**Review Article**

Lei Shi* and Xinran Li

# Some theoretical foundations for the design and analysis of randomized experiments

https://doi.org/10.1515/jci-2023-0067
received October 06, 2023; accepted April 22, 2024

**Abstract:** Neyman'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'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 experiments 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 we elaborate on 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, and cluster randomized experiments.

**Keywords:** causal inference, permutation, central limit theorem, Berry–Esseen bound, potential outcome

**MSC 2020:** 62K15, 62J05, 62G05

## 1 Review of the proposal in Neyman [1] and its influence

Neyman's seminal work [1] has been a cornerstone in the field of statistics over the last 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 such as agriculture, economics, biomedical research, social science, and beyond.

The main purpose of Neyman [1] is the analysis of field experiments conducted in order to compare a number of crop varieties. Suppose there are $m$ plots and $\nu$ varieties. Neyman [1] 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 $1 \le k \le \nu$; we use slightly different indices from Neyman to make them more intuitive. In Neyman's framework, the quantities $\{U_{ik}\}$ are fixed but may be 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, which is, in fact, an estimand representing the average yield in modern terminology. Neyman [1] then used an urn model as a thought experiment to depict the framework of sampling from a finite population. The $\nu$ types of varieties are treated as $\nu$ urns. Each urn contains $m$ balls, and each ball is associated with two labels: a plot label indexing the plots

*** Corresponding author: Lei Shi,** Division of Biostatistics, University of California, Berkeley, CA, USA, e-mail: leishi@berkeley.edu
**Xinran Li:** Department of Statistics, The University of Chicago, Chicago, IL, USA, e-mail: xinranli@uchicago.edu

and a yield label indicating the unknown potential yields on the plots for each of the varieties. Specifically, in the $k$th urn, there are $m$ balls with yield labels

$$U_{1k}, \ldots, U_{ik}, \ldots, U_{mk}.$$

Also, the urns have the property that "if one ball is taken from one of them, the balls having the same plot label disappear from all the other urns." Then, from each urn, a number of balls are drawn without replacement. With this model, 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 pioneering effort for studying the difference-in-means estimator in modern terminology. Notably, 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 [1] offered a series of groundbreaking and foundational insights. In the following, we outline three key facets of Neyman's [1] 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 [2–5]. The potential outcome paradigm serves as an impeccable model for a discussion in 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 [6–8], the prolonged impacts of interventions [9–11], and the causal analyses involving post-treatment variables such as instrumental variables [12] and mediation [13,14]. Moreover, the importance of potential outcomes transcends experimental settings and is also profound in observational studies, as highlighted by Rubin [15].

The second contribution of Neyman [1] 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 the 1920s, as commented by Rubin [16] citing Student [17] and Fisher and Mackenzie [18] as references. Neyman [1] 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 are viewed as fixed, and physical randomization emerges as the "reasoned basis" [19] for facilitating statistical testing and estimation [4,20–22]. Moreover, the proposal of sampling without replacement also inspires the pursuit of the parallels and linkages between survey sampling and randomized experiments [23–26].

The third contribution of Neyman [1] centers on the repeated sampling properties 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 the sharp null hypothesis of no causal effects for any units and finite-sample exact $p$-values [16]. Neyman [1] recognized from an empirical perspective that the asymptotic normality 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 populations.

Building upon the pioneering contribution of Neyman [1] 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 and permutations, such as central limit theorems (CLTs) and Berry–Esseen bounds (BEBs), which were developed over the century, and elaborate on 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.

*Notations.* We summarize a set of notations for the whole article. For an integer $N$, we use $[N]$ to denote the set of integers $\{1, \ldots, N\}$. For two positive semidefinite matrices $V_1$ and $V_2$, we use $V_1 \succeq V_2$ to indicate that $V_1$ dominates $V_2$, in the sense that $V_1 - V_2$ is positive semidefinite. For a random sequence $\{X_N\}_{N=1}^{\infty}$, we write $X_N \rightsquigarrow \mathcal{L}$ if $X_N$ converges weakly to the distribution $\mathcal{L}$ as $N \to \infty$, and $X_N \rightsquigarrow L_N$ if $X_N$ and $L_N$ converge weakly to the same distribution. When $X_N$ converges in probability, we use $\text{plim}_{N\to\infty} X_N$ to denote its probabilistic limit.

# 2 Design and analysis of CREs

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 2.2 presents several covariate-adjusted estimators, and Section 2.3 discusses rerandomization.

