - \hat{\tau}(\boldsymbol{n})| \right). \tag{8}$$

We can then reject $H_0(\delta)$ if $p(\boldsymbol{N}, \boldsymbol{n}) \leq \alpha$ and accept it otherwise. This procedure is commonly referred to as the Fisher randomized test or permutation test.

Returning to the topic of confidence intervals, we can obtain a confidence interval by inverting a series of permutation tests. In other words, for each possible value of δ, we perform a permutation test under $H_0(\delta)$ and include $\sum_{i=1}^{N} \delta_i / N$ in our confidence set for ATE if δ is accepted. This process results in a $1 - \alpha$ confidence set for ATE. In practice, our primary interest often lies in understanding the possible range of it. Therefore, rather than presenting a confidence set, we can also obtain a confidence interval with upper and lower limits determined by the largest and smallest elements in the confidence set. In the next section, we will again meet this issue about confidence sets and intervals, and we may use the terms interchangeably as the distinction between them is often not crucial for our purposes.

Since $\#(\mathcal{Z}(N, n)) = \binom{N}{n}$, it is computational inefficiency to derive the exact value of $p(\boldsymbol{N}, \boldsymbol{n})$ with large values of N. In such cases, it is advisable to employ Monte Carlo methods to approximate the p-value or use the (hit+1)/(replicate+1) trick to derive an exact p-value, as demonstrated in [Course notes-testing] (?).

It's worth noting that although we specify all the treatment effects when conducting a permutation test, the distribution of the test statistic is actually

fully determined by the potential outcome table $\boldsymbol{N}$. Therefore, the number of permutation tests that need to be performed is determined by the possible numbers of potential outcome tables, which is on the order of $O(N^3)$. To see this, we can observe that for a potential outcome table $\boldsymbol{N} = (N_{11}, N_{10}, N_{01}, N_{00})$ with the constraint $N_{11}+N_{10}+N_{01}+N_{00} = N$, there are at most $\binom{N+3}{3}$ potential tables that need to be tested.

To reduce the algorithm complexity,Li and Ding [2016] proposed a method that requires $O(N^2)$ permutation tests. Aronow et al. [2023] introduced a method that requires $O(N \log N)$ permutation tests when the experiment is balanced. In Section A, we will extend their methods to the stratified case.

4 Deriving Exact Confidence Interval

In this section, we extend the methods presented in Section 3 to the stratified randomized experiment.

4.1 Hypergeometric Test with Wendell& Schmee Method

For each stratum k, denote $N_{k1+} = N_{k10} + N_{k11}$ as the number of subjects in stratum k whose response would be 1 if assigned to active treatment and $N_{k+1} = N_{k01} + N_{k11}$ as the number of subjects in stratum k whose response would be 1 if assigned to control. Define $N_{1+} = \sum_{k=1}^{K} N_{k1+}$ and $N_{+1} = \sum_{k=1}^{K} N_{k+1}$. Then the treatment effect can be expressed as $\tau = (N_{1+} - N_{+1})/N$. Similarly to what was discussed in Section 3.1, we can obtain an exact confidence interval of τ by combining confidence intervals for N_{1+} and N_{+1} with the Bonferroni adjustment. The challenge lies in obtaining exact confidence intervals for N_{1+} and N_{+1}, as they involve sums of several hypergeometric parameters.

Wendell and Schmee [1996] studied this inference problem for sums of hypergeometric parameters: In this senario, a population of N items of which M are labeled "1" and $N - M$ are labeled "0" is allocated into K strata. Strata k contains N_k items of which M_k are labeled "1". We draw a simple random sample of size n_k items from stratum k, independently across strata. Let y_k denote the number of items labeled "1" in the sample from stratum k. To test the null hypothesis $H_0 : M = m$, Wendell and Schmee [1996] initially use the tail probability of the following statistics:

$$\hat{p} := \frac{1}{N} \sum_{k=1}^{K} \frac{N_k y_k}{n_k}$$

as the p-value for the hypotheses $M_k = m_k, k = 1, \ldots, K$. Subsequently, they consider the maximum of this tail probability over all possible values of $(m_1, \ldots, m_K)$ that satisfy the constraint $\sum_{k=1}^{K} m_k = m$. This maximum tail probability is then treated as the p-value for the hypothesis $H_0 : M = m$. Furthermore, Wendell and Schmee [1996] proceeds to construct $1 - \alpha$ confidence sets that encompass all values of m for which the null hypothesis $H_0 : M = m$ is not rejected at the significance level of α.

Since both N_{1+} and N_{+1} can be viewed as sums of several hypergeometric parameters, we can use their methods to calculate the confidence intervals for N_{1+} and N_{+1}. Denote $[N_{1+}^{L,ws}(1-\alpha), N_{1+}^{U,ws}(1-\alpha)]$ and $[N_{+1}^{L,ws}(1-\alpha), N_{+1}^{U,ws}(1-\alpha)]$ as the $1-\alpha$ confidence interval for N_{1+} and N_{+1} derived through their approach. Then, by applying the Bonferroni adjustment, we can determine that

$$\Big[\frac{N_{1+}^{L,ws}(1-\alpha/2) - N_{+1}^{U,ws}(1-\alpha/2)}{N}, \frac{N_{1+}^{U,ws}(1-\alpha/2) - N_{1+,}^{L,ws}(1-\alpha/2)}{N}\Big]$$

is an exact $(1-\alpha)$ confidence interval for the ATE.

The procedure described here bears a resemblance to the approach outlined in Section 5 of Rigdon and Hudgens [2015]. While it holds an intuitive appeal, it presents challenges in terms of computational complexity, as finding the optimal combination of parameters can be a demanding task. In our simulations, we have found that this method becomes impractical when dealing with more than a few strata and moderate sample sizes. Furthermore, Wendell and Schmee [1996] has demonstrated that the tail probability is not convex in its original parametrization, which complicates efforts to enhance computational efficiency. In the following section, we will introduce an alternative, more efficient approach that builds upon a similar concept.

4.2 Hypergeometric Test with Combining Function

Recently, Stark [Stark ??] presented a fast constructive method to find the optimal allocation with a specific test statistics: Fisher combining function (in fact, it works for any p-value combing function) applied to the p-values from the individual stratumwise tests of hypergeometric parameters. Then, similarly in the previous section, we can find the $1-\alpha$ confidence interval for N_{1+} and N_{+1} by their approach, denote as $[N_{1+}^{L,fast}(1-\alpha), N_{1+}^{U,fast}(1-\alpha)]$ and $[N_{+1}^{L,fast}(1-\alpha), N_{+1}^{U,fast}(1-\alpha)]$. By Bonferroni adjustment,

$$\Big[\frac{N_{1+}^{L,fast}(1-\alpha/2) - N_{+1}^{U,fast}(1-\alpha/2)}{N}, \frac{N_{1+}^{U,fast}(1-\alpha/2) - N_{1+,}^{L,fast}(1-\alpha/2)}{N}\Big]$$

is an exact $(1-\alpha)$ confidence interval for τ.

4.3 Extended Inverted Permutation Test

In the stratified setting, we can still use the idea in Section 3.2 to obtain a confidence interval. To make this precise, consider the following hypothesis:

$$H_0(\delta) : y_{kj}(1) - y_{kj}(0) = \delta_{kj}, \forall k = 1, \ldots, K, j = 1, \ldots, N_k \tag{9}$$

where $\delta = (\delta_{11}, \ldots, \delta_{1N_1}, \ldots, \delta_{k1}, \ldots, \delta_{kN_k})$ is a known vector. Under the null, the potential outcomes are fully determined and can be summarized in a table

$\boldsymbol{N} = (\boldsymbol{N}_k)_{k=1}^K$. Moreover, in a stratified randomized experiment, $\boldsymbol{Z}$ is uniformly distributed over the set:

$$\mathcal{Z}(N_1, \ldots, N_K, n_1, \ldots, n_K) := \Big\{ \sum_{j=1}^{N_k} Z_{kj} = n_k, \quad k = 1, 2, .., K. \Big\}$$

Thus, follow the discuss in Section 3, we can use the below probability as the p-value of $H_0(\delta)$:

$$p(\boldsymbol{N}, \boldsymbol{n}) = \mathbb{P}\Big(|\tau(\boldsymbol{N}) - \hat{\tau}(\bar{\tilde{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z}))| \geq |\tau(\boldsymbol{N}) - \hat{\tau}(\boldsymbol{n})| \Big), \tag{10}$$

where the probability is with respect to $\boldsymbol{Z}$. We can then reject $H_0(\delta)$ if $p(\boldsymbol{N}_0, \boldsymbol{n}) \leq \alpha$ and accept it otherwise. To derive a confidence interval for ATE, we can perform a permutation test under $H_0(\delta)$ for each possible value of δ. Then, we select the largest and smallest values of $\bar{\delta} = \sum_{k,j} \delta_{kj}/N$ obtained from these permutation tests as the upper and lower bounds of the confidence interval for ATE.

This method uses a similar idea in Section 5 in Rigdon and Hudgens [2015] and Chiba [2017]. However, as reported in both papers, the computational complexity is extremely high. Since for each stratum in the potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \ldots, \boldsymbol{N}_K)$, there are approximately $O((N_k)^3)$ possible values of $\boldsymbol{N}_k$ to consider. The number of permutation tests that need to be performedin total is on the order of $O(\prod_{k=1}^K (N_k)^3)$. In Section A, we will explore methods to reduce this high algorithm complexity and make the computations more manageable. In the upcoming section, we will also introduce an alternative method that demands less computational time.

4.4 Inverted Permutation Test with Combing Function

Besides the method in the previous section, we can also construct a confidence interval for τ by extending the ideas in 3.2 and 4.1. Let's first consider testing of a null hypothesis

$$H_0(\tau_{10}, \ldots, \tau_{k0}) : \tau_i = \tau_{i0}, \forall i \in [k],$$

This hypothesis can be expressed as an intersection of k hypotheses:

$$H_0(\tau_{10}, \ldots, \tau_{k0}) = \bigcap_{i=1}^{k} (H_{i0} : \tau_i = \tau_{i0})$$

To compute a p-value for an intersection hypothesis, we can first obtain p-values from the individual hypotheses and then use a p-value combining function to combine them across strata. From Section 3.2, the p-value for the hypothesis $H_{i0} : \tau_i = \tau_{i0}$ can be calculated as:

$$p(\tau_{k0}) = \max_{\tau(N_k) = \tau_{k0}} p(\boldsymbol{N}_k, \boldsymbol{n}_k)$$

By applying a p-value combining function to $p(\tau_{10}), \ldots, p(\tau_{k0})$, we can define a test statistic for $H_0(\tau_{10}, \ldots, \tau_{k0})$. For example, if we choose the Fisher combining function, the test statistic for $H_0(\tau_{10}, \ldots, \tau_{k0})$ becomes:

$$t_F(\tau_{10}, \ldots, \tau_{K0}) = -\sum_{k=1}^{K} \ln(p(\tau_{k0}))$$

Since t_F follows a chi-square distribution, we can reject $H_0(\tau_{10}, \ldots, \tau_{k0})$ if $t_F(\tau_{10}, \ldots, \tau_{k0}) \geq \chi^2_{2k}(\alpha)$, where $\chi^2_{2k}(\alpha)$ is the $1-\alpha$ quantile of the chi-square distribution with $2k$ degrees of freedom. Here, we have chosen the Fisher combing function as the default option. The rationale for this decision is explained in Appendix C.

To construct a confidence set for τ, we test the hypothesis $H_0(\tau_{10}, \ldots, \tau_{K0})$ for each possible combination of $\tau_{10}, \ldots, \tau_{K0}$ and include $\sum_{k=1}^{K} N_k \tau_{k0}/N$ in our $1-\alpha$ confidence set if $H_0(\tau_{10}, \ldots, \tau_{K0})$ is not rejected.

Note that, unlike the method described in the previous section, we do not need to perform a permutation test for every combination of potential outcome tables $(\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)$. Instead, we can conduct permutation tests in each stratum to obtain p-values for each possible value of τ_k and then combine them using p-value combining functions. The advantage of this approach is that there are only N_k possible values of τ_k in each stratum, resulting in a time complexity of $O(\prod_{k=1}^{K} N_k)$ for finding the confidence interval. Also, the time complexity for finding the p-values in stratum k is at most $O(AN_k^3)$ if we choose to use Monte Carlo permutation test with A samples (For the choice of A, more details can be found in Aronow et al. [2023]). Thus, the overall time complexity is $O(\max\{AN_{max}^3, \prod_{k=1}^{K} N_k\})$ where $N_{max} = \max\{N_1, \ldots, N_K\}$. This is much less than the complexity $O(\prod_{k=1}^{K} N_k^3)$ as the method in the previous section.

5 Illustration

5.1 Simulation study

In this section, the four presented methods are compared with the conventional Wald interval:

$$\hat{\tau} \pm z_{1-\alpha/2} \left\{ \sum_{k=1}^{K} \left(\frac{N_k}{N} \right)^2 \left(\frac{\hat{S}_k^2(1)}{n_k} + \frac{\hat{S}_k^2(0)}{N_k - n_k} \right) \right\}^{1/2} \tag{11}$$

where $S_k^2(1) = (n_k - 1)^{-1} \sum_{j=1}^{N_k} Z_{kj}(Y_{kj} - \hat{\bar{Y}}_k(1))$ and $S_k^2(0) = (N_k - n_k - 1)^{-1} \sum_{j=1}^{N_k} (1 - Z_{kj})(Y_{kj} - \hat{\bar{Y}}_k(0))$, and $\hat{\bar{Y}}_k(1) = 1/n_k \sum_{j=1}^{N_k} Z_{kj} Y_{kj}$, $\hat{\bar{Y}}_k(0) = 1/(N_k - n_k) \sum_{j=1}^{N_k} (1 - Z_{kj}) Y_{kj}$.$z_{1-\alpha/2}$ denote the $1-\alpha/2$ quantile of a standard normal distribution. As $N_k \to \infty$ with $n_k/N_k \to c_k \in (0,1)$ for all $k \in [K]$, the interval (11) will contain τ with probability $1-\alpha$.

Data were first simulated under a 'balanced' setting, where the data sizes in each stratum are equal, and the experiments are balanced in each stratum.

Simulations were carried out for strata size equals to 2, various size of N and τ_1, τ_2 using the following steps:

1. Potential outcomes in each strata k were generated by first letting $y_{kj}(1) = 1$ and $y_{kj}(0) = 0$ for subjects $j = 1, \ldots, \tau_k N_k$. Then for $j = \tau_k N_k + 1, \ldots, N_k$, the potential outcome $y_{kj}(1)$ was sampled from a Bernoulli distribution with mean 0.5. The potential outcome $y_{\tau_i N_k+1}(0), \ldots, y_{N_k}(0)$ were set equal to a random permutation of $y_{\tau_k N_k+1}(1), \ldots, y_{N_k}(1)$, in order to guarantee the average treatment effect equal to τ_k.

2. Observed data in each strata k were generated by randomly assigning $N_k/2$ subjects to treatment and the others to control. Observed outcomes were then generated based on the treatment assignment and the potential outcomes in step 1.

3. Compute the 95% confidence intervals using the four presented methods and the Wald method. A random sample with replacement of 100 randomizations was used to approximate permutation tests

4. Repeat 100 times. Report the average width of the confidence intervals.

The results in Figure **??** reveal that, on average, among the four exact methods, the extended permutation confidence set achieved the narrowest width. It was followed by the attributable effect method with the Wendell and Schmee approach, then the combining permutation method. The last one is the attributable effect with combing function method. The Wald method falls between the extended permutation and the Wendell and Schmee method.

We also conducted another simulation for a more general setting, where the data is not 'balanced' anymore. The results are shown in Table 1. Simulations were carried out similarly to the first setting, but instead of randomly producing potential outcomes, we provided the potential outcome table in advance. We also included the corresponding coverage rate and running time in the table.

The results in Table 1 reveal that, even in a non-balanced setting, the extended permutation confidence set also generates the narrowest confidence interval among these exact methods. Next is the Wendell & Schmee method, followed by the combing permutation method. The last is the combing attributable effect method. However, there are some extreme cases where the order does not follow this pattern; we ran additional simulation results in Appendix C.2 for more details. The Wald method sometimes produces the narrowest confidence interval but may fail to achieve the nominal coverage rate, even when the sharp null is true, indicating no treatment effect at all.

