Review Article **Open Access**

Lei Shi* and Xinran Li

A technical delving into the design and analysis of randomized experiments

DOI: DOI, Received ..; revised ..; accepted ..

Abstract: Neyman [79]'s seminal work in 1923 has been a milestone in statistics over the century, which has motivated many fundamental statistical concepts and methodology. In this review, we delve into Neyman [79]'s groundbreaking contribution and offer technical insights into the design and analysis of randomized experiments. We shall review the basic setup of completely randomized experiments and the classical approaches for inferring the average treatment effects. We shall in particular review more efficient design and analysis of randomized experiment by utilizing pretreatment covariates, which move beyond Neyman's original work without involving any covariate. We then summarize several technical ingredients regarding randomizations and permutations that have been developed over the century, such as permutational central limit theorems and Berry–Esseen bounds, and elaborate how these technical results facilitate the understanding of randomized experiments. The discussion is also extended to other randomized experiments including rerandomization, stratified randomized experiments, matched pair experiments, cluster randomized experiments, etc.

Keywords: Causal inference, Randomized experiments, Permutation, Central limit theorem, Berry-Esseen bound

MSC: 62K15, 62J05, 62G05

1 A review of Neyman [79]'s proposal and its influence

Neyman [79]'s seminal work in 1923 has been a cornerstone in the field of statistics over the century. It has laid foundational principles that have significantly shaped multiple research areas such as causal inference, experimental design, and survey sampling. Its influence has been profound across a diverse range of applications, encompassing sectors like agriculture, economics, biomedical research, social science, and beyond.

The main purpose of Neyman [79] is the analysis of field experiments conducted in order to compare a number of crop varieties. Suppose there are m plots and ν varieties. Neyman [79] introduced the notion of *potential yield* of the k-th variety being applied to the i-th plot, which is denoted as U_{ik}, for $1 \le i \le m$ and $1 \le k \le \nu$. In Neyman's framework, the quantities $\{U_{ik}\}$ are fixed but unknown. The number

$$a_k = \frac{1}{m} \sum_{i=1}^{m} U_{ik}$$

is called "the best estimate" of the yield from the k-th variety on the field. Neyman [79] then used an urn model as a thought experiment to depict the framework of sampling from a finite population. The ν types of varieties are treated as ν urns, whereas m plots are regarded as m balls to be drawn from the urns without replacement. Then Neyman studied the properties (in particular the means and variances) of the sample averages across all varieties as well as their difference under the randomization distribution. This marks the pioneer effort for studying the difference-in-means estimator in modern terminology. Notably,

***Corresponding Author: Lei Shi:** University of California, Berkeley; E-mail: leishi@berkeley.edu
Xinran Li: The University of Chicago; E-mail: xinranli@uchicago.edu

he was able to "determine empirically that the difference of partial averages of the plots sampled shows a fair agreement with the Gaussian law distribution". This corrects the "common misunderstanding" at that time that inference can be performed only if the yields from different plots follow the Gaussian law. Combined with a conservative variance estimation strategy, he suggested a confidence interval for the true difference between two varieties based on normal approximation.

Neyman [79] offered a series of groundbreaking and foundational insights. Below we outline three key facets of Neyman [79]'s contributions.

The first contribution is the introduction of the potential outcome model. This model has since become a standard framework for illustrating possible experimental outcomes, as referenced in works such as [60, 61, 86, 112]. The potential outcome paradigm serves as an impeccable model for discussion of causation within randomized experiments. Within this framework, researchers pose and address causal questions by analyzing causal effects that are defined as comparisons between potential outcomes, which represent various hypothetical scenarios or states of the world. This framework also elegantly facilitates the representation of interference between units [55, 107, 119] and the prolong impacts of interventions [57, 71, 101]. Moreover, the importance of potential outcomes transcends experimental settings and is also profound in observational studies, as highlighted by Rubin [92].

The second contribution of Neyman [79] lies in that it further highlights the importance of physical randomization or random selection when conducting experiments or performing sampling. Randomization has been in the air since 1920s, as commented by Rubin [94] citing Student [105] and Fisher and Mackenzie [34] as references. Neyman [79] contributed to the randomization world by introducing the potential outcome model and describing a finite population inference framework for randomization. Within this framework, potential outcomes remain constant, and physical randomization emerges as the "reasoned basis" [33] for facilitating statistical testing and estimation [51, 58, 60, 91]. Moreover, the proposal of sampling without replacement also inspires the pursuit of the parallels and linkages between survey sampling and randomized experiments [31, 80, 81, 102].

The third contribution of Neyman [79] centers on the repeated sampling property of statistics over their non-null randomization distribution. This viewpoint offers a new perspective on randomization-based or design-based inference, distinguishing it from Fisher's focus on sharp null hypothesis of no causal effects for any units and finite-sample exact p-values [94]. Neyman [79] recognized from an empirical perspective that the asymptotic normality property holds under the described sampling scheme, without requiring the outcomes to come from a Gaussian law. Moreover, he proposed to estimate the variance of an estimator conservatively in expectation, which can further lead to a conservative confidence interval. These efforts built up the foundation for large-sample randomization-based inference in finite population.

Building upon the pioneering contribution of Neyman [79] in randomization-based inference, there have been many new developments in the design and analysis of randomized experiments. In the following sections, we shall first review the basic setup of completely randomized experiments (CREs) and the classical approaches for analysis. We then present several technical ingredients regarding randomizations or permutations, such as central limit theorems (CLTs) and Berry–Esseen bounds (BEBs), that were developed over the century, and elaborate how these results enhance and expand our understanding of the design and analysis of CREs. We also extend the discussion to other randomized experiments and permutation-related technical tools.

2 Design and analysis of completely randomized experiments

In this section we introduce the basic setup for the design and analysis of CREs. Section 2.1 discusses the setup of a simple treatment-control CRE as well as strategies for estimation and inference. The results are extended to a more general multi-level CRE. We then consider more efficient design and analysis of randomized experiments by incorporating pretreatment covariates. In particular, Section 2.3 presents several covariate-adjusted estimators, and Section 2.4 discusses rerandomization.

2.1 Basic design and analysis of completely randomized experiments

2.1.1 Treatment-control completely randomized experiment

We start by considering a treatment-control completely randomized experiment that enrolls N units, with N_1 units in the treatment arm and N_0 in the control arm. Let Z_i denote the treatment assignment indicator for the i-th unit, for $1 \leq i \leq N$. The treatment assignment status for the entire experiment is vectorized as $Z = (Z_1, \ldots, Z_N)$. Under complete randomization,

$$\mathbb{P}\{Z = (z_1, \ldots, z_N)\} = 1/\binom{N}{N_1}, \quad \text{for any } (z_1, \ldots, z_N) \in \{0,1\}^N \text{ with } \sum_{i=1}^N z_i = N_1, \ \sum_{i=1}^N (1 - z_i) = N_0.$$

The potential outcomes for the i-th unit are $(Y_i(1), Y_i(0))$. This is essentially a special case of Neyman [79]'s setup with two interventional arms. The more general notions of experimental units, treatment/control arms and potential outcomes presented here correspond to Neyman [79]'s notions of plots, varieties and potential yields.

Rubin [95] called the $N \times 2$ matrix of potential outcomes in Table 1 as the Science Table. The observed

Table 1. Science Table for Treatment-control CRE

i	$Y_i(0)$	$Y_i(1)$
1	$Y_1(0)$	$Y_1(1)$
$\vdots$	$\vdots$	$\vdots$
N	$Y_N(0)$	$Y_N(1)$

outcome for the i-th unit is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$, one of the two potential outcomes. Importantly, the potential outcomes are fixed and the randomness comes merely from the random allocation of the treatment, reflected by the random vector $\boldsymbol{Z}$. Scheffé [97, Chapter 9] called it the *randomization model*. Under this model, it is conventional to call the resulting inference as *randomization-based inference*, *design-based inference* or *finite population inference*. It has become increasingly popular in both theory and practice [e.g. 2, 24, 26, 36, 39, 40, 42, 50, 58, 60, 68, 90, 91]. Define further the following finite-population mean and variance of potential outcomes for each arm, which are essentially summaries of the science table in Table 1:

$$\overline{Y}(0) = \frac{1}{N}\sum_{i=1}^N Y_i(0), \quad \overline{Y}(1) = \frac{1}{N}\sum_{i=1}^N Y_i(1);$$
$$S^2(0) = \frac{1}{N-1}\sum_{i=1}^N (Y_i(0) - \overline{Y}(0))^2, \quad S^2(1) = \frac{1}{N-1}\sum_{i=1}^N (Y_i(1) - \overline{Y}(1))^2. \tag{1}$$

Under the potential outcome framework, the i-th unit has individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$, for $1 \leq i \leq N$. The average treatment effect (ATE) over all units is then defined as:

$$\tau = \frac{1}{N}\sum_{i=1}^N \tau_i = \overline{Y}(1) - \overline{Y}(0).$$

Neyman [79] proposed to estimate the ATE τ by the difference-in-means estimator:

$$\hat{\tau} = \hat{Y}(1) - \hat{Y}(0), \text{ where } \hat{Y}(z) = \frac{1}{N_z}\sum_{i=1}^N Y_i \mathbf{1}\{Z_i = z\}. \tag{2}$$

He proved that $\hat{\tau}$ is an unbiased estimator for τ, i.e., $\mathbb{E}\{\hat{\tau}\} = \overline{Y}(1) - \overline{Y}(0) = \tau$, with true variance

$$\mathrm{Var}\{\hat{\tau}\} = \frac{1}{N_1}S^2(1) + \frac{1}{N_0}S^2(0) - \frac{1}{N}S^2(\tau),$$

where the variances $S^2(0)$ and $S^2(1)$ are defined in (1) and $S^2(\tau)$ is the variance of the individual treatment effects

$$S^2(\tau) = \frac{1}{N-1}\sum_{i=1}^{N}(\tau_i - \tau)^2. \tag{3}$$

Due to the fact that we are never able to jointly observe the two potential outcomes for any unit, the variance of individual effects in (3) is generally not estimable based on the observed data. Neyman [79] proposed the following variance estimator:

$$\hat{V} = \frac{1}{N_1}\hat{S}^2(1) + \frac{1}{N_0}\hat{S}^2(0), \text{ where } \hat{S}^2(z) = \frac{1}{N_z - 1}\sum_{i=1}^{N}(Y_i - \hat{Y}(z))^2\mathbf{1}\{Z_i = z\}. \tag{4}$$

which essentially circumvents the problem by dropping the unestimable component regarding $S^2(\tau)$. The variance estimator in (4) has expectation

$$\mathbb{E}\left\{\hat{V}\right\} = \frac{1}{N_1}S^2(1) + \frac{1}{N_0}S^2(0) \geq \mathrm{Var}\{\hat{\tau}\},$$

which suggests that $\hat{V}$ is in general not unbiased but conservative. A level-α confidence interval is then given by

$$\left[\hat{\tau} - z_{\alpha/2}\sqrt{\hat{V}},\ \hat{\tau} + z_{\alpha/2}\sqrt{\hat{V}}\right], \tag{5}$$

where $z_{\alpha/2}$ is the $\alpha/2$ upper quantile of a standard normal distribution. In Sections 3 and 4 we will discuss more technical results for the asymptotic validity of the confidence interval in (5).