## 2.1 Basic design and analysis of CREs

### 2.1.1 Treatment-control CRE

We start by considering a treatment-control CRE 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 $1 \le i \le N$. The treatment assignment status for the entire experiment is vectorized as $Z = (Z_1, \ldots, Z_N)$. Under complete randomization,

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

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

Rubin [27] called the $N \times 2$ matrix of potential outcomes in Table 1 as the science table. The observed outcome for the $i$-th unit is $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. Importantly, the potential outcomes are fixed and the randomness comes merely from the random allocation of the treatment, reflected by the random vector $Z$. Scheffé [28, 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., 4,20,22,29–38]. 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:

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

**Table 1:** Science table for treatment-control CRE

| $i$ | $Y_i(0)$ | $Y_i(1)$ |
|---|---|---|
| 1 | $Y_1(0)$ | $Y_1(1)$ |
| ⋮ | ⋮ | ⋮ |
| $N$ | $Y_N(0)$ | $Y_N(1)$ |

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

Under the potential outcome framework, the $i$th unit has individual treatment effect $\tau_i = Y_i(1) - Y_i(0)$, for $1 \le i \le N$. 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 [1] proposed to estimate the ATE $\tau$ by the difference-in-means estimator:

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

He proved that $\hat{\tau}$ is an unbiased estimator for $\tau$, 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 [1] proposed the following variance estimator:

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

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

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

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

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

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's [1] 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 [39]. In contrast, Fisher [19] proposed to test the following null hypothesis:

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

which is called the sharp null hypothesis by Rubin [40] or the strong null hypothesis by Wu an Ding [39]. The Fisherian perspective is fundamentally different, as it focuses on testing the hypothesis of no causal effects for any units whatsoever, whereas the Neymanian perspective focuses on testing no average causal effect [41]. Obviously, Fisher's null implies Neyman's null, but either of them can be practically relevant depending on the

application. Under (6), one can impute the unobserved potential outcomes and perform Fisher's randomization test (FRT) to deliver finite sample exact inference [19]. Fisher's test has the advantage of being finite-sample valid, while Neyman's requires large-sample approximation. Nevertheless, Neyman's asymptotic results can also help ease the computation for Fisher's null. We refer interested readers to references [42–44] for a unification of both perspectives, and to references [45–50] for extending FRT to nonsharp null hypotheses.

**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:

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

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 [32]. More concretely, the OLS-based variance estimator is

$$\hat{V}_{\mathrm{OLS}} = \frac{N(N_1 - 1)}{(N-2)N_1N_0}\hat{S}^2(1) + \frac{N(N_0 - 1)}{(N-2)N_1N_0}\hat{S}^2(0) \approx \frac{\hat{S}^2(1)}{N_0} + \frac{\hat{S}^2(0)}{N_1},$$

which can be very different from (13) if the number of units or the sample variances of observed outcomes differ a lot between the two arms. Instead, one can use the Eicker–Huber–White (EHW) variance estimator to obtain a robust estimation:

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

which is asymptotically equivalent to $\hat{V}$ in (4). Alternatively, the so-called HC2 variant of the EHW robust variance estimator is identical to $\hat{V}$ (see Chapter 4 of Ding [51] for a more detailed discussion on regression-based analyses for the ATE).

### 2.1.2 Multi-level CREs

Much 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 [35,52], conjoint analysis [53,54], partially nested experiment [55,56], and sampling-based randomized experiments [57,58].

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 $q = 1,\ldots,Q$ (see the science table in Table 2). Despite its simplicity, the multi-level CRE has been widely used in practice and has generated rich theoretical results. Definition 1

**Table 2:** Science table for multi-level CRE

| $i$ | $Y_i(1)$ | $Y_i(2)$ | $\cdots$ | $Y_i(Q)$ |
|---|---|---|---|---|
| 1 | $Y_1(1)$ | $Y_1(2)$ | $\cdots$ | $Y_1(Q)$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\ddots$ | $\vdots$ |
| $N$ | $Y_N(1)$ | $Y_N(2)$ | $\cdots$ | $Y_N(Q)$ |

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 $\mathcal{Z} \equiv \{z \in \{1, 2, \ldots, Q\}^N : \sum_{i=1}^{N} \mathbf{1}\{z_i = q\} = N_q, \text{ for } 1 \le q \le Q\}$.

Mathematically, Definition 1 implies that $\mathbb{P}(Z = z) = N_1! \ldots N_Q!/N!$ 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, $N_2$ 2'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's [1] framework, all potential outcomes are fixed and only the treatment indicators are random according to Definition 1.