However, there is a trade-off between statistical efficiency and computational efficiency. As demonstrated in Table 1, the Wendell & Schmee method and the extended permutation method become impractical when $K \geq 3$ and $N \geq 100$. Notably, the extended permutation method, which yields the narrowest confidence intervals, required up to 30 minutes to compute results for the case of $K = 3$ and $N = 60$. Although the algorithms proposed in Section A can improve its computational efficiency, our simulation (not shown) suggest that

the extended permutation method still remains unfeasible in many practical scenarios.

5.2 Case study

In a recent study of vedolizumab treatment effect of chronic pouchitis Travis et al. [2023], 102 patients were randomized to receive vedolizumab or placebo. Randomization was stratified according to continuous antibiotic use at baseline (yes vs. no). In the 'yes' strata, the incidence of mPDAI-defined remission at week 14 was 27.6%(8 of 29 patients) with vedolizumab and 12.0% (3 of 25 patients) with placebo; In the 'no' strata, the incidence of mPDAI-defined remission at week 14 was 36.4%(8 of 22 patients) with vedolizumab and 7.7% (2 of 26 patients) with placebo. The 95% confidence interval derived from the four methods we present, the Wald method, and the Cochran-Mantel-Haenszel test in their study is summarized in table 2.

Thus, for this example, the extended permutation confidence set is the narrowest of the four exact approaches. The permutation confidence set has the same width as the Wald interval and is slightly narrower than the Cochran-Mantel-Haenszel interval. However, unlike the Wald or CMH intervals, the permutation confidence set is guaranteed to cover at the nominal level.

6 Discussion

In this paper, we have introduced four methods for constructing exact confidence sets for the average treatment effect in stratified randomized experiments with binary outcomes. These methods are extensions of non-stratified approaches developed by Rigdon and Hudgens [2015] and Li and Ding [2016]. We extended their techniques by either directly extending the ideas or integrating their methods with p-value combining functions. All of these methods are non-parametric and provide exact confidence sets, guaranteeing a probability of at least $1 - \alpha$ of containing the true treatment effect.

The extended permutation method generally yields the narrowest confidence intervals and is recommended for situations with small data sizes (such as $K \leq 2$, $N \leq 200$, or $K \leq 3, N \leq 100$). When dealing with larger datasets, the permutation method and the Wendell & Schmee method may not be feasible. In such cases, we recommend using the fast method or the combining permutation method. Additionally, when the dataset is large, the Wald method can also be employed for inference, as it often offers greater statistical power compared to these exact methods. However, it's important to note that the Wald method may fail to achieve the nominal coverage rate in extreme cases, such as when $|\tau_i|$ is close to 1 for some strata i or when the sharp null hypothesis is nearly true.

Several potential avenues for future research in this field exist. Firstly, improving the computational efficiency of the extended permutation method could be a focus, making it more feasible for larger and more complex data sets. An-

N	n	τ	$\boldsymbol{N}$	Wald	Fast	Ws	Comb	Perm
(40,40)	(10,10)	(0,0)	[10,10,10,10], [10,10,10,10]	0.51, 96%, 0s	0.67, 100%, 0.01s	0.57, 100%, 12.78s	0.57, 100%, 0.46 s	0.5, 99%, 181.27s
(20,20)	(15,15)	(0.2,0.9)	[3,8,4,5], [0,19,1,0]	0.53, 90%, 0s	0.75, 98%, 0s	0.64, 100%, 0.9s	0.63, 100%, 0.05 s	0.54, 100%, 3.97s
(30,40)	(5,30)	(0.7,-0.7)	[3,23,2,2], [4,2,30,4]	0.4, 99%, 0s	0.7, 100%, 0.01s	0.68, 100%, 6.1s	0.61, 100%, 0.19s	0.54, 100%, 21.78s
(30,30)	(5,25)	(0.8,0.8)	[2,24,0,4], [1,26,2,1]	0.38, 65%, 0s	0.72, 100%, 0.01s	0.62, 100%, 0.82s	0.53, 100%, 0.08s	0.43, 99%, 17.54s
(10,40)	(5,20)	(-0.9,1)	[1,0,9,0], [0,40,0,0]	0.09, 56%, 0s	0.48, 100%, 0.01s	0.51, 100%, 1.15s	0.35, 100%, 0.03s	0.4, 100%, 1.89s
(20,80)	(15,60)	(0,0.6)	[5,5,5,5], [20,50,2,8]	0.4, 96%, 0s	0.54, 100%, 0.02s	0.47, 100%, 13.24s	0.51, 100%, 1.04s	0.4, 99%, 438.26s
(15,60)	(10,40)	(0.8,0.9)	[2,12,0,1], [2,55,1,2]	0.22, 95%, 0s	0.35, 100%, 0.02s	0.28, 100%, 1.04s	0.31, 99%, 0.23s	0.22, 98%, 43.41s
(20,70)	(5,60)	(-0.5,0.9)	[2,2,12,4], [3,64,1,2]	0.28, 97%, 0s	0.59, 100%, 0.01s	0.53, 100%, 4.66s	0.54, 100%, 0.16s	0.44, 100%, 60.12s
(20,20,20)	(5,10,15)	(0.8,0.4,0)	[0,16,0,4], [3,9,1,7], [5,5,5,5]	0.49, 98%, 0s	0.76, 100%, 0.01s	0.64, 100%, 287.23s	0.65, 100%, 0.08s	0.51, 100%, 1277.17s
(15,20,25)	(10,10,10)	(0.8,0.9,0.8)	[1,13,1,0], [0,18,0,2], [0,20,0,5]	0.3, 97%, 0s	0.51, 100%, 0.01s	0.36, 100%, 59.4s	0.43, 100%, 0.09s	0.29, 99%, 1729.77s
(20,15,20)	(5,5,5)	(0.9,0,-0.8)	[0,19,1,0], [3,4,4,4], [0,2,18,0]	0.41, 99%, 0s	0.81, 100%, 0.01s	0.65, 100%, 244.09s	0.69, 100%, 0.07s	0.55, 100%, 81.74s
(10,20,30)	(5,5,25)	(0,0,0)	[5,0,0,5], [6,0,0,14], [18,1,1,10]	0.61, 92%, 0s	0.88, 100%, 0.01s	0.77, 100%, 197.29s	0.74, 100%, 0.09s	0.58, 98%, 469.31s
(30,40,50)	(10,10,10)	(0.5,0.5,0.5)	[8,15,0,7], [9,21,1,9], [12,26,1,11]	0.36 97%, 0s	0.55, 99%, 0.02s	*	0.48, 99%, 0.6s	*
(40,40,40)	(20,30,10)	(0.5,-0.6,0)	[10,20,0,10], [7,1,25,7], [12,8,8,12]	0.36 97%, 0s	0.58, 100%, 0.02s	*	0.51, 100%, 0.53s	*
(50,50, 50,80)	(25,25, 25,40)	(0,0,0,0)	[5,0,0,45], [10,0,0,40], [10,0,0,40] [5,0,0,75]	0.17 91%, 0s	0.29, 100%, 0.05s	*	0.34, 100%, 6.11s	*

Table 1: Simulation results for unbalanced setting. The three numbers in the last 5 cells represents (average width, coverage rate, average time). * indicates this method can't finish running in 30 minutes.

Methods	95 % Confidence Interval for ATE
WS	[0.04,0.39]
fast	[-0.01,0.41]
perm	[0.06,0.38]
perm comb	[0.02,0.40]
Wald	[0.06,0.38]
CMH	[0.05,0.38]

Table 2: Confidence intervals for vedolizumab treatment effect of chronic pouchits [nejm2023] from different methods

other promising research direction is enhancing the statistical efficiency of other computationally manageable methods like the combing permutation method and the fast method. Additionally, the application of these methods to observational studies and the handling of missing data represent intriguing areas for further exploration.

6.1 Missing Data

Treating in the least favorable way versus imputation / multiple imputation. Least favorable treatment is computationally tractable whenever the basic method is: to find an upper bound on the treatment effect, treat missing control cases as 0 and missing treated cases as 1. For the lower bound, do the opposite.

6.2 Confidence Intervals for Relative Risk

Same approach works, but the choice of test statistics matters. If the same test statistic is used for ATE and RR, the confidence bounds are simultaneous.

References

Ronald A. Fisher. The design of experiments. 1935.

J. Wu and P. Ding. Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, 116(536):1898–1913, 2021. doi: 10.1080/01621459.2020.1750415.

Pierre Charles, Bruno Giraudeau, Agnes Dechartres, Gabriel Baron, and Philippe Ravaud. Reporting of sample size calculation in randomised controlled trials: review. *BMJ*, 338, 2009. ISSN 0959-8138. doi: 10.1136/bmj.b1732. URL `https://www.bmj.com/content/338/bmj.b1732`.

Yasutaka Chiba. Exact tests for the weak causal null hypothesis on a binary out come in randomized trials. *Journal of Biometrics & Biostatistics*, 06, 01 2015. doi: 10.4172/2155-6180.1000244.

Joseph Rigdon and Michael G. Hudgens. Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine*, 34(6):924–935, 2015. doi: https://doi.org/10.1002/sim.6384. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6384`.

Xinran Li and Peng Ding. Exact confidence intervals for the average causal effect on a binary outcome. *Statistics in Medicine*, 35(6):957–960, 2016. doi: https://doi.org/10.1002/sim.6764. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.6764`.

P. M. Aronow, Haoge Chang, and Patrick Lopatto. Fast computation of exact confidence intervals for randomized experiments with binary outcomes. In *Proceedings of the 24th ACM Conference on Economics and Computation*, EC '23, page 120, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701047. doi: 10.1145/3580507.3597750. URL `https://doi.org/10.1145/3580507.3597750`.

Cydney Bruce, Edmund Juszczak, Reuben Ogollah, Christopher Partlett, and Alan Montgomery. A systematic review of randomisation method use in rcts and association of trial design characteristics with method selection. *BMC Medical Research Methodology*, 22, 12 2022. doi: 10.1186/s12874-022-01786-4.

Yasutaka Chiba. Stratified exact tests for the weak causal null hypothesis in randomized trials with a binary outcome. *Biometrical Journal*, 59(5):986–997, 2017. doi: https://doi.org/10.1002/bimj.201600085. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/bimj.201600085`.

Weizhen Wang. Exact optimal confidence intervals for hypergeometric parameters. *Journal of the American Statistical Association*, 110(512):1491–1499, 2015. doi: 10.1080/01621459.2014.966191. URL `https://doi.org/10.1080/01621459.2014.966191`.

John P. Wendell and Josef Schmee. Exact inference for proportions from a stratified finite population. *Journal of the American Statistical Association*, 91 (434):825–830, 1996. ISSN 01621459. URL http://www.jstor.org/stable/2291677.

Simon Travis, Mark S. Silverberg, Silvio Danese, Paolo Gionchetti, Mark Löwenberg, Vipul Jairath, Brian G. Feagan, Brian Bressler, Marc Ferrante, Ailsa Hart, Dirk Lindner, Armella Escher, Stephen Jones, and Bo Shen. Vedolizumab for the treatment of chronic pouchitis. *New England Journal of Medicine*, 388(13):1191–1200, 2023. doi: 10.1056/NEJMoa2208450. URL https://doi.org/10.1056/NEJMoa2208450. PMID: 36988594.

A Reduce the Computational Complexity for Extended Inverted Permutation Test

A.1 A General Approach

In this section, we present a general approach to reduce the computational complexity of the method outlined in Section 4.4. Given that we do not make any specific assumptions about the data, the result may not be surprising: Roughly speaking, if all the strata contain a similar number of subjects, then the number of permutation tests we need to perform is of the order $O(N^{3k-1})$. However, if some or all of the strata are balanced, we can achieve significantly better computational efficiency, as will be demonstrated in the following two subsections.

Theorem A.1. *Let $\boldsymbol{n}$ be an observed table. Let*

$$\begin{aligned}\Delta_{LD} = \Big\{ e_k \cdot \delta_k \Big| k \in [K], \delta_k \in \{(-1,1,0,0),(0,0,-1,1)\} \text{ if } n_k < N_k/2, \\ \delta_k \in \{(0,1,0,-1),(1,0,-1,0)\} \text{ if } n_k > N_k/2 \\ \delta_k \in \{(0,1,0,-1),(1,0,-1,0), \\ (-1,1,0,0),(0,0,-1,1)\} \text{ if } n_k = N_k/2 \Big\} \quad (12)\end{aligned}$$

where $\{e_1, e_2, \ldots, e_K\}$ is the standard basis in $\mathbb{R}^K$. Consider two potential outcome table $\boldsymbol{N}$ and $\boldsymbol{N}'$, assume $\boldsymbol{N}' = \boldsymbol{N} + \delta_{LD}, \delta_{LD} \in \Delta_{LD}$. If $\tau(\boldsymbol{N}') \leq \hat{\tau}(\boldsymbol{n})$, then $p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$. If $\tau(\boldsymbol{N}) \geq \hat{\tau}(\boldsymbol{n})$, then $p(\boldsymbol{N}', \boldsymbol{n}) \leq p(\boldsymbol{N}, \boldsymbol{n})$.

This theorem extends the findings of Lemma A.3 in [LD16]. Intuitively, it shows that if a potential outcome table exhibits a value of τ that is closer to the observed $\hat{\tau}$ than another potential outcome table, then the probability of the data originating from the former potential outcome table is higher than that of the latter.

Based on Theorem A.1, we now propose a faster algorithm of the method in Section 4.3. In this algorithm, we suppose $K \geq 2$ since the case about $K = 1$ is already well-studied, see [Li and Ding, 2016, Aronow et al., 2023].

Algorithm A.1. (Obtain a confidence set in a faster way)

Input: An observed table $\boldsymbol{n}$ and the significance level α

Output: An $1-\alpha$ confidence set S for the average treatment effect.

1. Initialize $S=\emptyset$, $N_k=\sum_{ab} n_{kab}$, $N=\sum_k N_k$, $n=n_{k10}+n_{k11}$, $n=\sum_k n_k$.
2. For every element in the following set:

$$\Big\{(M_{101},\ldots,M_{K01},M_{100},\ldots,M_{K00})\Big|0\le M_{k01}\le N_k, M_{k01}\in\mathbb{Z},\\ 0\le M_{k00}\le N_k-M_{k01}, M_{k00}\in\mathbb{Z}\Big\},$$

   do the following thing:

   (a) Use Algorithm A.2 to find the $1-\alpha$ confidence set S' of τ among the following set of potential outcome tables:

$$\Big\{\boldsymbol{N}\Big|N_{k00}=M_{k00},N_{k01}=M_{k01}k\in[K]\Big\}$$

   (b) Add all the values in S' that found in (a) to S.

3. Return S as the $1-\alpha$ confidence set.

Theorem A.2. *Algorithm A.1 provides the same $1-\alpha$ confidence set as in section 4.3. It needs at most $O((N_1+N_2)*\prod_{k=3}^{K}N_k*\prod_{k=1}^{K}(N_k)^2)$ permutation tests.*

The key point is that by using Theorem A.1, the set

$$\Big\{\boldsymbol{N}\Big|N_{k00}=M_{k00},N_{k01}=M_{k01}k\in[K]\Big\}$$

has a monotonic p-value within each strata, so we can construct an efficient way to search in this set. Algorithm A.2 is such a way that requires $O((N_1+N_2)*\prod_{k=3}^{K}N_k)$ times of searching. Now we state Algorithm A.2. Suppose $n_1\le N_1/2$ and $n_2\le N_2/2$ for convenience. The other case is analogous so we omit here.

Algorithm A.2. (Obtain a confidence set in the set in Algorithm A.1(a))

Input: An observed table $\boldsymbol{n}$, the significance level α and a set of potential outcome tables:

$$A=\Big\{\boldsymbol{N}\Big|N_{k00}=M_{k00},N_{k01}=M_{k01}k\in[K]\Big\}$$

Output: An $1-\alpha$ confidence set S for the average treatment effect in the given set.

