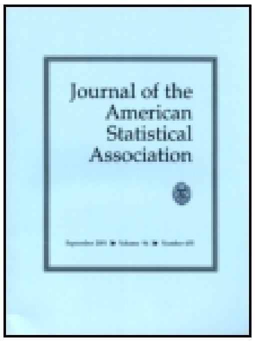

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

General forms of finite population central limit theorems with applications to causal inference

Xinran Li & Peng Ding

To cite this article: Xinran Li & Peng Ding (2017): General forms of finite population central limit theorems with applications to causal inference, Journal of the American Statistical Association, DOI: 10.1080/01621459.2017.1295865

To link to this article: http://dx.doi.org/10.1080/01621459.2017.1295865

General forms of finite population central limit theorems with applications to causal inference

Xinran Li and Peng Ding [*]

Abstract

Frequentists' inference often delivers point estimators associated with confidence intervals or sets for parameters of interest. Constructing the confidence intervals or sets requires understanding the sampling distributions of the point estimators, which, in many but not all cases, are related to asymptotic Normal distributions ensured by central limit theorems. Although previous literature has established various forms of central limit theorems for statistical inference in super population models, we still need general and convenient forms of central limit theorems for some randomization-based causal analysis of experimental data, where the parameters of interests are functions of a finite population and randomness comes solely from the treatment assignment. We use central limit theorems for sample surveys and rank statistics to establish general forms of the finite population central limit theorems that are particularly useful for proving asymptotic distributions of randomization tests under the sharp null hypothesis of zero individual causal effects, and for obtaining the asymptotic repeated sampling distributions of the causal effect estimators. The new central limit theorems hold for general experimental designs with multiple treatment levels, multiple treatment factors and vector outcomes, and are immediately applicable for studying the asymptotic properties of many methods in causal inference, including instrumental variable, regression adjustment, rerandomization, clustered randomized experiments, and so on. Previously, the asymptotic properties of these problems are often based on heuristic arguments, which in fact rely on general forms of finite population central limit

[*] Xinran Li is Doctoral Candidate, Department of Statistics, Harvard University, Cambridge, MA 02138 (E-mail: xinranli@fas.harvard.edu). Peng Ding is Assistant Professor, Department of Statistics, University of California, Berkeley, CA 94720 (E-mail: pengdingpku@berkeley.edu). The authors thank Mr. Yotam Shem-Tov and Ms. Lo-Hua Yuan for comments, and Professor Peter J. Bickel for the references of finite population inference. The comments from the Editor and three reviewers have greatly improved the quality of the paper.

theorems that have not been established before. Our new theorems fill in this gap by providing more solid theoretical foundation for asymptotic randomization-based causal inference.

Key Words: Conservative confidence set; Fisher randomization test; Potential outcome; Randomization inference; Repeated sampling property; Sharp null hypothesis

1 Introduction

Central limit theorems (CLTs) are central pillars of many frequentists' inferential procedures. Most CLTs assume that the observations are samples from a hypothetical infinite super population model (e.g. Lehmann 1999; Van der Vaart 2000). In sample surveys and randomized experiments, however, the infinite super population seems contrived, and the parameters of interests are functions of the attributes of well-defined finite units. Finite population inference requires no assumptions on the data generating process of the units, and quantifies the uncertainty based on the randomness from the study design. In sample surveys, the population has fixed quantities of interest, and the sampling process induces randomness in the estimators (cf. Cochran 1977); in randomized experiments, the potential outcomes (Neyman 1923; Rubin 1974) of the experimental units are fixed, and the physical randomization acts as the "reasoned basis" (Fisher 1935) for conducting statistical testing and estimation (cf. Kempthorne 1952; Hinkelmann and Kempthorne 2008; Rosenbaum 2002b; Abadie et al. 2014; Imbens and Rubin 2015). This is sometimes called randomization-based or design-based inference, dating back to the classical analysis of sample surveys (Splawa-Neyman 1925; Neyman 1934) and randomized experiments (Neyman 1923, 1935; Fisher 1935).

For simple random sampling, Erdös and Rényi (1959), Hájek (1960) and Madow (1948) obtained various forms of CLTs, with a convenient form presented in Lehmann (1975, Appendix 4, Theorem 6) and Lehmann (1999, Theorem 2.8.2). In fact, these theorems are special cases of the CLTs for rank statistics (Wald and Wolfowitz 1944; Noether 1949; Fraser 1956; Hájek 1961). In randomized experiments, because the treatment and control groups are simple random samples from the finite experimental units, the CLTs for sampling surveys are sometimes adequate for establishing asymptotic distributions of the causal effect estimators (e.g. Liu and Hudgens 2014; Ding 2016; Ding and Dasgupta 2016). Unfortunately, however, these CLTs do not immediately apply to estimators beyond the difference-in-means in treatment-control experiments. For instance, Freedman (2008a,b) provided only an informal proof for the asymptotic Normality of the regression estimator in randomized experiments based on Hoeffding (1951) and Höglund (1978). Many other randomization-based causal inferences invoked CLTs implicitly without a formal proof, e.g., rerandomization in Morgan and Rubin (2012), factorial experiments in Dasgupta et al. (2015) and Ding (2016), and clustered randomized experiments in Middleton and Aronow (2015).

Therefore, causal inference needs general forms of CLTs that apply to more than two treatment levels, more complex designs than completely randomized experiments, and more complex estimators than difference-in-means. We first recall a deep connection between sample surveys and randomized experiments (cf. Neyman 1923; Splawa-Neyman 1925; Neyman 1934, 1935; Rubin 1990; Fienberg and Tanur 1996), and then utilize a CLT for rank statistics (Fraser 1956) to establish the CLTs that are particularly useful for causal analysis of randomized experiments. The salient feature of the new CLTs is that the asymptotic variances and covariances depend on the correlation structure among the potential outcomes under different treatment levels. This feature did not appear in any CLTs for sample surveys or rank statistics, but did appear in the variance formula of the difference-in-means estimator in Neyman (1923). Because of the generality of the new CLTs, they are readily applicable to many existing causal inference problems, including instrumental variable estimation, randomization tests with more than two treatment levels, multiple randomization tests, rerandomization to ensure covariate balance (Morgan and Rubin 2012, 2015; Li, Ding, and Rubin 2016), regression adjustment for completely randomized experiments (Freedman 2008a,b; Lin 2013), clustered randomized experiments (Middleton and Aronow 2015), and unbalanced factorial experiments (Dasgupta et al. 2015), etc. The new CLTs not only justify the asymptotic properties of some existing procedures, but also help to establish new results that did not appear in the previous literature. They will become useful tools for studying asymptotic properties of many randomization-based inferential procedures in causal inference.

Under the sharp null hypothesis with zero or general known unit-level causal effects, all the potential outcomes are known, and the randomization distribution of any test statistics can be computed exactly or at least simulated by Monte Carlo. In this case, the role of the CLTs is to give convenient approximations of the null distributions and provide statistical insights with explicit formulas. More importantly, without imposing the sharp null hypothesis as in the repeated sampling evaluations (Neyman 1923), the randomization distributions of the causal effect estimators depend on unknown values of the potential outcomes. In this case, the role of the CLTs is then not only to give convenient approximations but also allow for asymptotic statistical inference without knowing all the values of the potential outcomes. As shown in Neyman (1923), this type of inference is often statistically conservative even asymptotically, which will be clearer with our general finite population CLTs in Section 3.

Below we first review some classical finite population CLTs for simple random sampling (Hájek 1960) and for random partition (cf. Lehmann 1975, Appendix 8, Theorem 19), and then establish new finite population CLTs that apply to general randomized experiments with multiple treatment levels. Throughout the paper, we use important causal inference problems to illustrate the CLTs. All technical details are in the Supplementary Material.

2 Classical finite population CLTs

2.1 Simple random sampling

Consider a finite population $\Pi_N = \{y_{N1}, y_{N2}, \ldots, y_{NN}\}$ with N units. The population mean, $\bar{y}_N = (y_{N1} + \cdots + y_{NN})/N$, is often of interest. A sample is a subset of Π_N represented by the vector of inclusion indicators $(Z_1, \ldots, Z_N) \in \{0, 1\}^N$, where $Z_i = 1$ if the sample contains unit i and $Z_i = 0$ otherwise. In simple random sampling, the probability that the inclusion indicator vector $(Z_1, \ldots, Z_N)$ takes a particular value $(z_1, \ldots, z_N)$ is $n!(N-n)!/N!$, where $\sum_{i=1}^{N} z_i = n$ and $\sum_{i=1}^{N}(1 -$

$z_i) = N - n$. The sample average $\bar{y}_{\mathrm{S}} = \sum_{i=1}^{N} Z_i y_{Ni}/n$ is an intuitive estimator for the population mean. In the formula of $\bar{y}_{\mathrm{S}}$, randomness comes from $(Z_1, \ldots, Z_N)$, and all the y_{Ni}'s are fixed population quantities. Because of this feature, it is straightforward to show that $\bar{y}_{\mathrm{S}}$ has mean $\bar{y}_N$ and variance

$$\mathrm{Var}(\bar{y}_{\mathrm{S}}) = \left(\frac{1}{n} - \frac{1}{N}\right) v_N, \tag{1}$$

depending on the finite population variance of Π_N (cf. Cochran 1977):

$$v_N = \frac{1}{N-1} \sum_{i=1}^{N} (y_{Ni} - \bar{y}_N)^2 . \tag{2}$$

To conduct statistical inference of $\bar{y}_N$ based on $\bar{y}_{\mathrm{S}}$, we need to characterize the sampling distribution of $\bar{y}_{\mathrm{S}}$ induced by simple random sampling. Although the exact distribution of $\bar{y}_{\mathrm{S}}$ is complex, some asymptotic techniques help to use its first two moments to describe its asymptotic distribution. The finite population asymptotic scheme embeds Π_N into a hypothetical infinite sequence of finite populations with increasing sizes, and the asymptotic distribution of any sample quantity is its limiting distribution along this hypothetical infinite sequence (cf. Lehmann 1999, 1975; Ding 2016; Aronow et al. 2014; Middleton and Aronow 2015). Similar to the classical Lindeberg–Feller CLT (Durrett 2010), the asymptotic behavior of $\bar{y}_{\mathrm{S}}$ depends crucially on the maximum squared distance of the y_{Ni}'s from the population mean $\bar{y}_N$:

$$m_N = \max_{1 \le i \le N} (y_{Ni} - \bar{y}_N)^2. \tag{3}$$

The following theorem due to Hájek (1960) states that, under some regularity conditions on the sequence of finite populations $\{\Pi_N\}_{N=1}^{\infty}$ and sizes of simple random samples, the sample average is asymptotically Normal.

