Exact and Conservative Inference for the Average Treatment Effect in Blocked Experiments with Binary Outcomes

Jiaxun Li and Jacob Spertus and Philip B. Stark

DRAFT: April 6, 2024
Rough and Incomplete

Contents

Abstract

We extend methods for inference about the treatment effect in fully randomized binary experiments with binary outcomes to accommodate blocking. Three methods are presented within the potential outcome framework to assess treatment effects, each differing in computational and statistical efficiency. Our first method calculates confidence intervals for treatment effects in both treatment and control groups, incorporating the Bonferroni adjustment to control for multiple comparisons. The second method inverts a series of permutation tests. Our third approach similarly inverts permutation tests across each stratum to derive confidence sets for treatment effects per stratum, subsequently integrating these sets using combining functions. Through numerical simulations and a case study, we compare these methods in terms of their statistical and computational performance. Our findings indicate that the second approach has the highest statistical efficiency. It requires a computational complexity of $O(\Pi_{k=1}^K N_k^4)$ permutation tests across K strata with sample sizes $N_k, k = 1, \ldots, K$. We present new theorems that significantly reduce this computational demand to $O(\Pi_{k=1}^K N_k^2)$ if all the strata are balanced and to at most $O(\Pi_{k=1}^K N_k^3)$ otherwise.

1 Introduction

A collection of N subjects is randomized into two groups, one of size n and one of size $m := N - n$. The first group is assigned to (active) “treatment” and the second to “control.” For each subject, a binary response is measured, e.g., survival to time t. Such *randomized, controlled trials* (RCTs) with binary treatments and binary outcomes have been studied at least since Fisher’s seminal work Fisher [1935]; see also **?**.

A convenient and powerful way to model 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 [Neyman, 1923, Rubin, 1974]. The numbers $\{y_j(k)\}_{j=1}^N, k = 0, 1$ are considered fixed before the randomization. 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 observed 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), denoted τ, is the mean of the responses that would have resulted if every subject had been assigned to treatment, minus the mean of the responses that would have resulted result if every subject had been assigned to control:

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

The ATE measures the effectiveness of treatment.[2] Many applications involve tests and confidence sets for the ATE in binary experiments with binary outcomes, including agricultural experiments, medical experiments, marketing, and many others PBS: add citations. Those tests and confidence sets often rely on asymptotic theory, resulting in tests and confidence sets that can be anti-conservative in practice [Rigdon and Hudgens, 2015, Li and Ding, 2016].

A number of methods have been proposed for making exact or conservative inferences about the ATE in *completely randomized experiments* with binary treatments and binary outcomes, wherein the assignment to treatment is by simple random sampling so that every possible subset of n of the N subjects is equally likely to be given the active treatment [Santner and Snell, 1980, Branson and Bind, 2019, Chiba, 2015, Rigdon and Hudgens, 2015, Li and Ding, 2016].

A simple, computationally efficient, but statistically over-conservative approach uses the fact that under simple random sampling, the number of responses that are equal to 1 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 responses equal to 1 in the control group has a hypergeometric distribution with parameters $N = N$, $n = N - n$, and $G = \sum_{j=1}^N y_j(0)$. The number of 1s among the treated and among the controls 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 [Chiba, 2015, 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

[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 disagreement between Fisher and Neyman about the "correct" null hypothesis to test [Fienberg and Tanur, 1996, Wu and Ding, 2021]. Fisher advocated testing the "strong" null that treatment has no effect whatsoever on any individual, i.e., that $y_j(0) = y_j(1)$ for all j. Neyman advocated testing the "weak" null hypothesis that $\tau = 0$.

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 every table with $\tau = \tau_0$ 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); indeed, Bruce et al. [2022] estimates that almost two-thirds of clinical trials use some form of stratification.

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 experiments we know of are those of Rigdon and Hudgens [2015], Chiba [2017]. Those 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. The methods are illustrated 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 whom 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 assignment table* $\boldsymbol{Z} := (\boldsymbol{Z}_k)_{k=1}^{K}$. 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 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 the potential outcomes using $2 \times 2 \times K$ integers: $\boldsymbol{N} := (\boldsymbol{N}_k)_{k=1}^{K}$. The stratum-wise average treatment effect τ_k and the average treatment effect τ 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 outcome table* is $\boldsymbol{n} := (\boldsymbol{n}_k)_{k=1}^{K}$. The the estimated stratum-wise average treatment effect $\hat{\tau}_k$ and the estimated average treatment effect $\hat{\tau}$ can be written as functions of $\boldsymbol{n}$:

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

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

Let $\bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z})$ denote the observed outcome table that would result from the treatment assignment table $\boldsymbol{Z}$ 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 observed outcome table $\boldsymbol{n}$ constrains the potential table $\boldsymbol{N}$ algebraically: $\boldsymbol{N}$ is *algebraically compatible* with $\boldsymbol{n}$ if there is some treatment assignment table $\boldsymbol{Z}$ for which $\boldsymbol{n} = \bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z})$.

3 Existing Methods for Constructing Confidence Intervals for the ATE

Rigdon and Hudgens [2015], Li and Ding [2016] present two basic approaches to finding confidence intervals for the ATE for completely randomized binary experiments (i.e., experiments without blocking or stratification, in which subjects are assigned to treatment by simple random sampling). One is based on hypergeometric confidence sets as described above; the other is based on inverting a series of permutation tests.

In this section, we suppress the stratum subscript since there is only one stratum. For example, $\boldsymbol{N} := (N_{11}, N_{10}, N_{01}, N_{00})$ denotes the potential outcome table and $\boldsymbol{n} := (n_{11}, n_{10}, n_{01}, n_{00})$ denotes the observed outcome table.

3.1 Combining hypergeometric confidence bounds

Let $N_{1\bullet} := N_{10} + N_{11}$ be the number of subjects whose response would be 1 if assigned to the active treatment and $N_{\bullet 1} := N_{01} + N_{11}$ be the number of subjects whose response would be 1 if assigned to control. The ATE is

$$\tau(\boldsymbol{N}) := (N_{10} - N_{01})/N = (N_{1\bullet} - N_{\bullet 1})/N. \tag{7}$$

One way to construct a conservative confidence interval for the ATE combines simultaneous confidence intervals for N_{1+} and $N_{\bullet 1}$. In a completely randomized experiment, the treatment and control groups are simple random samples of the N subjects, so

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

Standard methods can be used to find $1-\alpha/2$ confidence intervals for the hypergeometric parameters $N_{1\bullet}$ and $N_{\bullet 1}$. The union bound ensures that the pair of intervals has simultaneous confidence level $1-\alpha$. Let $N_{1\bullet}^{\mathrm{L}}$ be the resulting

[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.

lower confidence bound for $N_{1\bullet}$, $N_{1\bullet}^{\mathrm{U}}$ be the resulting upper confidence bound for $N_{1\bullet}$ and define $N_{\bullet 1}^{\mathrm{L}}$ and $N_{\bullet 1}^{\mathrm{U}}$ analogously. Then

$$[(N_{1\bullet}^{\mathrm{L}} - N_{\bullet 1}^{\mathrm{U}})/N, (N_{1\bullet}^{\mathrm{U}} - N_{\bullet 1}^{\mathrm{L}})/N]$$

is a conservative $1 - \alpha$ confidence interval for τ. The method we propose in section 4 extends this idea to stratified (blocked) randomization. Other methods for finding confidence intervals for the ATE using simultaneous hypergeometric confidence intervals are given in Rigdon and Hudgens [2015], Li and Ding [2016].