1. Initialize $S=\emptyset$. Initialize N_k, n_k, N and n.
2. For every element in the set:

$$\Big\{(M_{311},\ldots,M_{K11})\Big|0\le M_{k11}\le N_k-M_{k00}-M_{k01},3\le k\le K\Big\},$$

   do the following thing:

(a) Find all potential outcome tables in A such that $N_{k11} = M_{k11}, 3 \leq k \leq K$. Denote the set as B. For a given non-negative integer l such that $l \leq N_2 - M_{201} - M_{200}$, let

$$\underline{N_{110}}(l) = \arg\max_j \{\boldsymbol{N} | \boldsymbol{N} \in B, N_{210} = l, N_{110} = j, \tau(\boldsymbol{N}) < \hat{\tau}(\boldsymbol{n})\}$$

If the set in the right hand side is empty, set $\underline{N_{110}}(l) = -1$. Let

$$\overline{N_{110}}(l) = \arg\max_j \{\boldsymbol{N} | \boldsymbol{N} \in B, N_{210} = l, N_{110} = j, p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha \text{ and } \tau(\boldsymbol{N}) \geq \hat{\tau}(\boldsymbol{n})\}$$

If the set in the right hand side is empty, set $\overline{N_{110}}(l) = \underline{N_{110}}(l)$. We now find $\overline{N_{110}}(l)$ for all $0 \leq l \leq N_2 - M_{201} - M_{200}$.

(b) When $N_{210} = 0$, we find $\overline{N_{110}}(0)$ by performing randomization tests starting from the maximum value of N_{110}: $N_1 - M_{101} - M_{100}$, and then working our way downwards until we find a potential table $\boldsymbol{N}$ such that $p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha$ or $\tau(\boldsymbol{N}) < \hat{\tau}(\boldsymbol{n})$. When N_{210} increases to 1, we find $\overline{N_{110}}(1)$ by performing randomization tests starting from $\overline{N_{110}}(0)$. Sequentially, when N_{210} increase by 1, we find $\overline{N_{110}}(N_{210})$ by performing randomization tests starting from $N_{110} = \overline{N_{110}}(N_{210} - 1)$. We repeat this process until N_{210} increases to $N_2 - M_{201} - M_{200}$. The logic can be found in Section B.5.

(c) Find all possible values of τ such that

$$\exists \boldsymbol{N} \in B, s.t. \tau(\boldsymbol{N}) = \tau, \boldsymbol{N} \text{ is compatible with } \boldsymbol{n}, \text{ and } \underline{N_{110}}(N_{210}) < N_{110} \leq \overline{N_{110}}(N_{210})$$

This can be done in at most $O(N_2 - M_{201} - M_{200})$ time as shown in section B.5.

(d) Similarly in (a) to (c), we can find all possible values of τ such that

$$\exists \boldsymbol{N} \in B, s.t. \tau(\boldsymbol{N}) = \tau, \boldsymbol{N} \text{ is compatible with } \boldsymbol{n}, p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha \text{ and } \tau(\boldsymbol{N}) \leq \hat{\tau}(\boldsymbol{n})$$

(e) Add all values τ found in (c) and (d) to S

3. Return S as the $1 - \alpha$ confidence set.

Remark. *This algorithm is the same idea as in Section 3 in [Li and Ding, 2016], or the "saddleback search" in computer science.*

Note that we only relied on the monotonicity of p-values in the first two strata in this algorithm. However, the monotonicity of p-values exists in every strata. Therefore, there may be room for improvement if one can discover a more efficient algorithm or unearth alternative theories regarding the behavior of p-values.

Our simulations using Algorithm A.1 shows improvements over the original algorithm in Section 4.3. See Table 3 for example.

n	CI	A.1	4.3
$[3,10,1,1]$ $[1,1,1,13]$	$[-10,15]$	4,875	14,454
$[5,5,1,9]$ $[2,8,2,7]$	$[-3,17]$	85,193	209,880
$[3,2,2,3]$ $[1,1,1,4]$ $[0,9,1,5]$	$[-8,10]$	199,592	466,560

Table 3: 95% confidence intervals for $N * \tau$. The second column gives the confidence interval after scaling by N. The remaining columns indicate the number of permutation tests required for Algorithm A.1 and the original algorithm in Section 4.3. A random sample with replacement of 10,000 randomizations was used to approximate permutation tests.

A.2 Partially Balanced Data

Suppose we have a observed table $\boldsymbol{n}$. In some strata, we have $N_k = 2n_k$, that is, the number of the treatment group is the same as the number of the control group. Then, we say this observed table $\boldsymbol{n}$ is *partially balanced*. In this section, we suppose the first l strata ($1 \leq l \leq K$) in $\boldsymbol{n}$ is balanced. In this section, we will show that we only need to do $O((\sum_{k=1}^{l} N_k) * \prod_{k=1}^{l} (N_k)^2 * \prod_{k=l+1}^{K} (N_k)^3)$ permutation tests. If each strata has similar size of data, the number is about $O(N^{3K-l+1})$, which is much smaller than the original number $O(N^{3K})$

Lemma A.1. *Consider a potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$. The experiment $\boldsymbol{Z}$ are assigned such that the first l strata are balanced. Then, for any $i \in [l]$, let x_{kab} be the number of the subjects in the set $\{c : y_{kc}(1) = a, y_{kc}(0) = b\}$ who are assigned to treatment. Then,*

$$\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = \frac{\sum_{k=1}^{l}(2x_{k11} - N_{k11})}{N} - \frac{\sum_{k=1}^{l}(2x_{k00} - N_{k00})}{N} + \frac{\sum_{k=1}^{l}(N_{k10} - N_{k01})}{N} + \frac{1}{N}\sum_{k=l+1}^{K} N_k \hat{\tau_k}.$$

By this representation, one can see that the variation of the distribution of $\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ is affected by N_{k10} and $N_{k01} (1 \leq k \leq l)$ only through $\sum_{k=1}^{l}(N_{k10} - N_{k01})$. To be precise, see the next theorem:

Theorem A.3. *Consider an observed table $\boldsymbol{n}$ in which the first l strata are balanced. If two potential outcome table $\boldsymbol{N}_1$ and $\boldsymbol{N}_2$ satisfies*

1. $(N_1)_{k11} = (N_2)_{k11}, (N_1)_{k00} = (N_2)_{k00}, \forall k \in [K]$
2. $(N_1)_{k10} = (N_2)_{k10}, (N_1)_{k01} = (N_2)_{k01}, \forall k \in \{l+1, l+2, \ldots, K\}$

3. $\sum_{k=1}^{l}[(N_1)_{k10} - (N_1)_{k01}] = \sum_{k=1}^{l}[(N_2)_{k10} - (N_2)_{k01}]$

Then, we have

$$p(\boldsymbol{N}_1, \boldsymbol{n}) = p(\boldsymbol{N}_2, \boldsymbol{n})$$

Based on Theorem A.3, the next algorithm takes the significance level α and an observed table as input and outputs a confidence set which is the same as the one in Section 4.4.

Algorithm A.3. (Obtain a confidence set for partial balanced data)

Input: An observed table $\boldsymbol{n}$ and the significance level α

Output: An $1-\alpha$ confidence set S for the average treatment effect.

1. Initialize $S = \emptyset$, $N_k = \sum_{ab} n_{kab}$, $N = \sum_k N_k$, $n_k = n_{k10} + n_{k11}$, $n = \sum_k n_k$
2. For every element in the following set:

$$\Big\{(M_{111}, \ldots, M_{K11}, M_{100}, \ldots, M_{K00}, M_{(l+1)10}, \ldots, M_{K10}, a)\Big| \\ 0 \le M_{k11} \le N_k, M_{k11} \in \mathbb{Z}, k \in [K], \\ 0 \le M_{k00} \le N_k - M_{k11}, M_{k00} \in \mathbb{Z}, k \in [K], \\ 0 \le M_{k10} \le N_k - M_{k11} - M_{k00}, M_{k10} \in \mathbb{Z}, k \in \{l+1, \ldots, K\}, \\ |a| \le \sum_{k=1}^{l}(N_k - M_{k00} - M_{k11})\Big\},$$

do the following thing:

(a) Find a compatible potential outcome table $\boldsymbol{N}$ such that

$$N_{k11} = M_{k11}, k \in [k], N_{k00} = M_{k00}, k \in [k], N_{k10} = M_{k10}, k \in \{l+1, \ldots, k\}$$

and

$$\sum_{k=1}^{l}(N_{k10} - N_{k01}) = a.$$

In Section B.5, we showed that this can be done in a constant time. If no such compatible potential outcome table is found, go to the next loop.

(b) For the table in (a), if $\tau(\boldsymbol{N}) \in S$, go to the next loop. Otherwise, do permutation test with $\boldsymbol{N}$ and the given value of α. If $\boldsymbol{N}$ is accepted, put $\tau(\boldsymbol{N})$ in S.

3. Return S as the $1-\alpha$ confidence set.

Theorem A.4. *Algorithm A.3 provides the same $1-\alpha$ confidence set as in section 4.3. It needs at most $O((\sum_{k=1}^{l} N_k) * \prod_{k=1}^{l} N_k^2 * \prod_{k=l+1}^{K} N_k^3)$ permutation tests.*

Our simulations using Algorithm A.3 shows improvements over the original algorithm in Section 4.3 when some of the strata are balanced. See Table 4 for example.

n	CI	A.3	A.1	4.3
$[2,6,1,7]$ $[0,3,5,4]$	$[-12,5]$	30,240	17,320	30,240
$[2,3,1,4]$ $[2,7,5,4]$ $[0,10,1,9]$	$[-17,6]$	1,361,920	4,541,833	9,292,800
$[2,3,4,1]$ $[0,4,0,4]$ $[2,4,4,2]$ $[5,1,3,3]$	$[-15,8]$	6,301,809	35,393,259	56,800,800

Table 4: 95% confidence intervals for $N * \tau$. The second column gives the confidence interval after scaling by N. The remaining columns indicate the number of permutation tests required for Algorithm A.3, Algorithm A.1 and the original algorithm in Section 4.3. A random sample with replacement of 10,000 randomizations was used to approximate permutation tests.

A.3 Completely Balanced Data

Consider an observed table $\boldsymbol{n}$. In this section, we make the assumption that $N_k = 2n_k$ for all $k \in [K]$, meaning that the experiment is balanced within all strata. This balanced scenario significantly simplifies the analysis. For instance, we can establish that the confidence set we derive is, in fact, a confidence interval. Moreover, leveraging Theorem A.6 and A.7, which address the monotonicity of p-values, we can devise an algorithm that requires at most $O((\prod_{k=1}^{K} N_k)^2)$ permutations tests. In our simulation, the number is often much less than this order. More details can be found in the end of this section.

Theorem A.5. *Consider a observed table $\boldsymbol{n}$ with completely balanced experiment. The $1-\alpha$ confidence set found in Section 4.3 is actually a confident interval, in the sense that it must have the form*

$$\{\frac{l}{N}, \frac{l+1}{N}, \ldots, \frac{u-1}{N}\frac{u}{N}\}$$

for some integer l and u

Theorem A.6. *(An extension of Lemma 4.2 in [Aronow23])*
Fix observed table $\boldsymbol{n}$ and a potential table $\boldsymbol{N}$. Let

$$\delta_A := (+1, -1, -1, +1).$$

If an another potential outcome table $\boldsymbol{N}'$ satisfies

1. $\boldsymbol{N}' = \boldsymbol{N} + e_k \cdot \delta_A, k \in [K]$
2. $N * \tau(\boldsymbol{N}')$ *is odd, or* $N'_{k10} + N'_{k01} \geq 1, \forall k \in [K]$

Then we have

$$p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$$

Theorem A.7. *Fix observed table $\boldsymbol{n}$ and a potential table $\boldsymbol{N}$. Let δ_L^k be a function of the potential outcome table such that*

$$\delta_L^k(\boldsymbol{N}) = \begin{cases} (-2,0,0,+2) & \text{if } N_{k11} - 2 \geq N_{k00} + 2 \\ (+2,0,0,-2) & \text{if } N_{k11} + 2 \leq N_{k00} - 2 \\ (0,0,0,0) & \text{otherwise} \end{cases}$$

If an another potential outcome table $\boldsymbol{N}'$ satisfies $\boldsymbol{N}' = \boldsymbol{N} + e_k \cdot \delta_L^k(\boldsymbol{N}), k \in [K]$, we have

$$p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n}).$$

Basically, the above theorem says that if a potential table has a smaller difference between N_{k11} and N_{k00}, the p-value of the outcome table will be lager.

The above two theorems are all base on the intuition that if a potential table has a more 'spread out' distribution of $\hat{\tau}$, then $p(\boldsymbol{N}, \boldsymbol{n})$ will be larger. Following this idea, Aronow et al. [2023] and us found two different transformations of the potential table that can make the table become more 'spread out': If less subjects have a preference of the control or the treatment, or, if the number of the subjects who would always be cured and would never be cured are closed to each other, the distribution of $\hat{\tau}$ will be more spread out, as is shown in the below pictures:

[TBD]

We now give the steps for finding $U_\alpha(\boldsymbol{n})$. Finding $L_\alpha(\boldsymbol{n})$ is analogous so we omitted here.

Algorithm A.4. (Obtain a confidence interval for completely balanced data)

Input: An observed table $\boldsymbol{n}$ and the significance level α

Output: The upper bound $U_\alpha(\boldsymbol{n})$ of the $1-\alpha$ confidence interval for the average treatment effect.

1. Initialize N_k, N, n_k, n, $\hat{\tau}_i$ and $\hat{\tau}$. Initialize $U_\alpha(\boldsymbol{n}) = -1$.
2. For each possible ATE vector $\boldsymbol{T} = (\tau_1, \tau_2, \ldots, \tau_K)$ such that $\tau = \frac{1}{N}\sum_{k=1}^K N_k \tau_k \geq \hat{\tau}$, do the following thing:
   (a) If $\tau \leq U_\alpha(\boldsymbol{n})$, go to the next loop.
   (b) Find all elements in the following set. This can be done in $O(\prod_{k=1}^K N_k)$ time as shown in Section B.5:

$$\Big\{ \boldsymbol{N} \Big| \boldsymbol{N} \text{ is compatible}, \boldsymbol{T}(\boldsymbol{N}) = \boldsymbol{T}, \\ \boldsymbol{N} + e_k * \delta_A \text{ is not compatible for all } k \in [K], \\ \text{for all } k \in [K],\ \boldsymbol{N} + e_k * \delta_L^k(\boldsymbol{N}) \text{ is not compatible or } \delta_L^k(\boldsymbol{N}) = \boldsymbol{0}\ . \Big\}$$

(c) For every potential outcome $\boldsymbol{N}$ found in (b), do permutation test with $\boldsymbol{N}$ and the given value of α. If $\boldsymbol{N}$ is accepted, set $U_\alpha(\boldsymbol{n}) = \tau$ and go to 2 to the next loop.

(d) If $N * \tau$ is even, also find all elements in the following set. This can be done in $O(\prod_{k=1}^{K} N_k)$ time as shown in Section B.5. Note that this set is empty if $\tau_i \neq 0, \forall i \in [k]$

$$\Big\{\boldsymbol{N}\Big|\boldsymbol{N} \text{ is compatible}, \boldsymbol{T}(\boldsymbol{N}) = \boldsymbol{T},$$
$$\exists j \in [K], s.t. N_{j10} = N_{j01} = 0,$$
$$\text{for all } k \in [K],\ \boldsymbol{N} + e_k * \delta_L^k(\boldsymbol{N}) \text{ is not compatible or } \delta_L^k(\boldsymbol{N}) = \boldsymbol{0}.\Big\}$$
$$\bigcup \Big\{\boldsymbol{N}\Big|\boldsymbol{N} \text{ is compatible}, \boldsymbol{T}(\boldsymbol{N}) = \boldsymbol{T},$$
$$\forall j \in [k],\ \text{if } \tau_j = 0,\ \text{then } N_{j10} = N_{j01} = 1,$$
$$\text{if } \tau_j \neq 0,\ \text{then } N + e_j * \delta_A \text{ is not compatible},$$
$$\text{for all } k \in [K],\ \boldsymbol{N} + e_k * \delta_L^k(\boldsymbol{N}) \text{ is not compatible or } \delta_L^k(\boldsymbol{N}) = \boldsymbol{0}\ .\Big\}$$

(e) If $N * \tau$ is even, for every potential outcome $\boldsymbol{N}$ found in (d), do permutation test with $\boldsymbol{N}$ and the given value of α. If $\boldsymbol{N}$ is accepted, set $U_\alpha(\boldsymbol{n}) = \tau$ and go to 2 to the next loop.