Theorem 1. Let $\bar{y}_{\mathrm{S}}$ be the average of a simple random sample of size n from a finite population

$\Pi_N = \{y_{N1}, y_{N2}, \ldots, y_{NN}\}$. As $N \to \infty$, if

$$\frac{1}{\min(n, N-n)} \cdot \frac{m_N}{v_N} \to 0, \tag{4}$$

then $(\bar{y}_S - \bar{y}_N)/\sqrt{\mathrm{Var}(\bar{y}_S)} \xrightarrow{d} \mathcal{N}(0,1)$, recalling that v_N and m_N are defined in (2) and (3).

Lehmann (1975, Appendix 4) presented a special case of a theorem in Hájek (1961), requiring an equivalent form of Condition (4) and additionally $n \to \infty$ and $N - n \to \infty$. For the ease of notation and interpretation, we present Condition (4) in the main text, and give more technical details in the Supplementary Material. In fact, we can show that $m_N/v_N \geq 1 - N^{-1}$, and therefore Condition (4) implies $n \to \infty$ and $N - n \to \infty$. Because a weighted sum of the means of the sampled units and unsampled units is fixed at the population mean, their asymptotic behaviors must be exactly the same up to some scaling factors. This further explains the symmetry of n and $N - n$ in Condition (4).

Condition (4) needs further explanation. Importantly, it does not depend on the scale of the y_{Ni}'s of the finite population. We can simply standardize the y_{Ni}'s to ensure a constant finite population variance (e.g., $v_N = 1$ for all N), and the values of m_N for these finite populations change correspondingly. We further assume that the proportion of sampled units has a limiting value $n/N \to \gamma_\infty \in [0,1]$. The value of γ_∞ is usually positive in randomized experiments; we assign a proportion of units to the treatment group and a comparable proportion of units to the control group, and both groups are simple random samples from the finite units. The value of γ_∞ may be zero in survey sampling when the sample fraction is extremely small. When $0 < \gamma_\infty < 1$, Condition (4) is equivalent to $m_N/N \to 0$; when $\gamma_\infty = 0$, it is equivalent to $m_N/n \to 0$; when $\gamma_\infty = 1$, it is equivalent to $m_N/(N-n) \to \infty$. It is apparent that when all units of Π_N have bounded values, all equivalent forms must hold if both n and $N - n$ go to infinity. Moreover, if $0 < \gamma_\infty < 1$ and the units in Π_N are independent and identically distributed (i.i.d) draws from a super population with more than two moments and a nonzero variance, then v_N converges to the variance of the super population v_∞, and Condition (4) and its equivalent form $m_N/N \to 0$ hold with probability one, as commented in the Supplementary Material. In this case, the asymptotic Normality reduces to $\sqrt{n}(\bar{y}_{\mathrm{S}} - \bar{y}_N) \xrightarrow{d} \mathcal{N}(0, (1-\gamma_\infty)v_\infty)$.

Separately in the literature, Erdös and Rényi (1959) and Hájek (1960) established finite population CLTs for simple random sampling, and Wald and Wolfowitz (1944), Noether (1949), Hoeffding (1951), Motoo (1956), and Schneller (1988) established various forms of CLTs for rank statistics. Madow (1948) used a CLT for rank statistics to prove a version of finite population CLT for simple random sampling. Holst (1979) proved the finite population CLT by first considering the Bernoulli sampling with i.i.d. inclusion indicators and then conditioning on the sum of these indicators. Hájek (1960, 1961) and Robinson (1972) discussed different forms of sufficient and necessary conditions.

Theorem 1 suggests a strategy to construct a large-sample confidence interval for $\bar{y}_N$ based on the Normal approximation. This strategy requires us to consistently estimate the variance of $\bar{y}_{\mathrm{S}}$. The sample variance of the simple random sample

$$\widehat{v}_N = \frac{1}{n-1} \sum_{i:Z_i=1} (y_{Ni} - \bar{y}_{\mathrm{S}})^2$$

is unbiased for the population variance v_N, and therefore

$$\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{S}}) = \left(\frac{1}{n} - \frac{1}{N}\right)\widehat{v}_N$$

is unbiased for the variance of $\bar{y}_{\mathrm{S}}$.

Proposition 1. Under the conditions in Theorem 1, $\widehat{\mathrm{Var}}(\bar{y}_{\mathrm{S}})/\mathrm{Var}(\bar{y}_{\mathrm{S}}) = \widehat{v}_N/v_N \xrightarrow{p} 1$ as $N \to \infty$.

Therefore, Theorem 1 and Proposition 1 justify the usual confidence interval for $\bar{y}_N$ based on the Normal approximation. This confidence interval is standard in the survey sampling literature (eg. Cochran 1977), but to the best of our knowledge, the proof of the simple fact $\widehat{v}_N/v_N \xrightarrow{p} 1$ was neglected or derived under unnecessarily strong conditions in the literature (e.g., Lehmann 1999, page 259). In the randomization-based causal inference, Proposition 1 is crucial for consistency of the variance estimators, but previous literature provided only heuristic arguments without formal proofs (e.g., Liu and Hudgens 2014; Ding 2016).

Theorem 1 and Proposition 1 have numerous applications. We review two examples below. For more applications in nonparametric tests and randomization-based causal inference, please see Lehmann (1975), Liu and Hudgens (2014), Ding (2016), and Ding and Dasgupta (2016).

Example 1 (Normal approximation of the Hypergeometric distribution)**.** If all the units in the finite population Π_N take binary values with N_1 of them being 1, then $n\bar{y}_S$, the number of 1's in a simple random sample of size n, follows a Hypergeometric distribution. We verify in the Supplementary Material that, if $\text{Var}(n\bar{y}_S) \to \infty$ then Condition (4) holds. Therefore, $n\bar{y}_S$ is asymptotically Normal if its variance goes to infinity (Lehmann 1975; Vatutin and Mikhailov 1982). In fact, this is a sufficient and necessary condition (Kou and Ying 1996). Both Fisher's exact test and the randomization test for a binary outcome have null distributions depending on a Hypergeometric random variable (cf. Ding and Dasgupta 2016), and therefore can be efficiently computed using Normal approximations with large samples. Hannan and Harkness (1963) and Harkness (1965) discussed Normal approximation of the extended Hypergeometric distribution, and their regularity condition for the Hypergeometric distribution reduces to $\text{Var}(n\bar{y}_S) \to \infty$. □

Example 2 (Randomization-based instrumental variable estimation)**.** Consider a completely randomized experiment with N units, in which n_1 assigned to the treatment and n_0 assigned to the control. For unit i, let Z_i be the binary indicator for treatment assignment, D_i be the binary or continuous received dose of the treatment, and Y_i be the response. Because both the dose and response are affected by the treatment, we define $(D_i(1), D_i(0))$ as the potential outcomes for the dose, and $(Y_i(1), Y_i(0))$ as the potential outcomes for the response. Under the linear instrumental variable model, $Y_i(1) - Y_i(0) = \beta\{D_i(1) - D_i(0)\}$ for all unit i, where the coefficient β is a measure of the dose-response relationship (Rosenbaum 2002b; Imbens and Rosenbaum 2005). The model automatically satisfies the so-called exclusion restriction assumption, because $D_i(1) = D_i(0)$

implies $Y_i(1) = Y_i(0)$. Define the adjusted outcome as $A_i \equiv Y_i - \beta D_i$ with potential outcomes $A_i(z) = Y_i(z) - \beta D_i(z)$ under treatment z. Because $Y_i(1) - \beta D_i(1) = Y_i(0) - \beta D_i(0)$, the adjusted outcome satisfies $A_i = A_i(1) = A_i(0)$, i.e., the treatment does not affect the observed value of A_i at the true value of β. Therefore, we can construct a confidence interval for β by inverting randomization tests. Although general test statistics such as rank statistics are useful in practice (Rosenbaum 2002b; Imbens and Rosenbaum 2005), we use the difference-in-means of A as the test statistic for simplicity. Let $(\widehat{\tau}_A, \widehat{\tau}_Y, \widehat{\tau}_D)$ be the difference-in-means between treatment and control groups, $(\bar{A}, \bar{Y}, \bar{D})$ the pooled means, and (s_A^2, s_Y^2, s_D^2) the finite population variances of the pooled observed values of A, Y and D. Let s_{YD} be the finite population covariance between the pooled observed values of Y and D. Then the test statistic has the following equivalent forms:

$$\widehat{\tau}_A = \frac{1}{n_1}\sum_{i=1}^{N} Z_i(Y_i - \beta D_i) - \frac{1}{n_0}\sum_{i=1}^{N}(1 - Z_i)(Y_i - \beta D_i) = \widehat{\tau}_Y - \beta\widehat{\tau}_D = \frac{N}{n_0}\left(\frac{1}{n_1}\sum_{i=1}^{N} Z_i A_i - \bar{A}\right). \quad (5)$$

Under the null hypothesis that β is the true value, if $\{A_i : i = 1, \dots, N\}$ satisfy Condition (4) in Theorem 1, then according to the last equivalent form of the test statistic $\widehat{\tau}_A$ in (5), it converges to a Normal distribution with mean 0 and variance

$$\mathrm{Var}_0(\widehat{\tau}_A) = \frac{N^2}{n_0^2}\left(\frac{1}{n_1} - \frac{1}{N}\right) s_A^2 = \frac{N}{n_1 n_0} s_A^2 = \frac{N}{n_1 n_0}(s_Y^2 + \beta^2 s_D^2 - 2\beta s_{YD}).$$