3.2 Inverting hypothesis tests

We will construct confidence intervals for the ATE using the general duality between confidence sets and hypothesis tests. Consider the simple null hypothesis

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

where $\boldsymbol{\delta} = (\delta_1, \delta_2, \ldots, \delta_N)$ is a known vector. This is an example of a *sharp* null hypothesis [Rubin, 1980]: it allows us to impute the full set of potential outcomes $\boldsymbol{y}$ from the observed outcomes $\boldsymbol{Y}$, thereby completely determining the randomization distribution of any test statistic $T(\boldsymbol{Y}, \boldsymbol{Z})$.

When larger values of T are stronger evidence against the null, the natural way to define a P-value from T is

$$\mathbb{P}_{\boldsymbol{\delta}} \left(T(\boldsymbol{Y}, \boldsymbol{Z}) \geq T(\boldsymbol{Y}, \boldsymbol{Z}^{\mathrm{obs}}) \right), \tag{8}$$

where $\boldsymbol{Z}^{\mathrm{obs}}$ is the observed treatment vector, and the probability on the right hand side is calculated with respect to the random assignment $\boldsymbol{Z}$, on the assumption that the full set of potential outcomes are those implied by the observed outcomes and the sharp null hypothesis $H_0(\boldsymbol{\delta})$.

Following Rigdon and Hudgens [2015], Li and Ding [2016], Aronow et al. [2023], we take $T(\boldsymbol{Y}, \boldsymbol{Z}) := |\hat{\tau}(\boldsymbol{Y}, \boldsymbol{Z}) - \tau(\boldsymbol{y})|$. (There are other sensible choices, but this one works well in practice; see section 6.1.) Recall that observed binary outcomes can be summarized by $\boldsymbol{n}$. Under $H_0(\boldsymbol{\delta})$, there is a unique potential outcome table $\boldsymbol{N} = \boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}$ for which the treatment effects are $\boldsymbol{\delta}$ and $\bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}, \boldsymbol{Z}^{\mathrm{obs}}) = \boldsymbol{n}$. Thus, for this test statistic, the probability (8) can be rewritten:

$$p(\boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}, \boldsymbol{n}) := \mathbb{P}_{\boldsymbol{\delta}} \left(|\tau(\boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}) - \hat{\tau}(\bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}, \boldsymbol{Z}))| \geq |\tau(\boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}) - \hat{\tau}(\boldsymbol{n})| \right) \tag{9}$$

We reject $H_0(\boldsymbol{\delta})$ if $p(\boldsymbol{N}_{\boldsymbol{\delta},\boldsymbol{n}}, \boldsymbol{n}) \leq \alpha$. This procedure is often called the *Fisher randomization test* (FRT) or *permutation test* of the sharp null $H_0(\boldsymbol{\delta})$.

We can obtain a confidence interval by inverting a family of permutation tests. Explicitly, for each possible value of $\boldsymbol{\delta}$, we perform a permutation test of $H_0(\boldsymbol{\delta})$. The value τ is in the $1-\alpha$ confidence set for the ATE iff there is at least one $\boldsymbol{\delta}$ for which $H_0(\boldsymbol{\delta})$ is not rejected and $\tau = \sum_{i=1}^{N} \delta_i/N$. In practice, upper and lower confidence bounds are more useful than a possibly disconnected confidence set, so we focus on finding the smallest and largest and smallest values of τ in the confidence set.

Since $\#(\mathcal{Z}(N,n)) = \binom{N}{n}$, it is computationally intractable to compute the "full randomization" P-value when N is large and n is not close to 0 or N. Instead, we shall use an exact Monte Carlo P-value for a related randomized test: we draw r treatment allocations at random, uniformly, according to the experimental design, creating r sets of observed responses for a given hypothesized potential outcome table. For each, we compute the value of the test statistic. Let h denote the number of allocations among those r for which the test statistic is greater than or equal to the observed value for the original data. Then the ratio $(h+1)/(r+1)$ is an exact P-value for the hypothesis $H_0(\boldsymbol{\delta})$ [Dwass, 1957, Ramdas et al., 2023]. One may take r to be arbitrarily small to reduce the computational burden, but as r shrinks, the smallest attainable P-value grows (it is $1/(r+1)$).

While the null hypothesis for the permutation test specifies the individual treatment effects, the permutation distribution of the test statistic is fully determined by the potential outcome table $\boldsymbol{N}$. Therefore, the number of permutation tests that need to be performed is at most the number of potential outcome tables tested. (Each test can be conducted using a fixed number r of randomly generated treatment allocations; indeed, the same set of r allocations can be used to test every table, to reduce the computational burden.) There are at most $\binom{N+3}{3}$ potential outcome tables $\boldsymbol{N} = (N_{11}, N_{10}, N_{01}, N_{00})$ with $N_{11} + N_{10} + N_{01} + N_{00} = N$. Hence, the computational complexity of the overall confidence procedure is at most $O(N^3)$. Li and Ding [2016] showed that the number of potential outcome tables that must be considered is at most $O(N^2)$. Aronow et al. [2023] showed that for balanced experiments, testing $O(N \log N)$ potential outcome tables suffices. Section A extends their methods to stratified experiments.

4 Constructing Exact Confidence Intervals

This section extends the methods presented in Section 3 to stratified experiments.

4.1 Extending the hypergeometric confidence bound approach to stratified samples

We describe a procedure analogous to that outlined in Section 5 of Rigdon and Hudgens [2015]. For each stratum k, let $N_{k1\bullet} := N_{k10} + N_{k11}$ denote the number of subjects in stratum k whose response would be 1 if assigned to active treatment, and let $N_{k\bullet 1} := N_{k01} + N_{k11}$ denote the number of subjects in stratum k whose response would be 1 if assigned to control. Define $N_{1\bullet} := \sum_{k=1}^{K} N_{k1\bullet}$ and $N_{\bullet 1} := \sum_{k=1}^{K} N_{k\bullet 1}$. Then the average treatment effect can be written $\tau = (N_{1\bullet} - N_{\bullet 1})/N$. We can obtain a conservative confidence interval for τ in the stratified case using a method analogous to that in Section 3.1 by combining simultaneous confidence intervals for $N_{1\bullet}$ and $N_{\bullet 1}$ from stratified samples. Specifically, suppose that $[N_{1\bullet}^L, N_{1\bullet}^U]$ and $[N_{\bullet 1}^L, N_{\bullet 1}^U]$ are simultaneous $1-\alpha$ confidence intervals for $N_{1\bullet}$ and $N_{\bullet 1}$, respectively. Then

$$\left[\frac{N_{1\bullet}^L - N_{\bullet 1}^U}{N}, \frac{N_{1\bullet}^U - N_{\bullet 1}^L}{N}\right] \tag{10}$$

is a conservative $1-\alpha$ confidence interval for the ATE τ.