We consider a general contrast matrix $F \in \mathbb{R}^{Q \times H}$ of full column rank, i.e., $F^\top 1_Q = 0_H$ and $\text{rank}(F) = H$, and a set of individual treatment effects defined as the linear contrasts of the potential outcomes:

$$\tau_i = F^\top Y_i(\cdot), \tag{8}$$

where $Y_i(\cdot)$ is the vectorized potential outcome for unit $i$:

$$Y_i(\cdot) = (Y_i(1), \ldots, Y_i(Q))^\top.$$

The average effect is defined as

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i = F^\top \overline{Y}(\cdot), \tag{9}$$

where $\overline{Y}(\cdot)$ is the vectorized average potential outcome:

$$\overline{Y}(\cdot) = \frac{1}{N}\sum_{i=1}^{N} Y_i(\cdot) = (\overline{Y}(1), \ldots, \overline{Y}(Q))^\top.$$

When $Q = 2$ and $F = (1, -1)^\top$, $\tau$ in (9) reduces to the ATE in the treatment-control setting. Moreover, we can estimate $\tau$ by the following generalization of the difference-in-means estimator:

$$\hat{\tau} = F^\top \hat{Y}(\cdot), \tag{10}$$

where $\hat{Y}(\cdot) = (\hat{Y}(1), \ldots, \hat{Y}(Q))^\top$ is the vectorized sample average of observed outcomes for all treatment arms, with $\hat{Y}(q) = N_q^{-1}\sum_{i=1}^{N} Y_i \mathbf{1}\{Z_i = q\}$. The estimator in (10) has variance [59]

$$\text{Var}\{\hat{\tau}\} = F^\top \text{Diag}\left\{\frac{1}{N_q} S(q, q)\right\}_{q=1}^{Q} F - \frac{1}{N} F^\top S F, \tag{11}$$

where $S \in \mathbb{R}^{Q \times Q}$ is a covariance matrix for the potential outcomes with the $(q, q')$th entry given by

$$S(q, q') = \frac{1}{N-1}\sum_{i=1}^{N} (Y_i(q) - \overline{Y}(q))(Y_i(q') - \overline{Y}(q')), \quad q, q' = 1, \ldots, Q, \tag{12}$$

and $F^\top S F$ is essentially the finite population covariance of the individual effects $\tau_i$'s in (8). A variance estimator for (11) is

$$\hat{V} = F^\top \text{Diag}\left\{\frac{1}{N_q} \hat{S}(q, q)\right\}_{q=1}^{Q} F, \tag{13}$$

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

$$\hat{S}(q, q) = \frac{1}{N_q - 1}\sum_{i=1}^{N} (Y_i - \hat{Y}(q))^2 \mathbf{1}\{Z_i = q\}. \tag{14}$$

Using (10) and (13), a Wald-type confidence region for $\tau$ is given by

$$\{\tau : (\hat{\tau} - \tau)^\top \hat{V}^{-1}(\hat{\tau} - \tau) \le q_{H,\alpha}\}, \tag{15}$$

where $q_{H,\alpha}$ is the upper-$\alpha$ quantile of the $\chi^2_H$ distribution. (15) can be proved to be asymptotically valid under mild regularity conditions. More details are deferred to Sections 3 and 4.

In the following, we give two remarks in parallel with Remarks 1 and 2. First, FRT has also been a popular tool for analyzing multiple-level randomized experiments, which can be used to test sharp nulls and deliver finite-sample exact $p$-values [22] (see also [39,43,44] for the unification of Neyman's and Fisher's approaches). Second, similar to the treatment-control case, we can perform analysis with the regression-based approach. Zhao and Ding [60] studied general regression-based analyses in multi-level experiments.

## 2.2 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 $\mathbb{R}^p$. Covariate adjustment has become a standard approach for analyzing randomized experiments and has been widely adopted in many fields. As one example, in 2023, US Food and Drug Administration issued the final guidance on *Adjusting for Covariates in Randomized Clinical Trials for Drugs and Biological Products Final Guidance for Industry*. This guidance describes the agency's current recommendations regarding adjusting for covariates in the statistical analysis of randomized clinical trials in drug and biological product development programs. A natural question is how to optimally adjust covariates for inference? The problem is nontrivial in several aspects: (i) the true relation between outcomes and covariates is usually unknown and (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 while being robust to model misspecification [34].