3. Return $U_\alpha(\boldsymbol{n})$

Theorem A.8. *Algorithm A.4 provides the same $1 - \alpha$ confidence set as in section 4.3. It needs at most $O((\prod_{k=1}^{K} N_k)^2)$ permutation tests.*

Our simulations using Algorithm A.4 shows significant improvements over the other algorithms and the original algorithm in Section 4.3. See Table 5 for example.

We end this section by pointing out that the actual algorithmic complexity of Algorithm A.4 can sometimes be much less than $O((\prod_{k=1}^{K} N_k)^2)$. In an ideal scenario, the actual order would be $O(\prod_{k=1}^{K} N_k)$. There are two intuitions for this efficiency. First, in Step 2b, the number of elements can be as low as $O(1)$ because, in each stratum, there are three degrees of freedom for potential outcome tables, and the constraints in Step 2b eliminate two of them. Second, it's rarely necessary to execute Step 2d because the constraints in this step are highly restrictive. For instance, it would require conditions like $\tau_k = 0$ for some k or that $N\tau$ is even. This efficiency is particularly evident in the one-stratum case when we compare our algorithm to previous methods that require $O(N \log N)$ permutation tests [Aronow, 2023], see Table 6 for example. By using a binary search to find the upper and lower bound of the confidence sets, our algorithm only need to produce $O(\log N)$ permutation tests in some cases.

n	CI	A.4	A.3	A.1	4.3
$[2, 6, 8, 0]$ $[3, 7, 1, 9]$	$[-10, 15]$	619	44,247	64,326	85,239
$[10, 10, 10, 10]$ $[1, 19, 0, 20]$	$[-13, 17]$	9,284	706,440	1,694,995	3,898,440
$[2, 6, 4, 4]$ $[2, 6, 0, 8]$ $[2, 4, 4, 2]$	$[-15, 8]$	3,923	1,665,657	4,780,515	10,132,857

Table 5: 95% confidence intervals for $N * \tau$. The second column gives the confidence interval after scaling by n. The remaining columns indicate the number of permutation tests required for Algorithm A.4, Algorithm A.3 and Algorithm A.1 and the original algorithm in Section 4.3 A random sample with replacement of 10,000 randomizations was used to approximate permutation tests.

n	CI	A.4	Aronow 23
$[18, 82, 11, 89]$	[-7,35]	399	492
$[50, 50, 50, 50]$	[-26,26]	18	744
$[67, 33, 10, 90]$	[93,131]	23	491
$[18, 82, 70, 30]$	[-122,-82]	16	461

Table 6: 95% confidence intervals for $N * \tau$. The second column gives the confidence interval after scaling by N. The remaining columns indicate the number of permutation tests required for Algorithm A.4 and Algorithm 4.3 in [Aronow et al., 2023]. A random sample with replacement of 10,000 randomizations was used to approximate permutation tests. Binary search was used in Algorithm A.4 to find the upper and lower bound of confidence intervals.

B Proof for Section A

B.1 Proof for Section A.1 and A.2

In this section, we prove all the theories in Section A.1 and Section A.2, except for Theorem A.2 and Theorem A.4 regarding the algorithm complexities. We leave the proof of these two theorems in Section B.5.

Proof of Theorem A.1. For convenience, suppose $k = 1$ and $n_1 \leq N_1/2$. Let $\delta = (-1, 1, 0, 0), \boldsymbol{N}' = \boldsymbol{N} + e_1 \cdot \delta$. We omit the other cases since it follows a similar logic. Let $\boldsymbol{Y}$ be a potential outcome vector with potential outcome table $\boldsymbol{N}$ such that $y_{11} = (1, 1)$. Let $\boldsymbol{y}'$ be a potential outcome vector such that $y'_{11} = (1, 0)$ and $y'_{ij} = y_{ij}$ for all $i > 1, j > 0$ and $i = 1, j > 1$. Then, $\boldsymbol{y}'$ has the potential outcome table $\boldsymbol{N}'$.

Now, for an experiment vector $\boldsymbol{Z}$, we consider the difference between $\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ and $\hat{\tau}(\boldsymbol{N}', \boldsymbol{Z})$ by "changing" the potential outcome vector $\boldsymbol{y}$ to $\boldsymbol{y}'$. There are two cases:

1. $Z_{11} = 1$. Then the change $(1, 1) \to (1, 0)$ leaves $\hat{\tau}$ invariant.
2. $Z_{11} = 0$. Then the change $(1, 1) \to (1, 0)$ increases $\hat{\tau}$ by $N_1/(N*(N_1-n_1))$.

To conclude, we have

$$\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \leq \hat{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \leq \hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) + \frac{N_1}{N(N_1 - n_1)} \leq \hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) + \frac{2}{N}$$

The last inequality holds because of our assumption $n_1 \leq N_1/2$. Thus

$$2\tau(\boldsymbol{N}') - \hat{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \geq 2\tau(\boldsymbol{N}) + \frac{2}{N} - \hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) - \frac{2}{N} = 2\tau(\boldsymbol{N}) - \hat{\tau}(\boldsymbol{N}, \boldsymbol{Z})$$

Then, for a observed table $\boldsymbol{n}$, if $\hat{\tau}(\boldsymbol{n}) \geq \tau(\boldsymbol{N}')$, then

$$\begin{aligned}
p(\boldsymbol{N}', \boldsymbol{n}) =& \mathbb{P}\Big(|\hat{\tau}(\boldsymbol{N}', \boldsymbol{Z}) - \tau(\boldsymbol{N}')| \geq \hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N}')\Big) \\
=& \mathbb{P}\Big(\max\{\hat{\tau}(\boldsymbol{N}', \boldsymbol{Z}), 2\tau(\boldsymbol{N}') - \hat{\tau}(\boldsymbol{N}', \boldsymbol{Z})\} \geq \hat{\tau}(\boldsymbol{n})\Big) \\
\geq& \mathbb{P}\Big(\max\{\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}), 2\tau(\boldsymbol{N}) - \hat{\tau}(\boldsymbol{N}, \boldsymbol{Z})\} \geq \hat{\tau}(\boldsymbol{n})\Big) \\
=& \mathbb{P}\Big(|\hat{\tau}(\boldsymbol{N}', \boldsymbol{Z}) - \tau(\boldsymbol{N}')| \geq \hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})\Big) \\
=& p(\boldsymbol{N}, \boldsymbol{n})
\end{aligned}$$

The logic is the same if $\hat{\tau}(\boldsymbol{n}) \leq \tau(\boldsymbol{N})$, so we omit here. □

Proof of Lemma A.1. The proof is direct by computing:

$$\begin{aligned}
\hat{\tau} - \frac{1}{N}\sum_{k=l+1}^{K} N_k \hat{\tau_k} &= \frac{1}{N}\sum_{k=1}^{l} N_k * \Big(\frac{\sum_{k=1}^{l}(x_{k11}+x_{k10})}{n_k} - \frac{\sum_{k=1}^{l}(N_{k11}-x_{k11}+N_{k01}-x_{k01})}{n_k}\Big) \\
&= \frac{\sum_{k=1}^{l}(x_{k11}+x_{k10})}{N} + \frac{\sum_{k=1}^{l}(n_k - x_{k00} - x_{k01})}{N} \\
&\quad - \frac{\sum_{k=1}^{l}(N_{k11}-x_{k11}+N_{k01}-x_{k01})}{N} - \frac{\sum_{k=1}^{l}(n_k-(N_{k00}-x_{k00}+N_{k10}-x_{k10}))}{N} \\
&= \frac{\sum_{k=1}^{l}(N_{k10}-N_{k01})}{N} + \frac{\sum_{k=1}^{K}(2x_{k11}-N_{k11})}{N} - \frac{\sum_{k=1}^{K}(2x_{k00}-N_{k00})}{N}
\end{aligned}$$

□

Proof of Theorem A.3. By the representation of $p(\boldsymbol{N}, \boldsymbol{n})$ in (10), we only need to prove that $\tau(\boldsymbol{N}_1) = \tau(\boldsymbol{N}_2)$ and the distribution of $\hat{\tau}(\boldsymbol{N}_1, \boldsymbol{Z})$ and $\hat{\tau}(\boldsymbol{N}_2, \boldsymbol{Z})$ are the same. Recall that for a potential outcome table $\boldsymbol{N}$, $\tau(\boldsymbol{N}) = \frac{1}{N}\sum_{k=1}^{K}(N_{k10} - N_{k01})$, then from 2 and 3 in our assumption, we know that $\tau(\boldsymbol{N}_1) = \tau(\boldsymbol{N}_2)$. Also, from the assumption and Lemma A.1, we know that the distribution of $\hat{\tau}(\boldsymbol{N}_1, \boldsymbol{Z})$ and $\hat{\tau}(\boldsymbol{N}_2, \boldsymbol{Z})$ are the same. The proof is done. □

B.2 Proof of Theorem A.5

Remark. *This proof structure is nearly the same of Li and Ding's Theorem A.4[LD,2016]. We just extended it to the stratified case.*

Lemma B.1. *For any potential outcome table which is compatible for the observed table $\boldsymbol{n}$, if $\tau(\boldsymbol{N}) < \sum_{k=1}^{K}(n_{k11}+n_{k00})/N$, there exists a potential outcome table $\boldsymbol{N}'$ such that $\boldsymbol{N}'$ is compatible for the observed table and $\boldsymbol{N}' = \boldsymbol{N} + \delta_{LD}$ with $\delta_{LD} \in \Delta_{LD}$. Similarly, if $\tau(\boldsymbol{N}) > \sum_{k=1}^{K}(-n_{k11}-n_{k00})/N$, there exists a potential outcome table $\boldsymbol{N}'$ such that $\boldsymbol{N}'$ is compatible for the observed table and $\boldsymbol{N}' = \boldsymbol{N} - \delta_{LD}$ with $\delta_{LD} \in \Delta_{LD}$. Note that we do not need to assume the experiment is completely balanced.*

Proof. We only prove the first part of the lemma because the second part can be solved by the first part and switching labels of the treatment and control. Because $\boldsymbol{N}$ is compatible for the observed table $\boldsymbol{n}$, there exists a potential outcome vector $\boldsymbol{y}$, summarized by $\boldsymbol{N}$, that give the observed table $\boldsymbol{n}$ under the treatment assignment $\boldsymbol{Z}$. We construct a potential outcome vector $\boldsymbol{y}'$ different from $\boldsymbol{y}$ by only one unit, say $\exists k \in [K], r \in [N_k]$ such that $(y_{kr}(1), y_{kr}(0)) \neq (y'_{kr}(1), y'_{kr}(0))$ and $(y_{lj}(1), y_{lj}(0)) \neq (y'_{lj}(1), y'_{lj}(0))$ for all $(l,j) \neq (k,r)$. We construct $\boldsymbol{y}'$ such that under the same treatment assignment $\boldsymbol{Z}$, $\boldsymbol{y}'$ gives the observed table $\boldsymbol{n}$, and $\boldsymbol{N}' - \boldsymbol{N} \in \Delta_{LD}$, where $\boldsymbol{N}'$ is the corresponding potential outcome table summarized by $\boldsymbol{y}'$. We show the construction in Table 7

We only need to show that (k,r) exists if $\tau(\boldsymbol{N}) < \sum_{i=1}^{K}(n_{i11}+n_{i00})/N$. For any $k \in [K]$, let x_{kab} to be the number of the subjects in the set $\{l : y_{kl}(1) =$

Z_{kr}	$(y_{kr}(1), y_{kr}(0))$	$(y'_{kr}(1), y'_{kr}(0))$	$\boldsymbol{N}' - \boldsymbol{N}$
0	(0,0)	(1,0)	e_k·(0,1,0,-1)
1	(1,1)	(1,0)	e_k·(-1,1,0,0)
0	(0,1)	(1,1)	e_k·(1,0,-1,0)
1	(0,1)	(0,0)	e_k·(0,0,-1,1)

Table 7: Constructing potential table $\boldsymbol{N}'$

$a, y_{kl}(0) = b\}$ who are assigned to treatment. If (k, r) does not exist, then the following must be true:

$$N_{k00} - x_{k00} = 0, \quad x_{k11} = 0, \quad N_{k01} - x_{k01} = 0, x_{k01} = 0, \forall k \in [K].$$

However, this implies $\boldsymbol{N}_k = (n_{k01}, n_{k11}+n_{k00}, 0, n_{k10}), \forall k \in [K]$ and $\tau(\boldsymbol{N}) = (\sum_{k=1}^{K}(n_{k11} + n_{k00}))/N$, which contradicts our assumption. Therefore, (k, r) much exist and our proof is done. □

Proof of Theorem A.5. First we observe that

$$\hat{\tau}(\boldsymbol{n}) = \frac{2\sum_{k=1}^{K}(n_{k11} - n_{k01})}{N} <= \frac{\sum_{k=1}^{K}(n_{k11} - n_{k01})}{N} + \frac{1}{2} = \frac{\sum_{k=1}^{K}(n_{k11} + n_{k00})}{N}$$

and

$$\hat{\tau}(\boldsymbol{n}) = \frac{2\sum_{k=1}^{K}(n_{k11} - n_{k01})}{N} >= \frac{\sum_{k=1}^{K}(n_{k11} - n_{k01})}{N} - \frac{1}{2} = \frac{\sum_{k=1}^{K}(-n_{k10} - n_{k00})}{N}.$$

For any $\tau < \hat{\tau}(\boldsymbol{n})$, if there exists a potential outcome table $\boldsymbol{N}$ such that $\tau(\boldsymbol{N}) = \tau$ and $p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha$, then according to Lemma A.1 and Lemma B.1 , there exists a potential table $\boldsymbol{N}'$ such that $\tau(\boldsymbol{N}') = \tau + 1/N \leq \hat{\tau}(\boldsymbol{n})$ and $p(\boldsymbol{N}', \boldsymbol{n}) \geq \alpha$. Similarly, for any $\tau > \hat{\tau}(\boldsymbol{n})$, if there exists a potential outcome table $\boldsymbol{N}$ such that $\tau(\boldsymbol{N}) = \tau$ and $p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha$, then there exists a potential table $\boldsymbol{N}'$ such that $\tau(\boldsymbol{N}') = \tau - 1/N \geq \hat{\tau}(\boldsymbol{n})$ and $p(\boldsymbol{N}', \boldsymbol{n}) \geq \alpha$. □

B.3 Proof of Theorem A.6

To prove Theorem A.6, we will first introduce a useful lemma. For the sake of clarity in our argument, we introduce a notation: Let $\boldsymbol{N}$ be a potential outcome table, and $\boldsymbol{Z}$ be an experimental vector. Then, we define:

$$\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N[\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) - \tau(\boldsymbol{N})] = 2\sum_{k=1}^{K}(n_{k11} - n_{k01}) - \sum_{k=1}^{K}(N_{k10} - N_{k01}) \quad (13)$$

Furthermore, denote $\tilde{\tau_k}(\boldsymbol{N}, \boldsymbol{Z}) = N_k(\hat{\tau_k}(\boldsymbol{N}, \boldsymbol{Z}) - \tau_k(\boldsymbol{N}))$.

Definition B.1. *For a real-valued random variable X support on $\mathbb{Z}$, we say X is symmetric decreasing, denoted as SD if the pmf of X satisfies:*

1. $\mathbb{P}(X = k) = \mathbb{P}(X = -k), \forall k \in \mathbb{Z}$

2. $\mathbb{P}(X = k) \geq \mathbb{P}(X = k + m), \forall k \in \mathbb{Z}_{\geq 0}$

3. *Either* $\mathbb{P}(X = 0) = 0$, *or* $\mathbb{P}(X = 1) = 0$.

Furthermore, if $\mathbb{P}(X = 1) > 0$, *we classify* X *as symmetric decreasing of type 1, denoted as SD(1). Otherwise, we categorize* X *as symmetric decreasing of type 2, denoted as SD(2). This notation indicates the parity of the support of* X.

The third condition is equivalent to stating that X has support either on odd numbers or on even numbers. We will show that $\tilde{\tau}$ defined in (13)(If $\boldsymbol{N}$ is not too extreme) is SD.