Remark 1. *Neyman [79]'s approach can also be used to test the following null hypothesis:*

$$H_{0\mathrm{N}} : \tau = \overline{Y}(1) - \overline{Y}(0) = 0,$$

which is often called the weak null hypothesis [115]. In contrast, Fisher [33] proposed to test the following null hypothesis:

$$H_{0\mathrm{F}} : Y_i(1) = Y_i(0) \text{ for all units } i = 1, \ldots, N, \tag{6}$$

which is called the sharp null hypothesis by Rubin [93] or the strong null hypothesis by Wu and Ding [115]. Under (6)*, one can impute the unobserved potential outcomes and perform Fisher's randomization test to deliver finite sample exact inference [33]. Fisher's test has the advantage of being finite-sample valid, while Neyman's requires large-sample approximation. However, the tested null hypothesis by Fisher is more restrictive than that by Neyman, since the sharp null requires speculation of all individual treatment effects. A unification of both perspectives is discussed in [23, 29, 116]. As a side note, Neyman's asymptotic results can also help ease the computation for Fisher's null.*

Remark 2. *For analysis, practitioners usually prefer regression-based inference for the average causal effect. The standard approach is to run the ordinary least squares (OLS) of the outcomes on the treatment indicators with an intercept:*

$$(\hat{\gamma}, \hat{\tau}) = \underset{\gamma,\tau\in\mathbb{R}}{\arg\min}\sum_{i=1}^{N}(Y_i - \gamma - Z_i\tau)^2. \tag{7}$$

As implicitly written in (7)*, the point estimator from the OLS for the treatment effect is identical to the difference-in-means estimator in* (2)*. However, the usual variance estimation based on the OLS usually fails (in the sense of either underestimating or overestimating the truth by possibly a quite large factor), due to heteroskedasticity in potential outcomes. Instead, one can use the Eicker-Huber-White (EHW) variance estimator to obtain a robust estimation:*

$$\hat{V}_{\text{EHW}} = \frac{\hat{S}^2(1)}{N_1}\frac{N_1-1}{N_1} + \frac{\hat{S}^2(0)}{N_0}\frac{N_0-1}{N_0},$$

which is asymptotically equivalent to $\hat{V}$ *in* (4)*. Alternatively, the so-called HC2 variant of the EHW robust variance estimator is identical to* $\hat{V}$*.*

2.2 Multi-level completely randomized experiments

Many efforts have been devoted to extending the treatment-control CRE to multi-level scenarios, which caters for many practical problems and designs such as (fractional) factorial experiments [26, 114], conjoint analysis [43, 44], partially nested experiment [7, 48], etc.

In a multi-level randomized experiment, there are N units and Q treatment arms, where the number of units under treatment q equals N_q, with $\sum_{q=1}^{Q} N_q = N$. Corresponding to treatment level q, unit i has the potential outcome $Y_i(q)$, where $i = 1, \ldots, N$ and $q = 1, \ldots, Q$. Despite its simplicity, the completely randomized experiment has been widely used in practice and has generated rich theoretical results. Definition 1 below characterizes the joint distribution of $Z = (Z_1, \ldots, Z_N)$ under complete randomization, where $Z_i \in \{1, \ldots, Q\}$ is the treatment indicator for unit i.

Definition 1 (Complete randomization)**.** *Fix integers* $N_1, \ldots, N_Q$ *with* $\sum_{q=1}^{Q} N_q = N$*. The treatment vector* Z *is uniformly distributed over* $\mathcal{Z} \equiv \{z \in \{1, 2, \ldots, Q\}^N : \sum_{i=1}^{N} \mathbf{1}\{z_i = q\} = N_q,$ *for* $1 \le q \le Q\}$*.*

Mathematically, Definition 1 implies that $\mathbb{P}(Z = z) = N_1! \cdots N_Q!/N!$ for all possible values of z in $\mathcal{Z}$. Computationally, Definition 1 implies that Z is from a random permutation of N_1 1's, ..., N_Q Q's. The observed outcome is $Y_i = \sum_{q=1}^{Q} Y_i(q)\mathbf{1}\{Z_i = q\}$ for each unit i.

Similar to the two-arm setting discussed in Section 2.1, in Neyman [79]'s framework, all potential outcomes are fixed and only the treatment indicators are random according to Definition 1.

Table 2. Science Table for Multi-level CRE

i	$Y_i(1)$	$Y_i(2)$	...	$Y_i(Q)$
1	$Y_1(1)$	$Y_1(2)$	...	$Y_1(Q)$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
N	$Y_N(1)$	$Y_N(2)$	...	$Y_N(Q)$

Neyman's weak null hypothesis is also extended to the multi-level experiments [115]. Let $F \in \mathbb{R}^{Q \times H}$ be a full column rank contrast matrix, i.e. $F^\top 1_Q = 0_H$. Then a general set of individual treatment effects can be defined as

$$\tau_i = F^\top Y_i(\cdot),$$

with the average effect

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i = F^\top \overline{Y}(\cdot). \tag{8}$$

Neyman's weak null can be defined as follows:

$$H_{0\mathrm{N}} : \tau = F^\top \overline{Y}(\cdot) = 0.$$

$F = (1, -1)^\top$ corresponds to the ATE in the treatment-control setting. Moreover, using the plug-in estimator

$$\hat{\tau} = F^\top \hat{Y}(\cdot), \tag{9}$$

where $\hat{Y}(\cdot) = (\hat{Y}(1), \ldots, \hat{Y}(Q))^\top$ is the aggregated sample averages over each treatment arm, with $\hat{Y}(q) = N_q^{-1} \sum_{i=1}^N Y_i \mathbf{1}\{Z_i = q\}$. (9) has variance [64]

$$\mathrm{Var}\{\hat{\tau}\} = F^\top \mathrm{Diag}\left\{\frac{1}{N_q} S(q,q)\right\}_{q=1}^{Q} F - \frac{1}{N} F^\top S F, \tag{10}$$

where $S \in \mathbb{R}^{Q \times Q}$ is a covariance matrix for the potential outcomes with entries

$$S(q,q') = \frac{1}{N-1} \sum_{i=1}^N (Y_i(q) - \overline{Y}(q))(Y_i(q') - \overline{Y}(q')), \quad q, q' \in [Q].$$

A variance estimator for (10) is

$$\hat{V} = F^\top \mathrm{Diag}\left\{\frac{1}{N_q} \hat{S}(q,q)\right\}_{q=1}^{Q} F, \tag{11}$$

where $\hat{S}(q,q)$ is the sample variance within treatment level q:

$$\hat{S}(q,q) = \frac{1}{N_q - 1} \sum_{i=1}^N (Y_i - \hat{Y}(q))^2 \mathbf{1}\{Z_i = q\}.$$

Using (9) and (11), a Wald-type confidence region for τ is given by

$$\left\{\tau : (\hat{\tau} - \tau)^\top \hat{V}^{-1} (\hat{\tau} - \tau) \le q_{H,\alpha}\right\}, \tag{12}$$

where $q_{H,\alpha}$ is the upper-α quantile of the χ^2_H distribution. (12) is proved to be asymptotically valid under mild regularity conditions. More details are deferred to Section 3 and 4.

Similar to the treatment-control case, we can perform analysis with the regression-based approach. Zhao and Ding [117] studied general regression adjustment results in multi-level experiments.

2.3 Covariate adjustment

In many randomized experiments, there are pre-treatment covariates $X_1, \ldots, X_N$ for the N units, where X_i's are encoded as vectors in $\mathbb{R}^p$. A natural question is: whether adjusting for covariates is helpful for inference? The problem is nontrivial in several aspects: (i) the true relation between outcomes and covariates is usually unknown; (ii) the potential outcomes under different treatment levels are in general heterogeneous. Many research works explored covariate adjustment from both practical and theoretical perspectives. It has become a standard practice to use a model-assisted method for covariate adjustment to gain efficiency for inference even while being robust to model misspecification [68].

2.3.1 Fisher's analysis of covariance (ANCOVA)

Historically, Fisher [32] proposed to use the analysis of covariance (ANCOVA) to improve estimation efficiency. This remains a standard strategy in many fields. He suggested running the OLS of Y_i on (Z_i, X_i)

and obtaining the coefficient of Z_i as an estimator for τ. Mathematically, let $\overline{X}$ be the mean of the covariates: $\overline{X} = N^{-1}\sum_{i=1}^{N} X_i$. Fisher's ANCOVA estimator $\hat{\tau}$ is given by the following OLS output:

$$(\hat{\tau}_{\mathrm{F}}, \hat{\alpha}_{\mathrm{F}}, \hat{\beta}_{\mathrm{F}}) = \underset{\alpha,\tau\in\mathbb{R},\beta\in\mathbb{R}^p}{\arg\min} \frac{1}{2N}\sum_{i=1}^{N}\{Y_i - \alpha - Z_i\tau - (X_i - \overline{X})^\top\beta\}^2, \tag{13}$$

noting that the centering of covariates in (13) is not necessary and will not affect the OLS estimator $\hat{\tau}_{\mathrm{F}}$.

Freedman [39, 40] showed some negative results for Fisher's ANCOVA estimator. First, it can be biased in finite sample. Second, the asymptotic variance of $\hat{\tau}_{\mathrm{F}}$ can be even larger than the simple difference-in-means estimator $\hat{\tau}$. Third, the standard error estimator from OLS is not consistent for the true standard error of $\hat{\tau}_{\mathrm{F}}$ under the CRE.

2.3.2 Lin's estimator

In response to Freedman's negative findings, Lin [68] proposed a remedy, which is called "Lin's estimator" nowadays. Concretely speaking, he proposed to run OLS of Y_i on Z_i and X_i as well as their interaction term $Z_i \times X_i$:

$$(\hat{\tau}_{\mathrm{L}}, \hat{\alpha}_{\mathrm{L}}, \hat{\beta}_{\mathrm{L}}, \hat{\eta}_{\mathrm{L}}) = \underset{\alpha,\tau\in\mathbb{R},\beta,\eta\in\mathbb{R}^p}{\arg\min} \frac{1}{2N}\sum_{i=1}^{N}\{Y_i - \alpha - Z_i\tau - (X_i - \overline{X})^\top\beta - Z_i \times (X_i - \overline{X})^\top\eta\}^2. \tag{14}$$

Importantly, unlike (13), the centering of covariates here is critical since it will affect the OLS estimator $\hat{\tau}_{\mathrm{L}}$.

Lin's estimator demonstrates several benefits. First, it is consistent when the sample size N goes to infinity. Second, the asymptotic variance of $\hat{\tau}_{\mathrm{L}}$ is no larger than that of both $\hat{\tau}$ and $\hat{\tau}_{\mathrm{F}}$. Third, the EHW standard error estimator for (14) is asymptotically conservative for the true variance of $\hat{\tau}_{\mathrm{L}}$. As a side note, the EHW standard error estimator for (13) is also asymptotically conservative for the true variance of $\hat{\tau}_{\mathrm{F}}$. See Lin [68] for a more formal presentation of the theoretical results.

Besides the regression proposal, a second perspective for understanding Lin's estimator is based on minimizing the true or estimated variance of linearly adjusted estimators [21]. Consider the following class of linearly covariate-adjusted estimators:

$$\begin{aligned}\hat{\tau}(\beta_1, \beta_0) &= N_1^{-1}\sum_{i=1}^{N} Z_i\{Y_i - (X_i - \overline{X})^\top\beta_1\} - N_0^{-1}\sum_{i=1}^{N}(1 - Z_i)\{Y_i - (X_i - \overline{X})^\top\beta_0\} \\ &= \{\hat{Y}(1) - (\hat{X}(1) - \overline{X})^\top\beta_1\} - \{\hat{Y}(0) - (\hat{X}(0) - \overline{X})^\top\beta_0\} = \hat{\tau} - \delta^\top\hat{\tau}_X,\end{aligned} \tag{15}$$

where $\hat{X}(1)$ and $\hat{X}(0)$ denote the average covariates in treatment and control groups, $\hat{\tau}_X \equiv \hat{X}(1) - \hat{X}(0)$ denotes the difference-in-means of covariates, and $\delta = N_0/N \cdot \beta_1 + N_1/N \cdot \beta_0$ is a weighted average of the two linear adjustment coefficients. Obviously, the true variance of the covariate-adjusted estimator in (15) is minimized when δ is the least squares coefficient from the linear projection of the difference-in-means estimator $\hat{\tau}$ on the difference-in-means of covariates $\hat{\tau}_X$ under the CRE. Li and Ding [64] showed that this is further achieved when β_1 and β_0 are the least squares coefficients from the linear projection of treatment and control potential outcomes on covariates, respectively. Moreover, since the potential outcomes cannot be fully observed, we can estimate the least squares coefficients by their sample analogues $\hat{\beta}_1$ and $\hat{\beta}_0$, which are the least squares coefficients from the linear projection of observed outcomes on covariates in treatment and control groups, respectively. The resulting covariate-adjusted estimator $\hat{\tau} - \hat{\delta}^\top\hat{\tau}_X$ with $\hat{\delta} = N_0/N \cdot \hat{\beta}_1 + N_1/N \cdot \hat{\beta}_0$ is actually identical to Lin's estimator.