Simultaneous confidence intervals for $N_{1\bullet}$ and $N_{\bullet 1}$ can be constructed from the stratified experimental data in a variety of ways. We consider two: the method of Wendell and Schmee [1996] with the Bonferroni adjustment for simultaneity, and the method of **?** with the Bonferroni adjustment. The method of Wendell and Schmee [1996] is computationally intractable when the number of strata is greater than about 3; the method of **?** is still efficient when there are many strata.

4.2 Stratified Permutation Method

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:

$$H_0(\boldsymbol{\delta}) : y_{kj}(1) - y_{kj}(0) = \delta_{kj},\ k = 1, \ldots, K;\ j = 1, \ldots, N_k \tag{11}$$

where $\boldsymbol{\delta} = (\delta_{11}, \ldots, \delta_{1N_1}, \ldots, \delta_{k1}, \ldots, \delta_{kN_k})$ 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:

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

Thus, a P-value for $H_0(\delta)$ is:

$$p(\boldsymbol{N}, \boldsymbol{n}) = \mathbb{P}\left(|\tau(\boldsymbol{N}) - \hat{\tau}(\bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z}))| \geq |\tau(\boldsymbol{N}) - \hat{\tau}(\boldsymbol{n})|\right), \tag{12}$$

where the probability is with respect to $\boldsymbol{Z}$. We can then reject $H_0(\boldsymbol{\delta})$ 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 $H_0(\boldsymbol{\delta})$ for each possible value of $\boldsymbol{\delta}$. Then, we select the largest and smallest values of $\bar{\bar{\delta}} = \mathbf{1}^T\boldsymbol{\delta}/N$ obtained from these permutation tests as the upper and lower bounds of the confidence interval for the ATE.

This method is related to the method in Chiba [2017] and the method in Section 5 of Rigdon and Hudgens [2015]. As those papers mention, the computational complexity of this approach 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 performed in total is on the order of $O(\prod_{k=1}^K N_k^3)$. The next section introduces an alternative approach that is more computationally efficient.

4.3 Inverted Permutation Test with Combining Function

This section extends the ideas in 3.2 and 3.1. Consider testing the intersection null hypothesis

$$H_0(\tau_{10}, \ldots, \tau_{K0}) : \tau_k = \tau_{k0}, k = 1, \ldots, K.$$

We can compute a P-value for this hypothesis by finding P-values for the individual hypotheses, then combining those P-values, for instance using Fisher's combining function. (The P-values are independent across strata because the random treatment assignment is independent across strata.) A P-value for the composite hypothesis hypothesis $H_{k0} : \tau_k = \tau_{k0}$ can be constructed by finding the largest P-value over all simple hypotheses (potential outcome tables for stratum k) that have that value of the ATE, i.e., by finding the maximum P-value over a multidimensional nuisance parameter, the potential outcome table:

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

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

For the Fisher combining function, a conservative P-value for $H_0(\tau_{10}, \ldots, \tau_{K0})$ can be found by using the fact that under the intersection null, the distribution of the test statistic

$$t_F(\tau_{10}, \ldots, \tau_{K0}) = -2 \sum_{k=1}^{K} \ln(p(\tau_{k0}))$$

is dominated by the chi-square distribution with $2K$ degrees of freedom. Let $\chi^2_{2K}(\alpha)$ denote the $1-\alpha$ quantile of the chi-square distribution with $2K$ degrees of freedom. Then we can reject $H_0(\tau_{10}, \ldots, \tau_{k0})$ if $t_F(\tau_{10}, \ldots, \tau_{k0}) \geq \chi^2_{2K}(\alpha)$. The rationale for using Fisher's combining function is in Appendix C.

Unlike the method described in the previous section, this method based on testing intersection hypotheses by combining P-values does not require performing a permutation test for every combination of potential outcome tables $(\boldsymbol{N}_1, \boldsymbol{N}_2, \ldots, \boldsymbol{N}_K)$. Instead, one can conduct permutation tests in each stratum to obtain P-values for each possible value of τ_k (i.e., larger than $(n_{k11} + n_{k00} - N_k)/N_K$ and less than $(n_{k01} + n_{k10})/N_k$) and then combine those P-values using a P-value combining function. One advantage of this approach is that at most N_k values of τ_k are algebraically compatible with the data in stratum k, resulting in a time complexity of $O(\prod_{k=1}^{K} N_k)$ for finding the confidence interval for τ. Also, the number of operations to find a P-value for stratum k is at most $O(AN_k^3)$, where A is the number of Monte Carlo permutations. While Aronow et al. [2023] choose A to be proportional to $\log(N_k \log N_k)$, one may construct exact randomized P-values by picking A to be any constant [Dwass, 1957], which leads to an overall operation count $O(\max\{\sum_k N_k^3, \prod_k N_k\}$. This is much less than the complexity $O(\prod_{k=1}^{K} N_k^3)$ of the "global" permutation method in Section 4.2.

5 Illustration

5.1 Simulations

This section compares the four new methods with the approximate, asymptotic $1-\alpha$ Wald confidence interval based on the normal approximation:

$$\hat{\tau} \pm z_{1-\alpha/2} \left\{ \sum_{k=1}^{K} \left(\frac{N_k}{N} \right)^2 \left(\frac{\widehat{S}_k^2(1)}{n_k} + \frac{\widehat{S}_k^2(0)}{N_k - n_k} \right) \right\}^{1/2}, \tag{13}$$

where

$$\widehat{S}_k^2(1) := (n_k - 1)^{-1} \sum_{j=1}^{N_k} Z_{kj}(Y_{kj} - \hat{\bar{Y}}_k(1)),$$

$$\widehat{S}_k^2(0) := (N_k - n_k - 1)^{-1} \sum_{j=1}^{N_k} (1 - Z_{kj})(Y_{kj} - \hat{\bar{Y}}_k(0)),$$

$$\hat{\bar{Y}}_k(1) := 1/n_k \sum_{j=1}^{N_k} Z_{kj} Y_{kj},$$

$$\hat{\bar{Y}}_k(0) := 1/(N_k - n_k) \sum_{j=1}^{N_k} (1 - Z_{kj}) Y_{kj},$$

and $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard normal distribution. Table 1 reports the mean widths and empirical coverage rates of the confidence intervals for a variety of potential outcome tables.[4] The simulations were conducted as follows:

1. Set the number of strata K, the stratum sizes (N_k), and the potential outcomes $\boldsymbol{N}$, which specifies the treatment effect τ. Set the sizes of the treated groups in each stratum (n_k) For $r = 1, \ldots, 100$:

   - For $k = 1, \ldots, K$, randomly assign n_k subjects to treatment, generating an observed outcome table $\boldsymbol{n}$.
   - Compute the (nominal or actual) 95% confidence interval using each of the five methods; the permutation methods used 100 random permutations.

2. Report the average width, empirical coverage rate, and mean running time.

[4] Throughout this article, the simulations are done with a ROG14 laptop with AMD Ryzen 7 6800HS CPU, 4.00 GHz clock, about 580 MB memory, with python version 3.10.11 and visual studio code version 1.85.