### 2.2.1 Fisher's analysis of covariance (ANCOVA)

Historically, Fisher [61] proposed to use ANCOVA to improve estimation efficiency. This remains a standard strategy in many fields. He suggested running the OLS of $Y_i$ on $(1, Z_i, X_i)$ and using the coefficient of $Z_i$ as an estimator for $\tau$. 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:

$$(\hat{\tau}_{\mathrm{F}}, \hat{\alpha}_{\mathrm{F}}, \hat{\beta}_{\mathrm{F}}) = \mathop{\arg\min}_{\alpha,\tau\in\mathbb{R},\beta\in\mathbb{R}^p} \sum_{i=1}^N \{Y_i - \alpha - Z_i\tau - (X_i - \overline{X})^\top\beta\}^2, \tag{16}$$

noting that the centering of covariates in (16) will not affect the OLS estimator $\hat{\tau}_{\mathrm{F}}$.

Freedman [32,33] studied Fisher's ANCOVA estimator under the CRE. He showed that $\hat{\tau}_F$ can be biased in the finite sample, but is consistent for the true average effect as the sample size goes to infinity. Moreover, he showed some negative results for Fisher's ANCOVA estimator. First, the asymptotic variance of $\hat{\tau}_{\mathrm{F}}$ can be even larger than the simple difference-in-means estimator $\hat{\tau}$ without adjusting any covariates. Second, the standard error estimator from OLS can underestimate the true standard error of $\hat{\tau}_{\mathrm{F}}$ under the CRE.

### 2.2.2 Lin's estimator

In response to Freedman's negative findings, Lin [34] 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 $Z_i \times X_i$:

$$(\hat{\tau}_{\mathrm{L}}, \hat{\alpha}_{\mathrm{L}}, \hat{\beta}_{\mathrm{L}}, \hat{\eta}_{\mathrm{L}}) = \mathop{\arg\min}_{\alpha,\tau\in\mathbb{R},\beta,\eta\in\mathbb{R}^p} \frac{1}{2N}\sum_{i=1}^N \{Y_i - \alpha - Z_i\tau - (X_i - \overline{X})^\top\beta - Z_i \times (X_i - \overline{X})^\top\eta\}^2. \tag{17}$$

Importantly, unlike (16), the centering of covariates here is critical since it will affect the OLS estimator $\hat{\tau}_{\mathrm{L}}$.

Lin's estimator is also consistent when the sample size $N$ goes to infinity. Moreover, it enjoys several benefits. First, the asymptotic variance of $\hat{\tau}_{\mathrm{L}}$ is not larger than that of both $\hat{\tau}$ and $\hat{\tau}_{\mathrm{F}}$. Second, the EHW variance estimator for (17) is asymptotically conservative for the true variance of $\hat{\tau}_{\mathrm{L}}$. As a side note, the EHW standard error estimator for (16) is also asymptotically conservative for the true variance of $\hat{\tau}_{\mathrm{F}}$ (see Lin [34] 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 [62]. Consider the following class of linearly covariate-adjusted estimators:

$$\begin{aligned}\hat{\tau}(\beta_1, \beta_0) &= N_1^{-1}\sum_{i=1}^{N} Z_i\{Y_i - (X_i - \overline{X})^\top \beta_1\} - N_0^{-1}\sum_{i=1}^{N}(1 - Z_i)\{Y_i - (X_i - \overline{X})^\top \beta_0\} \\ &= \{\hat{Y}(1) - (\hat{X}(1) - \overline{X})^\top \beta_1\} - \{\hat{Y}(0) - (\hat{X}(0) - \overline{X})^\top \beta_0\} \\ &= \hat{\tau} - \delta^\top \hat{\tau}_X,\end{aligned} \tag{18}$$

where $\hat{X}(1)$ and $\hat{X}(0)$ denote the averages of 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 (18) is minimized when $\delta$ is the least-squares coefficient from regressing the difference-in-means estimator $\hat{\tau}$ on the difference-in-means of covariates $\hat{\tau}_X$ under the CRE. Li and Ding [59] showed that this is further achieved when $\beta_1$ and $\beta_0$ are the least-squares coefficients from projecting the 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 analog $\hat{\beta}_1$ 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 (18). 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 (18) can be

$$\hat{V}(\beta_1, \beta_0) = \{N_1(N_1 - 1)\}^{-1}\sum_{i=1}^{N} Z_i\{Y_i - \hat{\gamma}_1 - (X_i - \overline{X})^\top \beta_1\}^2 + \{N_0(N_0 - 1)\}^{-1}\sum_{i=1}^{N}(1 - Z_i)\{Y_i - \hat{\gamma}_0 - (X_i - \overline{X})^\top \beta_0\}^2, \tag{19}$$