We consider then the estimated variance for the covariate-adjusted estimator in (15). We can essentially view the covariate-adjusted estimator as the difference-in-means estimator but with the adjusted potential outcomes $Y_i(1) - (X_i - \overline{X})^\top\beta_1$ and $Y_i(0) - (X_i - \overline{X})^\top\beta_0$. From the discussion in Section 2.1, a conservative

variance estimator for (15) can be

$$\hat{V}(\beta_1,\beta_0)=N_1^{-1}\sum_{i=1}^{N}Z_i\{Y_i-\hat{\gamma}_1-(X_i-\overline{X})^\top\beta_1\}^2+N_0^{-1}\sum_{i=1}^{N}(1-Z_i)\{Y_i-\hat{\gamma}_0-(X_i-\overline{X})^\top\beta_0\}^2. \quad (16)$$

This formulation suggests minimizing the variance estimator $\hat{V}(\beta_1,\beta_0)$ to obtain a plug-in estimator for β_1 and β_0, which is equivalent to solving the following two regression problems for treated and control groups, respectively:

$$\min_{\gamma_1,\beta_1}\sum_{i=1}^{N}Z_i\{Y_i-\gamma_1-(X_i-\overline{X})^\top\beta_1\}^2 \text{ and } \min_{\gamma_0,\beta_0}\sum_{i=1}^{N}(1-Z_i)\{Y_i-\gamma_0-(X_i-\overline{X})^\top\beta_0\}^2. \quad (17)$$

It is not difficult to see that the least squares estimators for β_1 and β_0 from (17) are actually $\hat{\beta}_1$ and $\hat{\beta}_0$ defined before. Consequently, the resulting covariate-adjusted estimator $\hat{V}(\hat{\beta}_1,\hat{\beta}_0)$ is equivalent to Lin's estimator.

From the above, the Lin's estimator not only achieves the minimum true variance among all linearly covariate-adjusted estimators in (15), but also achieves the minimum estimated variance when we use the conservative variance estimator of form (16). A subtle issue here is that Lin's estimator uses estimated coefficients rather than fixed one. With the technical tools discussed later, we can prove that the difference with Lin's estimator and the one with the oracle adjustment coefficients are asymptotically equivalent; see, e.g., Li and Ding [64].

2.3.3 Further extensions

There are a variety of extensions of covariate adjustment beyond treatment-control completely randomized experiments.

First, it is natural to consider generalization to multiple treatment levels. Lu [72] studied covariate adjustment in 2^K factorial designs by extending (17) to multi-level settings. Zhao and Ding [117] considered covariate adjustment in general multi-level experiments and made comprehensive comparison among the unadjusted estimator, Fisher's ANCOVA and Lin's estimator. The generalization of Fisher's ANCOVA is given by the following *additive treatment regression*:

$$(\hat{\gamma}_{1,\mathrm{F}},\ldots,\hat{\gamma}_{Q,\mathrm{F}},\hat{\eta}_{\mathrm{F}})=\mathop{\arg\min}_{\gamma_1,\ldots,\gamma_Q,\eta}\sum_{i=1}^{N}\left\{Y_i-\sum_{q=1}^{Q}\gamma_q\mathbf{1}\{Z_i=q\}-(X_i-\overline{X})^\top\eta\right\}^2.$$

Meanwhile, Lin's estimator can be generalized from either the regression with interaction perspective (14) or the variance estimation minimization perspective (17). Here we present the former one, which applies the following fully interacted treatment-based regressions:

$$(\hat{\gamma}_{1,\mathrm{L}},\ldots,\hat{\gamma}_{Q,\mathrm{L}},\hat{\eta}_{1,\mathrm{L}},\ldots,\hat{\eta}_{Q,\mathrm{L}})=\mathop{\arg\min}_{\gamma_1,\ldots,\gamma_Q,\eta_1,\ldots,\eta_Q}\sum_{i=1}^{N}\left\{Y_i-\sum_{q=1}^{Q}\gamma_q\mathbf{1}\{Z_i=q\}-\sum_{q=1}^{Q}\mathbf{1}\{Z_i=q\}(X_i-\overline{X})^\top\eta_q\right\}^2.$$

With the vectorized slope estimates $\hat{\gamma}_*=(\hat{\gamma}_{1,*},\ldots,\hat{\gamma}_{Q,*})^\top$, where $*=\mathrm{F},\mathrm{L}$, an estimator for the target effects (8) is given by the plug-in estimator

$$\hat{\tau}_*=F^\top\hat{\gamma}_*,\quad *=\mathrm{F},\mathrm{L}.$$

Besides, we can obtain EHW variance estimators $\hat{V}_{\mathrm{EHW},*}$, which is conservative in large samples. In multi-level CRE, Lin's estimator is also guaranteed to be at least as efficient as Fisher's ANCOVA and Neyman's difference-in-means estimator.

Second, covariate adjustment has also been discussed in treatment-control trials when the dimension of the covariates is diverging or high-dimensional. For example, Lei and Ding [62] proposed a debiased estimator that allows the dimension of the covariates p to grow with the sample size N. In high-dimensional regimes, Bloniarz et al. [11] considered LASSO estimator for covariate adjustment and showed its benefits for inference.

Third, some works explored the other variants of Lin's estimators. For example, Zhao and Ding [117] studied restricted least squares (RLS) and established for the first time its properties for inferring ATE from the design-based perspective. Guo and Basse [42] considered generalized oaxaca blinder estimators and extend the covariate adjustment framework from linear models to nonlinear ones; see also [22].

2.4 Rerandomization

Neyman [79] focuses on the completely randomized experiments, which can balance all potential confounding factors, no matter observed or unobserved, on average and justifies the intuitive difference-in-means estimator for estimating the average treatment effect. In practice, in the design stage of an experiment, we often have access to a (rich) set of pretreatment covariates, and it has been a routine to check whether these covariates are balanced between different treatment groups. As commented by Morgan and Rubin [77], for a realized treatment allocation, the covariates are likely to be imbalance; for example, with 10 independent covariates, at least one of the t-statistics for checking the imbalance of these covariates will exceed 2 with probability about 40%. It is then natural to incorporate the pretreatment covariate information into the design, aiming to get more balanced treatment groups as well as more efficient inference for treatment effects.

Blocking is a classical and popular approach that can balance a few discrete covariates, but its implementation is not obvious once we have many continuous covariates. Rerandomization, a design recently formally proposed by Morgan and Rubin [77], provides a general approach to improve covariate imbalance, although its idea has existed for a long time in the literature and dates back to many earlier works [15, 25, 74, 96, 103, 113]. Under a general rerandomization design, for a randomly drawn treatment allocation, we will check the covariate balance between different treatment groups and see whether it satisfies a prespecified covariate balance criterion; if the balance criterion is met, we proceed to the actual experiment with this treatment allocation; otherwise, we redraw the treatment allocation and will keep redrawing until the balance criterion is met.

Although the balance criterion can be general, in the context of a treatment-control experiment, Morgan and Rubin [77] suggested a balance criterion based on the Mahalanobis distance:

$$M = \hat{\tau}_X^\top \{\mathrm{Cov}(\hat{\tau}_X)\}^{-1} \hat{\tau}_X = \frac{N_1 N_0}{N} \hat{\tau}_X^\top (S_X^2)^{-1} \hat{\tau}_X,$$

recalling that N_1 and N_0 are the treated and control group sizes, $\hat{\tau}_X$ is the difference-in-means of covariates defined as in Section 2.3.2, and S_X^2 is the finite population covariance matrix of covariates defined as follows:

$$S_X^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \overline{X})(X_i - \overline{X})^\top.$$

Under rerandomization using the Mahalanobis distance, denote by ReM, we will repeatedly draw random treatment assignment from the CRE until getting an acceptable one with the corresponding Mahalanobis distance bounded by a prespecified threshold a.

Importantly, the analysis for rerandomization needs to take into account the selection step in its design. This is often ignored in practice, and rerandomization is often analyzed as if it was a CRE. Morgan and Rubin [77] proposed randomization tests for sharp null hypotheses, employing assignments generated randomly in accordance with the rerandomization protocol. More recently, Li et al. [66] conducted Neyman-type large-sample inference for rerandomization, considering also the intuitive difference-in-means estimator. They demonstrated that, asymptotically, the difference-in-means estimator is more concentrated

around the true average treatment effect with smaller asymptotic variance and shorter asymptotic quantile ranges, and proposed accurate confidence intervals for the average effect, which are always shorter than Neyman's intervals for the CRE while remaining valid asymptotically.

In recent years, rerandomization has been extended to more general experiments, such as factorial experiments [14, 67] and blocked experiments [59, 110], and can also be combined with covariate adjustment discussed in Section 2.3 [65].

3 Permutational central limit theorems

In this section we focus on the technical aspect of completely randomized experiments. The main question to answer is how to deliver valid inference with different estimators. Permutational/combinatorial CLTs and BEBs are core to the technical development of randomization-based inference. We summarize the theoretical results regarding permutational CLTs and BEBs in Sections 3 and 4, respectively, and discuss their application in analyzing randomized experiments.

3.1 Sample average from simple random sampling

We start with the simple random sampling (SRS) from a finite population [21]. Let $\{a_N(i)\}_{i=1}^N$ be a sequence of real numbers. Suppose we sample N_1 elements *without replacement* from the population and use a binary variable Z_i to indicate the sampling status of the i-th element, i.e., $Z_i = 1$ indicates $a_N(i)$ being sampled while $Z_i = 0$ not sampled. Write $N_0 = N - N_1$. Consider the sample average obtained from the above procedure:

$$\Gamma = \frac{1}{N_1}\sum_{i=1}^{N} a_N(i)\mathbf{1}\left\{Z_i = 1\right\}. \tag{18}$$

Γ has mean and variance

$$\mathbb{E}\left\{\Gamma\right\} = \overline{a}_N, \quad V_N \triangleq \mathrm{Var}\left\{\Gamma\right\} = (\frac{1}{N_1} - \frac{1}{N})S_N^2,$$

where

$$\overline{a}_N \triangleq \frac{1}{N}\sum_{i=1}^{N} a_N(i), \quad S_N^2 \triangleq \frac{1}{N-1}\sum_{i=1}^{N}(a_N(i) - \overline{a}_N)^2.$$

One fundamental technical question is to establish central limit theorems for Γ to characterize its asymptotic distribution. Paul and Rényi [84] established the following CLT for (18):

Proposition 1. *If for any $\epsilon > 0$,*

$$\lim_{N\to\infty} \frac{\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2 \mathbf{1}\left\{|a_N(i) - \overline{a}_N| > \epsilon N_1 \sqrt{V_N}\right\}}{\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2} \to 0, \tag{19}$$

then as $N \to \infty$,

$$\frac{\Gamma - \mathbb{E}\left\{\Gamma\right\}}{\sqrt{V_N}} \rightsquigarrow \mathcal{N}(0,1).$$

Hájek [45] further proved that Condition (19) is not only sufficient but also necessary provided that $N_1, N_0 \to \infty$. Moreover, Theorem 1 covers some other works on finite population CLTs. For example, Madow [75] proved asymptotic normality under the moment condition:

$$\frac{N^{-1}\sum_{i=1}^N (a_N(i) - \overline{a}_N)^r}{\{N^{-1}\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2\}^{r/2}} = O(1), \text{ for all integers } r > 2.$$

which is stronger than (19). David [27] established a CLT for the hypergeometric distribution, which is a special case of Madow [75] thus also stronger than (19). Li and Ding [64, Section 2.1] also provided a thorough exposition of CLT under the simple random sampling scheme with a sufficient condition based on the maximum squared distance.