The results in Table 1 show that the Wald method does not achieve its nominal coverage rate even when the sharp null is true. Among the exact/conservative methods, the stratified permutation method generally produced the narrowest confidence intervals, followed by the Bonferroni-adjusted Wendell & Schmee method, combining stratumwise permutation P-values, and finally the Bonferroni-adjusted fast stratified inference method. For some potential outcome tables, the methods do not follow this pattern; see Appendix C.2 for the results of additional simulations.

In both types of approach—permutation tests versus Bonferroni-adjusted population confidence bounds—the method that generally leads to the narrowest confidence bounds is computationally intractable when there are more than a few strata of moderate size. Thus, there is a trade-off between statistical efficiency and computational efficiency. In particular, Table 1 shows that Bonferroni-adjusted Wendell & Schmee and the stratified permutation method are impractical when $K \geq 3$ and $N \geq 100$. Computing confidence intervals with the stratified permutation method, which produced the narrowest confidence intervals on average, required up to 30 minutes for $K = 3$ and $N = 60$. Algorithms proposed in Section A improve the computational efficiency of the stratified permutation method, but it remains intractable for problems of sizes that arise frequently in practice.

We also ran another simulation involved "balanced" experiments: in each stratum, the number of subjects assigned to treatment is equal to the number assigned to control. The true value of the ATE in stratum k, τ_k, is always a multiple of $1/N_k$ so that $\tau_k N_k$ is an integer. Simulations were carried out for two strata and various values of N, τ_1, and τ_2, as follows:

1. For $r = 1, \ldots, 100$:
    - For $k = 1, \ldots, K$, potential outcomes in stratum k were generated by setting $y_{kj}(1) = 1$ 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 outcomes $y_{\tau_i N_k+1}(0), \ldots, y_{N_k}(0)$ were set to a random permutation of $y_{\tau_k N_k+1}(1), \ldots, y_{N_k}(1)$, guaranteeing that the average treatment effect is τ_k.
    - Data for stratum k were generated by randomly assigning $N_k/2$ subjects to treatment and the others to control.
    - The 95% confidence intervals were computed using all five methods; for the permutation methods, 100 random permutations were used.
2. Report the average width of the confidence intervals.

The results in Figure 1 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 combining function method. The Wald method falls between the extended permutation and the Wendell and Schmee method.

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

Table 1: Mean confidence interval widths, empirical coverage rates, and mean running times for confidence bounds for τ in unbalanced experiments. The three rows of numbers in the last five columns represent average width, coverage rate, average time).Scenarios in which the method did not complete within 30 minutes are marked with an asterisk.

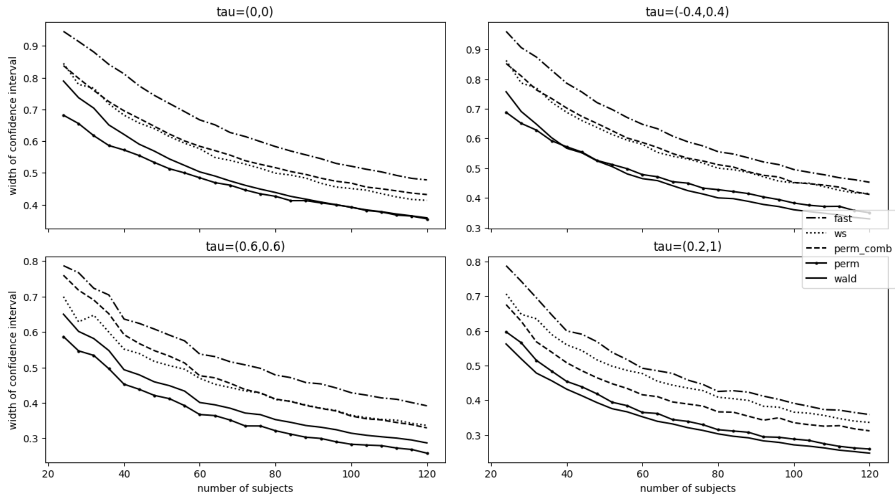

Figure 1: Simulation results for balanced experiments. The X-axis represents the number of subjects. The Y-axis represents the average confidence width given by different methods.

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 Travis et al. [2023] from different methods

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' stratum, 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' stratum, 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

We introduced four methods for constructing exact or conservative confidence sets for the average treatment effect in stratified randomized experiments with binary outcomes. These methods extend 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 or conservative 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. Another 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 Other combining functions and other test statistics

Section 4.3, uses Fisher's P-value combing function. We found that Fisher's combining function performed better in simulations across diverse scenarios, described in Appendix C, which compares the power of various combination functions. When strata are large, Tippett's combination function may have a lower computational burden than Fisher's function because it decouples the stratumwise tests; see Appendix C.

To test the hypothesis that the observed table $\boldsymbol{n}$ arose from the potential outcome table $\boldsymbol{N}$, we used (8), the P-value based on the absolute difference test statistic $|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|$. We tried two other approaches: a test based on the

$\boldsymbol{n}$	Absolute difference	Studentized absolute difference
[10, 10, 10, 10]	[-11,11]	[-12,12]
[10, 10, 9, 10]	[-10,12]	[-10,13]
[10, 10, 1, 19]	[7,26]	[9,26]
[19, 1, 1, 19]	[29,37]	[30,37]
[6, 6, 10, 10]	[-9,9]	[-11,11]
[6, 6, 1, 19]	[5,21]	[4,24]
[11, 1, 1, 19]	[20,29]	[21,29]
[2, 2, 15, 15]	[-13,13]	[-15,15]
[2, 2, 2, 28]	[0,28]	[-2,29]
[3, 1, 2, 28]	[8,30]	[0,30]

Table 3: Confidence interval for the ATE, multiplied by N, for a variety of unstratified experiments. Column 1: observed outcome table. Column 2: confidence interval when P-values are based on the absolute value of the difference between the sample ATE and the true ATE. Column 3: confidence interval when P-values are based on the studentized absolute value of the difference between the sample ATE and the true ATE.

studentized absolute difference between the sample ATE and the true ATE, and Sterne's method for the signed estimated ATE, $\hat{\tau}(\boldsymbol{n})$ (rather than a symmetric test).

The studentized absolute difference is $|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|/\widehat{V}(\boldsymbol{n})^{1/2}$, where $\widehat{V}(\boldsymbol{n})$ is the estimated variance (see (13)):

$$\widehat{V}(\boldsymbol{n}) = \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 \left(\frac{\widehat{S}_k^2(1)}{n_k} + \frac{\widehat{S}_k^2(0)}{N_k - n_k}\right)$$

Table 3 compares confidence intervals based on the 'raw' and studentized absolute difference between the sample ATE and the true ATE. Table 3 suggests that the absolute difference statistic performs better, especially when the experiment is unbalanced. This could be because, for some potential outcome tables, the variability of $\widehat{V}(\boldsymbol{n})$ increases the variability of the test statistic. Moreover, when strata are small, there may be a large probability that the estimated variance is 0, making the statistic infinite.