Moreover, as commented in the Supplementary Material, Condition (4) for $\{A_i : i = 1, \dots, N\}$ holds for any β if

$$\frac{1}{\min(n, N-n)}\left\{\frac{\max_{1\le i\le N}(Y_i - \bar{Y})^2}{s_Y^2 - s_{YD}^2/s_D^2} + \frac{\max_{1\le i\le N}(D_i - \bar{D})^2}{s_D^2 - s_{YD}^2/s_Y^2}\right\} \to 0. \quad (6)$$

Let $\Phi(\cdot)$ be the cumulative distribution function of the standard Normal random variable. Based on the Normal approximation, the $1-\alpha$ confidence interval for β is the values satisfying $|\widehat{\tau}_A / \sqrt{\mathrm{Var}_0(\widehat{\tau}_A)}| \le$

$|\Phi^{-1}(\alpha/2)|$, or equivalently the solution of the following inequality:

$$(\widehat{\tau}_Y - \beta\widehat{\tau}_D)^2 \leq \{\Phi^{-1}(\alpha/2)\}^2 \times \frac{N}{n_1 n_0}(s_Y^2 + \beta^2 s_D^2 - 2\beta s_{YD}). \tag{7}$$

Note that for different observed data, the solution of the quadratic inequality in (7) can be an interval, an empty set, or two disjoint sets, a phenomenon that also occured in the classical Fieller–Creasy problem (Fieller 1954; Creasy 1954). Let $\eta = N/(n_1 n_0) \cdot \{\Phi^{-1}(\alpha/2)\}^2$ and $\Delta = 4(\widehat{\tau}_D\widehat{\tau}_Y - \eta s_{YD})^2 - 4(\widehat{\tau}_D^2 - \eta s_D^2)(\widehat{\tau}_Y^2 - \eta s_Y^2)$. Table 1 shows four possible forms of the confidence sets for β. We give more detailed discussion in the Supplementary Material. As pointed out by Rosenbaum (2002b, pp 185), infinite confidence sets suggest little or no identification due to weak instruments, and empty confidence sets suggest possible violations of the linear instrumental variable model. □

2.2 Random partition

Theorem 1 is useful for deriving asymptotic distributions in survey sampling and treatment-control experiments. However, it is not adequate for treatments with more than two levels. In this section, we present a CLT for random partition as an extension of Lehmann (1975, Theorem 19).

Again let $\Pi_N = \{y_{N1}, \ldots, y_{NN}\}$ be a finite population. Similar to Section 2.1, we define $\bar{y}_N$, v_N, and m_N as the finite population mean, variance, and the maximum squared distance of the y_{Ni}'s from the population mean. We consider a random partition of Π_N: units are partitioned into Q groups of size $(n_1, \ldots, n_Q)$, where $\sum_{q=1}^{Q} n_q = N$. Let L_i be the group number, where $L_i = q$ if unit i belongs to group q. The group number vector is $(L_1, \ldots, L_N)$, and the probability that $(L_1, \ldots, L_N)$ takes a particular value $(l_1, \ldots, l_N)$ is $n_1! \cdots n_Q!/N!$, where $\sum_{i=1}^{N} 1\{l_i = q\} = n_q$ for all q. For any $1 \leq q \leq Q$, the sample average in group q is $\bar{y}_{Sq} = \sum_{i:L_i=q} y_{Ni}/n_q$. Because group q is a simple random sample with size n_q, the sample average has mean $\bar{y}_N$ and variance $(n_q^{-1} - N^{-1})v_N$. Instead of focusing on only one sample average as in Theorem 1, we consider the joint distribution of Q

standardized sample averages

$$\boldsymbol{t}_N = \left(\frac{\bar{y}_{S1} - E(\bar{y}_{S1})}{\sqrt{\mathrm{Var}(\bar{y}_{S1})}}, \ldots, \frac{\bar{y}_{SQ} - E(\bar{y}_{SQ})}{\sqrt{\mathrm{Var}(\bar{y}_{SQ})}} \right)^{\top}. \tag{8}$$

Proposition 2. $\boldsymbol{t}_N$ has mean zero and covariance matrix

$$\mathrm{Cov}(\boldsymbol{t}_N) = \begin{pmatrix} 1 & -\sqrt{\frac{n_1 n_2}{(N-n_1)(N-n_2)}} & \cdots & -\sqrt{\frac{n_1 n_Q}{(N-n_1)(N-n_Q)}} \\ -\sqrt{\frac{n_2 n_1}{(N-n_2)(N-n_1)}} & 1 & \cdots & -\sqrt{\frac{n_2 n_Q}{(N-n_2)(N-n_Q)}} \\ \vdots & \vdots & \ddots & \vdots \\ -\sqrt{\frac{n_Q n_1}{(N-n_Q)(N-n_1)}} & -\sqrt{\frac{n_Q n_2}{(N-n_Q)(N-n_2)}} & \cdots & 1 \end{pmatrix}. \tag{9}$$

Proposition 2 appeared in Lehmann (1975, page 393). Furthermore, $\boldsymbol{t}_N$ is asymptotically Normal under the regularity condition below.

Theorem 2. Let $(\bar{y}_{S1}, \ldots, \bar{y}_{SQ})$ be the Q sample averages of a random partition of sizes $(n_1, \ldots, n_Q)$ for a finite population $\Pi_N = \{y_{N1}, \ldots, y_{NN}\}$. As $N \to \infty$, if (i) $\mathrm{Cov}(\boldsymbol{t}_N)$ in (9) has a limiting value $\boldsymbol{V} \in \mathbb{R}^{Q \times Q}$, and (ii)

$$\frac{1}{\min_{1 \le q \le Q} n_q} \cdot \frac{m_N}{v_N} \to 0, \tag{10}$$

then $\boldsymbol{t}_N \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V})$.

Because the components of $\boldsymbol{t}_N$ are linearly dependent, the rank of its covariance matrix is $Q-1$. When $Q = 2$, Theorem 2 reduces to Theorem 1. We use Fraser (1956)'s vector CLT for rank statistics to prove Theorem 2. Lehmann (1975, Theorem 19, page 393) presented a slightly weaker form and gave a different proof, requiring an equivalent form of Condition (10) and additionally that $n_q \to \infty$ and n_q/N has a limiting value less than 1. Recall that $m_N/v_N \geq 1 - N^{-1}$, and therefore Condition (10) implies $n_q \to \infty$ for all q.

Theorem 2 is particularly useful for studying the asymptotic properties of randomization tests in completely randomized experiments with multiple arms. Consider a completely randomized experiment with N units and Q treatments. For each unit i, the Q dimensional vector $(Y_i(1), \ldots, Y_i(Q))$ denotes its potential outcomes under all treatments. Let L_i be the treatment number for unit i, where $L_i = q$ if it is assigned to treatment q. Therefore, $Y_i = Y_i(L_i)$ is the observed outcome of unit i. Fisher's sharp null hypothesis states that

$$H_0 : Y_i(1) = Y_i(2) = \cdots = Y_i(Q) \quad (i = 1, \ldots, N). \tag{11}$$

Under the sharp null hypothesis that the treatment does not affect any units, all the observed outcomes are fixed numbers, and the randomization of the treatment numbers are the only source of randomness. Because the joint distribution of the L_i's is known, the distribution of any test statistic under H_0 is also known and can often be approximated by simple distributions with large sample sizes. We review three examples for testing the sharp null hypothesis using the ranks of the pooled observed outcomes. Assuming no ties, let R_i be the rank of Y_i among all units, $\bar{R}_{(q)} = \sum_{i:L_i=q} R_i/n_q$ be the average rank of units under treatment q, and

$$\widetilde{R}_{(q)} = \frac{\bar{R}_{(q)} - E(\bar{R}_{(q)})}{\sqrt{\mathrm{Var}(\bar{R}_{(q)})}} = \sqrt{\frac{12n_q}{(N+1)(N-n_q)}}\left(\bar{R}_{(q)} - \frac{N+1}{2}\right)$$

be the standardized rank average.

Corollary 1. Under the sharp null hypothesis in (11), as $N \to \infty$, if for each $1 \le q \le Q$, $n_q \to \infty$ and $n_q/N \to \gamma_q < 1$, then

$$\left(\widetilde{R}_{(1)}, \widetilde{R}_{(2)}, \ldots, \widetilde{R}_{(Q)}\right)^\top \xrightarrow{d} \mathcal{N}(\mathbf{0}, \boldsymbol{V}_R), \tag{12}$$

where $\boldsymbol{V}_R$ is a correlation matrix with the (q, r)th entry $-\sqrt{\gamma_q\gamma_r/\{(1-\gamma_q)(1-\gamma_r)\}}$.

Corollary 1 plays a crucial role in nonparameteric tests based on ranks. Below we discuss three examples.