3.2 Simple linear rank statistics

The sample average in (18) from a SRS is a special case of a more general type of permutational statistics, called *simple linear rank statistics*. Formally, let $\{a_N(i)\}_{i=1}^N$ and $\{b_N(i)\}_{i=1}^N$ be two sequences of real numbers. A simple linear rank statistics is defined as

$$\Gamma = \sum_{i=1}^N a_N(i) b_N(\pi(i)), \tag{20}$$

which has mean and variance

$$\mathbb{E}\{\Gamma\} = N\overline{a}_N \cdot \overline{b}_N, \quad V_N \triangleq \mathrm{Var}\{\Gamma\} = \frac{1}{N-1} \cdot \sum_{i=1}^N (a_N(i) - \overline{a}_N)^2 \cdot \sum_{i=1}^N (b_N(i) - \overline{b}_N)^2.$$

The statistic in (20) has been studied by many researchers. Wald and Wolfowitz [109] established CLT under the following condition: for all integers $r > 2$,

$$\frac{N^{-1}\sum_{i=1}^N (a_N(i) - \overline{a}_N)^r}{\{N^{-1}\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2\}^{r/2}} = O(1), \quad \frac{N^{-1}\sum_{i=1}^N (b_N(i) - \overline{b}_N)^r}{\{N^{-1}\sum_{i=1}^N (b_N(i) - \overline{b}_N)^2\}^{r/2}} = O(1).$$

Noether [82] proved CLT under the following slightly weakened condition than Wald and Wolfowitz [109]: for all integers $r > 2$,

$$\frac{N^{-1}\sum_{i=1}^N (a_N(i) - \overline{a}_N)^r}{\{N^{-1}\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2\}^{r/2}} = O(1), \quad \frac{\sum_{i=1}^N (b_N(i) - \overline{b}_N)^r}{\{\sum_{i=1}^N (b_N(i) - \overline{b}_N)^2\}^{r/2}} = o(1),$$

which, however, is not symmetric for $a_N(i)$'s and $b_N(i)$'s. Hoeffding [53] further proved CLT under a weaker and symmetric condition: for all integers $r > 2$,

$$N^{\frac{1}{2}r-1} \frac{\sum_{i=1}^N (a_N(i) - \overline{a}_N)^r}{\{\sum_{i=1}^N (a_N(i) - \overline{a}_N)^2\}^{r/2}} \cdot \frac{\sum_{i=1}^N (b_N(i) - \overline{b}_N)^r}{\{\sum_{i=1}^N (b_N(i) - \overline{b}_N)^2\}^{r/2}} = o(1). \tag{21}$$

Motoo [78] proved that CLT holds under an even weaker Lindeberg type condition:

Proposition 2. *Suppose that for any $\epsilon > 0$,*

$$\lim_{N\to\infty} \frac{\sum_{i,j=1}^N (a_N(i) - \overline{a}_N)^2 (b_N(j) - \overline{b}_N)^2 \mathbf{1}\left\{|(a_N(i) - \overline{a}_N)(b_N(j) - \overline{b}_N)| > \epsilon\sqrt{V_N}\right\}}{\sum_{i,j=1}^N (a_N(i) - \overline{a}_N)^2 (b_N(j) - \overline{b}_N)^2} = 0. \tag{22}$$

Then

$$\frac{\Gamma - \mathbb{E}\{\Gamma\}}{\sqrt{\mathrm{Var}\{\Gamma\}}} \rightsquigarrow \mathcal{N}(0,1).$$

Hájek [46] proved further that Condition (22) is not only sufficient but also necessary, and presented a comprehensive comparison of the CLT conditions introduced in the literature. There are also several multidimensional extensions based on the Cramér-Wold device. See for example Fraser [37], Hájek [46, Section 7], DiCiccio and Romano [28, Lemma S.3.3], etc.

3.3 General univariate linear permutational statistics

Taking one step further from the simple linear rank statistics, the permutational central limit theorems are proposed for the following *linear permutational statistic*:

$$\Gamma = \sum_{i=1}^{N} M_N(i, \pi(i)), \tag{23}$$

where $\{M_N(i,j)\}_{i,j\in[N]}$ is a matrix in $\mathbb{R}^{N\times N}$. Hoeffding [53] computed the mean and variance of (23):

$$\mathbb{E}\{\Gamma\} = \frac{1}{N}\sum_{i,j=1}^{N} M_N(i,j), \quad V_N = \mathrm{Var}\{\Gamma\} = \frac{1}{N-1}\sum_{i,j=1}^{N} \tilde{M}_N(i,j)^2.$$

Here $\tilde{M}_N(i,j)$ is the centered array based on the following rule:

$$\tilde{M}_N(i,j) = M_N(i,j) - \frac{1}{N}M_N(i,+) - \frac{1}{N}M_N(+,j) + \frac{1}{N^2}M_N(+,+). \tag{24}$$

where "+" means summation over the corresponding index.

Moreover, Hoeffding [53] showed that the asymptotic normality of Γ in (23) holds under the following condition:

$$\lim_{N\to\infty} \frac{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_N(i,j)^r}{\{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_N(i,j)^2\}^{r/2}} = 0, \quad \text{for all integers } r > 2. \tag{25}$$

Condition (25) is equivalent to Condition (21) in the simple linear rank statistics setting. A more compact sufficient condition for (25) is also provided in Hoeffding [53]:

$$\lim_{N\to\infty} \frac{\max_{i,j\in[N]} \tilde{M}_N(i,j)^2}{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_N(i,j)^2} = 0. \tag{26}$$

Motoo [78] weakened Hoeffding [53]'s condition (25) to the following Lindeberg type condition:

Proposition 3 (Main theorem of Motoo [78])**.** *Suppose for any $\epsilon > 0$,*

$$\lim_{N\to\infty} \frac{\sum_{i,j=1}^{N} \tilde{M}_N(i,j)^2 \mathbf{1}\left\{|\tilde{M}_N(i,j)| > \epsilon\sqrt{V_N}\right\}}{\sum_{i,j=1}^{N} \tilde{M}_N(i,j)^2} = 0.$$

Then

$$\frac{\Gamma - \mathbb{E}\{\Gamma\}}{\sqrt{\mathrm{Var}\{\Gamma\}}} \rightsquigarrow \mathcal{N}(0,1).$$

3.4 General multivariate linear permutational statistics

We now discuss the generalization of (25) to a multivariate case. Concretely, define the multivariate linear permutational statistics:

$$\Gamma = (\Gamma_1, \ldots, \Gamma_H)^\top, \quad \Gamma_h = \sum_{i=1}^{N} M_{N,h}(i, \pi(i)), \tag{27}$$

where $\{M_{N,h}(i,j)\}_{i,j\in[N]}$, $h = 1, \ldots, H$ are H matrices in $\mathbb{R}^{N\times N}$. Shi and Ding [99, Appendix A.1.] presented many basic facts about (27), including its mean and covariance calculation and its standardization. Fraser [37] extended Hoeffding [53] to the multi-dimensional setting by applying the Cramér-Wold device

to establish a multivariate central limit theorem. More concretely, define the centered version of $\tilde{M}_{N,h}$ in a similar way as (24). Fraser [37] proposed the following condition for CLT as an extension to (25):

$$\lim_{N\to\infty} \frac{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_{N,h}(i,j)^r}{\{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_{N,h}(i,j)^2\}^{r/2}} = 0, \quad \text{for all integers } r > 2 \text{ and } h \in [H]. \tag{28}$$

Similarly, Fraser [37] also provided a sufficient condition for (28):

$$\lim_{N\to\infty} \frac{\max_{i,j\in[N]} \tilde{M}_{N,h}(i,j)^2}{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_{N,h}(i,j)^2} = 0, \quad \text{for } h \in [H]. \tag{29}$$

The condition in (29) is further utilized by Li and Ding [64] to build asymptotic normality results for many permutational related problems.

3.5 Application of permutational CLT in randomization-based inference

In this subsection, we collect several theoretical arguments in randomization inference that applies permutational CLTs to deliver technical justification for studying ATE.

3.5.1 Wald-type inference in CRE

For the confidence region (12), Li and Ding [64] proved the following result to justify its asymptotic validity under several regularity conditions:

Proposition 4 (Theorem 5 and Proposition 3 of Li and Ding [64])**.** *Let Q be fixed and N go to infinity. If the covariance matrix S have limiting values, N_q/N has positive limiting value, and*

$$\max_{1\le q\le Q} \max_{1\le i\le N} |Y_i(q) - \overline{Y}(q)|^2/N \to 0,$$

then the following conclusions hold:

(i) Asymptotic normality. $N\mathrm{Var}\{\hat{\tau}\}$ *has a limiting value V_∞, and*

$$\sqrt{N}(\hat{\tau} - \tau) \rightsquigarrow \mathcal{N}(0, V_\infty).$$

(ii) Variance estimation. *The sample variance $\hat{S}(q,q)$ is consistent for $S(q,q)$.*

(iii) Wald-type inference. *If the limit of $NF^\top \mathrm{Diag}\left\{N_q^{-1}S(q,q)\right\}F$ is nonsingular, then $\hat{V}$ is nonsingular with probability converging to one, and the Wald-type confidence region* (12) *has asymptotic coverage rate at least $1-\alpha$. Moreover, the asymptotic coverage rate equals $1-\alpha$ if and only if the causal effects are asymptotically additive, in the sense that $\lim_{N\to\infty} F^\top SF = 0$.*

We briefly comment on the technical details behind Proposition 4. Proposition 4(i) utilized the permutational CLT for general linear permutational statistics. In particular, Condition (26) and (29) are the main technical tools that motivate the bounded scale conditions on potential outcomes and justify the multivariate asymptotic normality of the estimators. Proposition 4(ii) applied the Chebyshev inequality to the sample variance estimators and showed their consistency. Proposition 4(iii) combined (i) and (ii) and formally established the asymptotic validity of (12).

Proposition 4 covers the treatment-control experiments as special cases. In other words, under certain regularity conditions, the difference-in-means estimator $\hat{\tau}$ in (2) is asymptotically normal, and the variance estimator $\hat{V}$ in (4) is consistent for a limit that is no less than the asymptotic variance of $\hat{V}$. These then justify the asymptotic validity of the level-α confidence interval in (5).

3.5.2 Analyzing covariate adjustment estimators

For covariate adjustment, we present a result by Zhao and Ding [117]. Let $\hat{Y}_*$ ($* = \text{N}, \text{F}, \text{L}$) be the estimates for the averaged potential outcomes across each treatment level from three estimation strategies: Neyman's approach, Fisher's ANCOVA and Lin's estimator (see more detailed definition in [117]).

Proposition 5 (Lemma 1 of Zhao and Ding [117])**.** *Let $N \to \infty$. Assume that, for $q \in [Q]$, $e_q = N_q/N$ has a limit in $(0,1)$. Assume that the first two finite population moments of $Y_i(q), X_i, X_iY_i(q) : q \in [Q]$ have finite limits, and both $S_x^2 = (N-1)^{-1}\sum_{i=1}^N X_iX_i^\top$ and its limit are nonsingular. Also assume that $N^{-1}\sum_{i=1}^N Y_i^4(q) = O(1)$, $N^{-1}\sum_{i=1}^N \|X_i\|_4^4 = O(1)$, and $N^{-1}\sum_{i=1}^N \|X_iY_i(q)\|_4^4 = O(1)$. Then the following results hold:*

(i) Asymptotic normality. $\sqrt{N}(\hat{Y}_* - \overline{Y}) \rightsquigarrow \mathcal{N}(0, V_*)$ *for some* $V_* \succeq 0$, $* = \text{N}, \text{F}, \text{L}$.
(ii) Conservative variance estimation. $\operatorname{plim}_{N\to\infty} N\hat{V}_{*,\text{EHW}} \succeq V_*$, $* = \text{N}, \text{F}, \text{L}$.
(iii) Efficiency comparison. $V_\text{L} \preceq V_\text{N}$ and $V_\text{L} \preceq V_\text{F}$.

Proposition 5(i) established the asymptotic normality property of Neyman's difference-in-means, Fisher's ANCOVA and Lin's estimator. Together with the conservative variance estimation in Proposition 5(ii), one can justify the asymptotic validity of Wald-type confidence regions constructed from covariate-adjusted estimators. Proposition 5(iii) indicates that Lin's estimator guarantees at least as much asymptotic efficiency as difference-in-means estimator and Fisher's ANCOVA.