where $\hat{\gamma}_1$ and $\hat{\gamma}_0$ are the sample mean of the adjusted outcomes for the treatment and control arm, respectively:

$$\hat{\gamma}_1 = \frac{1}{N_1}\sum_{i=1}^{N} Z_i\{Y_i - (X_i - \overline{X})^\top \beta_1\} \quad \text{and} \quad \hat{\gamma}_0 = \frac{1}{N_0}\sum_{i=1}^{N}(1 - Z_i)\{Y_i - (X_i - \overline{X})^\top \beta_0\}.$$

This formulation suggests choosing $\beta_1$ and $\beta_0$ to minimize the variance estimator $\hat{V}(\beta_1, \beta_0)$ to obtain a plug-in estimator for $\beta_1$ and $\beta_0$, which is equivalent to solving the following two regression problems for treated and control groups, respectively, with intercept terms $\gamma_1$ and $\gamma_0$ [51]:

$$\min_{\gamma_1, \beta_1} \sum_{i=1}^{N} Z_i\{Y_i - \gamma_1 - (X_i - \overline{X})^\top \beta_1\}^2 \quad \text{and} \quad \min_{\gamma_0, \beta_0} \sum_{i=1}^{N}(1 - Z_i)\{Y_i - \gamma_0 - (X_i - \overline{X})^\top \beta_0\}^2. \tag{20}$$

It is not difficult to see that the least-squares estimators for $\beta_1$ and $\beta_0$ from (20) are actually $\hat{\beta}_1$ and $\hat{\beta}_0$ defined before. Consequently, the resulting covariate-adjusted estimator $\hat{\tau}(\hat{\beta}_1, \hat{\beta}_0)$ is equivalent to Lin's estimator $\hat{\tau}_{\mathrm{L}}$. In addition, the corresponding variance estimator constructed as in (19) is asymptotically equivalent to the EHW variance estimator suggested by Lin [34].

From the above, Lin's estimator not only achieves the minimum true variance among all linearly covariate-adjusted estimators in (18), but also achieves the minimum estimated variance when we use the conservative variance estimator of form (19). A subtle issue here is that Lin's estimator uses estimated coefficients rather than fixed ones. With the technical tools discussed later, we can prove that the difference between Lin's estimator and the one with the oracle adjustment coefficients is asymptotically equivalent (see, e.g., Li and Ding [59]).

### 2.2.3 Further extensions

There are a variety of extensions of covariate adjustment beyond treatment-control CREs.

First, it is natural to consider generalization to multiple treatment levels. Lu [63] studied covariate adjustment in $2^K$ factorial designs by extending (20) to multi-level settings. Zhao and Ding [60] considered covariate adjustment in general multi-level experiments and made comprehensive comparison among the unadjusted estimator, Fisher's ANCOVA, and Lin's estimator. The unadjusted estimator is given by the regression:

$$(\hat{\gamma}_{1,\mathrm{N}}, \ldots, \hat{\gamma}_{Q,\mathrm{N}}) = \arg\min_{\gamma_1,\ldots,\gamma_Q} \sum_{i=1}^{N} \left\{ Y_i - \sum_{q=1}^{Q} \gamma_q \mathbf{1}\{Z_i = q\} \right\}^2. \tag{21}$$

The generalization of Fisher's ANCOVA is given by the following *additive treatment regression*:

$$(\hat{\gamma}_{1,\mathrm{F}}, \ldots, \hat{\gamma}_{Q,\mathrm{F}}, \hat{\eta}_{\mathrm{F}}) = \arg\min_{\gamma_1,\ldots,\gamma_Q,\eta} \sum_{i=1}^{N} \left\{ Y_i - \sum_{q=1}^{Q} \gamma_q \mathbf{1}\{Z_i = q\} - (X_i - \overline{X})^\top \eta \right\}^2. \tag{22}$$