To use Sterne's method, we first identified the distribution of $\hat{\tau}(\bar{\bar{\boldsymbol{n}}}(\boldsymbol{N}, \boldsymbol{Z}))$. We approximated the acceptance region for Sterne's method by the narrowest range of values of $\hat{\tau}$ that contained $(1-\alpha) \times 100\%$ of the samples.

Table 4 shows the confidence intervals for the ATE multiplied by N for various observed tables, along with the number of potential outcome tables that were rejected for each observed table. The result shows that Sterne's method usually rejects more potential outcome tables than the test based on the absolute difference, yet does not produce narrower confidence intervals for the ATE.

Overall, we recommend using the absolute difference between the sample

$\boldsymbol{n}$	Absolute difference		Sterne's method	
	interval	rejected	interval	rejected
[10, 10, 10, 10]	[-11, 11]	1133	[-10, 10]	1298
[10, 10, 9, 10]	[-10, 12]	1245	[-10, 12]	1313
[10, 10, 1, 19]	[7, 26]	1564	[8, 25]	1671
[19, 1, 1, 19]	[29, 37]	1125	[30, 37]	1135
[6, 6, 10, 10]	[-9, 9]	509	[-10, 10]	538
[6, 6, 1, 19]	[5, 21]	520	[5, 22]	533
[11, 1, 1, 19]	[20, 29]	644	[26, 31]	649
[2, 2, 15, 15]	[-13, 13]	121	[-13, 13]	116
[2, 2, 2, 28]	[0, 28]	118	[0, 28]	116
[3, 1, 2, 28]	[8, 30]	167	[17, 33]	197

Table 4: Confidence intervals (adjusted by a factor of N) and the number of potential outcome tables that been rejected, for a variety of observed tables for unstratified experiments.

ATE and the true ATE as the test statistic, as described in Section 3.2.

6.2 Missing Data

Missing observations can be accounted for conservatively by replacing them with the least favorable value for each endpoint of the confidence interval: when computing the upper confidence bound on the treatment effect, treat missing control cases as 0 and missing treated cases as 1. When computing the lower confidence bound, do the opposite. This approach is computationally tractable whenever the full-data problem is tractable.

6.3 Confidence Intervals for Relative Risk

The methodology described in Section 4 also works to find confidence intervals for relative risk (RR), the rate of ones if all subjects were assigned to the active treatment, $N_{1\bullet}/N$, divided by the rate of ones if all subjects were assigned to control, $N_{\bullet 1}/N$. An inexpensive approach is to find simultaneous $1-\alpha$ confidence intervals for $N_{1\bullet}$ and $N_{\bullet 1}$, for instance by finding level $1-\alpha/2$ confidence intervals using the methods outlined in Section 4.1.

A sharper approach is to find the smallest and largest values of RR in a confidence set for the potential outcome table derived using methods in Section 4.2.

While confidence bounds for RR can be based on any test statistic, different statistics generally yield confidence intervals with different widths. Instead of basing the confidence set on the absolute difference between the estimated and true ATE, using the absolute difference between the estimated and true RR might yield shorter intervals on average.

However, if the same test statistic (and the same set of permutations) is used to construct confidence intervals for ATE and for RR, the confidence bounds for the two quantities have simultaneous coverage probability: whenever the CI for the ATE covers, the CI for RR also covers. Such a "strict bounds" approach [Evans and Stark, 2002] makes it possible to make valid confidence intervals for more than one quantity without the need to adjust for multiplicity.

A Reducing the Computational Cost of Inverting Permutation Tests

A.1 A General Approach

This section presents a general approach to reduce the computational cost of the method outlined in Section 4.2. The approach does not requite the experiment to be balanced: roughly speaking, if all the strata contain a similar number of subjects, the method requires testing on the order of $O(N^{3k-1})$ potential outcome tables, which can be conducted using a single set of random allocations. (In sections A.2 and A.3 below, we show that if some or all of the strata are balanced, the number of tests can be reduced substantially.)

We represent the outcome table $\boldsymbol{N}$ with K strata as a matrix with 4 columns and K rows; its kth row is $[N_{k11}, N_{k10}, N_{k01}, N_{k00}]$. If K is a positive integer, let $[K]$ denote the set $1, \ldots, K$. For $k \in [K]$, let $\boldsymbol{e}^k$ denote the "one-hot" column vector of length K, all of whose components are zero except the kth, which is 1.

We will use the symbol $\boldsymbol{d}$ to denote a column vector of length 4 with integer entries that sum to 0. Adding $\boldsymbol{d}^T$ to the kth row of $\boldsymbol{N}$ corresponds to changing the potential outcomes of some number of subjects in stratum k. For instance, let $\boldsymbol{d} := (-1, 1, 0, 0)^T$ and let $\boldsymbol{e}^k\boldsymbol{d}^T$ be the K by 4 matrix whose (i, j)th entry is $e_i^k \boldsymbol{d}_j$. Adding $\boldsymbol{e}^k\boldsymbol{d}^T$ to $\boldsymbol{N}$ corresponds to changing the responses of a subject in stratum k whose response to treatment and to control would be 1 instead to have a response of 1 to treatment and 0 to control. If $\boldsymbol{d} := (-2, 0, 0, 2)^T$, then adding $\boldsymbol{e}^k\boldsymbol{d}^T$ to $\boldsymbol{N}$ corresponds to changing the responses of two subjects in stratum k whose responses both to treatment and to control were originally 1 instead to have responses of 0 both to treatment and to control.

Recall that N_k is the number of subjects in stratum k and n_k is the number of subjects in stratum k who were assigned to treatment.

Theorem A.1. *Define*

$$\Delta_k := \begin{cases} \{(-1,1,0,0), (0,0,-1,1)\}, & n_k < N_k/2 \\ \{(0,1,0,-1), (1,0,-1,0)\}, & n_k > N_k/2 \\ \{(0,1,0,-1), (1,0,-1,0), (-1,1,0,0), (0,0,-1,1)\}, & n_k = N_k/2. \end{cases}$$

Let

$$\Delta := \{\boldsymbol{e}^k \boldsymbol{d}_k^T : \boldsymbol{d}_k \in \Delta_k, k \in [K]\}. \tag{14}$$

Consider two potential outcome tables $\boldsymbol{N}$ and $\boldsymbol{N}'$, and let $\boldsymbol{n}$ be an observed table. Suppose $\boldsymbol{N}' = \boldsymbol{N} + \boldsymbol{d}$ for some $\boldsymbol{d} \in \Delta$. Then $\tau(\boldsymbol{N}') > \tau(\boldsymbol{N})$. Moreover, if $\tau(\boldsymbol{N}') \leq \hat{\tau}(\boldsymbol{n})$, then $p(\boldsymbol{N}', \boldsymbol{n}) \geq p(\boldsymbol{N}, \boldsymbol{n})$, but if $\tau(\boldsymbol{N}) \geq \hat{\tau}(\boldsymbol{n})$, then $p(\boldsymbol{N}', \boldsymbol{n}) \leq p(\boldsymbol{N}, \boldsymbol{n})$.

This theorem extends Li and Ding [2016, Lemma A.3]. It shows that if the ATE for one potential outcome table is closer to the unbiased estimate $\hat{\tau}$ than the ATE for another potential outcome table is, then the P-value for the first potential outcome table is larger than that of the second table. PBS: Fix this:

only applies if the tables are related in that particular way, not for any tables whatsoever.