In terms of technical details, Proposition 5(i) utilized a finite population central limit theorem proved by Li and Ding [64, Theorem 4], which is a result motivated by Hoeffding & Fraser's permutational CLT (more specifically, Condition (26) and (29)) and can accommodate vector outcomes and multi-level randomized experiments. Proposition 5(ii) utilized the Chebyshev inequality under the bounded moment conditions. Proposition 5(iii) involves some delicate analysis of the population variance structure V_* where $* = \text{N}, \text{F}, \text{L}$, which has closed-form expressions [117].

3.5.3 Analyzing rerandomization

For rerandomization using the Mahalanobis distance discussed in Section 2.4, we present the following result by Li et al. [67].

Proposition 6. *Consider ReM with a fixed positive threshold a, and assume that, as $N \to \infty$, (a) the proportions of units under treatment and control have positive limits, (b) the finite population variances and covariances for potential outcomes and covariates have limits, and the limit of S_X^2 is nonsingular; and (c) $N^{-1}\max_{1\le i\le N}|Y_i(z) - \bar{Y}(z)|^2 \to 0$ and $N^{-1}\max_{1\le i\le N}\|X_i - \bar{X}\|_2^2 \to 0$.*

(i) Asymptotic distribution. $\sqrt{N}(\hat{\tau} - \tau) \mid M \le a \,\dot\sim\, V^{1/2}(\sqrt{1-R^2}\cdot\varepsilon_0 + \sqrt{R^2}\cdot L_{K,a})$.
(ii) Conservative inference. *As $N \to \infty$,* $\operatorname{plim}_{N\to\infty}(\hat{V} - V) \ge 0$ *and* $\operatorname{plim}_{N\to\infty}(\hat{V}\hat{R}^2 - VR^2) = 0$.
(iii) Efficiency comparison. *The asymptotic distribution under ReM has a smaller (or equal) variance and narrower (or equal) symmetric quantile ranges compared to that under the CRE.*

Proposition 6(i) shows that the distribution of the difference-in-means estimator under rerandomization converges in distribution to the same limit as $V^{1/2}(\sqrt{1-R^2}\cdot\varepsilon_0 + \sqrt{R^2}\cdot L_{K,a})$, which is a convolution of a Gaussian and a constrained Gaussian random variables. In particular, V is the variance of $\sqrt{N}\hat{\tau}$ under the CRE, R^2 is an R^2-type measure for the association between potential outcomes and covariates, $\varepsilon_0 \sim \mathcal{N}(0,1)$, $L_{K,a} \sim D_1 \mid D'D \le a$ with $D = (D_1, \ldots, D_K)^\top \sim \mathcal{N}(0, I_K)$, and $\varepsilon_0 \perp\!\!\!\perp L_{K,a}$; see Li et al. [66] for more details. Interestingly, unlike that under the CRE, the asymptotic distribution of $\hat{\tau}$ under rerandomization is non-Gaussian in general, while it is still symmetric and unimodal around zero [66]. In

addition, when $a = \infty$ or $R^2 = 0$, the asymptotic distribution reduces to that for the CRE. The former is not surprising, because ReM without rejecting any assignment is essentially the CRE. The latter is also intuitive, implying that ReM using covariates that are irrelevant to potential outcomes is asymptotically equivalent to the CRE without using any covariates. Proposition 6(ii) shows that we can consistently estimate the coefficient of $L_{K,a}$ in the asymptotic distribution, while only conservatively estimate the coefficient of ε_0. Fortunately, due to the symmetric and unimodal property of the asymptotic distribution, these will lead to conservative variance estimation and confidence intervals. Proposition 6(iii) demonstrates the advantage of rerandomization over the CRE. In particular, the stronger the association between covariates and potential outcomes, as measured by R^2, the larger the gain from ReM [66].

We now discuss the technical aspects of Proposition 6. A key for its derivation is to notice that the distribution of $\hat{\tau}$ under rerandomization is the same as its conditional distribution under the CRE given that the covariate balance criterion is satisfied. This is emphasized by the conditioning on $M \leq a$ in Proposition 6(i). Thus, to understand this conditional distribution, it suffices to study the joint distribution of the difference-in-means vector $(\hat{\tau}, \hat{\tau}_X^\top)$ for both outcome and covariates, noting that M is a deterministic function of $\hat{\tau}_X$. Such a joint distribution will be asymptotically normal, which can be derived using Proposition 4 by viewing $(Y(z), X)$ as a "potential outcome" vector. For the asymptotic conservative inference in Proposition 6(ii), we can study the probability limits of the estimators $\hat{V}$ and $\hat{R}^2$ again utilizing their properties under the CRE through the conditioning argument; see details at [66]. Proposition 6(iii) involves careful analysis of the non-Gaussian distribution.

4 Permutational Berry-Esseen bounds

4.1 Several univariate and multivariate permutational Berry-Esseen bounds

Recently, permutational Berry-Esseen bounds (permutational BEBs, also called combinatorial BEBs) start to raise attention from the randomization world. Berry-Esseen bounds depict the distance between the sampling distribution of a statistic and a target, often normal, distribution. Theoretically speaking, it measures the convergence rate of central limit theorems. In general, the distance between two probability distributions is based on a class of metric of the following form:

$$d(\mathbb{P}_1, \mathbb{P}_2) = \sup_{h \in \mathcal{H}} \left| \int h \mathrm{d}\mathbb{P}_1 - \int h \mathrm{d}\mathbb{P}_2 \right|.$$

In particular, Berry-Esseen bounds consider H to be the class of indicator functions over a family of sets. For univariate distributions, BEBs study the upper bound based on the Kolmogorov metric, where $\mathcal{H}$ contains half-line indicator functions:

$$\sup_{t \in \mathbb{R}} \left| \mathbb{P}_1\{X \leq t\} - \mathbb{P}_2\{X \leq t\} \right|.$$

In the multivariate case, there are many choices of sets for different purposes, such as Euclidean balls [8], rectangular sets [19], measurable convex sets [9, 10], etc.

Permutational Berry-Esseen bounds are particularly important for analyzing randomized experiments in finite population. There have been many theoretical progresses.

4.1.1 Univariate case

We consider the univariate linear permutational statistics in (23) and adopt the notation from Section 3.3. We will summarize BEB results for Γ upon standardization. The standardized version of Γ can be

expressed as

$$\mathrm{Var}\{\Gamma\}^{-1/2}(\Gamma - \mathbb{E}\{\Gamma\}) = \frac{\sum_{i=1}^N \tilde{M}_N(i,\pi(i))}{\{\frac{1}{N-1}\sum_{i,j=1}^N \tilde{M}_N(i,j)^2\}^{1/2}} \triangleq \sum_{i=1}^N \check{M}_N(i,\pi(i)),$$

where

$$\check{M}(i,j) = \frac{\tilde{M}_N(i,j)}{\{(N-1)^{-1}\sum_{i,j=1}^N \tilde{M}_N(i,j)^2\}^{1/2}}.$$

Therefore, without loss of generality, we assume the following condition:

Condition 1 (Normalizing Γ)**.** Γ *in Definition* (23) *is defined with* M_N *satisfying the following normalizing condition:*

$$M_N(i,+) = M_N(+,j) = M_N(+,+) = 0, \text{ for all } i,j \in [N]; \quad \sum_{i,j\in[N]} M_N(i,j)^2 = N-1.$$

von Bahr [108] and Ho and Chen [52] established some early results, which require strong conditions such as boundedness to achieve a $O(N^{-1/2})$ convergence rate. Bolthausen [12] applied one version of Stein's method [104] to establish the following result that greatly relaxed the requirement to some conditions concerning the third moment of the matrix M_N.

Proposition 7 (Main theorem of Bolthausen [12])**.** *Assume Condition 1. There exists some universal constant* $C > 0$*, such that*

$$\sup_{t\in\mathbb{R}} |\mathbb{P}\{\Gamma \le t\} - \Phi(t)| \le CN^{-1} \sum_{i,j\in[N]} |M_N(i,j)|^3. \tag{30}$$

The upper bound on the right-hand-side of (30) achieves the rate of $O(N^{-1/2})$ if

$$N^{-1/2} \sum_{i,j\in[N]} |M_N(i,j)|^3 = O(1). \tag{31}$$

The following boundedness condition is provided by von Bahr [108] and Ho and Chen [52], which further gives a sufficient condition for (31) and thus (30):

$$\sup_{i,j\in[N]} |M_N(i,j)| = O(N^{-1/2}).$$

As a side note, Chen et al. [18, Chapter 6.1] also presented a thorough discussion about the univariate permutational BEB.

4.1.2 Multivariate case

We now consider the multivariate linear permutational statistics (27). For the ease of presentation, we focus on results upon standardization. Specifically, Shi and Ding [99, Lemma S2] proved that the standardized version of (27) can still be written as a multivariate linear permutational statistics with a different set of $\check{M}_{N,h}$'s:

$$\mathrm{Var}\{\Gamma\}^{-1/2}(\Gamma - \mathbb{E}\{\Gamma\}) = \left(\sum_{i=1}^N \check{M}_{N,1}(i,\pi(i)), \ldots, \sum_{i=1}^N \check{M}_{N,H}(i,\pi(i))\right)^\top,$$

where $\check{M}_{N,h}$'s satisfy the following normalizing conditions:

$$\check{M}_{N,h}(i,+) = \check{M}_{N,h}(+,j) = \check{M}_{N,h}(+,+) = 0, \text{ for all } i,j \in [N] \text{ and } h \in [H]; \tag{32}$$

$$\sum_{i,j\in[N]} \check{M}_{N,h}(i,j)^2 = N-1, \text{ for all } h \in [H]; \tag{33}$$

$$\sum_{i,j\in[N]} \check{M}_{N,h}(i,j)\check{M}_{N,h'}(i,j) = 0, \text{ for all } h \neq h' \in [H]. \tag{34}$$

$\check{M}_{N,h}$'s can be constructed from $M_{N,h}$'s by performing the centering step as in (24) then applying a linear combination using the matrix $\operatorname{Var}\{\Gamma\}^{-1/2}$. Therefore, without loss of generality, we assume the following condition.

Condition 2 (Normalizing Γ in the multivariate case)**.** *Γ in (27) is defined with $M_{N,h}$'s satisfying the normalizing conditions in (32), (33) and (34), which guarantees that $\mathbb{E}\{\Gamma\} = 0_H$ and $\operatorname{Var}\{\Gamma\} = I_H$.*

Bolthausen and Gotze [13] extended the univariate result in Proposition 7 to the multivariate, possibly nonlinear setting. In particular, for the multivariate linear case, Bolthausen and Gotze [13, Theorem 1] established the following BEB.