Example 3 (Kruskal–Wallis test)**.** Conducting analysis of variance on the ranks results in the Kruskal–Wallis test statistic

$$H = (N-1)\frac{\sum_{q=1}^{Q} n_q\{\bar{R}_{(q)} - \bar{R}\}^2}{\sum_{i=1}^{N}(R_i - \bar{R})^2} = \sum_{q=1}^{Q}\frac{N-n_q}{N}\widetilde{R}^2_{(q)},$$

which is a quadratic form of the standardized ranks in (12). Corollary 1 and the properties of quadratic forms of multivariate Normal distributions guarantee that H converges to a χ^2_{Q-1} random variable, as shown in Lehmann (1975, Appendix 8, Example 31). □

Example 4. Besides the Kruskal–Wallis test statistic, Lehmann (1975) suggested using $\max_{1\leq q\leq Q}\bar{R}_{(q)}$ for testing (11). Another reasonable test statistic is $\max_q \bar{R}_{(q)} - \min_q \bar{R}_{(q)} = \max_{q,r}\{\bar{R}_{(q)} - \bar{R}_{(r)}\}$. After proper standardization, the asymptotic distributions of both test statistics can be approximated by the distribution functions of multivariate Normal distributions. □

Example 5 (Rank test with dose)**.** Assume that each level of the treatment represents a dose z_q for $q = 1,\ldots,Q$. If we anticipate a monotonic dose-response relationship, then it seems more reasonable to use the following test statistic $\sum_{q=1}^{Q} z_q\bar{R}_{(q)}$, weighting the average ranks by the corresponding dose (c.f. Page 1963; Rosenbaum 2003). Because it is a linear combination of (12), Corollary 1 implies that this test statistic has an asymptotic Normal distribution, based on which we can conduct one sided or two sided tests. □

3 Finite population CLTs in randomized experiments

In this section, we will establish finite population CLTs in completely randomized experiments without assuming the sharp null hypothesis. These CLTs play crucial roles in repeated sampling evaluations of causal effect estimators in randomization-based causal inference. These finite population CLTs deal with vector outcomes under multiple treatments, and work for general causal estimators. Consider an experiment with N units and Q treatments, where n_q units receive treatment q, and $\sum_{q=1}^{Q} n_q = N$. For any treatment $1 \leq q \leq Q$, let $\boldsymbol{Y}_i(q) \in \mathbb{R}^p$ be the p dimensional potential outcome vector of unit i, and $\bar{\boldsymbol{Y}}(q) = \sum_{i=1}^{N} \boldsymbol{Y}_i(q)/N \in \mathbb{R}^p$ be the average potential outcome vector of the N units. Causal estimands of interest are often linear combinations of the potential outcomes. For any $K \geq 1$ and coefficient matrices $\boldsymbol{A}_q \in \mathbb{R}^{K\times p}$ $(q = 1, 2, \ldots, Q)$, we consider individual causal effect of the form $\boldsymbol{\tau}_i(\boldsymbol{A}) = \sum_{q=1}^{Q} \boldsymbol{A}_q \boldsymbol{Y}_i(q)$, and the population average causal effect of the form

$$\boldsymbol{\tau}(\boldsymbol{A}) = (\tau_{(1)}(\boldsymbol{A}), \ldots, \tau_{(K)}(\boldsymbol{A}))^\top = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{\tau}_i(\boldsymbol{A}) = \sum_{q=1}^{Q} \boldsymbol{A}_q \bar{\boldsymbol{Y}}(q). \tag{13}$$

The general causal estimand (13) covers many important cases, including vector outcomes and joint effects. For example, if $\boldsymbol{A}_1 = \boldsymbol{I}_{p\times p}, \boldsymbol{A}_2 = -\boldsymbol{I}_{p\times p}, \boldsymbol{A}_3 = \cdots = \boldsymbol{A}_Q = \boldsymbol{0}_{p\times p}$, then $\boldsymbol{\tau}(\boldsymbol{A}) = \bar{\boldsymbol{Y}}(1) - \bar{\boldsymbol{Y}}(2)$ is the average causal effect comparing treatments 1 and 2. If

$$\boldsymbol{A}_1 = \begin{pmatrix} \boldsymbol{I}_{p\times p} \\ \boldsymbol{I}_{p\times p} \end{pmatrix}, \quad \boldsymbol{A}_2 = \begin{pmatrix} -\boldsymbol{I}_{p\times p} \\ \boldsymbol{0}_{p\times p} \end{pmatrix}, \quad \boldsymbol{A}_3 = \begin{pmatrix} \boldsymbol{0}_{p\times p} \\ -\boldsymbol{I}_{p\times p} \end{pmatrix}, \quad \boldsymbol{A}_4 = \cdots = \boldsymbol{A}_Q = \begin{pmatrix} \boldsymbol{0}_{p\times p} \\ \boldsymbol{0}_{p\times p} \end{pmatrix},$$

then $\boldsymbol{\tau}(\boldsymbol{A}) = \left(\bar{\boldsymbol{Y}}(1) - \bar{\boldsymbol{Y}}(2), \bar{\boldsymbol{Y}}(1) - \bar{\boldsymbol{Y}}(3)\right)$ is the average causal effects comparing treatment 1 to treatments 2 and 3. Many applications are interested in jointly estimating multiple causal effects. We will discuss more examples intensively in Sections 4 and 5.

We consider again a completely randomized experiment with Q treatment groups of sizes $(n_1, \ldots, n_Q)$, as introduced in Section 2.2. Let $\boldsymbol{Y}_i = \boldsymbol{Y}_i(L_i)$ be the observed outcome of unit i. The average observed outcome under treatment q is $\widehat{\bar{\boldsymbol{Y}}}(q) = \sum_{i:L_i=q} \boldsymbol{Y}_i/n_q$, and an intuitive estimator for $\boldsymbol{\tau}(\boldsymbol{A})$ is

$$\widehat{\boldsymbol{\tau}}(\boldsymbol{A}) = (\widehat{\tau}_{(1)}(\boldsymbol{A}), \ldots, \widehat{\tau}_{(K)}(\boldsymbol{A}))^\top = \sum_{q=1}^{Q} \boldsymbol{A}_q \widehat{\bar{\boldsymbol{Y}}}(q),$$

by replacing $\bar{\boldsymbol{Y}}(q)$ by $\widehat{\bar{\boldsymbol{Y}}}(q)$ in (13). A central question is to study the repeated sampling properties of $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$ over all randomizations.

Extending Neyman (1923), the following theorem shows that $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$ is unbiased for $\boldsymbol{\tau}(\boldsymbol{A})$, with sampling covariance depending on the finite population covariances of the potential outcomes

$$\boldsymbol{S}_q^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\boldsymbol{Y}_i(q)-\bar{\boldsymbol{Y}}(q)\right\}\left\{\boldsymbol{Y}_i(q)-\bar{\boldsymbol{Y}}(q)\right\}^\top, \quad (q=1,\dots,Q)$$

the finite population covariances between the potential outcomes

$$\boldsymbol{S}_{qr} = \frac{1}{N-1}\sum_{i=1}^{N}\left\{\boldsymbol{Y}_i(q)-\bar{\boldsymbol{Y}}(q)\right\}\left\{\boldsymbol{Y}_i(r)-\bar{\boldsymbol{Y}}(r)\right\}^\top, \quad (q,r=1,\dots,Q; q\neq r)$$

and the finite population covariance of the individual causal effects

$$\boldsymbol{S}_{\tau(A)}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\{\boldsymbol{\tau}_i(\boldsymbol{A})-\boldsymbol{\tau}(\boldsymbol{A})\}\{\boldsymbol{\tau}_i(\boldsymbol{A})-\boldsymbol{\tau}(\boldsymbol{A})\}^\top.$$

Theorem 3. In a completely randomized experiment with n units and Q groups, let $\boldsymbol{Y}_i(q)\in\mathbb{R}^p$ be unit i's potential outcome under treatment q. Over all $N!/(n_1!\cdots n_Q!)$ randomizations, the estimator $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$ has mean $\boldsymbol{\tau}(\boldsymbol{A})$ and covariance

$$\mathrm{Cov}\{\widehat{\boldsymbol{\tau}}(\boldsymbol{A})\} = \sum_{q=1}^{Q}\frac{1}{n_q}\boldsymbol{A}_q\boldsymbol{S}_q^2\boldsymbol{A}_q^\top - \frac{1}{N}\boldsymbol{S}_{\tau(A)}^2.$$

To construct a confidence set for the causal estimand $\boldsymbol{\tau}(\boldsymbol{A})$, we need to establish CLTs under complete randomization. Below we use $[\boldsymbol{Y}]_{(k)}$ to denote the k-th coordinate of a vector $\boldsymbol{Y}$. Analogous to Theorems 1 and 2, the asymptotic behavior of $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$ depends on the maximum square distance of the k-th coordinate of the $\boldsymbol{A}_q\boldsymbol{Y}_i(q)$'s from their population mean

$$m_q(k) = \max_{1\leq i\leq N}\left[\boldsymbol{A}_q\boldsymbol{Y}_i(q)-\boldsymbol{A}_q\bar{\boldsymbol{Y}}(q)\right]_{(k)}^2, \quad (1\leq k\leq K)$$

the finite population variance of the k-th coordinate of the $\boldsymbol{A}_q\boldsymbol{Y}_i(q)$'s

$$v_q(k) = \frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{A}_q\boldsymbol{Y}_i(q) - \boldsymbol{A}_q\bar{\boldsymbol{Y}}(q)\right]^2_{(k)}, \quad (1 \le k \le K)$$

and the finite population variance of the k-th coordinate of the $\boldsymbol{\tau}_i(\boldsymbol{A})$'s

$$v_\tau(k) = \frac{1}{N-1}\sum_{i=1}^{N}\left[\boldsymbol{\tau}_i(\boldsymbol{A}) - \boldsymbol{\tau}(\boldsymbol{A})\right]^2_{(k)}, \quad (1 \le k \le K).$$

Theorem 4. Under the setting of Theorem 3, as $N \to \infty$, if

$$\max_{1\le q\le Q}\max_{1\le k\le K}\frac{1}{n_q^2}\frac{m_q(k)}{\sum_{r=1}^{Q} n_r^{-1}v_r(k) - N^{-1}v_\tau(k)} \to 0, \tag{14}$$

and the correlation matrix of $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$ has a limiting value $\boldsymbol{V}$, then

$$\left(\frac{\widehat{\tau}_{(1)}(\boldsymbol{A}) - \tau_{(1)}(\boldsymbol{A})}{\sqrt{\mathrm{Var}\left\{\widehat{\tau}_{(1)}(\boldsymbol{A})\right\}}}, \ldots, \frac{\widehat{\tau}_{(K)}(\boldsymbol{A}) - \tau_{(K)}(\boldsymbol{A})}{\sqrt{\mathrm{Var}\left\{\widehat{\tau}_{(K)}(\boldsymbol{A})\right\}}}\right) \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}). \tag{15}$$

Although Condition (14) is general, it is not intuitive and needs more explanation. We consider two special cases and provide more easy-to-check conditions for the CLT of $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$. First, we assume that the causal effects are additive, i.e., $\boldsymbol{\tau}_i(\boldsymbol{A})$ is a constant vector for all unit i. The following corollary is useful for randomization inference with the additive causal effects assumption, including randomization tests under the sharp null hypothesis.