Based on Theorem A.1, we now propose a faster algorithm of the method in Section 4.2. 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:

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

   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 4.2. 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 $n_1 \leq N_1/2$ and $n_2 \leq N_2/2$ 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:

$$\left\{(M_{311}, \ldots, M_{K11}) \middle| 0 \leq M_{k11} \leq N_k - M_{k00} - M_{k01} \; 3 \leq k \leq K\right\},$$

   do the following:

   (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

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

   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

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

   (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.*

n	CI	A.1	4.2.
$[3, 10, 1, 1]$ $[1, 1, 1, 13]$	$[-10, 15]$	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 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.1 and the original algorithm in Section 4.2 A random sample with replacement of 10,000 randomizations was used to approximate permutation tests.

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 4.2. See Table 5 for example.

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. 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, we have:

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 4.2

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:

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

   do the following:

   (a) Find an algebraically 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 algebraically compatible potential outcome table is found, go to the next loop.

n	CI	A.3	A.1	4.2.
$[2,6,1,7]$ $[0,3,5,4]$	$[-12,5]$	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 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.3, Algorithm A.1 and the original algorithm in Section 4.2. A random sample with replacement of 10,000 randomizations was used to approximate permutation tests.

(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 4.2. 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 4.2 when some of the strata are balanced. See Table 6 for example.

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 4.2 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 [Aronow et al., 2023])*
Fix observed table $\boldsymbol{n}$ and a potential table $\boldsymbol{N}$. Let

$$\boldsymbol{d}_A := (+1, -1, -1, +1)^T.$$

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

1. $\boldsymbol{N'} = \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_A^T, k \in [K]$
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 $\boldsymbol{d}_L^k$ be a function of the potential outcome table such that*

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

If an another potential outcome table $\boldsymbol{N'}$ satisfies $\boldsymbol{N'} = \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T, k \in [K]$, 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:
   (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:
   $$\Big\{\boldsymbol{N}\Big|\boldsymbol{N} \text{ is algebraically compatible}, \boldsymbol{T}(\boldsymbol{N}) = \boldsymbol{T}, \\ \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_A^T \text{ is not algebraically compatible for all } k \in [K], \\ \text{for all } k \in [K],\ \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T \text{ is not algebraically compatible or } \boldsymbol{d}_L^k(\boldsymbol{N}) = \boldsymbol{0}\ .\Big\}$$
   (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 $\tau_i \neq 0, \forall i \in [k]$
   $$\Big\{\boldsymbol{N}\Big|\boldsymbol{N} \text{ is algebraically compatible}, \boldsymbol{T}(\boldsymbol{N}) = \boldsymbol{T}, \\ \exists j \in [K], s.t. N_{j10} = N_{j01} = 0, \\ \text{for all } k \in [K],\ \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T \text{ is not algebraically compatible or } \boldsymbol{d}_L^k(\boldsymbol{N}) = \boldsymbol{0}.\Big\} \\ \bigcup \Big\{\boldsymbol{N}\Big|\boldsymbol{N} \text{ is algebraically compatible}, \boldsymbol{T}(\boldsymbol{N}) = \boldsymbol{T}, \\ \forall j \in [k],\ \text{if } \tau_j = 0,\ \text{then } N_{j10} = N_{j01} = 1, \\ \text{if } \tau_j \neq 0,\ \text{then } N + e_j * \boldsymbol{d}_A \text{ is not algebraically compatible}, \\ \text{for all } k \in [K],\ \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T \text{ is not algebraically compatible or } \boldsymbol{d}_L^k(\boldsymbol{N}) = \boldsymbol{0}\ .\Big\}$$
   (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 4.2. 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 4.2. See Table 7 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

n	CI	A.4	A.3	A.1	4.2
$[2,6,8,0]$ $[3,7,1,9]$	$[-10,15]$	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 7: 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 4.2. 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 8: 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.

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 et al., 2023], see Table 8 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.

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 $\boldsymbol{d} = (-1, 1, 0, 0), \boldsymbol{N}' = \boldsymbol{N} + e_1 \cdot \boldsymbol{d}$. 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

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

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

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

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

□

Proof of Theorem A.3. By the representation of $p(\boldsymbol{N}, \boldsymbol{n})$ in (12), 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[Li and Ding, 2016]. We just extended it to the stratified case.*

Lemma B.1. *For any potential outcome table which is algebraically 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 algebraically compatible with the observed table and $\boldsymbol{N}' = \boldsymbol{N} + \boldsymbol{d}_{LD}$ with $\boldsymbol{d}_{LD} \in \Delta_{LD}$. 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 algebraically compatible for the observed table and $\boldsymbol{N}' = \boldsymbol{N} - \boldsymbol{d}_{LD}$ with $\boldsymbol{d}_{LD} \in \Delta_{LD}$. 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 algebraically compatible with 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 9

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)	$\boldsymbol{e}^k$∗(0, 1, 0, -1)
1	(1, 1)	(1, 0)	$\boldsymbol{e}^k$∗(-1, 1, 0, 0)
0	(0, 1)	(1, 1)	$\boldsymbol{e}^k$∗(1, 0,-1, 0)
1	(0, 1)	(0, 0)	$\boldsymbol{e}^k$∗(0, 0, -1, 1)

Table 9: 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. $\square$

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

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:

$$\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N[\hat{\tau}(\boldsymbol{N}, \boldsymbol{Z}) - \tau(\boldsymbol{N})] = 2\sum_{k=1}^{K}(n_{k11} - n_{k01}) - \sum_{k=1}^{K}(N_{k10} - N_{k01}) \quad (15)$$

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 (15) (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* $\boldsymbol{N}_{(k)} = \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_A^T$*. Set*

$$\boldsymbol{N}_{\{k\}} = \boldsymbol{N} + \boldsymbol{e}^k * (0, -1, -1, 0). \tag{16}$$

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

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

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

$$\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right) \geq \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right) \tag{17}$$

and

$$\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) \leq -N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right) \geq \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) \leq -N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right) \tag{18}$$

Here, we will only establish the proof of (17) since the proof of (18) 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, $\forall a \in \mathbb{Z}$,

$$\begin{aligned}\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}', \boldsymbol{Z}) = a\right) = &\, \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a - 2, (Z_{11}, Z_{12}) = (1,0)\right) \\ &+\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a, Z_{11} = Z_{12}\right) \\ &+\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a + 2, (Z_{11}, Z_{12}) = (0,1)\right)\end{aligned} \tag{19}$$

Summing (19), we have

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

Here we used the fact that

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

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 (17) is equivalent to

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

which rearranges to

$$\begin{aligned}&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| - 2 \middle| (Z_{11}, Z_{12}) = (1,0)\right) \mathbb{P}((Z_{11}, Z_{12}) = (1,0)) \\ \geq&\, \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})| \middle| (Z_{11}, Z_{12}) = (0,1)\right) \mathbb{P}((Z_{11,12}) = (0,1))\end{aligned} \tag{20}$$