Proposition 8. *Let $\mathcal{A}$ be the family of all measurable convex sets. Let Γ_Z be a random Gaussian vector that follows $\mathcal{N}(0_H, I_H)$. Assume Condition 2. Then there exists a constant C_H that only depends on the dimension H such that*

$$\sup_{A\in\mathcal{A}} |\mathbb{P}\{\Gamma \in A\} - \mathbb{P}\{\Gamma_Z \in A\}| \leq \frac{C_H}{N} \sum_{i,j\in[N]} \left(\sum_{h=1}^{H} M_{N,h}(i,j)^2 \right)^{3/2}. \tag{35}$$

The BEB in (35) covers the univariate case in Proposition 7 as a special case with $H = 1$. However, Bolthausen and Gotze [13] did not give a closed form expression for C_H, whose dependence on the dimension H is unknown. Raic [87] conjectured the following result:

$$\sup_{A\in\mathcal{A}} |\mathbb{P}\{\Gamma \in A\} - \mathbb{P}\{\Gamma_Z \in A\}| \leq C\frac{H^{1/4}}{N} \sum_{i,j\in[N]} \left(\sum_{h\in[H]} M_{N,h}(i,j)^2 \right)^{3/2},$$

where C_H can be an absolute constant that does not depend on the dimension H. However, no formal proof is provided by the author. Chatterjee and Meckes [16] made one step forward to reveal the dimensional dependence using Stein's method with multivariate exchangeable pairs. In Chatterjee and Meckes [16, Section 3.2], the authors established a bound for the following distance:

$$\sup_{g\in C^2(\mathbb{R}^H)} |\mathbb{E}\{g(\Gamma)\} - \mathbb{E}\{g(\Gamma_Z)\}|,$$

where $C^2(\mathbb{R}^H)$ stands for the class of 2-times continuously differentiable functions on $\mathbb{R}^H$. We state a special case of Chatterjee and Meckes [16]'s result:

Proposition 9. *Under Condition 2 and the condition of bounded entries:*

$$\sup_{i,j\in[N],h\in[H]} |M_{N,h}(i,j)| = O(N^{-1/2}), \tag{36}$$

we have

$$\sup_{g\in C^2(\mathbb{R}^H)} |\mathbb{E}\{g(\Gamma)\} - \mathbb{E}\{g(\Gamma_Z)\}| = O\left(\frac{H^3}{N^{1/2}}\right). \tag{37}$$

Nevertheless, (37) does not translate directly into a Berry-Esseen bound under the Kolmogorov metric because the indicator functions are not members of $C^2(\mathbb{R}^H)$. Shi and Ding [99, Theorem S2] made use of one key result established by Fang and Röllin [30] regarding Stein's coupling and established the following multivariate permutational BEB with explicit dependence on the dimension:

Proposition 10. *Under Condition 2 and the condition of bounded entries* (36)*, we have*

$$\sup_{A\in\mathcal{A}} |\mathbb{P}\{\Gamma \in A\} - \mathbb{P}\{\Gamma_Z \in A\}| = O\left(\frac{H^{13/4}}{N^{1/2}}\right).$$

4.2 Application of permutational BEBs to randomization-based inference

In this section, we present several applications of permutational BEBs for randomization-based inference.

4.2.1 Completely randomized experiments with possibly varying group sizes and diverging treatment levels

Shi and Ding [99] and Shi et al. [100] discussed general CREs where the number of treatment levels Q and the treatment group sizes N_q's follow a variety of asymptotic regimes beyond the classical setup. Table 3 presented several possible regimes that are of interest both technically and practically.

Table 3. Theoretical results for multi-level experiments under the randomization model. The column title "Q" and "N_q" stand for the number of treatment levels and the number of replications within the treatment levels, respectively. The last column summarizes how well each of the regimes is studied in the literature regarding CLT, variance estimation and BEB.

Regime	Q	N_q	CLT, variance estimation, and BEB
(R1)	Small	Large	CLT and variance estimation established; no BEB
(R2)	Large	Large	Seems similar to (R1) but not studied
(R3)	Large	Small but $N_q \geq 2$	Not studied
(R4)	Large	$N_q = 1$	Not studied; variance estimation is nontrivial
(R5)	Mixture of the above		Not studied

Most of the regimes in Table 3 are less visited by literature and lacks scientific justification. Shi and Ding [99] utilized permutational BEBs to characterize the normal approximation for sampling distributions of statistics in general CREs, and managed to present a unified discussion of all the regimes listed in Table 3. We elaborate the usage of permutational BEBs with a canonical example in factorial experiments from Shi and Ding [99].

In a 2^K factorial design with K binary factors, there are $Q = 2^K$ possible treatment levels. Index the potential outcomes $Y_i(q)$'s also as $Y_i(z_1, \ldots, z_K)$'s, where $q = 1, \ldots, Q$ and $z_1, \ldots, z_K = 0, 1$. The parameter of interest $\gamma = F^\top \overline{Y}$ may consist of a subset of the factorial effects. The contrast matrix F has orthogonal columns and entries of $\pm(Q/2)^{-1}$; see Dasgupta et al. [26] for precise definitions of main effects and interactions. The factorial design is called *nearly uniform* if the sizes of each arm, N_q's, are approximately of the same order. More rigorously, we assume that there exists a positive integer $N_0 > 0$ and absolute constants $\underline{c} \leq \overline{c}$, such that $N_q = c_q N_0$ with $\underline{c} \leq c_q \leq \overline{c}$, for all $q = 1, \ldots, Q$. Such a setup can cover many cases in regimes (R1)-(R4) in Table 3. Shi and Ding [99] established the following result:

Proposition 11 (Shi and Ding [99], Example 6, nearly uniform factorial design)**.** *Consider a nearly uniform 2^K factorial experiment. Let $\tilde{\tau} = \mathrm{Var}\{\hat{\tau}\}^{-1/2}(\hat{\tau} - \tau)$ be the standardized version of $\hat{\tau}$. Let $F \in \mathbb{R}^{Q\times H}$ with $H = K + K(K-1)/2 = K(K+1)/2$ be the contrast matrix for all main effects and two-way inter-*

actions. Under some mild regularity conditions, we have

$$\sup_{b\in\mathbb{R}^H,\|b\|_2=1}\sup_{t\in\mathbb{R}}\left|\mathbb{P}\{b^\top\tilde{\tau}\le t\}-\Phi(t)\right|\le C\sigma_F\frac{\max_{q\in[Q],i\in[N]}|Y_i(q)-\overline{Y}(q)|}{\{\min_{q\in[Q]}S(q,q)\}^{1/2}}\sqrt{\frac{K^2}{N}}. \tag{38}$$

Proposition 11 is established based on the permutational BEB from Bolthausen [12] (presented in Proposition 7 in Section 4.1). From Proposition 11, we can obtain a sufficient condition for the the upper bound (38) to converge to 0, which implies a CLT of $\tilde{\tau}$. The general requirement for CLT is that the population size N is large enough in the study, regardless of how many treatment levels there are or how large each individual treatment arm is. Shi and Ding [99] also established design-based properties of Wald-type inference under general CREs, which attributes to multivarite permutational BEBs such as Proposition 10.

4.2.2 Rerandomization with diminishing covariate imbalance and diverging number of covariates

Li et al. [66] studied the asymptotic theory of rerandomization with a fixed covariate imbalance threshold that does not vary with the sample size, as discussed in Sections 2.4 and 3.5.3. The theory there suggests that the smaller the threshold, the more improvement we can gain from rerandomization over the complete randomization. Although intuitive, such a conclusion is not precise. When the covariate balance criterion is too stringent, there may be no acceptable assignments, and, even if there are acceptable ones, the asymptotic approximation may work poorly due to the small and even diminishing acceptance probability, i.e., the probability that a complete randomization is acceptable under rerandomization. Specifically and technically, the derivation for properties of rerandomization is through analyzing conditional distributions under the CRE, which will involve the acceptance probability in the denominator. The resulting convergence analysis will then encounter a ratio between two quantities of order $o(1)$ when we allow the acceptance probability (or the imbalance threshold) to diminish with the sample size. In such cases, BEBs are crucial for conducting convergence analysis.

In the context of simple random sampling, Wang and Li [111] derived a multivariate BEB for the sample average using Hájek [45]'s coupling and the BEB for sums of independent random vectors [88] with explicit dependence on the dimension. The bound, although weaker than that implied by the conjecture in Raic [88], is sufficient for studying rerandomization with diminishing covariate imbalance (or equivalently acceptance probability) and diverging number of covariates. With the derived BEB bounds, Wang and Li [111] presents the following asymptotic theory for ReM, which is stronger than Proposition 6. We adopt the same notation from Section 3.5.3, and denote the covariance imbalance threshold by a_n and the number of covariates by K_n, allowing them to vary with the sample size. Let $r_1=N_1/N$, $r_0=N_0/N$, $u_i=(r_0\cdot Y_i(1)+r_1\cdot Y_i(0),X_i^\top)^\top$, $\bar{u}$ and S_u^2 be the finite population mean and covariance of u_i's, and

$$\gamma_n\equiv\frac{(K_n+1)^{1/4}}{\sqrt{Nr_1r_0}}\frac{1}{N}\sum_{i=1}^N\|S_u^{-1}(u_i-\bar{u})\|_2^3,$$

where S_u^{-1} is the inverse of the positive semidefinite square root of S_u^2. We have the following BEB under ReM:

Proposition 12. *As $N\to\infty$, if $\gamma_n\to 0$ and $p_n/\gamma_n^{1/3}\to\infty$ with $p_n\equiv\Pr(\chi^2_{K_n}\le a_n)$, then*

$$\sup_{c\in\mathbb{R}}\left|\Pr\{V^{-1/2}(\hat{\tau}-\tau)\le c\mid M\le a_n\}-\Pr(\sqrt{1-R^2}\ \varepsilon_0+\sqrt{R^2}\ L_{K_N,a_N}\le c)\right|\to 0.$$

Wang and Li [111] further studied additional conditions such that the constrained Gaussian random variable L_{K_N,a_N} becomes ignorable as $N\to\infty$, under which $V^{-1/2}(\hat{\tau}-\tau)$ can asymptotically follow the Gaussian distribution $\mathcal{N}(0,1-R^2)$ under rerandomization. This is the ideally optimal precision that one can expect under rerandomization, since the remaining variation is due to the part of potential outcomes that cannot be linearly explained by the covariates. Moreover, the Gaussian asymptotic distribution is the same as that of

Lin's regression-adjusted estimator under the CRE. Intuitively, rerandomization and covariate adjustment are dual of each other, where the former is at the design stage while the latter is at the analysis stage. Wang and Li [111] also proposed large-sample valid confidence intervals for the average treatment effect under rerandomization, allowing sample-size-varying imbalance threshold and number of covariates.

5 Extensions

Neyman [79] has motivated many important extensions for the design and analysis of randomized experiments, and the technical tools regarding permutations have been evolving during the past century. In this section, we discuss some other extensions beyond the Neyman [79]'s proposal.

5.1 Other randomized experiments

In this section we discuss several other widely used and studied randomized experiments.

5.1.1 Stratified (block) randomized experiments

Stratified randomized experiments (SRE) have been used widely in many fields, including agricultural study [85], biomedical study [41], social science [20], etc. A SRE combines several different CREs according to the levels of a stratum indicator. Concretely speaking, consider an experiment with K strata. Denote the number and proportion of units in stratum k as $N_{[k]}$ and $\pi_{[k]} = N_{[k]}/N$, respectively, where $k = 1, \ldots, K$. Within stratum k, $N_{[k]1}$ units are randomized to receive treatment and $N_{[k]0} = N_{[k]} - N_{[k]1}$ units are assigned to control. Across strata, the randomization is conducted independently. The treatment assignment distribution is uniform over all possible randomizations.

Analogous to CRE, in SRE, for unit i in stratum k, we have potential outcomes $Y_{ki}(1)$ and $Y_{ki}(0)$ and individual causal effect $\tau_{ki} = Y_{ki}(1) - Y_{ki}(0)$. For stratum k, we have stratum-specific average causal effect

$$\tau_{[k]} = N_{[k]}^{-1} \sum_{i=1}^{N_{[k]}} \tau_{ki},$$

and the average causal effect is

$$\tau = N^{-1} \sum_{k=1}^{K} \sum_{i=1}^{N_{[k]}} \tau_{ki} = \sum_{k=1}^{K} \pi_{[k]} \tau_{[k]}.$$

which is also the weighted average of the stratum-specific average causal effects. For Neyman-type analysis, a point estimator can be obtained by taking a weighted combination of stratum-specific difference-in-means estimators:

$$\hat{\tau}_{\mathrm{S}} = \sum_{k=1}^{K} \pi_{[k]} \hat{\tau}_{[k]}, \tag{39}$$

where $\hat{\tau}_{[k]}$ is the difference-in-means estimator for stratum k. It has variance

$$\mathrm{Var}\left\{\hat{\tau}_{\mathrm{S}}\right\} = \sum_{k=1}^{K} \pi_{[k]}^2 \mathrm{Var}\left\{\hat{\tau}_{[k]}\right\},$$

which motivates the variance estimator

$$\hat{V}_{\mathrm{S}} = \sum_{k=1}^{K} \pi_{[k]}^2 \left\{ \frac{\hat{S}_{[k]}^2(1)}{N_{[k]1}} + \frac{\hat{S}_{[k]}^2(0)}{N_{[k]0}} \right\},$$

with $\hat{S}^2_{[k]}(1)$ and $\hat{S}^2_{[k]}(0)$ being the stratum-specific sample variance estimators for the treatment and control arms. A Wald-type confidence interval can then be constructed for τ.