Corollary 2. Under the setting of Theorem 3 with the additive causal effect assumption, as $N \to \infty$, if

$$\max_{1\le q\le Q}\max_{1\le k\le K}\frac{1}{n_q}\frac{m_q(k)}{v_q(k)} \to 0, \tag{16}$$

and the correlation matrix of $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$ has a limiting value $\boldsymbol{V}$, then (15) holds.

Corollary 2 is similar to Theorem 1 in the sense that the regularity condition involves the ratio between the maximum squared distance and the finite population variance of certain populations. It is directly implied by Theorem 4 by noticing that $v_\tau(k) = 0$ under the additive causal effect assumption. If we use the ranks of the observed outcomes in nonparametric tests for the sharp null hypothesis, then $[\boldsymbol{A}_q \boldsymbol{Y}_i(q)]_{(k)}$'s are a permutation of $\{1, 2, \dots, N\}$, and the corresponding regularity condition holds automatically, because as $n_q \to \infty$ for all $1 \leq q \leq Q$,

$$\frac{m_q(k)}{n_q v_q(k)} = \frac{3(N-1)^2}{n_q N(N+1)} \to 0. \tag{17}$$

Second, we assume that the finite population of experimental units has limiting covariances, and the proportions of units receiving all treatments have positive limiting values.

Theorem 5. Under the setting of Theorem 3, if for any $1 \leq q \neq r \leq Q$, $\boldsymbol{S}_q^2$ and $\boldsymbol{S}_{qr}$ have limiting values, n_q/N has positive limiting value, and $\max_{1\leq q\leq Q} \max_{1\leq i\leq N} \left\| \boldsymbol{Y}_i(q) - \bar{\boldsymbol{Y}}(q) \right\|_2^2 / N \to 0$, then $N\text{Var}\{\widehat{\boldsymbol{\tau}}(\boldsymbol{A})\}$ has a limiting value, denoted by $\boldsymbol{V}$, and

$$\sqrt{N}\{\widehat{\boldsymbol{\tau}}(\boldsymbol{A}) - \boldsymbol{\tau}(\boldsymbol{A})\} \xrightarrow{d} \mathcal{N}(\boldsymbol{0}, \boldsymbol{V}).$$

Note that by properly scaling the potential outcomes, the diagonal elements of the finite population covariances $\boldsymbol{S}_q^2$'s can always have limits. Thus the condition in Theorem 5 that the $\boldsymbol{S}_q^2$'s and $\boldsymbol{S}_{qr}$'s have limits essentially require only the convergence of the correlations between different coordinates of potential outcomes. The regularity conditions in Theorem 5 also involve the restriction on the order of the maximum squared distance of the $\boldsymbol{Y}_i(q)$'s from the population mean $\bar{\boldsymbol{Y}}(q)$. When the coordinates of the $\boldsymbol{Y}_i(q)$'s are bounded, or i.i.d draws from a superpopulation with more than two moments, the regularity condition $\max_{1\leq i\leq N} \left\| \boldsymbol{Y}_i(q) - \bar{\boldsymbol{Y}}(q) \right\|_2^2 / N \to 0$ holds with probability one.

For a scalar outcome, Freedman (2008a,b) established a finite population CLT under stronger conditions, requiring that the fourth moments of the potential outcomes are finite. Theorem 4 deals with a vector outcome and requires weaker conditions.

Now we consider estimation of the covariance of $\widehat{\boldsymbol{\tau}}(\boldsymbol{A})$. Because treatment arm q is a simple random sample of the finite population, the sample covariance of $\boldsymbol{Y}_i$ under treatment q,

$$\boldsymbol{s}_q^2 = \frac{1}{n_q - 1} \sum_{L_i = q} \left\{ \boldsymbol{Y}_i - \widehat{\bar{\boldsymbol{Y}}}(q) \right\} \left\{ \boldsymbol{Y}_i - \widehat{\bar{\boldsymbol{Y}}}(q) \right\}^\top, \quad (1 \leq q \leq Q)$$

is unbiased for the population covariance $\boldsymbol{S}_q^2$ (Cochran 1977). However, $\boldsymbol{S}_{\tau(A)}^2$ is generally not estimable, because the potential outcomes $\boldsymbol{Y}_i(1), \ldots, \boldsymbol{Y}_i(Q)$ cannot be jointly observed. On average, the covariance estimator $\widehat{\boldsymbol{V}}_A = \sum_{q=1}^Q n_q^{-1} \boldsymbol{A}_q \boldsymbol{s}_q^2 \boldsymbol{A}_q^\top$ over estimates the sampling variance by $\boldsymbol{S}_{\tau(A)}^2 / N$. Let $q_{K,1-\alpha}$ be the $(1-\alpha)$th quantile of a χ^2 distribution with degrees of freedom K.

Proposition 3. Under the regularity conditions in Theorem 5, $\boldsymbol{s}_q^2 - \boldsymbol{S}_q^2 \xrightarrow{p} 0$ for each $1 \leq q \leq Q$. If the limits of $\boldsymbol{S}_q^2$'s are not all zero, then the probability that $\widehat{\boldsymbol{V}}_A$ is nonsingular converges to one, and the Wald-type confidence region for $\boldsymbol{\tau}(\boldsymbol{A})$,

$$\left\{ \boldsymbol{\mu} : \{\widehat{\boldsymbol{\tau}}(\boldsymbol{A}) - \boldsymbol{\mu}\}^\top \widehat{\boldsymbol{V}}_A^{-1} \{\widehat{\boldsymbol{\tau}}(\boldsymbol{A}) - \boldsymbol{\mu}\} \leq q_{K,1-\alpha} \right\},$$

has asymptotic coverage rate at least as large as $1-\alpha$, and the asymptotic coverage rate equals $1-\alpha$ if and only if the causal effects are asymptotically additive, i.e., $\lim_{N\to\infty} \boldsymbol{S}_{\tau(A)}^2 = \boldsymbol{0}$.

Theorems 3–5 and Proposition 3 generalize Neyman (1923). For a binary treatment and a scalar outcome, his results (c.f. Imbens and Rubin 2015) ensure that the difference-in-means estimator $\widehat{\tau}$ is unbiased for $\tau = \sum_{i=1}^N \{Y_i(1) - Y_i(0)\}/N$ with variance $S_1^2/n_1 + S_0^2/n_0 - S_\tau^2/N$, where S_1^2 and S_0^2 are the finite population variances of the treatment and control potential outcomes, and S_τ^2 is the finite population variance of the individual causal effects. Based on the Normal approximation, a conservative $1-\alpha$ confidence interval for τ is $\widehat{\tau} \pm \Phi^{-1}(1-\alpha/2)(s_1^2/n_1 + s_0^2/n_0)^{1/2}$, where s_1^2 and s_0^2 are the sample variances of the outcomes under treatment and control. Aronow et al. (2014) proposed a consistent estimator of sharp bounds on the sampling variance of difference-in-means estimator in this setting with a scalar outcome. It will be interesting to extend their result to general experiments with general outcomes.

4 Applications to treatment-control experiments

The generality of Theorems 4 and 5 allows us to prove asymptotics for many causal inference problems. Below we review five important examples in treatment-control experiments. Previous literature provided intuitive arguments for asymptotic Normalities in these examples, but our proofs based on Theorems 4 and 5 are more rigorous and provide more general results.

In Examples 6–9, we consider completely randomized treatment-control experiments. For descriptive simplicity, we unify the notation in these four examples. Consider a completely randomized experiment with N units, among which n_1 units receive treatment and n_0 receive control. For each unit i, let Z_i be the treatment assignment indicator ($Z_i = 1$ if treatment; $Z_i = 0$ if control), $\boldsymbol{X}_i = (X_{1i}, \ldots, X_{Ki})$ the K dimensional pretreatment covariates, $Y_i(z)$ the potential outcome under treatment arm z, $\tau_i = Y_i(1) - Y_i(0)$ the individual causal effect, and $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$ the observed outcome. We use $\bar{Y}(z)$ and $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_K)$ to denote finite population means of the potential outcomes and covariates, and $\tau = \bar{Y}(1) - \bar{Y}(0)$ to denote the average causal effect. To facilitate the discussion, we center the covariates with zero finite population means ($\bar{\boldsymbol{X}} = \boldsymbol{0}$). As $N \to \infty$, we assume the proportions of units receiving both treatments have positive limiting values, the finite population variances and covariances among potential outcomes and covariates have limiting values, and

$$\frac{1}{N} \max_{1 \leq i \leq N} \left\{ Y_i(z) - \bar{Y}(z) \right\}^2 \to 0, \quad \frac{1}{N} \max_{1 \leq i \leq N} X_{ki}^2 \to 0, \quad (z = 0, 1; k = 1, \ldots, K). \tag{18}$$