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 (20) as

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

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 $\boldsymbol{N}_1$:*

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

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

$$\mathbb{P}(X = k) = \sum_{i=-\infty}^{+\infty} \mathbb{P}(X_1 = k - i)\mathbb{P}(X_2 = i) \tag{21}$$

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

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

□

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*

$$\begin{cases} \mathbb{P}(X = k) = \mathbb{P}(X = -k), & \forall k \in \mathbb{N} \\ \mathbb{P}(X = k) \geq \mathbb{P}(X = k+4), & \forall k \in \mathbb{N} \end{cases} \tag{22}$$

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

Proof. The proof of (22) 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

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

$\square$

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*

$$\begin{cases} \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = k) = \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = -k), & \forall k \in \mathbb{N} \\ \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = k) \geq \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = k+4), & \forall k \in \mathbb{N} \\ \textit{Either } \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = 1) = 0 \textit{ or } \mathbb{P}(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = 0) = 0 \end{cases} \tag{23}$$

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 $\square$

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 (15), we have

$$\tilde{\tau}_k(\boldsymbol{N}, \boldsymbol{Z}) = [x_{k11} - (N_{k11} - x_{k11})] - [x_{k00} - (N_{k00} - x_{k00})]$$

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

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. $\square$

Corollary B.1. *Consider two potential outcome tables,*

$$\boldsymbol{N}_1 := ((\boldsymbol{N}_1)_1, (\boldsymbol{N}_1)_2, \ldots, (\boldsymbol{N}_1)_K)^T$$

and

$$\boldsymbol{N}_2 := ((\boldsymbol{N}_2)_1, (\boldsymbol{N}_2)_2, \ldots, (\boldsymbol{N}_1)_K)^T.$$

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. $\square$

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*

$$\boldsymbol{y}_k = \left(\underbrace{(1,1), (0,0), (1,1), (0,0), \ldots, (1,1), (0,0)}_{2N_{k00}}, \ldots \right)^T \tag{24}$$

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*,*

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

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

In this chapter, unless specified, all the potential outcome vectors are in the order (24)

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

$$\boldsymbol{N}_{[k]} = \boldsymbol{N} + \boldsymbol{e}^k * (-1, 0, 0, 1) \tag{25}$$

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 $\boldsymbol{N}^{\circ}$, $\boldsymbol{N} + \boldsymbol{d}_L^1(\boldsymbol{N})$ 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

$$\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right) \leq \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}^{\circ\circ},\boldsymbol{Z}) \geq N|\hat{\tau}(\boldsymbol{n}) - \tau(\boldsymbol{N})|\right), \tag{26}$$

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 (26) 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, $\forall a \in \mathbb{Z}$,

$$\begin{aligned} \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})\right) =& \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}^{\circ},\boldsymbol{Z}) = a-2, Z_{12} = 1\right) \\ &+ \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}^{\circ},\boldsymbol{Z}) = a+2, Z_{12} = 0\right) \end{aligned} \tag{27}$$

Summing (27), we get

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

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,

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

Thus (26) is equivalent to

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

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*

$$
\boldsymbol{N}(j) = \boldsymbol{N} + \boldsymbol{e}^k * (j, 0, 0, j) \tag{28}
$$

If all of the following three statement hold true:

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

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,

$$
\begin{aligned}
&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0\right)\\
=&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=0, Z_{1,2s+2}=0\right)\\
+&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=1, Z_{1,2s+2}=1\right)\\
+&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=0, Z_{1,2s+2}=1\right)\\
+&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0, Z_{1,2s+1}=1, Z_{1,2s+2}=0\right)\\
=&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s+2}=0\right)\\
+&\mathbb{P}\left(Z_{1,2s-1}=Z_{1,2s}=0, Z_{1,2s+1}=Z_{1,2s+2}=1\right)\\
&*\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s-2}=0|Z_{1,2s-1}=Z_{1,2s}=0, Z_{1,2s+1}=Z_{1,2s+2}=1\right)\\
+&\mathbb{P}\left(Z_{1,2s+1}=0, Z_{1,2s+2}=1\right)\\
&*\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0|Z_{1,2s+1}=0, Z_{1,2s+2}=1\right)\\
+&\mathbb{P}\left(Z_{1,2s+1}=1, Z_{1,2s+2}=0\right)\\
&*\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0|Z_{1,2s+1}=1, Z_{1,2s+2}=0\right)\\
=&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s+2}=0\right)\\
+&\mathbb{P}\left(Z_{1,2s-1}=Z_{1,2s}=0, Z_{1,2s+1}=Z_{1,2s+2}=1\right)\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}(-2),\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s-2}=0\right)\\
+&\mathbb{P}\left(Z_{1,2s+1}=0, Z_{1,2s+2}=1\right)\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a+2, Z_{11}=\cdots=Z_{1,2s}=0\right)\\
+&\mathbb{P}\left(Z_{1,2s+1}=1, Z_{1,2s+2}=0\right)\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a-2, Z_{11}=\cdots=Z_{1,2s}=0\right).
\end{aligned}
\tag{29}
$$

Similarly, we have

$$
\begin{aligned}
&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=1\right)\\
=&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s+2}=1\right)\\
+&\mathbb{P}\left(Z_{1,2s-1}=Z_{1,2s}=1, Z_{1,2s+1}=Z_{1,2s+2}=0\right)\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}(-2),\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s-2}=1\right)\\
+&\mathbb{P}\left(Z_{1,2s+1}=0, Z_{1,2s+2}=1\right)\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a+2, Z_{11}=\cdots=Z_{1,2s}=1\right)\\
+&\mathbb{P}\left(Z_{1,2s+1}=1, Z_{1,2s+2}=0\right)\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}(-1),\boldsymbol{Z})=a-2, Z_{11}=\cdots=Z_{1,2s}=1\right).
\end{aligned}
\tag{30}
$$

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 (29), (30) 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 $2s \geq n_k - 1$,

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

□

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*

$$
\begin{aligned}
&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=0\right)\\
\geq&\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N},\boldsymbol{Z})=a, Z_{11}=\cdots=Z_{1,2s}=1\right)
\end{aligned}
\tag{31}
$$

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

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

thus we only need to prove that

$$\mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = a \middle| Z_1 = \cdots = Z_{2s} = 0\right) \geq \mathbb{P}\left(\tilde{\tau}(\boldsymbol{N}, \boldsymbol{Z}) = -a \middle| Z_1 = \cdots = Z_{2s} = 0\right) \tag{32}$$

We assume the right hand side of (32) 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, (32) can be rearranged to

$$\begin{aligned} &\binom{g}{N_{11} - N_{00}} \binom{n-g}{N_{01} + N_{10}} \Big/ \binom{n}{N - 2N_{00}} \\ \geq &\binom{g'}{N_{11} - N_{00}} \binom{n-g'}{N_{01} + N_{10}} \Big/ \binom{n}{N - 2N_{00}} \end{aligned} \tag{33}$$

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 (33) 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 $0 \leq 2N_{00} \leq n - (a - N_{11} - N_{00})/2 = n - g \leq N_{01} + N_{10}$. Thus the left hand side of (33) make sense and is not 0. Now we rewrite (33), then it becomes