Lemma B.2. *Fix observed table* $\boldsymbol{n}$, *and a potential outcome table* $\boldsymbol{N}$. *Let* $\boldsymbol{N}_{(k)}$ *be a potential outcome table satisfies* $\boldsymbol{N}_{(k)} = \boldsymbol{N} + e_k \cdot \delta_A$. *Set*

$$\boldsymbol{N}_{\{k\}} = \boldsymbol{N} + e_k \cdot (0, -1, -1, 0). \tag{14}$$

Suppose the pmf of $\tilde{\tau}(\boldsymbol{N}_{\{k\}}, \boldsymbol{Z}'')$ *is SD, where* $\boldsymbol{Z}''$ *is uniformly distributed on*

$$\mathcal{Z}(N_1, N_2, \ldots, N_{k-1}, N_k - 2, N_{k+1}, \ldots, N_K, n_1, n_2, \ldots, n_{k-1}, n_k - 1, n_{k+1}, \ldots, n_K).$$

Then

$$p(\boldsymbol{N}_{(k)}, \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$$

Proof. For convenience, suppose $k = 1$ and denote $\boldsymbol{N}_{(1)}$ as $\boldsymbol{N}'$. Let $\boldsymbol{y}$ be a potential outcome vector with potential outcome table $\boldsymbol{N}$. Let $y_{11} = (1, 0)$,$y_{12} = (0, 1)$. Let $\boldsymbol{y}'$ be potential outcome vector such that $y'_{11} = (1, 1)$, $y'_{12} = (0, 0)$ and $y'_{ij} = y_{ij}$ for all $i > 1, j > 0$ and $i = 1, j > 2$. Then $\boldsymbol{y}'$ has the potential outcome table $\boldsymbol{N}'$. By definition,

$$p(\boldsymbol{N}, \boldsymbol{n}) = \mathbb{P}\left(|\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) - \tau(\boldsymbol{N})| \geq |\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right)$$

We assume $\hat{\tau}(\boldsymbol{n}) \neq \tau(\boldsymbol{N})$ otherwise $p(\boldsymbol{N}, \boldsymbol{n}) = p(\boldsymbol{N}', \boldsymbol{n}) = 1$ and the claim holds trivially. Then

$$\begin{aligned} p(\boldsymbol{N}, \boldsymbol{n}) = &\{\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \\ &+\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \leq -N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \end{aligned}$$

Thus from the above equation, we only need to prove that

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \geq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \tag{15}$$

and

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \leq -N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \geq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \leq -N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \tag{16}$$

Here, we will only establish the proof of (15) since the proof of (16) is analogous. In fact, it is identical if one observes that $\tilde{\tau}$ is symmetric around 0.

Now, fix $\boldsymbol{Z}$, we consider the difference between $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ and $\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z})$ by "changing" the potential outcome vector $\boldsymbol{y}$ to $\boldsymbol{y}'$. There are three cases (recall $\boldsymbol{y}$ and $\boldsymbol{y}'$ differ in only two elements):

1. $Z_{11} = Z_{12}$. Then the change $(1,0) \to (1,1)$ and $(0,1) \to (0,0)$ leaves $\tilde{\tau}$ invariant no matter $Z_{11} = Z_{12} = 1$ or $Z_{11} = Z_{12} = 0$.
2. $Z_{11} = 1,_{12} = 0$. The change of $(1,0) \to (1,1)$ for the first subject does not affect $\tilde{\tau}$ and the change $(0,1) \to (0,0)$ for the second subject increases $\tilde{\tau}$ by 2.
3. $Z_{11} = 0, Z_{12} = 1$. The change of $(1,0) \to (1,1)$ for the first subject decreases $\tilde{\tau}$ by 2. $(0,1) \to (0,0)$ for the second subject does not affect $\tilde{\tau}$.

To conclude, we have, $\forall a \in \mathbb{Z}$,

$$\begin{aligned}\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) = a\Big) = \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a - 2, (Z_{11}, Z_{12}) = (1,0)\Big) \\ + \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{11} = Z_{12}\Big) \\ + \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a + 2, (Z_{11}, Z_{12}) = (0,1)\Big) \qquad (17)\end{aligned}$$

Summing (17), we have

$$\begin{aligned}&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \\ =\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| + 2\Big) \\ +\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{11} = Z_{12}\Big) \\ +\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, (Z_{11}, Z_{12}) = (1,0)\Big) \\ +\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2, (Z_{11}, Z_{12}) = (1,0)\Big)\end{aligned}$$

Here we used the fact that

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| + 2a + 1\Big) = 0, a \in Z,$$

since for all possible $\boldsymbol{Z}$, the parity of $\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z})$ and $N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|$ are the same. Using the above equality, we see (15) is equivalent to

$$\begin{aligned}&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{11} = Z_{11}\Big) \\ +\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, (Z_{11}, Z_{12}) = (1,0)\Big) \\ +\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2, (Z_{11}, Z_{12}) = (1,0)\Big) \\ \geq\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big)\end{aligned}$$

which rearranges to

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=N|\hat{\tau}(\boldsymbol{n})-\tau(\boldsymbol{N})|-2\Big|(Z_{11},Z_{12})=(1,0)\Big)\mathbb{P}((Z_{11},Z_{12})=(1,0))$$
$$\geq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=N|\hat{\tau}(\boldsymbol{n})-\tau(\boldsymbol{N})|\Big|(Z_{11},Z_{12})=(0,1)\Big)\mathbb{P}((Z_{11,12})=(0,1)) \quad (18)$$

Let $\boldsymbol{Z}''=(\boldsymbol{Z}_1'',\boldsymbol{Z}_2,\ldots,\boldsymbol{Z}_k)^T$ where $\boldsymbol{Z}_1''=(Z_{13},\ldots,Z_{1N_1})$. Then, conditional on either $(Z_{11},Z_{12})=(1,0)$ or $(Z_{11},Z_{12})=(0,1)$, $\boldsymbol{Z}''$ is uniformly distributed on $\mathcal{Z}(N_1-2,N_2,\ldots,N_k,n_1-1,n_2,\ldots,n_k)$. Recall the definition of $\boldsymbol{N}_{\{1\}}$, note that $\mathbb{P}((Z_{11},Z_{12})=(1,0))=\mathbb{P}((Z_{11},Z_{12})=(0,1))$, and also, since $\boldsymbol{y}_{11}=(1,0),\boldsymbol{y}_{12}=(0,1)$, we can rewrite (18) as

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}_{\{1\}},\boldsymbol{Z}'')=N|\hat{\tau}(\boldsymbol{n})-\tau(\boldsymbol{N})|-2\Big)\geq\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\{1\}},\boldsymbol{Z}'')=N|\hat{\tau}(\boldsymbol{n})-\tau(\boldsymbol{N})|\Big)$$

Since $N|\hat{\tau}(\boldsymbol{n})-\tau(\boldsymbol{N})|\geq 1$, using the hypothesis that the pmf of $\hat{\tau}(\boldsymbol{N}_{\{1\}},\boldsymbol{Z}'')$ is SD, we see the above inequality holds naturally, which completes the proof. □

Remark. *Our proof is nearly the same as the proof of Lemma B.4 in [Aronow et al., 2023]*

One can see that Lemma B.2 is very close to Theorem A.6. The last step is just to prove the SD property of $\tilde{\tau}$. Fortunately, Aronow et al. [2023] have already proven the SD property of $\tilde{\tau}$ for the unstratified case, and we summarize their findings in the following lemma:

Lemma B.3. *For a potential outcome table $\boldsymbol{N}$ which only has 1 strata $\boldsymbol{N}_1$:*

$$\begin{cases} \tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) \text{ is SD(1)} & \text{if } N_{110}+N_{101}>0, \text{ and } N_{110}-N_{101} \text{ is odd} \\ \tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) \text{ is SD(2)} & \text{if } N_{110}+N_{101}>0, \text{ and } N_{110}-N_{101} \text{ is even} \\ \tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})/2 \text{ is SD(1)} & \text{if } N_{110}+N_{101}=0, \text{ and } N_{111} \text{ is odd} \\ \tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})/2 \text{ is SD(2)} & \text{if } N_{110}+N_{101}=0, \text{ and } N_{111} \text{ is even} \end{cases}$$

Proof. See Lemma B.7 and Lemma B.12 in [Aronow et al., 2023]. □

We now extend the SD property to the stratified case.

Lemma B.4. *If $X_1,X_2,\ldots,X_N$ are N independent random variables supported on $\mathbb{Z}$, each of them is SD around 0, then*

$$X:=\sum_{k=1}^{N}X_i$$

is SD around 0.

Proof. Using induction, we only need to prove the case when $N=2$. In this case the pmf of X satisfies

$$\mathbb{P}(X=k)=\sum_{i=-\infty}^{+\infty}\mathbb{P}(X_1=k-i)\mathbb{P}(X_2=i) \quad (19)$$

Using the above formula, it is easy to show that X is symmetric around 0 and satisfies the third condition. Now we prove the decreasing property of X. Suppose X_1 and X_2 both support on even numbers (the logic will be the same for other cases). It suffices to show that if $k \in 2\mathbb{Z}_{\geq 0}$, then $\mathbb{P}(X = k+2) - \mathbb{P}(X = k) \leq 0$. For convenience, for $i \in \mathbb{Z}$, denote $\mathbb{P}_1(i) := \mathbb{P}(X_1 = i)$ and $\mathbb{P}_2(i) := \mathbb{P}(X_2 = i)$. We have

$$\begin{aligned}
\mathbb{P}(X = k+2) - \mathbb{P}(X = k) &= \sum_{i=-\infty}^{+\infty} \mathbb{P}_1 k + 2 - i)\mathbb{P}_2(i) - \sum_{i=-\infty}^{\infty} \mathbb{P}_1(k-i)\mathbb{P}_2(i) \\
&= \sum_{i=2}^{+\infty} \mathbb{P}_1(k+2-i)\mathbb{P}_2(i) - \sum_{i=0}^{\infty} \mathbb{P}_1(k-i)\mathbb{P}_2(i) \\
&\quad + \sum_{i=-\infty}^{0} \mathbb{P}_1(k+2-i)\mathbb{P}_2(i) - \sum_{i=-\infty}^{-2} \mathbb{P}_1(k-i)\mathbb{P}_2(i) \\
&= \sum_{i=0}^{+\infty} \mathbb{P}_1(k-i)[\mathbb{P}_2(i+2) - \mathbb{P}_2(i)] \\
&\quad + \sum_{i=0}^{+\infty} \mathbb{P}_1(k+2+i)[\mathbb{P}_2(i) - \mathbb{P}_2(i+2)] \\
&= \sum_{i=0}^{+\infty} [\mathbb{P}_1(k-i) - \mathbb{P}_1(k+2+i)][\mathbb{P}_2(i+2) - \mathbb{P}_2(i)] \\
&\leq 0.
\end{aligned}$$

□

Lemma B.5. *For any potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_k)^T$, suppose for all $k \in [K]$, $\boldsymbol{N}_k$ satisfies*

$$N_{k10} + N_{k10} > 0$$

Then the pmf of $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ is SD.

Proof. This is an immediate result by combing Lemma B.3 and Lemma B.4. □

Lemma B.6. *Suppose X_1 and X_2 are independent random variables. If X_1 is SD and $X_2/2$ is SD, then $X := X_1 + X_2$ satisfies*

$$\begin{cases} \mathbb{P}(X = k) = \mathbb{P}(X = -k), & \forall k \in \mathbb{N} \\ \mathbb{P}(X = k) \geq \mathbb{P}(X = k+4), & \forall k \in \mathbb{N} \end{cases} \tag{20}$$

Furthermore, if X_1 is SD(1), then X is also SD(1).

Proof. The proof of (20) follows a similar approach to that of Lemma B.4, so we omit it here. Now we show that if X_1 is SD(1), then X is also SD(1). We

will focus on the case when $X_2/2$ is SD(2), as the other case is similar. Denote $\mathbb{P}_1(i) = \mathbb{P}(X_1 = i), \mathbb{P}_2(i) = \mathbb{P}(X_2 = i), i \in \mathbb{Z}$. Suppose $k \in \mathbb{N}$, then

$$\begin{aligned}\mathbb{P}(X = 4k+1) = &\sum_{i=-\infty}^{+\infty} \mathbb{P}_1(1-4i)\mathbb{P}_2(4k+4i) \\ = &\sum_{i=1}^{+\infty} \mathbb{P}_1(4i-1)\mathbb{P}_2(4k+4i) + \sum_{i=0}^{+\infty} \mathbb{P}_1(4i+1)\mathbb{P}_2(4k-4i),\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X = 4k+3) = &\sum_{i=-\infty}^{+\infty} \mathbb{P}_1(-1-4i)\mathbb{P}_2(4k+4+4i) \\ = &\sum_{i=1}^{+\infty} \mathbb{P}_1(4i-1)\mathbb{P}_2(4k+4-4i) + \sum_{i=0}^{+\infty} \mathbb{P}_1(4i+1)\mathbb{P}_2(4k+4+4i),\end{aligned}$$

$$\begin{aligned}\mathbb{P}(X = 4k+5) = &\sum_{i=-\infty}^{+\infty} \mathbb{P}_1(1-4i)\mathbb{P}_2(4k+4+4i) \\ = &\sum_{i=1}^{+\infty} \mathbb{P}_1(4i-1)\mathbb{P}_2(4k+4+4i) + \sum_{i=0}^{+\infty} \mathbb{P}_1(4i+1)\mathbb{P}_2(4k+4-4i),\end{aligned}$$

Thus

$$\begin{aligned}\mathbb{P}(X = 4k+1) - \mathbb{P}(X = 4k+3) = &\sum_{i=1}^{+\infty} \mathbb{P}_1(4i-1)[\mathbb{P}_2(4k+4i) - \mathbb{P}_2(4k+4-4i)] \\ &+ \sum_{i=0}^{+\infty} \mathbb{P}_1(4i+1)[\mathbb{P}_2(4k-4i) - \mathbb{P}_2(4k+4+4i)] \\ = &\sum_{i=1}^{+\infty} [\mathbb{P}_1(4i-1) - \mathbb{P}_1(4i-3)] * \\ &[\mathbb{P}_2(4k+4i) - \mathbb{P}_2(4k+4-4i)] \\ \geq &0,\end{aligned}$$

and

$$\begin{aligned}\mathbb{P}(X = 4k+3) - \mathbb{P}(X = 4k+5) = &\sum_{i=1}^{+\infty} \mathbb{P}_1(4i-1)[\mathbb{P}_2(4k+4-4i) - \mathbb{P}_2(4k+4+4i)] \\ &+ \sum_{i=1}^{+\infty} \mathbb{P}_1(4i+1)[\mathbb{P}_2(4k+4+4i) - \mathbb{P}_2(4k+4-4i)] \\ = &\sum_{i=1}^{+\infty} [\mathbb{P}_1(4i+1) - \mathbb{P}_1(4i-1)] * \\ &[\mathbb{P}_2(4k+4+4i) - \mathbb{P}_2(4k+4-4i)] \\ \geq &0.\end{aligned}$$

□

Lemma B.7. *Consider a potential outcome table* $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$. *The pmf of* $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ *satisfies*

$$\begin{cases} \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = k) = \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = -k), & \forall k \in \mathbb{N} \\ \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = k) \geq \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = k+4), & \forall k \in \mathbb{N} \\ \textit{Either } \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = 1) = 0 \textit{ or } \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = 0) = 0 \end{cases} \tag{21}$$

Moreover, if $N * \tau(\boldsymbol{N})$ *is odd, or equivalently, the pmf of* $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ *has support on odd numbers, then the pmf of* $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ *is SD(1).*

Proof. This is a immediate result by combing Lemma B.3, Lemma B.4, Lemma B.5 and Lemma B.6 □

Now Theorem A.6 is a immediate result by Lemma B.5, Lemma B.7 and Lemma B.2.