Example 6 (Combining test statistics in a randomization test)**.** For testing the sharp null hypothesis that $Y_i(1) = Y_i(0)$ for all i, we need to choose a test statistic, and can use the finite population central limit theorem to determine the rejection region. Two commonly used statistics are the difference-in-means statistic

$$T = \frac{1}{n_1} \sum_{i=1}^{N} Z_i Y_i - \frac{1}{n_0} \sum_{i=1}^{N} (1 - Z_i) Y_i,$$

and the Wilcoxon rank sum statistic

$$W = \frac{1}{n_1}\sum_{i=1}^{N} Z_i R_i - \frac{1}{n_0}\sum_{i=1}^{N}(1-Z_i)R_i,$$

where R_i is the rank of Y_i among all units. Combining T and W can sometimes leads to a more powerful test than using only one of them. To determine the rejection region, it is important to derive the asymptotic joint distribution of (T, W) under complete randomization. Under the sharp null hypothesis, all the Y_i's and R_i's are fixed quantities unaffected by the treatment, and the sampling distributions of T and W are determined by the distribution of $(Z_1, \dots, Z_N)$. Let $\overline{Y}$ and $\overline{R}$ be the averages, and s_Y^2 and s_R^2 be the variances of pooled Y_i's and R_i's. Let $\boldsymbol{V}_\rho$ be a two dimensional correlation matrix with off-diagonal element ρ. Using Corollary 2 with potential outcomes $(Y_i, R_i)^\top$ under both treatment conditions and coefficient matrices $\boldsymbol{A}_1 = \boldsymbol{I}_{2\times 2}$ and $\boldsymbol{A}_0 = -\boldsymbol{I}_{2\times 2}$, if

$$\frac{\max_{1\le i\le N}\left|Y_i - \overline{Y}\right|^2}{n_z s_Y^2} \to 0, \quad \frac{\max_{1\le i\le N}\left|R_i - \overline{R}\right|^2}{n_z s_R^2} \to 0, \quad (z = 0, 1) \tag{19}$$

and the finite population correlation between Y_i and R_i, ρ_N, has a limiting value ρ_∞, then

$$\left(\frac{T}{\sqrt{\mathrm{Var}(T)}}, \frac{W}{\sqrt{\mathrm{Var}(W)}}\right) = \sqrt{\frac{n_1 n_0}{N}}\,(T/s_Y, W/s_R) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\rho_\infty}\right).$$

Note that the first condition in (19) is ensured by (18), and the second condition in (19) holds automatically as argued in (17). To determine the rejection region for (T, W), similar to Rosenbaum (2012), we introduce $1-\gamma_\rho(c)$, the probability of the 2-dimensional lower orthant $(-\infty, c]\times(-\infty, c]$ for a 2-dimensional Normal distribution with mean zero and covariance matrix $\boldsymbol{V}_\rho$. Let c_α be a constant such that $\gamma_{\rho_N}(c_\alpha) = \alpha$. The rejection region with significance level α is $\sqrt{n_1 n_0/N}\max\{T/s_Y, W/s_R\} > c_\alpha$. We can easily generalize the above results to more than two test statistics. □

Example 7 (Multiple randomization tests)**.** Assuming for each unit i under treatment assignment z, there is a two dimensional potential outcome vector $\boldsymbol{Y}_i(z) = (Y_{1i}(z), Y_{2i}(z))$. We are interested in testing two sharp null hypotheses that $Y_{1i}(1) = Y_{1i}(0)$ and $Y_{2i}(1) = Y_{2i}(0)$ for all i, simultaneously. Let Y_{1i} and Y_{2i} be the observed outcomes for unit i. We can use the difference-in-means statistics to test both sharp null hypotheses:

$$T_1 = \frac{1}{n_1}\sum_{i=1}^{N} Z_i Y_{1i} - \frac{1}{n_0}\sum_{i=1}^{N}(1 - Z_i)Y_{1i},$$
$$T_2 = \frac{1}{n_1}\sum_{i=1}^{N} Z_i Y_{2i} - \frac{1}{n_0}\sum_{i=1}^{N}(1 - Z_i)Y_{2i}.$$

In Example 6, the two test statistics are used to test the same sharp null hypothesis, but T_1 and T_2 here are used to test two different null hypotheses. One way to control the type one error is the Bonferroni correction, using only the marginal asymptotic distribution of T_1 and T_2. A more efficient way is to use the joint distribution of (T_1, T_2). For $k = 1, 2$, let $\bar{Y}_k$ and $s^2_{Y_k}$ be the mean and variance of the pooled observed values of the Y_{ki}'s. We use Corollary 2 with potential outcomes $(Y_{1i}(z), Y_{2i}(z))^\top$ and coefficient matrices $\boldsymbol{A}_1 = \boldsymbol{I}_{2\times 2}$ and $\boldsymbol{A}_0 = -\boldsymbol{I}_{2\times 2}$. According to Corollary 2, under the sharp null hypothesis, if $\max_{1\le i\le N}(Y_{ki} - \bar{Y}_k)^2/(n_z s^2_{Y_k}) \to 0$ for $k = 1, 2$ and $z = 0, 1$, and the finite population correlation between Y_{1i} and Y_{2i}, η_N, has a limiting value η_∞, then

$$\left(\frac{T_1}{\sqrt{\mathrm{Var}(T_1)}}, \frac{T_2}{\sqrt{\mathrm{Var}(T_2)}}\right) = \sqrt{\frac{n_1 n_0}{N}}\,(T_1/s_{Y_1}, T_2/s_{Y_2}) \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{V}_{\eta_\infty}\right),$$

recalling that $\boldsymbol{V}_{\eta_\infty}$ is a two dimensional correlation matrix with off-diagonal element η_∞. Let c_α be a constant satisfying $\gamma_{\eta_N}(c_\alpha) = \alpha$ defined in Example 6, and we then choose the rejection region as $\sqrt{n_1 n_0/N}\max\{T_1/s_{Y_1}, T_2/s_{Y_2}\} > c_\alpha$. □

The previous examples in this section dealt with cases under the sharp null hypotheses that the individual outcomes are unaffected by the treatment. More interestingly, Theorems 3–5 are useful for obtaining the repeated sampling properties of many causal estimators under different designs. We illustrate this angle with Examples 8–11 below.

Example 8 (Rerandomization)**.** Over complete randomization, the difference-in-means of covariates, $\widehat{\boldsymbol{\tau}}_X = \sum_{i=1}^N Z_i \boldsymbol{X}_i/n_1 - \sum_{i=1}^N(1 - Z_i)\boldsymbol{X}_i/n_0$, has expectation zero. However, for a realized randomization, as pointed by Morgan and Rubin (2012), it is very likely that some covariates are not balanced in means between two treatment groups. Therefore, it is reasonable to accept only those randomizations satisfying certain balancing criterion, such as, a certain norm of the difference-in-means of covariates is less than or equal to a pre-determined threshold. This is rerandomizaton. Note that the covariates $\boldsymbol{X}$ are "outcomes" unaffected by the treatment, with known finite population covariance $\boldsymbol{S}_X^2$. Using Theorem 5 with potential outcome $\boldsymbol{X}_i$ for both treatment conditions and coefficient matrices $\boldsymbol{A}_1 = \boldsymbol{I}_{K\times K}$ and $\boldsymbol{A}_0 = -\boldsymbol{I}_{K\times K}$, over complete randomization,

$$\widehat{\boldsymbol{\delta}} \equiv \left(\frac{N}{n_1 n_0}\boldsymbol{S}_X^2\right)^{-1/2} \widehat{\boldsymbol{\tau}}_X \xrightarrow{d} \mathcal{N}(0, \boldsymbol{I}_{K\times K}),$$

and therefore, $\widehat{\boldsymbol{\delta}}^\top \widehat{\boldsymbol{\delta}} \xrightarrow{d} \chi_K^2$. Morgan and Rubin (2012) suggested a rerandomization criterion such that $\widehat{\boldsymbol{\delta}}^\top \widehat{\boldsymbol{\delta}}$ is smaller than a threshold, and obtained theoretical results assuming that $\widehat{\boldsymbol{\delta}}$ follows an exact Normal distribution. More importantly, we need to study the properties of the difference-in-means estimator $\widehat{\tau} = \sum_{i=1}^N Z_i Y_i/n_1 - \sum_{i=1}^N(1-Z_i)Y_i/n_0$ under rerandomization, which is conceptually the same as the conditional distribution of $\widehat{\tau}$ given that $\widehat{\boldsymbol{\delta}}$ satisfies the rerandomization criterion over complete randomization. Therefore, it is crucial to obtain the joint asymptotic distribution of $(\widehat{\tau}, \widehat{\boldsymbol{\tau}}_X)$ over complete randomization. According to Theorem 5 with potential outcomes $(Y(z), \boldsymbol{X}_i)$ and coefficient matrices $\boldsymbol{A}_1 = \boldsymbol{I}_{(K+1)\times(K+1)}$ and $\boldsymbol{A}_0 = -\boldsymbol{I}_{(K+1)\times(K+1)}$, the joint distribution of $(\widehat{\tau}, \widehat{\boldsymbol{\tau}}_X)$ is asymptotically Normal if (18) holds, a result utilized by Li, Ding, and Rubin (2016) to prove asymptotic properties of rerandomization. Cochran (1965, 1982) discussed a related problem in observational studies. □

Example 9 (Regression adjustment in completely randomized experiments)**.** Although the simple difference-in-means estimator of the average causal effect is unbiased, carefully utilizing the covariates often improves the efficiency (Freedman 2008a,b; Lin 2013). A class of regression adjusted estimators for the average causal effect is

$$\widehat{\tau}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_0) = \frac{1}{n_1} \sum_{i=1}^{N} Z_i \, (Y_i - \boldsymbol{\beta}_1^\top \boldsymbol{X}_i) - \frac{1}{n_0} \sum_{i=1}^{N} (1 - Z_i) \, (Y_i - \boldsymbol{\beta}_0^\top \boldsymbol{X}_i) \,, \tag{20}$$

where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$ are any vectors that do not depend on the treatment indicators but can implicitly depend on N with limiting values as $N \to \infty$. When $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_0 = \boldsymbol{0}$, it reduces to the simple difference-in-means estimator. According to (20), $\widehat{\tau}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_0)$ is essentially the difference-in-means estimator with "adjusted" treatment potential outcome $Y_i(1) - \boldsymbol{\beta}_1^\top \boldsymbol{X}_i$ and "adjusted" control potential outcome $Y_i(0) - \boldsymbol{\beta}_0^\top \boldsymbol{X}_i$ for unit i. Note that the average causal effect based on the adjusted potential outcomes remains the same as the average causal effect based on the original potential outcomes τ. Therefore, the classical result (Neyman 1923) guarantees that over complete randomization, $\widehat{\tau}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_0)$ is unbiased for τ, and the CLT in Theorem 5 with potential outcomes $Y_i(z) - \boldsymbol{\beta}_z^\top \boldsymbol{X}_i$ and coefficients $A_1 = 1$ and $A_0 = -1$ guarantees its asymptotic Normality. We can construct a conservative large-sample confidence interval based on the "adjusted" observed outcome data $Y_i - \boldsymbol{\beta}_1^\top \boldsymbol{X}_i$ for treated units with $Z_i = 1$ and $Y_i - \boldsymbol{\beta}_0^\top \boldsymbol{X}_i$ for control units with $Z_i = 0$.