Meanwhile, Lin's estimator can be generalized from either the regression with interaction perspective (17) or the (estimated) variance minimization perspective (18) or (20). Here, we present the former one, which applies the following fully interacted regressions:

$$(\hat{\gamma}_{1,\mathrm{L}}, \ldots, \hat{\gamma}_{Q,\mathrm{L}}, \hat{\eta}_{1,\mathrm{L}}, \ldots, \hat{\eta}_{Q,\mathrm{L}}) = \arg\min_{\gamma_1,\ldots,\gamma_Q,\eta_1,\ldots,\eta_Q} \sum_{i=1}^{N} \left\{ Y_i - \sum_{q=1}^{Q} \gamma_q \mathbf{1}\{Z_i = q\} - \sum_{q=1}^{Q} \mathbf{1}\{Z_i = q\}(X_i - \overline{X})^\top \eta_q \right\}^2. \tag{23}$$

With the vectorized slope estimates $\hat{\gamma}_* = (\hat{\gamma}_{1,*}, \ldots, \hat{\gamma}_{Q,*})^\top$, where * = N, F, L, an estimator for the target average effect (9) is given by the plug-in estimator

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

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 [64] proposed the following debiased estimator in treatment-control experiments:

$$\hat{\tau}_{\mathrm{adj}}^{\mathrm{de}} = \hat{\tau}_{\mathrm{L}} - \left( \frac{N_1}{N_0} \hat{\Delta}_0 - \frac{N_0}{N_1} \hat{\Delta}_1 \right),$$

where $\hat{\Delta}_z = N_z^{-1} \sum_{Z_i = z} \hat{e}_i H_{ii}$, $z = 0, 1$, where $\hat{e}_i$ is the $i$th residual from Lin's estimator (17) and $H_{ii}$ is the $i$th diagonal element of the hat matrix $H = X(X^\top X)^{-1} X^\top$, where $X$ is an $N \times p$ matrix with rows consisting of the covariates for the $N$ units. Under some structural conditions, the estimator $\hat{\tau}_{\mathrm{adj}}^{\mathrm{de}}$ achieves asymptotic normality if the following condition holds:

$$\kappa^2 p \log p = o(1), \quad \text{where } \kappa = \max_{i=1,\ldots,N} H_{ii}.$$

In the favorable case where $\kappa = O(p/N)$, the dimension $p$ is allowed to grow as fast as $o(N^{2/3}/\log(N)^{1/3})$, which is a strictly weaker restriction than that of $\hat{\tau}_{\mathrm{L}}$ (see also the study of Lu et al. [65] for some recent development that allows $p$ to be in the same order as $N$). As another example, Bloniarz et al. [66] considered LASSO estimator for covariate adjustment in the high-dimensional regime. Under a sparse linear model and some regularity conditions, the LASSO-adjusted regression estimator is asymptotically normal and the asymptotic variance is not greater than that of the difference-in-means estimator.

Third, some works explored the other variants of Lin's estimators. For example, Zhao and Ding [60] studied restricted least-squares and established for the first time its properties for inferring ATE from the design-based perspective. Guo and Basse [38] considered generalized Oaxaca–Blinder estimators and extend the covariate adjustment framework from linear models to nonlinear ones (see also the study of Cohen and Fogarty [67]).

## 2.3 Rerandomization

Neyman [1] focused on the CREs, 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 ATE. 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 [68], for a realized treatment allocation, the covariates are likely to be imbalanced; for example, with ten independent covariates, at least one of the $t$-statistics for checking the imbalance of these covariates will exceed 2 with a probability of 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 [68], 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 [69–74]. In a recent survey of researchers conducting randomized experiments in developing countries [75], the authors discovered that rerandomization has been commonly used in practice. For example, Lee et al. [76] conducted a rerandomized experiment to study the effect of mobile banking for rural households and their migrated family members.

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 [68] 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 2.2.2, 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, denoted by rerandomization with the Mahalanobis distance (ReM), we will repeatedly draw random treatment assignment from the CRE until obtaining 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 [68] proposed randomization tests for sharp null hypotheses, employing assignments generated randomly in accordance with the rerandomization protocol. More recently, Li et al. [77] 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 ATE 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 under ReM.

In recent years, rerandomization has been extended to more general experiments, such as factorial experiments [78,79], blocked experiments [80,81], and survey experiments [58], and it can also be combined with covariate adjustment discussed in Section 2.2 [82]. Zhao and Ding [83] studied the procedure of conducting rerandomization directly based on $p$-values from covariate balance tests, which is a general strategy that works for many basic designs. An alternative rerandomization scheme that randomizes treatment assignments multiple times and chooses the one with the best covariate balance has also been used in practice [75], and its property has recently been studied in the work of Wang and Li [84].