Under certain regularity conditions, the point estimator (39) is asymptotically normal and Wald-type inference is proved to be asymptotically valid (see for example [69]). The random assignment mechanism requires studying a convolution of independent and permutational distributions, which also motivates new theoretical tools. When the total number of strata K is small and the sizes of the strata are large, the permutational CLTs (e.g., Hoeffding's condition (25)) play a central role in the analysis. When K is large and the sizes of the strata are small, CLTs for independent summations play a crucial role instead. With a mixture of large and small strata, there are also theoretical results in the literature; see for example [69]. Moreover, Liu et al. [70] and Wang et al. [110] further investigated covariate adjustment and rerandomization in SREs.

5.1.2 Matched-pairs experiments

The matched-pairs experiment (MPE) is another popular experimental design in practice [6, 33, 56]. The MPE is the most extreme version of the SRE with only one treated unit and one control unit within each stratum, which is called a *pair*. We can adopt the same sets of notations as SRE to define potential outcomes, causal effects, stratum-specific difference-in-means estimator (denoted again as $\hat{\tau}_{[k]}$) and the aggregated difference-in-means estimator (denoted as $\hat{\tau}_{\mathrm{M}}$) in the MPE. However, the variance estimation strategy discussed in Section 5.1.1 is no longer applicable for the MPE, since it implicitly requires at least two treated and control units within each matched sets so that we can calculate the stratum-specific sample variances. Imai [56] proposed the following variance estimator by instead considering the sample variance of the stratum-specific difference-in-means estimators:

$$\hat{V}_{\mathrm{M}} = \frac{1}{n(n-1)} \sum_{k=1}^{n} (\hat{\tau}_{[k]} - \hat{\tau}_{\mathrm{M}})^2,$$

and he showed that it is conservative in expectation for the true variance of $\hat{\tau}_{\mathrm{M}}$. We can then construct the Wald-type confidence interval

$$\left[\hat{\tau}_{\mathrm{M}} - z_{\alpha/2}\hat{V}_{\mathrm{M}}^{1/2},\ \hat{\tau}_{\mathrm{M}} + z_{\alpha/2}\hat{V}_{\mathrm{M}}^{1/2}\right],$$

which can be asymptotically valid under certain regularity conditions. Moreover, regression adjustment can be applied to improve efficiency when baseline covariates are available, as shown by Fogarty [36].

In general stratified experiments with possibly one treated or one control unit in some strata, Fogarty [35] and Pashley and Miratrix [83] discussed general strategies to conservatively estimate the variance of the aggregated difference-in-means estimator.

5.1.3 Cluster randomized experiments

Cluster randomized experiments are widely used due to their logistical convenience and policy relevance. In a cluster randomized experiment, the treatment is assigned at the cluster level instead of the individual level. Consider a study with N units with M clusters. Cluster i has n_i units ($i = 1, \ldots, M$). Let (i, j) index the j-th unit within cluster i for $i = 1, \ldots, M$ and $j = 1, \ldots, n_i$. Unit (i, j) has covariates x_{ij}, and cluster i has covariates c_i. The experimenter randomly assigns M_1 clusters to receive the treatment and M_0 clusters to receive the control, where $0 < e < 1$ is a fixed number denoting the proportion of treated clusters. Let Z_i be the treatment indicator for cluster i and Z_{ij} be the treatment indicator for unit (i, j). In a cluster-randomized experiment, units within a cluster receive identical treatment levels. So if cluster i receives treatment, then $Z_{ij} = Z_i = 1$ for all j; if cluster i receives control, then $Z_{ij} = Z_i = 0$ for all j. Let $Y_{ij}(1)$ and $Y_{ij}(0)$ be the potential outcomes under treatment and control, respectively, for unit (i, j). The

observed outcome is then $Y_{ij} = Z_{ij}Y_{ij}(1) + (1 - Z_{ij})Y_{ij}(0)$. The average treatment effect over all units is

$$\tau = \frac{1}{N}\sum_{i=1}^{M}\sum_{j=1}^{n_i}\{Y_{ij}(1) - Y_{ij}(0)\}.$$

There are different strategies for inferring τ, including individual-level estimators and cluster-level estimators [106], both enjoying desirable asymptotic properties implied by permutational CLTs. We refer interested readers to a collection of works on analyzing cluster-randomized experiments [1, 3, 73, 76, 98, 106].

5.2 Other technical tools for permutations

Permutation is a core element in the design and analysis of Neyman [79]'s CRE. In this section we discuss other technical tools for analyzing permutation related questions.

5.2.1 Hajek's coupling

Hajek's coupling is one technique developed in [45] for proving central limit theorems in simple random sampling. The idea is based on constructing a coupling between simple random sampling and Bernoulli random sampling so that CLTs for i.i.d. sampling can be applied. Wang and Li [111] utilized Hajek's coupling together with a multivariate Berry–Esseen-type bound for sum of independent random vectors [88] to study rerandomization with diminishing covariate imbalance. The techniques are potentially useful to establish theories for a wider range of permutational statistics.

5.2.2 Double and multiple permutations

Nowadays, there are many new variants of of permutations in designing randomized experiments. For example, Fredrickson and Chen [38] and Chen and Friedman [17] discussed permutation and randomization test for analyzing network data. Bajari et al. [4, 5] proposed multiple randomization designs for marketplaces in which multiple populations interact and causal questions regarding interference are of particular interest. In terms of technical tools that are potentially useful for analyzing double or multiple permutations, Chen et al. [18], Reinert and Röllin [89], Zhao et al. [118], among others, used Stein's method [104] to study the asymptotic properties of the so-called *doubly indexed permutation statistics* (DIPS).

5.2.3 Concentration inequalities

Another technical tool that has been recognized by the literature is permutational/combinatorial concentration inequalities. Lei and Ding [62] used permutational concentration inequalities to analyze regression adjustment in CREs when the dimension of the covariates is diverging. Bloniarz et al. [11] used permutational concentration inequalities to analyze regression adjustment in CREs with high dimensional covariates. It is interesting to explore related potential research questions that involve delicate analysis of the finite sample properties of permutational designs and inspire the use of concentration inequalities.

6 Conclusion

In this review, we revisited the fundamental contributions of Neyman [79]'s seminal work regarding the introduction of potential outcomes, the promotion of physical randomization and the emphasis of the

repeated sampling property of statistics over the randomization distribution. These contributions lay down the foundation for the design and analysis of randomized experiments. We also reviewed permutational central limit theorems and Berry–Esseen bounds in great detail, and listed applications of these technical results in randomization-based inference.

Beyond what we have covered in the review, there are also many research questions that are closely related to Neyman [79]. From a technical point of view, many theoretical tools are not fully covered in the discussion. For example, when analyzing stratified randomized experiments, we need central limit theorems and Berry-Esseen bounds that combine the permutational and the independent distributions [69, 70]. As another example, for the design and analysis of adaptive experiments, a general martingale structure typically exists which requires a martingale central limit theorem or Berry-Esseen result [47, 54]. From a practical point of view, there are many real-world examples that can motivate new designs, assumptions and causal estimands. For example, interference among units is a common phenomenon for many experimental and observational study. The study of inference has motivated a lot of new designs and methods, such as designing and analyzing bipartite experiments [49, 119], multiple randomization [4], network interference [63], etc.

References

[1] A. Abadie, S. Athey, G. W. Imbens, and J. M. Wooldridge. When should you adjust standard errors for clustering? *The Quarterly Journal of Economics*, 138(1):1–35, 2023.

[2] S. Athey and G. W. Imbens. The Econometrics of Randomized Experiments. In A. Banerjee and E. Duflo, editors, *Handbook of Economic Field Experiments*, volume 1, chapter 3, pages 73–140. North-Holland, Amsterdam, 2017.

[3] S. Athey and G. W. Imbens. The econometrics of randomized experiments. In *Handbook of economic field experiments*, volume 1, pages 73–140. Elsevier, 2017.

[4] P. Bajari, B. Burdick, G. W. Imbens, L. Masoero, J. McQueen, T. Richardson, and I. M. Rosen. Multiple randomization designs, 2021.

[5] P. Bajari, B. Burdick, G. W. Imbens, L. Masoero, J. McQueen, T. S. Richardson, and I. M. Rosen. Experimental design in marketplaces. *Statistical Science*, 1(1):1–19, 2023.

[6] S. Ball et al. Reading with Television: An Evaluation of The Electric Company. A Report to the Children's Television Workshop. Volumes 1 and 2. 1973.

[7] D. J. Bauer, S. K. Sterba, and D. D. Hallfors. Evaluating group-based interventions when control participants are ungrouped. *Multivariate behavioral research*, 43(2):210–236, 2008.

[8] V. Bentkus. On the dependence of the Berry–Esseen bound on dimension. *Journal of Statistical Planning and Inference*, 113(2):385–402, 2003.

[9] V. Bentkus. A Lyapunov-type bound in $\mathbb{R}^d$. *Theory of Probability & Its Applications*, 49(2):311–323, 2005.

[10] R. N. Bhattacharya and R. R. Rao. *Normal Approximation and Asymptotic Expansions*. SIAM, 2010.

[11] A. Bloniarz, H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu. Lasso adjustments of treatment effect estimates in randomized experiments. *Proceedings of the National Academy of Sciences*, 113(27):7383–7390, 2016.

[12] E. Bolthausen. An estimate of the remainder in a combinatorial central limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 66(3):379–386, 1984.

[13] E. Bolthausen and F. Gotze. The rate of convergence for multivariate sampling statistics. *The Annals of Statistics*, pages 1692–1710, 1993.

[14] Z. Branson, T. Dasgupta, and D. B. Rubin. Improving covariate balance in 2 k factorial designs via rerandomization with an application to a new york city department of education high school study. *The Annals of Applied Statistics*, pages 1958–1976, 2016.

[15] M. Bruhn and D. McKenzie. In pursuit of balance: Randomization in practice in development field experiments. *American economic journal: applied economics*, 1(4):200–232, 2009.

[16] S. Chatterjee and E. Meckes. Multivariate normal approximation using exchangeable pairs. *arXiv preprint math/0701464*, 2007.

[17] H. Chen and J. H. Friedman. A new graph-based two-sample test for multivariate and object data. *Journal of the American statistical association*, 112(517):397–409, 2017.

[18] L. H. Chen, L. Goldstein, and Q.-M. Shao. *Normal Approximation by Stein's Method*, volume 2. Springer, 2011.

[19] V. Chernozhukov, D. Chetverikov, and K. Kato. Central limit theorems and bootstrap in high dimensions. *The Annals of Probability*, 45(4):2309, 2017.