In practice, how to choose $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_0$? Let $\widetilde{\boldsymbol{\beta}}_z$ be the finite population least squares coefficient of $Y_i(z)$ on $\boldsymbol{X}_i$ for units $i = 1, 2, \dots, N$ (cf. Cochran 1977). We show in the Supplementary Material that the sampling variance of any regression adjusted estimator has the following decomposition:

$$\text{Var}\{\widehat{\tau}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_0)\} = \text{Var}\{\widehat{\tau}(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_0)\} + \text{Var}\{\widehat{\tau}(\boldsymbol{\beta}_1, \boldsymbol{\beta}_0) - \widehat{\tau}(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_0)\},$$

demonstrating that $\widehat{\tau}(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_0)$ is optimal in the sense of having the smallest sampling variance among all regression adjusted estimators defined in (20). However, because $\widetilde{\boldsymbol{\beta}}_1$ and $\widetilde{\boldsymbol{\beta}}_0$ are both unknown, in practice we instead use the sample least squares coefficient of Y_i on $\boldsymbol{X}_i$ for units in treatment arm z, $\widehat{\boldsymbol{\beta}}_z$, to replace $\widetilde{\boldsymbol{\beta}}_z$. In the Supplementary Material, we show that the two regression adjusted estimators with true and estimated least squares coefficients, $\widehat{\tau}(\widetilde{\boldsymbol{\beta}}_1, \widetilde{\boldsymbol{\beta}}_0)$ and $\widehat{\tau}(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_0)$, have the same asymptotic Normal distribution, and the difference between them is of order $o_p(1/\sqrt{N})$, without requiring any further regularity conditions beyond (18). Furthermore, we show that treating $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_0$ as if they were pretreatment vectors does not affect the asymptotic coverage rate of the confidence interval based on the Normal approximation.

Recently, Lin (2013) established a connection between regression adjusted estimator (20) and linear regression with full treatment-covariate interactions using the Huber–White variance estimator. We supplement his result with "optimality" and a rigorous proof of the asymptotic Normality. Samii and Aronow (2012) focused on comparison of regression-based and randomization-based standard errors. Abadie et al. (2014) gave a proof of the asymptotic Normality of the regression estimator under independently assigned treatment indicators. Rosenbaum (2002a) discussed alternative forms of covariate adjustment in randomized experiments. □

Example 10 (Cluster-randomized experiments)**.** We consider a cluster-randomized experiment, where individuals within the same cluster receive the same treatment condition. Assume there are M clusters, and among them, m_1 clusters receive treatment and m_0 clusters receive control. Let $\widetilde{Y}_j(z)$ be the total potential outcome of units in cluster j under treatment arm z, $\widetilde{\boldsymbol{X}}_j$ a K dimensional cluster-level covariate (including the total number of units and aggregate covariates in each cluster), and $\overline{\widetilde{\boldsymbol{X}}}$ be the finite population average of cluster-level covariates. To faciliate the discussion, we center the cluster-level covariates at zero ($\overline{\widetilde{\boldsymbol{X}}} = \boldsymbol{0}$). Let N be the total number of units, and $\tau = \sum_{j=1}^{M}\{\widetilde{Y}_j(1) - \widetilde{Y}_j(0)\}/N$ be the average causal effect over all units. For cluster j, let $\widetilde{Z}_j$ be the treatment assignment indicator, and $\widetilde{Y}_j = \widetilde{Z}_j\widetilde{Y}_j(1) + (1 - \widetilde{Z}_j)\widetilde{Y}_j(0)$ be the observed outcome. We consider the following class of adjusted estimators for the average causal effect τ:

$$\widehat{\Delta}(\gamma_1, \gamma_0) = \frac{M}{N}\left[\frac{1}{m_1}\sum_{j=1}^{M}\widetilde{Z}_j\left(\widetilde{Y}_j - \gamma_1^\top\widetilde{\boldsymbol{X}}_j\right) - \frac{1}{m_0}\sum_{j=1}^{M}(1 - \widetilde{Z}_j)\left(\widetilde{Y}_j - \gamma_0^\top\widetilde{\boldsymbol{X}}_j\right)\right], \tag{21}$$

where γ_z's are any vectors that do not depend on the treatment indicators but can depend implicitly on M with limiting values as $M \to \infty$. When $\gamma_1 = \gamma_0 = \boldsymbol{0}$, (21) reduces to the simple difference-in-means estimator. Middleton and Aronow (2015) showed that $\widehat{\Delta}(\gamma_1, \gamma_0)$ is unbiased for τ when

$\gamma_1 = \gamma_0$ and they are both predetermined constant vectors. Note that if we view each cluster as an experimental unit in a completely randomized experiment, $\widehat{\Delta}(\gamma_1, \gamma_0)$ is essentially the regression adjusted estimator discussed in Example 9, up to a scale constant M/N. Therefore, all results of Example 9 apply here. For instance, we can choose the “optimal” adjustment coefficients as $\widehat{\gamma}_z$, the sample least squares coefficient of $\widetilde{Y}_j$ on $\widetilde{\boldsymbol{X}}_j$ for units in treatment arm $z = 1, 0$. Due to the similarity to Example 9, we relegate the details about the asymptotic distribution and confidence interval construction to the Supplementary Material. □

5 Application to factorial experiments

Our final example is about unbalanced 2^K factorial designs with multiple treatment factors and causal effects, extending Dasgupta et al. (2015)’s discussion of finite sample properties of balanced 2^K factorial designs.

Example 11. Consider a factorial design with K factors, where each factor has two levels $+1$ and -1, and in total there are $Q = 2^K$ treatment combinations. For each treatment combination $1 \leq q \leq Q$, let $\boldsymbol{z}(q) = (z_1(q), z_2(q), \ldots, z_K(q))$ be the levels of the K factors, and n_q be the number of units. Let $N = \sum_{q=1}^{Q} n_q$ be the total number of units, $Y_i(q)$ the potential outcome of unit i and $\bar{Y}(q) = \sum_{i=1}^{N} Y_i(q)/N$ the average potential outcome under treatment combination q. Let $\boldsymbol{Y}_i(1{:}Q) = (Y_i(1), Y_i(2), \ldots, Y_i(Q))$ be the Q dimensional row vector consisting of unit i’s potential outcomes under all treatment combinations, and $\bar{\boldsymbol{Y}}(1{:}Q) = (\bar{Y}(1), \bar{Y}(2), \ldots, \bar{Y}(Q))$ be the row vector consisting of all average potential outcomes. Following the notation in Dasgupta et al. (2015), each factorial effect can be characterized by a column vector with half of its elements being $+1$

and the other half being -1. For example, the average main effect of factor 1 is

$$\tau_1 = \frac{1}{2^{K-1}} \sum_{q=1}^{Q} 1\{z_1(q) = 1\} \cdot \bar{Y}(q) - \frac{1}{2^{K-1}} \sum_{q=1}^{Q} 1\{z_1(q) = -1\} \cdot \bar{Y}(q)$$
$$= \frac{1}{2^{K-1}} \sum_{q=1}^{Q} z_1(q)\bar{Y}(q) = \frac{1}{2^{K-1}} \bar{\boldsymbol{Y}}(1{:}Q)\boldsymbol{g}_1,$$

where $\boldsymbol{g}_1 = (z_1(1), z_1(2) \ldots, z_1(Q))^\top$ characterizes the first factorial effect. In general, let $\boldsymbol{g}_k = (g_{k1}, \ldots, g_{kQ})^\top \in \{+1, -1\}^Q$ be the vector generating the kth factorial effect, and $\tau_k = 2^{-(K-1)}\bar{\boldsymbol{Y}}(1{:}Q)\boldsymbol{g}_k$ be the kth average factorial effect.

For each unit i, let L_i be the treatment assignment indicator ($L_i = q$ if unit i receives treatment combination q), and $Y_i = Y_i(L_i)$ be the observed outcome. Let $\widehat{\bar{Y}}(q) = \sum_{i=1}^{N} 1\{L_i = q\}Y_i/n_q$ be the average observed outcome under treatment combination q, and $\widehat{\bar{\boldsymbol{Y}}}(1{:}Q) = (\widehat{\bar{Y}}(1), \widehat{\bar{Y}}(2), \ldots, \widehat{\bar{Y}}(Q))$ the row vector consisting of all average observed outcomes. An unbiased estimator for τ_k is

$$\widehat{\tau}_k = 2^{-(K-1)}\widehat{\bar{\boldsymbol{Y}}}(1{:}Q)\boldsymbol{g}_k = 2^{-(K-1)} \sum_{q=1}^{Q} g_{kq}\widehat{\bar{Y}}(q) \quad (1 \leq k \leq Q-1).$$

We first consider the joint asymptotic distribution of the $\widehat{\tau}_k$'s under the sharp null hypothesis that $Y_i(1) = \cdots = Y_i(Q)$ for all units i. Let m_N be the maximum squared distance of the Y_i's from the average, and v_N be the finite population variance. From Theorem 3, under the sharp null hypothesis, for any $1 \leq k, m \leq Q-1$, the variance of $\widehat{\tau}_k$ is $\mathrm{Var}_0(\widehat{\tau}_k) = 2^{-2(K-1)} \sum_{q=1}^{Q} n_q^{-1} v_N$, and the covariance between $\widehat{\tau}_k$ and $\widehat{\tau}_m$ is $\mathrm{Cov}_0(\widehat{\tau}_k, \widehat{\tau}_m) = 2^{-2(K-1)} \sum_{q=1}^{Q} n_q^{-1} g_{kq} g_{mq} v_N$. According to Corollary 2, as $N \to \infty$, if for each $1 \leq q \leq Q$, $m_N/(n_q v_N)$ converges to zero, and n_q/N has a positive limit, then all the $\widehat{\tau}_k$'s are jointly Normal asymptotically. Without the sharp null hypothesis, the asymptotic Normality of the $\widehat{\tau}_k$'s over repeated sampling can be similarly established, and it is straightforward to extend it to regression adjustment in factorial experiments (Lu 2016a,b).