B.4 Proof of Theorem A.7

Lemma B.8. *Consider a potential outcome table* $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_k)^T$ *with a total-balanced treatment vector* $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_k)^T$. *Then,*

$$\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = \tau(\boldsymbol{N}) + \frac{\sum_{k=1}^K [x_{k11} - (N_{k11} - x_{k11})]}{N} - \frac{\sum_{k=1}^K [x_{k00} - (N_{k00} - x_{k00})]}{N},$$

where N *is the number of the subjects. Also, recall the definition of* $\tilde{\tau}$ *in (13), we have*

$$\begin{aligned} \tilde{\tau}_k(\boldsymbol{N}, \boldsymbol{Z}) &= [x_{k11} - (N_{k11} - x_{k11})] - [x_{k00} - (N_{k00} - x_{k00})] \\ \tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) &= \sum_{k=1}^K [x_{k11} - (N_{k11} - x_{k11})] - \sum_{k=1}^K [x_{k00} - (N_{k00} - x_{k00})] \end{aligned}$$

Lemma B.8 told us that the subjects with outcome (1,0) and (0,1) won't affect the variation of the distribution of $\hat{\tau}$, and the "direction" of the effort of the subjects with outcome (1,1) and (0,0) are opposite to each other. This intuition will play a key role in the proof of the theorem.

Proof. The proof is directly by computing. Let $n = N/2$ be the number of the subjects who are assigned to treatment, then

$$\begin{aligned} \hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) &= \frac{\sum_{k=1}^K (x_{k11} + x_{k10})}{n} - \frac{\sum_{k=1}^K (N_{k11} - x_{k11} + N_{k01} - x_{k01})}{n} \\ &= \frac{\sum_{k=1}^K (x_{k11} + x_{k10})}{N} + \frac{\sum_{k=1}^K (n - x_{k00} - x_{k01})}{N} \\ &\quad - \frac{\sum_{k=1}^K (N_{k11} - x_{k11} + N_{k01} - x_{k01})}{N} - \frac{\sum_{k=1}^K (n - N_{k00} - x_{k00} + N_{k10} - x_{k10})}{N} \\ &= \tau(\boldsymbol{N}) + \frac{\sum_{k=1}^K [x_{k11} - (N_{k11} - x_{k11})]}{N} - \frac{\sum_{k=1}^K [x_{k00} - (N_{k00} - x_{k00})]}{N} \end{aligned}$$

The other equations are similar so we omitted here.

□

Corollary B.1. *Consider two potential outcome tables* $\boldsymbol{N}_1 = ((\boldsymbol{N}_1)_1, (\boldsymbol{N}_1)_2, \ldots, (\boldsymbol{N}_1)_K)^T$, $\boldsymbol{N}_2 = ((\boldsymbol{N}_2)_1, (\boldsymbol{N}_2)_2, \ldots, (\boldsymbol{N}_1)_K)^T$. *Suppose* $\exists i \in [K]$, *such that*

$$(\boldsymbol{N_1})_j = (\boldsymbol{N_2})_j, j \neq i,$$

$$(N_1)_{i11} = (N_2)_{i00}, (N_1)_{i00} = (N_2)_{i11}, (N_1)_{i01} = (N_2)_{i01}, (N_1)_{i10} = (N_2)_{i10},$$

then, $\tilde{\tau}(\boldsymbol{N}_1, \boldsymbol{Z})$ *and* $\tilde{\tau}(\boldsymbol{N}_2, \boldsymbol{Z})$ *have the same distribution.*

Proof. Suppose $i = 1$. For a potential outcome $\boldsymbol{N}$, let $\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = \tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) - \tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z})$. Then, from Lemma B.8, we have

$$\tilde{\tau}_1(\boldsymbol{N}_1, \boldsymbol{Z}) = -\tilde{\tau}_1(\boldsymbol{N}_2, \boldsymbol{Z}), \quad \tilde{\tau}_{-1}(\boldsymbol{N}_1, \boldsymbol{Z}) = \tilde{\tau}_{-1}(\boldsymbol{N}_2, \boldsymbol{Z}).$$

Thus

$$\tilde{\tau}(\boldsymbol{N}_1, \boldsymbol{Z}) = \tilde{\tau}_1(\boldsymbol{N}_1, \boldsymbol{Z}) + \tilde{\tau}_{-1}(\boldsymbol{N}_1, \boldsymbol{Z}),$$
$$\tilde{\tau}(\boldsymbol{N}_2, \boldsymbol{Z}) = -\tilde{\tau}_1(\boldsymbol{N}_1, \boldsymbol{Z}) + \tilde{\tau}_{-1}(\boldsymbol{N}_1, \boldsymbol{Z})$$

Since both $\tilde{\tau}_1(\boldsymbol{N}_1, \boldsymbol{Z})$ and $\tilde{\tau}_{-1}(\boldsymbol{N}_1, \boldsymbol{Z})$ are symmetric around 0, and they are independent of each other, it is straightforward to deduce that the distributions of $\tilde{\tau}(\boldsymbol{N}_2, \boldsymbol{Z})$ and $\tilde{\tau}(\boldsymbol{N}_1, \boldsymbol{Z})$ are identical. □

Corollary B.1 establishes the symmetry between N_{11} and N_{00}. With this corollary in mind, we will concentrate on the case where $N_{i00} \leq N_{i11}$ in this chapter, as the other case can be readily deduced using this corollary.

The next definition may look weird at the first glance, but will be clear when we prove the next lemma.

Definition B.2. *For a potential outcome table* $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$ *with a total-balanced experiment* $\boldsymbol{Z}$ *and a given strata* k, *suppose* $N_{k11} \geq N_{k00}$. *Let the potential outcome vector* $\boldsymbol{y}_k$ *be in the order*

$$\boldsymbol{y}_k = \left(\underbrace{(1,1),(0,0),(1,1),(0,0),\ldots,(1,1),(0,0)}_{2N_{k00}}, \ldots \right)^T \tag{22}$$

That is to say, the first $2N_{k00}$ *subjects are all with the potential outcome (1,1) or (0,0), alternately. Then, for a given non-negative integer* $s \leq \min(N_{k00}, n_k/2)$, *if for each non-negative integer* a,

$$\begin{cases} \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a - 2 \text{ or } a, Z_{k1} = \cdots = Z_{k,2s} = 0\Big) \\ \geq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a - 2 \text{ or } a, Z_{k1} = \cdots = Z_{k,2s} = 1\Big) & \text{if } a \text{ is odd} \\ \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{k1} = \cdots = Z_{k,2s} = 0\Big) \\ \geq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{k1} = \cdots = Z_{k,2s} = 1\Big) & \text{if } a \text{ is even} \end{cases}$$

then we say $\boldsymbol{N}$ *satisfies* k, s*-regular condition.*

In this chapter, unless specified, all the potential outcome vectors are in the order (22)

Lemma B.9. *Let $\boldsymbol{N}$ and $\boldsymbol{N}'$ be defined in Theorem A.7. Denote $\boldsymbol{N}_{[k]}$ be*

$$\boldsymbol{N}_{[k]} = \boldsymbol{N} + e_k \cdot (-1, 0, 0, 1) \tag{23}$$

Suppose $N_{k11} - 2 \geq N_{k00} + 2$. Then if $\boldsymbol{N}^{[k]}$ satisfies k,1-regular condition, we have

$$p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, n)$$

Proof. For convenience, suppose $k = 1$ and denote $\boldsymbol{N}_{[1]}$ as $\boldsymbol{N}^{\circ}$, $\boldsymbol{N} + \delta_L^1(\boldsymbol{N})$ as $\boldsymbol{N}^{\circ\circ}$. Let $\boldsymbol{y}$ be a potential outcome vector with potential outcome table $\boldsymbol{N}$ such that $y_{11} = (1, 1)$,$y_{12} = (1, 1)$. Let $\boldsymbol{y}^{\circ}$ be a potential outcome vector such that $y^{\circ}_{11} = (1, 1)$, $y^{\circ}_{12} = (0, 0)$ and $y^{\circ}_{ij} = y_{ij}$ for all $i > 1, j > 0$ or $i = 1, j > 2$. Let $\boldsymbol{y}^{\circ\circ}$ be a potential outcome vector such that $y^{\circ\circ}_{11} = (0, 0)$, $y^{\circ\circ}_{12} = (0, 0)$ and $y^{\circ\circ}_{ij} = y_{ij}$ for all $i > 1, j > 0$ or $i = 1, j > 2$. Then $\boldsymbol{y}^{\circ}$, $\boldsymbol{y}^{\circ\circ}$ has the potential outcome table $\boldsymbol{N}^{\circ}$ and $\boldsymbol{N}^{\circ\circ}$, separately. Similarly in Lemma B.2, we only need to prove that

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \leq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ\circ}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big), \tag{24}$$

when $N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| \neq 0$. Since both $\boldsymbol{N}$ and $\boldsymbol{N}^{\circ\circ}$ differ from $\boldsymbol{N}^{\circ}$ in only one element, we now try to express (24) in terms of $\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z})$.

Now fix $\boldsymbol{Z}$, we consider the difference between $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ and $\tilde{(}\boldsymbol{N}^{\circ}, \boldsymbol{Z})$ by "changing" the potential outcome vector $\boldsymbol{y}$ to $\boldsymbol{y}^{\circ}$:

1. If $Z_{12} = 1$, the change $(1, 1) \to (0, 0)$ decreases $\tilde{\tau}$ by 2.
2. If $Z_{12} = 0$, the change $(1, 1) \to (0, 0)$ increases $\tilde{\tau}$ by 2.

Thus, we have, $\forall a \in \mathbb{Z}$,

$$\begin{aligned}\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})\right) =& \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z}) = a - 2, Z_{12} = 1\right) \\ &+ \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z}) = a + 2, Z_{12} = 0\right)\end{aligned} \tag{25}$$

Summing (25), we get

$$\begin{aligned}&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \\ =&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| + 2\Big) \\ +&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \text{ or } N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{12} = 1\Big)\end{aligned}$$

Here we used the fact that the parity of $\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z})$ and $N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|$ are the same. Similarly,

$$\begin{aligned}&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ\circ}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\Big) \\ =&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| + 2\Big) \\ +&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \text{ or } N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{11} = 0\Big)\end{aligned}$$

Thus (24) is equivalent to

$$\begin{aligned}
&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ},\boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \text{ or } N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{11} = 0\Big)\\
\geq&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ},\boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \text{ or } N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{12} = 1\Big)\\
\Longleftrightarrow&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ},\boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \text{ or } N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{11} = Z_{12} = 0\Big)\\
\geq&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}^{\circ},\boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \text{ or } N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|, Z_{11} = Z_{12} = 1\Big)
\end{aligned}$$

Compared to the definition of 1,1-regular condition, we are done. □

Now we set up an induction to show that all potential outcome tables satisfy the regular condition.

Lemma B.10. *For a given potential outcome table $\boldsymbol{N}$ and a strata k, suppose $2 \leq min(N_{k00}, N_{k11})$,$s < min(N_{k00}, N_{k11})$ and $s \leq n_k/2$. Let $\boldsymbol{N}(j)$ be*

$$\boldsymbol{N}(j) = \boldsymbol{N} + e_k \cdot (j, 0, 0, j) \tag{26}$$

If all of the following three statement hold true:

$$\begin{cases}
\boldsymbol{N}(-1) \text{ satisfies k,s-regular condition or } 2s = n_k\\
\boldsymbol{N}(-2) \text{ satisfies k,(s-1)-regular condition}\\
\boldsymbol{N} \text{ satisfies k,(s+1)-regular condition or } 2s \in \{n_k, n_k - 1\},
\end{cases}$$

then $\boldsymbol{N}$ satisfies k, s-regular condition.

Proof. Suppose $k = 1$ and $N_{111} \geq N_{100}$. Using Lemma B.8, we have

$$\begin{aligned}
&\mathbb{P}(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) = -1 \text{ or } 1, Z_{11} = \cdots = Z_{1,2s} = 0)\\
=&\mathbb{P}(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) = -1 \text{ or } 1, Z_{11} = \cdots = Z_{1,2s} = 1)
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) = 0, Z_{11} = \cdots = Z_{1,2s} = 0)\\
=&\mathbb{P}(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) = 0, Z_{11} = \cdots = Z_{1,2s} = 1).
\end{aligned}$$

These can be easily seen by switching the labels of the treatment and the control. Now suppose $a \in \mathbb{Z}$ and $a \geq 2$. Since $s < N_{100} \leq N_{111}$, we can assume $y_{1,2s+1} = (1, 1)$ and $y_{1,2s+2} = (0, 0)$. Then,

$$
\begin{aligned}
&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0\Big)\\
=&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=0, Z_{1,2s+2}=0\Big)\\
+&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=1, Z_{1,2s+2}=1\Big)\\
+&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=0, Z_{1,2s+2}=1\Big)\\
+&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=1, Z_{1,2s+2}=0\Big)\\
=&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s+2}=0\Big)\\
+&\mathbb{P}\Big(Z_{1,2s-1}=Z_{1,2s}=0, Z_{1,2s+1}=Z_{1,2s+2}=1\Big)\\
&*\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s-2}=0|Z_{1,2s-1}=Z_{1,2s}=0, Z_{1,2s+1}=Z_{1,2s+2}=1\Big)\\
+&\mathbb{P}\Big(Z_{1,2s+1}=0, Z_{1,2s+2}=1\Big)\\
&*\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0|Z_{1,2s+1}=0, Z_{1,2s+2}=1\Big)\\
+&\mathbb{P}\Big(Z_{1,2s+1}=1, Z_{1,2s+2}=0\Big)\\
&*\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0|Z_{1,2s+1}=1, Z_{1,2s+2}=0\Big)\\
=&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s+2}=0\Big)\\
+&\mathbb{P}\Big(Z_{1,2s-1}=Z_{1,2s}=0, Z_{1,2s+1}=Z_{1,2s+2}=1\Big)\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}(-2),\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s-2}=0\Big)\\
+&\mathbb{P}\Big(Z_{1,2s+1}=0, Z_{1,2s+2}=1\Big)\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a+2, Z_{11}=\cdots=Z_{1,2s}=0\Big)\\
+&\mathbb{P}\Big(Z_{1,2s+1}=1, Z_{1,2s+2}=0\Big)\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a-2, Z_{11}=\cdots=Z_{1,2s}=0\Big).
\end{aligned}
\tag{27}
$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=1\Big)\\
=&\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s+2}=1\Big)\\
+&\mathbb{P}\Big(Z_{1,2s-1}=Z_{1,2s}=1, Z_{1,2s+1}=Z_{1,2s+2}=0\Big)\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}(-2),\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s-2}=1\Big)\\
+&\mathbb{P}\Big(Z_{1,2s+1}=0, Z_{1,2s+2}=1\Big)\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a+2, Z_{11}=\cdots=Z_{1,2s}=1\Big)\\
+&\mathbb{P}\Big(Z_{1,2s+1}=1, Z_{1,2s+2}=0\Big)\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a-2, Z_{11}=\cdots=Z_{1,2s}=1\Big).
\end{aligned}
\tag{28}
$$

Now since we assume $a \geq 2$, we have $a-2 \geq 0$. When $2s \leq n_k - 2$, the proof is complete by combing (27), (28) and the definition of the regular condition.

If $2s = n_k$ or $2s = n_k - 1$, the proof is also done by noting that some of the probability in the equation will be 0. For example, if $2s \geq n_k - 1$,

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{11} = \cdots = Z_{1,2s+2} = 0\Big) = 0$$

□

Lemma B.11. *For a potential outcome table $\boldsymbol{N}$ which only has 1 strata $\boldsymbol{N}_1$, assume $N_{111} \geq N_{100}$. Then, $\boldsymbol{N}$ satisfies $1, s$-regular condition if $s = \min N_{100}$ and $2s \leq n_1$. In fact, we have*

$$\begin{aligned} &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{11} = \cdots = Z_{1,2s} = 0\Big) \\ \geq\ &\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{11} = \cdots = Z_{1,2s} = 1\Big) \end{aligned} \tag{29}$$

for every $a \in \mathbb{N}$.