[20] A. Chong, I. Cohen, E. Field, E. Nakasone, and M. Torero. Iron deficiency and schooling attainment in Peru. *American Economic Journal: Applied Economics*, 8(4):222–255, 2016.
[21] W. G. Cochran. *Sampling techniques*. John Wiley & Sons, 1977.
[22] P. L. Cohen and C. B. Fogarty. No-harm calibration for generalized oaxaca-blinder estimators. *arXiv preprint arXiv:2012.09246*, 2020.
[23] P. L. Cohen and C. B. Fogarty. Gaussian prepivoting for finite population causal inference. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(2):295–320, 2022.
[24] J. B. Copas. Randomization models for the matched and unmatched 2×2 tables. *Biometrika*, 60(3):467–476, 1973.
[25] D. Cox. Randomization in the design of experiments. *International Statistical Review*, 77(3):415–429, 2009.
[26] 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*, 77:727–753, 2015.
[27] F. David. Limiting distributions connected with certain methods of sampling human populations. *Stat. Res. Mem*, 2: 69–90, 1938.
[28] C. J. DiCiccio and J. P. Romano. Robust permutation tests for correlation and regression coefficients. *Journal of the American Statistical Association*, 112(519):1211–1220, 2017.
[29] P. Ding and T. Dasgupta. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, 105(1):45–56, 2018.
[30] X. Fang and A. Röllin. Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics. *Bernoulli*, pages 2157–2189, 2015.
[31] S. E. Fienberg and J. M. Tanur. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. *International Statistical Review/Revue Internationale de Statistique*, pages 237–253, 1996.
[32] R. A. Fisher. *Statistical Methods for Research Workers*. Edinburgh by Oliver and Boyd, 1st edition, 1925.
[33] R. A. Fisher. *The Design of Experiments*. Edinburgh, London: Oliver and Boyd, 1st edition, 1935.
[34] R. A. Fisher and W. A. Mackenzie. Studies in crop variation. II. The manurial response of different potato varieties. *The Journal of Agricultural Science*, 13(3):311–320, 1923.
[35] C. B. Fogarty. On mitigating the analytical limitations of finely stratified experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(5):1035–1056, 2018.
[36] C. B. Fogarty. Regression assisted inference for the average treatment effect in paired experiments. *Biometrika*, 105: 994–1000, 2018.
[37] D. Fraser. A vector form of the Wald-Wolfowitz-Hoeffding theorem. *The Annals of Mathematical Statistics*, pages 540–543, 1956.
[38] M. M. Fredrickson and Y. Chen. Permutation and randomization tests for network analysis. *Social Networks*, 59: 171–183, 2019.
[39] D. A. Freedman. On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193, 2008.
[40] D. A. Freedman. On regression adjustments in experiments with several treatments. *Annals of Applied Statistics*, 2: 176–96, 2008.
[41] M. G. Goldner, G. L. Knatterud, and T. E. Prout. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: Iii. clinical implications of ugdp results. *JAMA*, 218(9):1400–1410, 1971.
[42] K. Guo and G. Basse. The generalized Oaxaca-Blinder estimator. *Journal of the American Statistical Association*, pages 1–13, 2021.
[43] J. Hainmueller and D. J. Hopkins. The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. *American journal of political science*, 59(3):529–548, 2015.
[44] J. Hainmueller, D. J. Hopkins, and T. Yamamoto. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political analysis*, 22(1):1–30, 2014.
[45] J. Hájek. Limiting distributions in simple random sampling from a finite population . *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5:361–374, 1960.
[46] J. Hájek. Some extensions of the Wald-Wolfowitz-Noether theorem. *The Annals of Mathematical Statistics*, pages 506–523, 1961.
[47] P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic press, 2014.
[48] D. Hallfors, H. Cho, V. Sanchez, S. Khatapoush, H. M. Kim, and D. Bauer. Efficacy vs effectiveness trial results of an indicated "model" substance abuse program: implications for public health. *American journal of public health*, 96 (12):2254–2259, 2006.
[49] C. Harshaw, F. Sävje, D. Eisenstat, V. Mirrokni, and J. Pouget-Abadie. Design and analysis of bipartite experiments under a linear exposure-response model. *Electronic Journal of Statistics*, 17(1):464–518, 2023.
[50] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Introduction to Experimental Design*, volume 1. New York: John Wiley & Sons, 2007.
[51] K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Volume 1: Introduction to Experimental Design*, volume 1. John Wiley & Sons, 2007.

[52] S.-T. Ho and L. H. Chen. An L_p Bound for the Remainder in a Combinatorial Central Limit Theorem. *The Annals of Probability*, 6(2):231–249, 1978.
[53] W. Hoeffding. A combinatorial central limit theorem. *The Annals of Mathematical Statistics*, pages 558–566, 1951.
[54] F. Hu and W. F. Rosenberger. *The theory of response-adaptive randomization in clinical trials*. John Wiley & Sons, 2006.
[55] M. G. Hudgens and M. E. Halloran. Toward causal inference with interference. *Journal of the American Statistical Association*, 103(482):832–842, 2008.
[56] K. Imai. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. *Statistics in medicine*, 27(24):4857–4873, 2008.
[57] K. Imai, I. S. Kim, and E. H. Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2023.
[58] G. W. Imbens and D. B. Rubin. *Causal Inference in Statistics, Social, and Biomedical Sciences*. New York: Cambridge University Press, 2015.
[59] P. Johansson and M. Schultzberg. Rerandomization: A complement or substitute for stratification in randomized experiments? *Journal of Statistical Planning and Inference*, 218:43–58, 2022.
[60] O. Kempthorne. *The Design and Analysis of Experiments.* New York: Wiley, 1952.
[61] O. Kempthorne. The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50(271):946–967, 1955.
[62] L. Lei and P. Ding. Regression adjustment in completely randomized experiments with a diverging number of covariates. *Biometrika*, 108(4):815–828, dec 2020. 10.1093/biomet/asaa103. URL https://doi.org/10.1093%2Fbiomet%2Fasaa103.
[63] M. P. Leung. Causal inference under approximate neighborhood interference. *Econometrica*, 90(1):267–293, 2022.
[64] X. Li and P. Ding. General forms of finite population central limit theorems with applications to causal inference. *Journal of the American Statistical Association*, 112(520):1759–1769, 2017.
[65] X. Li and P. Ding. Rerandomization and regression adjustment. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):241–268, 2020.
[66] X. Li, P. Ding, and D. B. Rubin. Asymptotic theory of rerandomization in treatment–control experiments. *Proceedings of the National Academy of Sciences*, 115(37):9157–9162, 2018.
[67] X. Li, P. Ding, and D. Rubin. Rerandomization in 2^K factorial experiments. *Annals of Statistics*, 48(1):43–63, 2020.
[68] W. Lin. Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique. *The Annals of Applied Statistics*, 7(1):295 – 318, 2013. 10.1214/12-AOAS583. URL https://doi.org/10.1214/12-AOAS583.
[69] H. Liu and Y. Yang. Regression-adjusted average treatment effect estimates in stratified randomized experiments. *Biometrika*, 107(4):935–948, 2020.
[70] H. Liu, J. Ren, and Y. Yang. Randomization-based joint central limit theorem and efficient covariate adjustment in randomized block 2^K factorial experiments. *Journal of the American Statistical Association*, pages 1–15, 2022.
[71] L. Liu, Y. Wang, and Y. Xu. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 2022.
[72] J. Lu. Covariate adjustment in randomization-based causal inference for 2^K factorial designs. *Statistics & Probability Letters*, 119:11–20, 2016.
[73] X. Lu, T. Liu, H. Liu, and P. Ding. Design-based theory for cluster rerandomization. *Biometrika*, 110(2):467–483, 2023.
[74] M. Maclure, A. Nguyen, G. Carney, C. Dormuth, H. Roelants, K. Ho, and S. Schneeweiss. Measuring prescribing improvements in pragmatic trials of educational tools for general practitioners. *Basic & clinical pharmacology & toxicology*, 98(3):243–252, 2006.
[75] W. G. Madow. On the limiting distributions of estimates based on samples from finite universes. *The Annals of Mathematical Statistics*, pages 535–545, 1948.
[76] J. A. Middleton and P. M. Aronow. Unbiased estimation of the average treatment effect in cluster-randomized experiments. *Statistics, Politics and Policy*, 6(1-2):39–75, 2015.
[77] K. L. Morgan and D. B. Rubin. Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282, 2012.
[78] M. Motoo. On the Hoeffding's combinatrial central limit theorem. *Annals of the Institute of Statistical Mathematics*, 8:145–154, 1956.
[79] J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, pages 465–472, 1923/1990.
[80] J. Neyman. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 97(4):558–606, 1934.
[81] J. Neyman and K. Iwaszkiewicz. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2):107–180, 1935.
[82] G. E. Noether. On a theorem by Wald and Wolfowitz. *The Annals of Mathematical Statistics*, 20(3):455–458, 1949.

[83] N. E. Pashley and L. W. Miratrix. Insights on variance estimation for blocked and matched pairs designs. *Journal of Educational and Behavioral Statistics*, 46(3):271–296, 2021.
[84] E. Paul 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.
[85] R. G. Petersen. *Agricultural Field Experiments: Design and Analysis*. CRC Press, 1994.
[86] E. J. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.
[87] M. Raic. Multivariate normal approximation: Permutation statistics, local dependence and beyond, 2015.
[88] M. Raič. A multivariate berry–esseen theorem with explicit constants. 2019.
[89] G. Reinert and A. Röllin. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. *Annals of Probability*, 37(6), 2007.
[90] J. M. Robins. Confidence intervals for causal parameters. *Statistics in Medicine*, 7(7):773–785, 1988.
[91] P. R. Rosenbaum. *Observational Studies*. Springer-Verlag, 2002.
[92] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.
[93] D. B. Rubin. Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American statistical association*, 75(371):591–593, 1980.
[94] D. B. Rubin. Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, 5(4):472–480, 1990.
[95] D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.
[96] D. B. Rubin. Comment: The design and analysis of gold standard randomized experiments. *Journal of the American Statistical Association*, 103(484):1350–1353, 2008.
[97] H. Scheffé. *The Analysis of Variance*. New York: John Wiley & Sons, 1959.
[98] P. Z. Schochet, N. E. Pashley, L. W. Miratrix, and T. Kautz. Design-based ratio estimators and central limit theorems for clustered, blocked RCTs. *Journal of the American Statistical Association*, 117(540):2135–2146, 2022.
[99] L. Shi and P. Ding. Berry–Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes. *arXiv preprint arXiv:2209.12345*, 2022.
[100] L. Shi, J. Wang, and P. Ding. Forward screening and post-screening inference in factorial designs. *arXiv preprint arXiv:2301.12045*, 2023.
[101] A. Sjölander, T. Frisell, R. Kuja-Halkola, S. Öberg, and J. Zetterqvist. Carryover effects in sibling comparison designs. *Epidemiology*, 27(6):852–858, 2016.
[102] J. Splawa-Neyman. Contributions to the theory of small samples drawn from a finite population. *Biometrika*, pages 472–479, 1925.
[103] D. Sprott and V. Farewell. Randomization in experimental science. *Statistical Papers*, 34:89–94, 1993.
[104] C. Stein. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Volume 2: Probability Theory*, volume 6, pages 583–603. University of California Press, 1972.
[105] Student. On testing varieties of cereals. *Biometrika*, pages 271–293, 1923.
[106] F. Su and P. Ding. Model-assisted analyses of cluster-randomized experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(5):994–1015, 2021.
[107] E. J. T. Tchetgen and T. J. VanderWeele. On causal inference in the presence of interference. *Statistical methods in medical research*, 21(1):55–75, 2012.
[108] B. von Bahr. Remainder term estimate in a combinatorial limit theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 35(2):131–139, 1976.
[109] A. Wald and J. Wolfowitz. Statistical tests based on permutations of the observations. *The Annals of Mathematical Statistics*, 15(4):358–372, 1944.
[110] X. Wang, T. Wang, and H. Liu. Rerandomization in stratified randomized experiments. *Journal of the American Statistical Association*, 118(542):1295–1304, 2023.
[111] Y. Wang and X. Li. Rerandomization with diminishing covariate imbalance and diverging number of covariates. *The Annals of Statistics*, 50(6):3439–3465, 2022.
[112] B. L. Welch. On the z-test in randomized blocks and Latin squares. *Biometrika*, 29(1/2):21–52, 1937.
[113] J. Worrall. Evidence: philosophy of science meets medicine. *Journal of evaluation in clinical practice*, 16(2):356–362, 2010.
[114] C. J. Wu and M. S. Hamada. *Experiments: Planning, Analysis, and Optimization*. John Wiley & Sons, 2011.
[115] 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.
[116] A. Zhao and P. Ding. Covariate-adjusted fisher randomization tests for the average treatment effect. *Journal of Econometrics*, 225(2):278–294, 2021.

[117] A. Zhao and P. Ding. Covariate adjustment in multiarmed, possibly factorial experiments. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 85(1):1–23, 2023.

[118] L. Zhao, Z. Bai, C.-C. Chao, and W.-Q. Liang. Error bound in a central limit theorem of double-indexed permutation statistics. *The Annals of Statistics*, 25(5):2210–2227, 1997.

[119] C. M. Zigler and G. Papadogeorgou. Bipartite causal inference with interference. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 36(1):109, 2021.