If we consider only the marginal distribution of $\widehat{\tau}_k$ under the sharp null hypothesis, then the regularity condition for asymptotic Normality can be weakened. In practice, it is likely that both the total number of units N and the total number of treatment combinations $Q = 2^K$ are large, but the number of units in each treatment combination is moderate. Instead of assuming that the total number of treatments is fixed and does not change as N increases, we allow the number of total treatment combinations Q to increase as N becomes larger. Under the sharp null hypothesis, if the design is balanced with $n_q = N/Q$, then the $\widehat{\tau}_k$'s have exactly the same distribution by symmetry, although their realized numerical values can be different. Without loss of generality, we consider only $\widehat{\tau}_1$, which is essentially the difference-in-means of a random half versus the remaining half of the observed values of the outcomes. Its asymptotic Normality follows directly from Theorem 1, if $\{Y_i : i = 1, \dots, N\}$ satisfies Condition (4). The theory for unbalanced designs appears to be more technical, and we defer the discussion to the Supplementary Material. □

6 Discussion

We have established general forms of finite population CLTs, which were frequently invoked implicitly in randomization-based causal inference. We use them to study asymptotic properties of randomization-based causal inference in completely randomized experiments, cluster randomized experiments, and factorial experiments. For stratified experiments, each stratum is essentially a completely randomized experiment (Kempthorne 1952; Hinkelmann and Kempthorne 2008). Therefore, if the number of strata is small but the sample size is large within each stratum, then we can apply the finite population CLTs to each stratum, and average over the strata to obtain the finite population CLTs for the causal estimators; if the number of independent strata is large, then it suffices to use the classical CLTs for independent variables. Matched-pair experiments are special cases of stratified experiments with two units within each stratum, and the CLT requires a large number of pairs (Rosenbaum 2002b; Imai 2008). Our finite population CLTs may also be useful for other experimental designs–a prospect that needs further investigation and analyses (Kempthorne 1952; Hinkelmann and Kempthorne 2008).

References

A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge. Finite population causal standard errors. Technical Report 20325, National Bureau of Economic Research, 2014.

P. M. Aronow, D. P. Green, and D. K. K. Lee. Sharp bounds on the variance in randomized experiments. *The Annals of Statistics*, 42:850–871, 2014.

W. G. Cochran. The planning of observational studies of human populations. *Journal of the Royal Statistical Society. Series A (General)*, 128:234–266, 1965.

W. G. Cochran. *Sampling Techniques*. New York: Wiley, 3rd edition, 1977.

W. G. Cochran. Performance of a preliminary test of comparability in observational studies (Technical report No. 29 to the U.S. Office of Naval Research, February 4, 1970). In Betty I. M. Cochran, editor, *Contributions to Statistics*, volume 96, pages 96.0–96.14. New York: John Wiley & Sons, 1982.

M. A. Creasy. Limits for the ratio of means. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16:186–194, 1954.

T. Dasgupta, N. S. Pillai, and D. B. Rubin. Causal inference from 2^K factorial designs by using potential outcomes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77:727–753, 2015.

P. Ding. A paradox from randomization-based causal inference. *Statistical Science*, in press, 2016.

P. Ding and T. Dasgupta. A potential tale of two by two tables from completely randomized experiments. *Journal of American Statistical Association*, 111:157–168, 2016.

R. Durrett. *Probability: Theory and Examples*. Cambridge: Cambridge University Press, fourth edition, 2010.

P. Erdös and A. Rényi. On the central limit theorem for samples from a finite population. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 4:49–61, 1959.

E. C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, 16:175–185, 1954.

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.

R. A. Fisher. *The Design of Experiments*. Edinburgh, London: Oliver and Boyd, 1st edition, 1935.

D. A. S. Fraser. A Vector form of the Wald–Wolfowitz–Hoeffding theorem. *The Annals of Mathematical Statistics*, 27:540–543, 1956.

D. A. Freedman. On regression adjustments in experiments with several treatments. *The Annals of Applied Statistics*, 2:176–196, 2008a.

D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40:180–193, 2008b.

J. Hájek. Limiting distributions in simple random sampling from a finite population. *Publications of the Mathematics Institute of the Hungarian Academy of Science*, 5:361–74, 1960.

J. Hájek. Some extensions of the Wald–Wolfowitz–Noether theorem. *The Annals of Mathematical Statistics*, 32:506–523, 1961.

J. Hannan and W. Harkness. Normal approximation to the distribution of two independent binomials, conditional on fixed sum. *Ann. Math. Statist.*, 34:1593–1595, 1963.

W. L. Harkness. Properties of the extended hypergeometric distribution. *Ann. Math. Statist.*, 36: 938–945, 1965.

K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design*. New Jersey: John Wiley & Sons, Inc., second edition, 2008.

W. Hoeffding. A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, 22: 558–566, 1951.

T. Höglund. Sampling from a finite population. a remainder term estimate. *Scandinavian Journal of Statistics*, 5:69–71, 1978.

Lars Holst. Two conditional limit theorems with applications. *Ann. Statist.*, 7:551–557, 1979.

K. Imai. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in Medicine*, 27:4857–4873, 2008.

G. W. Imbens and P. R. Rosenbaum. Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 168:109–126, 2005.

G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge: Cambridge University Press, 2015.

O. Kempthorne. *The Design and Analysis of Experiments.* Wiley, 1952.

S. G. Kou and Z. Ying. Asymptotics for a 2×2 table with fixed margins. *Statistica Sinica*, 6: 809–829, 1996.

E. L. Lehmann. *Nonparametrics: Statistical Methods Based on Ranks*. California: Holden-Day, Inc., 1975.

E. L. Lehmann. *Elements of Large-Sample Theory*. New York: Springer, 1999.

X. Li, P. Ding, and D. B. Rubin. Asymptotic theory of rerandomization in treatment-control experiments. *http://arxiv.org/abs/1604.00698*, 2016.

W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7:295–318, 2013.

L. Liu and M. G. Hudgens. Large sample randomization inference of causal effects in the presence of interference. *Journal of the American Statistical Association*, 109:288–301, 2014.

J. Lu. Covariate adjustment in randomization-based causal inference for 2^K factorial designs. *Statistics and Probability Letters*, 119:11–20, 2016a.

J. Lu. On randomization-based and regression-based inferences for 2^K factorial designs. *Statistics and Probability Letters*, 112:72–78, 2016b.

W. G. Madow. On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics*, 19:535–545, 1948.

J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6:39–75, 2015.

K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *The Annals of Statistics*, 40:1263–1282, 2012.

K. L. Morgan and D. B. Rubin. Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, 110:1412–1421, 2015.

M. Motoo. On the Hoeffding's combinatrial central limit theorem. *Annals of the Institute of Statistical Mathematics*, 8:145–154, 1956.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles (with discussion). section 9 (translated). reprinted ed. *Statistical Science*, 5:465–472, 1923.

J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection with discussion. *Journal of the Royal Statistical Society*, 97:558–625, 1934.

J. Neyman. Statistical problems in agricultural experimentation (with discussion). *Supplement to the Journal of the Royal Statistical Society*, 2:107–180, 1935.

G. E. Noether. On a theorem by Wald and Wolfowitz. *The Annals of Mathematical Statistics*, 20: 455–458, 1949.

E. B. Page. Ordered hypotheses for multiple treatments: A significance test for linear ranks. *Journal of the American Statistical Association*, 58:216–230, 1963.

J. Robinson. A converse to a combinatorial limit theorem. *The Annals of Mathematical Statistics*, 43:2053–2057, 1972.

P. R. Rosenbaum. Covariance adjustment in randomized experiments and observational studies. *Statistical Science*, 17:286–327, 2002a.

P. R. Rosenbaum. *Observational Studies*. New York: Springer, 2nd edition, 2002b.

P. R. Rosenbaum. Does a dose-response relationship reduce sensitivity to hidden bias? *Biostatistics*, 4:1–10, 2003.

P. R. Rosenbaum. Testing one hypothesis twice in observational studies. *Biometrika*, 99:763–774, 2012.

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

D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5:472–480, 1990.

C. Samii and P. M. Aronow. On equivalencies between design-based and regression-based variance estimators for randomized experiments. *Statistics and Probability Letters*, 82:365–370, 2012.

W. Schneller. A short proof of Motoo's combinatorial central limit theorem using Stein's method. *Probability Theory and Related Fields*, 78:249–252, 1988.

J. Splawa-Neyman. Contributions to the theory of small samples drawn from a finite population. *Biometrika*, 17:472–479, 1925.

A. W. Van der Vaart. *Asymptotic Statistics*. Cambridge: Cambridge University Press, 2000.

V. A. Vatutin and V. G. Mikhailov. Limit theorems for the number of empty cells in an equiprobable scheme for group allocation of particles. *Theory of Probability and Its Applications*, 27:734–743, 1982.

A. Wald and J. Wolfowitz. Statistical tests based on permutations of the observations. *The Annals of Mathematical Statistics*, 15:358–372, 1944.

Table 1: Four possible forms of the confidence sets for β, where $c_1 < c_2$ denote two roots of (7) if they exist.

$\widehat{\tau}_D^2 - \eta s_D^2$	Δ	form of the confidence set for β
> 0	< 0	empty set
> 0	> 0	$[c_1, c_2]$
< 0	≤ 0	the whole real line
< 0	> 0	$(-\infty, c_1] \bigcup [c_2, \infty)$