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

which rearranges to

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

Then, we only need to prove that $n - g' \geq N_{01} + N_{10} - n + g$, which is equivalent to $N_{11} + N_{00} \geq N_{11} - N_{00}$. However, the last inequality holds naturally, thus (33) is true, and the proof is complete. $\square$

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 $s_k := \min(N_{k11}, N_{k00}) \leq n_k/2$, then $\boldsymbol{N}$ satisfies k, s_k-regular condition.*

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

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,

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

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

$$\begin{aligned} &\sum_{k=-\infty}^{\infty} \mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\right) \mathbb{P}\left(\tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z}) = k, Z_{11} = \cdots = Z_{1,2s_1} = 0\right) \\ \geq &\sum_{k=-\infty}^{\infty} \mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\right) \mathbb{P}\left(\tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z}) = k, Z_{11} = \cdots = Z_{1,2s_1} = 1\right). \end{aligned} \tag{34}$$

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 (34) is equivalent to

$$\begin{aligned} &\sum_{k=-\infty}^{\infty} \mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\right) \mathbb{P}\left(\tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z}) = k, Z_{11} = \cdots = Z_{1,2s_1} = 0\right) \\ \geq &\sum_{k=-\infty}^{\infty} \mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\right) \mathbb{P}\left(\tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z}) = -k, Z_{11} = \cdots = Z_{1,2s_1} = 0\right). \end{aligned} \tag{35}$$

Denote $\tilde{p}(j) = \mathbb{P}\left(\tilde{\tau}_1(\boldsymbol{N}, \boldsymbol{Z}) = j, Z_{11} = \cdots = Z_{1,2s_1} = 0\right)$ for convenience. By rearranging (35), we only need to prove that

$$\begin{aligned} &\sum_{k=-\infty}^{\infty} \mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\right) \left(\tilde{p}(k) - \tilde{p}(-k)\right) \geq 0 \\ \iff &\sum_{k=1}^{\infty} \left[\mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a - k\right) - \mathbb{P}\left(\tilde{\tau}_{-1}(\boldsymbol{N}, \boldsymbol{Z}) = a + k\right)\right] \left(\tilde{p}(k) - \tilde{p}(-k)\right) \geq 0 \end{aligned}$$

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 (23); 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

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

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 $a-k\equiv a+k-2(\bmod 4)$ for odd k and $a-k-2\equiv a+k(\bmod 4)$ for odd k, then the proof is done by (23); 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 conditions for algebraic compatibility of potential outcome tables with observed outcome tables. 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 algebraically 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 algebraically 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 algebraically compatible with 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 $p(\boldsymbol{N}, \boldsymbol{n}) < p(\boldsymbol{N}', \boldsymbol{n}) < \alpha$. 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 algebraically compatible by checking whether a is within the range bounded by the lower and upper limits of $\sum_{k=1}^{l}(N_{k10} - N_{k01})$. To identify an algebraically 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 an algebraically 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 an algebraically 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:

$$\exists k, \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_A^T \text{ is algebraically compatible, or } \exists k, \boldsymbol{d}_L^k(\boldsymbol{N}) \neq 0, \boldsymbol{N} + \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T \text{ is algebraically compatible.}$$

For each k, Let L^k be the largest integer such that $\boldsymbol{N} + L^k * \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T$ is algebraically compatible, A^k be the largest integer such that $\boldsymbol{N} + A^k * \boldsymbol{e}^k \boldsymbol{d}_A^T$ is algebraically compatible, then

$$\boldsymbol{N'} := \boldsymbol{N} + \sum_{k=1}^{K} L^k * \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T + \sum_{k=1}^{K} A^k * \boldsymbol{e}^k \boldsymbol{d}_A^T$$

is an algebraically 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 $e_i * \boldsymbol{d}_L^i(\boldsymbol{N})$ 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 $\boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N})^T$ and $\boldsymbol{e}^k \boldsymbol{d}_A^T$ to $\boldsymbol{N}$, we can create a potential outcome table $\boldsymbol{N'}$ such that

1. $\boldsymbol{N'}$ is algebraically compatible,
2. $\forall k \in [K]$, either $\boldsymbol{N'} + \boldsymbol{e}^k \boldsymbol{d}_A^T$ is not algebraically compatible or $N_{k10} = N_{k01} = 1$.
3. $\forall k \in [K]$, either $\boldsymbol{N'} + \boldsymbol{e}^k \boldsymbol{d}_L^k(\boldsymbol{N'})^T$ is not algebraically compatible or $\boldsymbol{d}_L^k(\boldsymbol{N'}) = \boldsymbol{0}$

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 algebraically 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 4.3, 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 Stouffer'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 10.

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 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 C, 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

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

Table 10: 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.

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].

References

P. M. Aronow, H. Chang, and P. 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`.

Z. Branson and MA. Bind. Randomization-based inference for bernoulli trial experiments and implications for observational studies. *Statistical Methods in Medical Research*, 28(5):1378–1398, 2019. doi: 10.1177/0962280218756689. URL `https://doi.org/10.1177/0962280218756689`. PMID: 29451089.

C. Bruce, E. Juszczak, R. Ogollah, C. Partlett, and A. 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.

Y. 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.

Y. 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`.

M. Dwass. Modified randomization tests for nonparametric hypotheses. *Ann. Math. Statist.*, 28:181–187, 1957. doi: 10.1214/aoms/1177707045.

S. N. Evans and P. B. Stark. Inverse problems as Statistics. *Inverse Problems*, 18:R55–R97, 2002. doi: 10.1088/0266-5611/18/4/201. URL `https://dx.doi.org/10.1088/0266-5611/18/4/201`.

S.E. Fienberg and J.M. Tanur. Reconsidering the fundamental contributions of fisher and neyman on experimentation and sampling. *International Statistical Review*, 64:237–253, 1996. URL `https://api.semanticscholar.org/CorpusID:32108972`.

R.A. Fisher. *The Design of Experiments*. Oliver & Boyd, Edinburgh, 1935.

X. Li and P. 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`.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. *Ann. Agricultural Sciences*, pages 1–51, 1923.

A. Ramdas, R. F. Barber, E.J. Candès, and R.J. Tibshirani. Permutation tests using arbitrary permutation distributions. *Sankhya A: The Indian Journal of Statistics*, 85:1156–1177, 2023. doi: 10.1007/s13171-023-00308-8.

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`.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688–701, 1974.

D.B. Rubin. Comment on randomization analysis of experimental data: The Fisher randomization test by b. basu. *Journal of the American Statistical Association*, 91:267, 1980.

T.J. Santner and M.K. Snell. Small-sample confidence intervals for p1 - p2 and p1/p2 in 2×2 contingency tables. *Journal of the American Statistical Association*, 75(370):386–394, 1980. ISSN 01621459. URL `http://www.jstor.org/stable/2287464`.

S. Travis, M. S. Silverberg, S. Danese, P. Gionchetti, M. Löwenberg, V. Jairath, B.G. Feagan, B. Bressler, M. Ferrante, A. Hart, D. Lindner, A. Escher, S. Jones, and B. 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.

J.P. Wendell and J. 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`.

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.