# 3 Permutational CLTs

With all the design and analysis strategies introduced earlier, one natural question is how to theoretically justify their statistical property. In the following two sections, we focus on the technical aspect of CREs. The main question to answer is how to deliver valid inference with different estimators for different designs. Permutational/combinatorial CLTs and BEBs are core to the technical development of randomization-based inference. In Sections 3 and 4, we summarize the theoretical results regarding permutational CLTs and BEBs and discuss their application in analyzing randomized experiments.

## 3.1 Sample average under simple random sampling

We start with the simple random sampling from a finite population [62]. Let $\{a_N(i)\}_{i=1}^N$ be a sequence of real numbers. Suppose we randomly 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 aforementioned procedure:

$$\Gamma = \frac{1}{N_1}\sum_{i=1}^{N} a_N(i)\mathbf{1}\{Z_i = 1\}. \tag{24}$$

$\Gamma$ has mean and variance

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

where

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

One fundamental technical question is to establish CLTs for $\Gamma$ to characterize its asymptotic distribution. Paul and Rényi [85] established the following CLT for (24):

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

$$\lim_{N\to\infty} \frac{\sum_{i=1}^{N}(a_N(i)-\bar{a}_N)^2\mathbf{1}\{|a_N(i)-\bar{a}_N| > \varepsilon N_1\sqrt{V_N}\}}{\sum_{i=1}^{N}(a_N(i)-\bar{a}_N)^2} \to 0, \tag{25}$$

*then as* $N \to \infty$,

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

Hájek [86] further proved that Condition (25) 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 [87] proved asymptotic normality under the conditions that $N_1 \to \infty$ and there exists $\delta \in (0,1)$ such that $N_1/N < 1-\delta$ when $N$ is sufficiently large and that the following moment condition holds:

$$\frac{N^{-1}\sum_{i=1}^{N}|a_N(i)-\bar{a}_N|^r}{\left\{N^{-1}\sum_{i=1}^{N}(a_N(i)-\bar{a}_N)^2\right\}^{r/2}} = O(1), \quad \text{for all integers } r > 2. \tag{26}$$

The aforementioned moment condition (26) is stronger than (25), because for any $r > 2$,

$$\begin{aligned}
&\frac{\sum_{i=1}^{N}(a_N(i)-\bar{a}_N)^2\mathbf{1}\{|a_N(i)-\bar{a}_N| > \varepsilon N_1\sqrt{V_N}\}}{\sum_{i=1}^{N}(a_N(i)-\bar{a}_N)^2} \\
&= \left(1-\frac{N_1}{N}\right)\frac{N_1\varepsilon^2}{N-1}\sum_{i=1}^{N}\left(\frac{a_N(i)-\bar{a}_N}{\varepsilon N_1\sqrt{V_N}}\right)^2\mathbf{1}\{|a_N(i)-\bar{a}_N| > \varepsilon N_1\sqrt{V_N}\} \\
&\le \left(1-\frac{N_1}{N}\right)\frac{N_1\varepsilon^{2-r}}{N-1}\sum_{i=1}^{N}\left|\frac{a_N(i)-\bar{a}_N}{N_1\sqrt{V_N}}\right|^r \\
&\lesssim \frac{1}{N_1^{\frac{r}{2}-1}}\cdot\frac{N^{-1}\sum_{i=1}^{N}|a_N(i)-\bar{a}_N|^r}{\{N^{-1}\sum_{i=1}^{N}(a_N(i)-\bar{a}_N)^2\}^{r/2}},
\end{aligned}$$

which converges to zero under (26).

David [88] established a CLT for the hypergeometric distribution, which is a special case of Madow [87] thus also stronger than (25). Li and Ding [59, 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 (24) from an simple random sampling (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. Let $\pi$ be a random permutation over the indices $1,\ldots,N$, with $\pi(i)$ denoting the permuted index of $i$. A simple linear rank statistic is defined as

$$\Gamma = \sum_{i=1}^{N} a_N(i)b_N(\pi(i)), \tag{27}$$

which has mean and variance

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

In particular, if we take $b_N(i) = N_1^{-1}$ for $i = 1, \ldots, N_1$ and $b_N(i) = 0$ for $i = N_1 + 1, \ldots, N$, then (27) gives the sample average (24) in SRS. The statistic in (27) has been studied by many researchers. Wald and Wolfowitz [89] established CLT under the following condition: for all integers $r > 2$,

$$\frac{N^{-1}\sum_{i=1}^{N}(a_N(i) - \bar{a}_N)^r}{\left\{N^{-1}\sum_{i=1}^{N}(a_N(i) - \bar{a}_N)^2\right\}^{r/2}} = O(1) \quad \text{and} \quad \frac{N^{-1}\sum_{i=1}^{N}(b_N(i) - \bar{b}_N)^r}{\left\{N^{-1}\sum_{i=1}^{N}(b_N(i) - \bar{b}_N)^2\right\}^{r/2}} = O(1).$$