Proof. For convenience, we omit all the strata-script "1" in this proof. For example, we denote N_{111} by N_{11} and N_{100} by N_{00}. By symmetry, we know that

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_1 = \cdots = Z_{2s} = 1\Big) = \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = -a, Z_1 = \cdots = Z_{2s} = 0\Big)$$

thus we only need to prove that

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a \Big| Z_1 = \cdots = Z_{2s} = 0\Big) \geq \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = -a \Big| Z_1 = \cdots = Z_{2s} = 0\Big) \tag{30}$$

We assume the right hand side of (30) is not 0, otherwise the inequality holds naturally. Recall the definition of x_{ij} in Lemma B.8, let x'_{11} be the number of the subjects in the set $\{l : y_l(0) = y_l(1) = 1, l \leq 2s\}$ who are assigned to treatment and $x''_{11} = x_{11} - x'_{11}$. Then, condition on $Z_1 = \cdots = Z_{2s} = 0$, and $N_{00} = s$ we have $x'_{11} = x_{00} = 0$ and

$$\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = 2x''_{11} - (N_{11} - N_{00}).$$

Note x'_{11} follows the hypergeometric distribution, then, (30) can be rearranged to

$$\begin{aligned} &\binom{g}{N_{11} - N_{00}} \binom{n-g}{N_{01} + N_{10}} \Big/ \binom{n}{N - 2N_{00}} \\ \geq &\binom{g'}{N_{11} - N_{00}} \binom{n-g'}{N_{01} + N_{10}} \Big/ \binom{n}{N - 2N_{00}} \end{aligned} \tag{31}$$

where $g = (a + N_{11} - N_{00})/2$ and $g' = (-a + N_{11} - N_{00})/2$. Note that we have assumed the right hand side of (31) is not 0, thus g' is an integer, which means g is also an integer. Since $0 \leq (-a + N_{11} - N_{00})/2$, we have $a \leq N_{11} - N_{00}$ and $0 \leq (a + N_{11} - N_{00})/2 = g \leq N_{11} - N_{00}$. Also, since $n - (-a + N_{11} - N_{00})/2 \leq N_{01} + N_{10}$, we have $a \leq N_{01} + N_{10} - 2N_{00}$, then $0 \leq 2N_{00} \leq n - (a - N_{11} -$

$N_{00})/2 = n - g \le N_{01} + N_{10}$. Thus the left hand side of (31) make sense and is not 0. Now we rewrite (31), then it becomes

$$\begin{aligned}&\frac{(N_{11}-N_{00})!}{((N_{11}-N_{00}-a)/2)!((N_{11}-N_{00}+a)/2)!}\frac{(N_{01}+N_{10})!}{(n-g)!(N_{01}+N_{10}-n+g)!}\\ \ge&\frac{(N_{11}-N_{00})!}{((N_{11}-N_{00}-a)/2)!((N_{11}-N_{00}+a)/2)!}\frac{(N_{01}+N_{10})!}{(n-g')!(N_{01}+N_{10}-n+g')!},\end{aligned}$$

which rearranges to

$$\frac{(n-g')!}{(n-g)!} \ge \frac{(N_{01}+N_{10}-n+g)!}{(N_{01}+N_{10}-n+g')!}.$$

Then, we only need to prove that $n - g' \ge N_{01} + N_{10} - n + g$, which is equivalent to $N_{11} + N_{00} \ge N_{11} - N_{00}$. However, the last inequality holds naturally, thus (31) is true, and the proof is complete. □

Lemma B.12. *For a potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$, for all $k \in [K]$, if $s_k := \min(N_{k11}, N_{k00}) \le n_k/2$, then $\boldsymbol{N}$ satisfies k, s_k-regular condition.*

Proof. For simplicity, suppose $k = 1$ and $N_{111} \ge N_{100}$.

Case 1: a is even.

In this case, we assume that $\tilde{\tau}$ is supported on even numbers. Otherwise the inequality is trivial. We only need to prove that for every non-negative even integer a,

$$\mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) = a, Z_{11} = \cdots = Z_{1,2s_1} = 0\Big) \ge \mathbb{P}\Big(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) = a, Z_{11} = \cdots = Z_{1,2s_1} = 1\Big)$$

Now let $\tilde{\tau}_{-1}(\boldsymbol{N},\boldsymbol{Z}) = \tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) - \tilde{\tau}_1(\boldsymbol{N},\boldsymbol{Z})$. We only need to prove that

$$\begin{aligned}&\sum_{k=-\infty}^{\infty}\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N},\boldsymbol{Z}) = a-k\Big)\mathbb{P}\Big(\tilde{\tau}_1(\boldsymbol{N},\boldsymbol{Z}) = k, Z_{11} = \cdots = Z_{1,2s_1} = 0\Big)\\ \ge&\sum_{k=-\infty}^{\infty}\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N},\boldsymbol{Z}) = a-k\Big)\mathbb{P}\Big(\tilde{\tau}_1(\boldsymbol{N},\boldsymbol{Z}) = k, Z_{11} = \cdots = Z_{1,2s_1} = 1\Big).\end{aligned} \tag{32}$$

Here we used the independence of $\tilde{\tau}_1$ and $\tilde{\tau}_{-1}$. By switching label of the treatment and the control in the first strata, we see (32) is equivalent to

$$\begin{aligned}&\sum_{k=-\infty}^{\infty}\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N},\boldsymbol{Z}) = a-k\Big)\mathbb{P}\Big(\tilde{\tau}_1(\boldsymbol{N},\boldsymbol{Z}) = k, Z_{11} = \cdots = Z_{1,2s_1} = 0\Big)\\ \ge&\sum_{k=-\infty}^{\infty}\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N},\boldsymbol{Z}) = a-k\Big)\mathbb{P}\Big(\tilde{\tau}_1(\boldsymbol{N},\boldsymbol{Z}) = -k, Z_{11} = \cdots = Z_{1,2s_1} = 0\Big).\end{aligned} \tag{33}$$

Denote $\tilde{p}(j) = \mathbb{P}\Big(\tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z}) = j, Z_{11} = \cdots = Z_{1,2s_1} = 0\Big)$ for convenience. By rearranging (33), we only need to prove that

$$\sum_{k=-\infty}^{\infty} \mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\Big)\Big(\tilde{p}(k) - \tilde{p}(-k)\Big) \geq 0$$

$$\Longleftrightarrow \sum_{k=1}^{\infty} \Big[\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\Big) - \mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k\Big)\Big]\Big(\tilde{p}(k) - \tilde{p}(-k)\Big) \geq 0$$

Note that we already proved $\tilde{p}(k) - \tilde{p}(-k) \geq 0$ in Lemma B.11. Thus, we only need to prove that $\mathbb{P}(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k) - \mathbb{P}(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k) \geq 0$, for all $k \geq 1$. First we observe that $|a - k| \leq |k + a|$. If $\tilde{\tau}_1$ is supported on even numbers, note that $a - k \equiv a + k(\mod 4)$ for even k, then the proof is done by (21); If $\tilde{\tau}_1$ is supported on odd numbers, note that $a - k$ is odd for odd k, then the proof is also done by Lemma B.7.

Case 2: a is odd.

In this case, we assume that $\tilde{\tau}$ is supported on odd numbers. Otherwise the inequality is trivial. Similarly, we only need to prove that

$$\sum_{k=1}^{\infty} \Big[\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\Big) - \mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k - 2\Big)\Big]\Big(\tilde{p}(k) - \tilde{p}(-k)\Big)$$

$$+ \sum_{k=1}^{\infty} \Big[\mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k - 2\Big) - \mathbb{P}\Big(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k\Big)\Big]\Big(\tilde{p}(k) - \tilde{p}(-k)\Big) \geq 0$$

Similarly, we prove that $\mathbb{P}(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k) - \mathbb{P}(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k - 2) \geq 0$ and $\mathbb{P}(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k - 2) - \mathbb{P}(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k) \geq 0$. First we observe that $|a - k| \leq |a + k - 2|$ and $|a - k - 2| \leq |a + k|$ for $a \geq 1$ and $k \geq 1$. If $\tilde{\tau}_1$ is supported on odd numbers, note that $a - k \equiv a + k - 2(\mod 4)$ for odd k and $a - k - 2 \equiv a + k(\mod 4)$ for odd k, then the proof is done by (21); If $\tilde{\tau}_1$ is supported on even numbers, note that $a - k$ is odd for even k, then the proof is also done by Lemma B.7. □

Lemma B.13. *For a potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$, $\boldsymbol{N}$ satisfies k, s-regular condition if $k \in [K]$ and $s \leq \min(N_{k00}, N_{k11}, n_k/2)$.*

Proof. Again, for simplicity, suppose $k = 1$ and $N_{111} \geq N_{100}$. Recall the definition of $\boldsymbol{N}(j)$,$j \in \mathbb{Z}$, we induct on j. If $j = -N_{100}$, which is the least possible value of j, then the only possible value for s is 0. By definition, $\boldsymbol{N}(j)$ (In fact, all potential outcome tables) satisfies $1, 0-$ regular-condition. If $j = -N_{100} + 1$, then the only possible value of s is 0 and 1. If $s = 0$, then $\boldsymbol{N}(j)$ satisfies $1, 0-$ regular-condition. If $n_1 \geq 2$ and $s = 1$, then by Lemma B.12, $\boldsymbol{N}(j)$ satisfies $1, 1-$regular condition. Thus, for $j \leq N_{100} + 1$, $\boldsymbol{N}(j)$ satisfies $1, s-$ regular condition for all s.

Now for a integer $l > -N_{100}+1$, suppose $\boldsymbol{N}(j)$ satisfies $1, s$-regular condition for all $j < l$ and all possible s. We consider $\boldsymbol{N}(l)$.

Case 1: $N_{100} + l \leq (n_i + l)/2$:

If we let $s' = N^i_{00} + l$, then Lemma B.12 informs us that $\boldsymbol{N}(l)$ satisfies $1, s'$-regular condition. Then, by applying Lemma B.10 and inducting based on the assumption, we can conclude that $\boldsymbol{N}(l)$ satisfies $1, s$-regular condition for all possible values of s.

Case 2: $N_{100} + l \geq (n_1 + l)/2$:

If we set $s' = (n_1+l)/2$ or $(n_1+l-1)/2$, then Lemma B.10 implies that $\boldsymbol{N}(l)$ satisfies $1, s'$-regular condition. Continuing to apply Lemma B.10 and inductive the assumption, we can conclude that $\boldsymbol{N}(l)$ satisfies $1, s$-regular condition for all possible values of s.

As a result, for all $j \geq -N_{100}$ (In particular, $j = 0$), $\boldsymbol{N}(j)$ satisfies the desired property, and we are done. □

Proof of Theorem A.7. The proof is a immediate result from Lemma B.9 and Lemma B.13. □

B.5 Proof for Algorithm Complexity

In this section, we establish the validity of Theorem A.2, A.4, and A.8. Within each proof, we will demonstrate three key aspects:

1. The algorithm guarantees the construction of a $1 - \alpha$ confidence set.
2. The required permutation tests align with the stated theorem.
3. Additional processes, such as searching, do not impose an undue computational burden.

We first derive some conditions for when potential outcome tables are compatible. These conditions will play a pivotal role in Step 3 of each proof. To initiate this exploration, we extend the lemma by Li and Ding [2016] to the stratified case.

Lemma B.14. *A potential table $\boldsymbol{N}$ is compatible with the observed table $\boldsymbol{n}$ if and only if within every block k,*

$$\begin{aligned} &\max\{0, n_{k11} - N_{k10}, N_{k11} - n_{k01}, N_{k01} + N_{k11} - n_{k10} - n_{k01}\} \\ \leq\ &\min\{N_{k11}, n_{k11}, N_{k01} + N_{k11} - n_{k01}, N_k - N_{k10} - n_{k01} - n_{k10}\} \end{aligned}$$

Li and Ding [2016] previously established the unstratified version of this lemma. Since there is no interference between strata, the proof of this lemma remains identical to that of Li and Ding's. Below are two rephrased versions of the lemma, omitting the proof details as they involve straightforward algebra.

Lemma B.15. *A potential table $\boldsymbol{N}$ is compatible with the observed table $\boldsymbol{n}$ if and only if within every block k,*

$$\begin{aligned}
&\max\{2n_{k11} - N_k - N_{k11} + N_{k00}, -N_k + N_{k11} + N_{k00}, \\
&\qquad N^i - N^i_{11} - N_{k00} - 2n_{k01} - 2n_{k10}, N_{k11} - N_k - N_{00} + 2n_{k00}\} \\
\leq\ & N_{k10} - N_{k01} \\
\leq\ & \min\{N_{k11} + N_k - N_{k00} - 2n_{k01}, N_k + N_{k11} + N_{k00} - 2n_{k01} - 2n_{k10}, \\
&\qquad N_k - N_{k11} - N_{k00}, N_k - N_{k11} + N_{k00} - 2n_{k10}\}.
\end{aligned}$$

and

$$0 \leq N_{k11} \leq n_{k01} + n_{k11} \leq N_k - N_{k00} \leq N_k$$

Lemma B.16. *A potential outcome table $\boldsymbol{N}$ is compatible for the observed table $\boldsymbol{n}$ if and only if within every block i,*

$$\max\{0, -N_k\tau_i\} \leq N_{k01} \leq \min\{n_{k10} + n_{k01}, N_k - N_k\tau_k - n_{k01} - n_{k10}\}$$

and

$$\begin{aligned}
&\max\{2N_{k01} - N_k + N_k\tau_k, 2n_{k01} - N_k + N_k\tau_k, 2n_{k11} - N_k - N_k\tau_k, \\
&2n_{k11} + 2n_{k01} - 2N_{k01} - N_k + N_k\tau_k\} \\
\leq& N_{k11} - N_{k00} \\
\leq& \min\{2n_{k01} + 2n_{k11} + 2N_{k01} - N_k + N_k\tau_k, N_k - N_k\tau_k - 2n_{k10}, \\
&N_k - N_k\tau_k - 2N_{k01}, 2n_k - 2n_{k00} - N_k + N_k\tau_k\}
\end{aligned}$$

Now we begin to prove the theorems.

Proof of Theorem A.2. Step 1: The algorithm guarantees the construction of a $1-\alpha$ confidence set.

We just need to show the algorithm correctly screens out the potential outcome tables with p-value greater than α and less than α. Specifically, if $\tau(\boldsymbol{N}) \geq \hat{\tau}(\boldsymbol{n})$, we need to show that $p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha$ if and only if $\underline{N_{110}}(N_{210}) < N_{110} \leq \overline{N_{110}}(N_{210})$, and the algorithm correctly finds $\overline{N_{110}}(N_{210})$. (We omit the case when $\tau(\boldsymbol{N}) \leq \hat{\tau}(\boldsymbol{n})$ as in the algorithm)

Consider a potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$ where $\boldsymbol{N}_1 = (N_{111}, N_{110}, N_{101}, N_{100})$, $\boldsymbol{N}_2 = (N_{211}, N_{110}, N_{201}, N_{200})$. By the definition of $\overline{N_{110}}(N_{210})$, if $N_{110} > \overline{N_{110}}(N_{210})$, then $p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha$; Otherwise, if we set

$$\boldsymbol{N}' = (\boldsymbol{N}'_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$$

where

$$\boldsymbol{N}'_1 = (N_1 - N_{101} - N_{100} - \overline{N_{110}}(N_{210}), \overline{N_{110}}(N_{210}), N_{101}, N_{100}).$$

Then, $\boldsymbol{N}' = \boldsymbol{N} + (\overline{N_{110}}(N_{210}) - N_{210}) * e_1 \cdot (-1, 1, 0, 0)$. By Theorem A.1, we know that $p(\boldsymbol{N}', \boldsymbol{n}) \leq p(\boldsymbol{N}, \boldsymbol{n})$. Since $p(\boldsymbol{N}', \boldsymbol{n}) \geq \alpha$, we have $p(\boldsymbol{N}, \boldsymbol{n}) \geq \alpha$.

The next thing is to prove the algorithm finds $\overline{N_{110}}(N_{210})$ correctly. We prove this by induction. If $N_{210} = 0$, it is trivial because we start from the maximum value of N_{110}. Suppose the algorithm can correctly find $\overline{N_{110}}(N_{210})$ for $N_{210} = j$. If a potential outcome table $\boldsymbol{N}$ satisfies $N_{210} = j + 1$ and $N_{110} > \overline{N_{110}}(j)$, then we much have $\tau(\boldsymbol{N}) > \hat{\tau}(\boldsymbol{n})$ by the definition of $\overline{N_{110}}(j)$ and the monotonicity of $\underline{N_{110}}(j)$. Set $\boldsymbol{N'} = \boldsymbol{N} - e_2 \cdot (-1, 1, 0, 0)$, then $(N')_{210} = j$ and $(N')_{110} = N_{110} > \overline{(N')_{110}}(j)$. By our inductive assumption, $p(\boldsymbol{N'}, \boldsymbol{n}) < \alpha$. By Theorem A.2, we have $\underline{p}(\boldsymbol{N}, \boldsymbol{n}) < p(\boldsymbol{N'}, \boldsymbol{n}) < \alpha$. Thus it is valid to look for $\overline{N_{110}}(j+1)$ starting from $\overline{N_{110}}(j)$. The proof is done.

Step 2: The required permutation tests align with the stated theorem.

In Algorithm A.1, there is a loop that iterates a total of $\prod_{k=1}^{K}(N_k)^2$ times. In Algorithm A.2, there is a loop that iterates a total of $\prod_{k=3}^{K} N_k$ times. In each loop, we need to do permutation tests at most $N_1 + N_2$ times. So the overall permutation tests that need to be done is $O((N_1+N_2) * \prod_{k=3}^{K} N_k * \prod_{k=1}^{K}(N_k)^2)$.

Step 3: Additional processes, such as searching, do not impose an undue computational burden.

The additional process that could potentially increase the computational burden is Algorithm A.2.2(c). However, for each possible value of N_{210}, we can efficiently determine the largest and smallest values of N_{110} that result in a feasible potential outcome using Lemma B.14. This computation can be performed in constant time for each possible value of N_{210}. Since there are a total of $N_2 - M_2 - M_{200}$ possible values for N_{210}, the algorithm's complexity remains within $O(N_2 - M_2 - M_{200})$, which does not exceed the computational complexity of 2(b) since it requires $O(N_2 + N_1)$ permutation tests.

□

Proof of Theorem A.4. Step 1: The algorithm guarantees the construction of a $1 - \alpha$ *confidence set.*

This is a direct result by Theorem A.3.

Step 2: The required permutation tests align with the stated theorem.

This is a direct consequence of the algorithm's structure, as it involves performing $O((\sum_{k=1}^{l} N_k) * \prod_{k=1}^{l} N_k^2 * \prod_{k=l+1}^{K} N_k^3)$ loops. Within each loop iteration, a permutation test is conducted.

Step 3: Additional processes, such as searching, do not impose an undue computational burden.

The potential increase in computational burden is associated with Step 2(a). However, this step can be efficiently executed in constant time. Using Lemma B.15, we can determine whether a potential outcome table in Step 2(a) is compatible, by verifying whether a falls within the range bounded by the lower and upper limits of $\sum_{k=1}^{l}(N_{k10} - N_{k01})$. To identify a compatible potential outcome table, we can set $N_{k10} - N_{k01}$ to their lower bounds in the first u strata and to their upper bounds in the subsequent $l-u-1$ strata. Finally, we set $N_{k10} - N_{k01}$ in the last stratum to ensure that $\sum_{k=1}^{l}(N_{k10} - N_{k01}) = a$. The value of u is undetermined but can be determined efficiently in $O(1)$ time by testing each possible value of $u \in [K]$. □

Proof of Theorem A.8. Step 1: The algorithm guarantees the construction of a $1-\alpha$ confidence set.

For a compatible potential outcome table $\boldsymbol{N}$, we claim that there exists a $\boldsymbol{N}'$ found in Step 2(b) or Step 2(d) such that $p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$ and $\tau(\boldsymbol{N}') = \tau(\boldsymbol{N})$. As a result of this claim, we do not need to test $\boldsymbol{N}$:If $p(\boldsymbol{N}', \boldsymbol{n}) < \alpha$, then it logically follows that $p(\boldsymbol{N}, \boldsymbol{n}) < \alpha$ as well. Consequently, there's no necessity to test $\boldsymbol{N}$ since its p-value is already known to be below α. On the other hand, if $p(\boldsymbol{N}', \boldsymbol{n}) \geq \alpha$, we can confidently include $\tau(\boldsymbol{N}')$ in our confidence set. Since $\tau(\boldsymbol{N}) = \tau(\boldsymbol{N}')$, there is no requirement to test $\boldsymbol{N}'$ either.

Now we prove the claim. Consider a compatible potential outcome table $\boldsymbol{N}$. If $N * \tau(\boldsymbol{N})$ is odd or $\tau_k(\boldsymbol{N}) \neq 0, \forall k \in [K]$, then we do not need to perform Step 2(d). Suppose $\boldsymbol{N}$ is not in the set found in Step 2(b) for any ATE vector $\boldsymbol{T}$, then $\boldsymbol{N}$ satisfies:

$$\exists k, \boldsymbol{N} + e_k * \delta_A \text{ is compatible, or } \exists k, \delta_L^k(\boldsymbol{N}) \neq 0, \boldsymbol{N} + e_k * \delta_L^k(\boldsymbol{N}) \text{ is compatible.}$$

For each k, Let L^k be the largest integer such that $\boldsymbol{N} + L^k * e_k * \delta_L^k(\boldsymbol{N})$ is compatible, A^k be the largest integer such that $\boldsymbol{N} + A^k * e_k * \delta_A$ is compatible, then

$$\boldsymbol{N}' := \boldsymbol{N} + \sum_{k=1}^{K} L^k * e_k * \delta_L^k(\boldsymbol{N}) + \sum_{k=1}^{K} A^k * e_k * \delta_A$$

is a compatible potential outcome table in the set in Step 2(b) such that $\tau(\boldsymbol{N}') = \tau(\boldsymbol{N})$. Furthermore, by Theorem A.6 and Theorem A.7, $p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$.

If $N * \tau(\boldsymbol{N})$ is even and $\exists i \in [k]$ such that $\tau_i(\boldsymbol{N}) = 0$, the proof is a bit more complicated but follows a similar logic. If $\exists j \in [K]$ such that $N_{j10} = N_{j01} = 0$, then the condition of Theorem A.6 is not satisfied, but we can still use Theorem A.7. By repeatedly adding $e_i * \delta_L^i(\boldsymbol{N})$ to $\boldsymbol{N}$, we can construct a potential outcome table $\boldsymbol{N}'$ in the set in Step 2(d) such that $p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$. If $\forall j \in [K]$ such that $N_{j10} + N_{j01} > 0$, we can still use Theorem A.6, but we need to be careful not to violate the condition of Theorem A.6. By incrementally adding $e_k * \delta_L^k(\boldsymbol{N})$ and $e_k * \delta_A^k$ to $\boldsymbol{N}$, we can create a potential outcome table $\boldsymbol{N}'$ such that

1. $\boldsymbol{N}'$ is compatible,
2. $\forall k \in [K]$, either $\boldsymbol{N}' + e_k * \delta_A$is not compatible, or $N_{k10} = N_{k01} = 1$.
3. $\forall k \in [K]$, either $\boldsymbol{N}' + e_k * \delta_L^k(\boldsymbol{N}')$ is not compatible or $\delta_L^k(\boldsymbol{N}') = \mathbf{0}$.

Then, we know N' is in the set in Step 2(b) or 2(d) and $p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$.

Step 2: The required permutation tests align with the stated theorem.

Step 2 provides a loop that iterates $O(\prod_{k=1}^{K} N_k)$ times. Thus we only need to show the set in Step 2(b) and 2(d) have at most $O(\prod_{k=1}^{K} N_k)$ elements. For each potential outcome table $\boldsymbol{N} = (\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)^T$, we can rewrite its component $\boldsymbol{N}_k$ as

$$\boldsymbol{N}_k = (c_{N_k} + \frac{N_{k11} - N_{k00}}{2}, N_{k01} + N_k * \tau_k(\boldsymbol{N}), N_{k01}, c_{N_k} - \frac{N_{k11} - N_{k00}}{2}),$$

where

$$c_N = \frac{1}{2}(N_{k11} + N_{k00}) = \frac{1}{2}(N_k - 2N_{k01} - N_k * \tau_k(\boldsymbol{N})).$$

Thus, if τ_k is fixed, there are two degrees of freedom in each stratum: N_{k01} and $\frac{N_{k11}-N_{k00}}{2}$. However, in each set in Step 2(b) and 2(d), one degree of freedom has already been fixed: $\frac{N_{k11}-N_{k00}}{2}$. It can be observed that if N_{k01} is fixed, there can only be two compatible values of $\frac{N_{k11}-N_{k00}}{2}$ (the smallest two) for each stratum. Consequently, in each set in Step 2(b) and 2(d), there are at most $2^K \prod_{k=1}^K N_k$ potential outcome tables.

Step 3: Additional processes, such as searching, do not impose an undue computational burden.

The additional process that may potentially impose an undue computational burden is the searching process in Step 2(b) and 2(d). However, we can demonstrate that this search can be executed in $O(\prod_{k=1}^K N_k)$ time, alleviating concerns about computational burden.

For each $k \in [K]$, when we fix N_{k01}, we can efficiently determine the lower bound l of $|N_{k11} - N_{k00}|$ using Lemma B.16, which requires only $O(1)$ time. Given the structure of the sets in Step 2(b) and 2(d), any potential outcome table $\boldsymbol{N}$ within these sets must satisfy either:

$$|N_{k11} - N_{k00}| = l \text{ or } l + 1$$

Thus, for each fixed combination of $(N_{101}, N_{201}, \ldots, N_{K01})$, we can identify at most 2^K potential outcome tables in constant time. After identifying these tables, we check whether they meet the conditions specified in Step 2(b) or 2(d) and decide whether to include them. This approach yields an algorithm with a time complexity of $O(\prod_{k=1}^K N_k)$, demonstrating that the search process will not impose an undue computational burden.

□

C More simulation Results

C.1 Choice of combing functions in combining permutation method

In Section 4.4, we introduced the combing permutation method and selected the Fisher combing function as our default choice. In this section, we present a numerical comparison to demonstrate that the Fisher combing function typically yields the narrowest confidence intervals among common choices of p-value combing functions. Additionally, we introduce an alternative option, the Tippet's combing function. This method offers the advantage of high computational efficiency and can produce the narrowest confidence intervals in certain extreme cases.

We selected five different p-value combing functions: Fisher's method, Pearson's method, Mudholkar's and George's method, Tippett's method and Stouf-

fer's Z-score method and compared their performance in the combing permutation method. These five functions are the default method in scipy.stats package in Python. The difference between the methods can be illustrated by their statistics :

- The statistics of Fisher's method is $-2\sum_i \log(p_i)$, which is equivalent to the product of individual p-values. Under the null, the statistic is dominated by a χ^2 distribution. This method emphasises small p-values.
- Pearson's method uses $-2\sum_i \log(1-p_i)$ as the test statistic. This emphasises large p-values.
- Mudholkar and George compromise between Fisher's and Pearson's method by averaging their statistics. This method emphasises extreme p-values, both close to 0 and 1.
- Stouffer's method uses $\sum_i \Phi^{-1}(p_i)$ as the test statistic, where Φ is the CDF of the standard normal distribution.
- Tippett's method uses the smallest p-value as a statistic.(Note that the minimum is not the combined p-value)

We used the potential outcome tables in Table 1 to generate the observed data. For each of these tables, we generated 100 sets of observed data. Subsequently, we calculated the average width of the 95% confidence intervals obtained using the combining permutation method with various combining functions, along with the generated observed data. To approximate the permutation tests, we employed a random sample with replacement of 100 randomizations. The result is shown in Table 8.

The table illustrates that, in general, the Fisher combining function is the most powerful method. This is likely because Fisher's method places greater emphasis on small p-values, making it more likely to reject a treatment effect vector if one of its components is strongly rejected. In contrast, other methods tend to reject a treatment effect vector only if some or all of its components are rejected. This is the reason of selecting Fisher's method as our default choice.

Another notable observation in the table is that Tippet's method generally produces the second narrowest confidence interval and sometimes even the narrowest confidence interval. This tends to happen when the treatment vector exhibits significant heterogeneity across the strata. The rationale for this behavior could be that Tippet's method simply uses the smallest p-value as the test statistic, which may be less influenced by the presence of heterogeneity across the strata.

Tippet's combining function has another advantage in terms of computational efficiency. To obtain a confidence interval using Tippet's method with inverted permutation tests, we don't need to test every combination of stratum-wise treatment effects. Instead, we can break it down into two steps. First, obtain individual $(1-\alpha)^{1/K}$ confidence intervals in each stratum using inverted permutation tests. Then, add these stratum-wise confidence intervals together

N	n	τ	$\boldsymbol{N}$	Fisher	Pearson	George	Tippet	Stouffer
(40,40)	(10,10)	(0,0)	[10,10,10,10], [10,10,10,10]	**0.57**	0.74	0.59	0.67	0.6
(20,20)	(15,15)	(0.2,0.9)	[3,8,4,5], [0,19,1,0]	**0.65**	0.84	0.73	0.7	0.74
(30,40)	(5,30)	(0.7,-0.7)	[3,23,2,2], [4,2,30,4]	0.61	0.93	0.86	**0.59**	0.87
(30,30)	(5,25)	(0.8,0.8)	[2,24,0,4], [1,26,2,1]	**0.53**	0.71	0.61	0.62	0.62
(10,40)	(5,20)	(-0.9,1)	[1,0,9,0], [0,40,0,0]	0.36	1	1	**0.27**	1
(20,80)	(15,60)	(0,0.6)	[5,5,5,5], [20,50,2,8]	**0.51**	0.9	0.83	0.52	0.84
(15,60)	(10,40)	(0.8,0.9)	[2,12,0,1], [2,55,1,2]	**0.32**	0.87	0.83	0.34	0.84
(20,70)	(5,60)	(-0.5,0.9)	[2,2,12,4], [3,64,1,2]	0.56	0.97	0.93	**0.5**	0.93
(20,20,20)	(5,10,15)	(0.8,0.4,0)	[0,16,0,4], [3,9,1,7], [5,5,5,5]	**0.66**	0.87	0.79	0.79	0.79
(15,20,25)	(10,10,10)	(0.8,0.9,0.8)	[1,13,1,0], [0,18,0,2], [0,20,0,5]	**0.42**	0.81	0.68	0.56	0.68
(20,15,20)	(5,5,5)	(0.9,0,-0.8)	[0,19,1,0], [3,4,4,4], [0,2,18,0]	**0.69**	0.98	0.96	0.71	0.96
(10,20,30)	(5,5,25)	(0,0,0)	[5,0,0,5], [6,0,0,14], [18,1,1,10]	**0.74**	0.91	0.86	0.85	0.86
(30,40,50)	(10,10,10)	(0.5,0.5,0.5)	[8,15,0,7], [9,21,1,9], [12,26,1,11]	**0.47**	0.8	0.67	0.64	0.67
(40,40,40)	(20,30,10)	(0.5,-0.6,0)	[10,20,0,10], [7,1,25,7], [12,8,8,12]	**0.52**	0.9	0.78	0.64	0.78
(50,50, 50,80)	(25,25, 25,40)	(0,0,0,0)	[5,0,0,45], [10,0,0,40], [10,0,0,40] [5,0,0,75]	**0.34**	0.83	0.63	0.49	0.64

Table 8: Simulation results for different potential outcome tables. The numbers in each cell represent the average width of the 95% confidence intervals. The results highlighted in bold font indicate the method produced the narrowest confidence intervals.

to create an overall $1-\alpha$ confidence interval. This approach is equivalent to combining the stratum-wise confidence intervals using inverted permutation tests with Sidak's correction. To see this, one only need to notice that the p-value obtained from Tippett's method is given by

$$1-(1-\min(p_1,p_2,\ldots,p_K)^K).$$

As such, we recommend Tippett's method as an alternative choice of the combining function used in the permutation combining method, particularly when dealing with substantial heterogeneity across strata or larger datasets.

C.2 Extended permutation method is not always the best

In Section 5, we showed that extended permutation method is generally the most powerful method among the exact methods. However, when there is significant heterogeneity across strata, the "extend" methods are outperformed by the "combining" methods. Specifically, in cases where the treatment effect differs substantially across strata, the combining permutation method tends to yield the narrowest confidence interval, followed by the extended inverted permutation method and the fast method. The Wendell & Schmee method consistently lags behind. To illustrate this, we conducted a simulation with the same setup as in Figure **??** but with different τ_1 and τ_2 values that are far apart from each other. The results are presented in Figure **??**. Note that in this simulation we used Fisher combining function as the choice of combining method. If using Tippett's method instead, one may get a narrower confidence interval, as illustrated in the previous section. This finding is consistent as in [Stark,2023].