Noether [90] proved CLT under the following condition that is slightly weaker than Wald and Wolfowitz [89]: for all integers $r > 2$,

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

which, however, is not symmetric for $a_N(i)$'s and $b_N(i)$'s. Hoeffding [91] further proved CLT under a weaker and symmetric condition: for all integers $r > 2$,

$$N^{\frac{r}{2}-1}\frac{\sum_{i=1}^{N}(a_N(i) - \bar{a}_N)^r}{\left\{\sum_{i=1}^{N}(a_N(i) - \bar{a}_N)^2\right\}^{r/2}} \cdot \frac{\sum_{i=1}^{N}(b_N(i) - \bar{b}_N)^r}{\left\{\sum_{i=1}^{N}(b_N(i) - \bar{b}_N)^2\right\}^{r/2}} = o(1). \tag{28}$$

Motoo [92] proved that CLT holds under an even weaker Lindeberg-type condition:

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

$$\lim_{N\to\infty} \frac{\sum_{i,j=1}^{N}(a_N(i) - \bar{a}_N)^2(b_N(j) - \bar{b}_N)^2 \mathbf{1}\{|(a_N(i) - \bar{a}_N)(b_N(j) - \bar{b}_N)| > \varepsilon\sqrt{V_N}\}}{\sum_{i,j=1}^{N}(a_N(i) - \bar{a}_N)^2(b_N(j) - \bar{b}_N)^2} = 0. \tag{29}$$

*Then,*

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

Hájek [93] proved further that Condition (29) is not only sufficient but also necessary, and presented a comprehensive comparison of the CLT conditions introduced in the literature. There are also several multi-dimensional extensions based on the Cramér-Wold device (see, for example, refs [94], [93, Section 7], and [95, Lemma S.3.3]).

## 3.3 General univariate linear permutational statistics

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

$$\Gamma = \sum_{i=1}^{N} M_N(i, \pi(i)), \tag{30}$$

where $\{M_N(i,j)\}_{i,j\in[N]}$ is a matrix in $\mathbb{R}^{N\times N}$. In particular, if we take $M_N(i,j) = a_N(i)b_N(j)$, (30) recovers (27). Hoeffding [91] computed the mean and variance of (30):

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

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

$$\tilde{M}_N(i,j) = M_N(i,j) - \frac{1}{N}M_N(i,+) - \frac{1}{N}M_N(+,j) + \frac{1}{N^2}M_N(+,+), \tag{31}$$

where "+" means summation over the corresponding index.

Moreover, Hoeffding [91] showed that the asymptotic normality of $\Gamma$ in (30) holds under the following condition:

$$\lim_{N\to\infty} \frac{N^{-1}\sum_{i,j=1}^N \tilde{M}_N(i,j)^r}{\{N^{-1}\sum_{i,j=1}^N \tilde{M}_N(i,j)^2\}^{r/2}} = 0, \quad \text{for all integers } r > 2. \tag{32}$$

Condition (32) is equivalent to Condition (28) in the simple linear rank statistics setting. A more compact sufficient condition for (32) is also provided in Hoeffding [91]:

$$\lim_{N\to\infty} \frac{\max_{i,j\in[N]}\tilde{M}_N(i,j)^2}{N^{-1}\sum_{i,j=1}^N \tilde{M}_N(i,j)^2} = 0. \tag{33}$$

Motoo [92] weakened Hoeffding's [91]'s condition (32) to the following Lindeberg-type condition:

**Proposition 3.** (Main theorem of Motoo [92]) *Suppose for any* $\varepsilon > 0$,

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

*Then*,

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

**Remark 3.** Although Proposition 3 gives the weakest condition for permutational CLT in the literature, it is not very convenient for use in many concrete examples. On the contrary, Condition (33) involves the maximum of the centered matrices and is simpler for use and interpretation. Condition (33) and its multivariate generalization (presented in (36) in Section 3.4) are frequently applied to investigation of the properties of various analysis and design strategies in randomized experiments presented in Section 2; for example, they can faciliate the proof for the convergence of variance estimation. We will have more discussion in Section 3.5.

## 3.4 General multivariate linear permutational statistics

We now discuss the generalization of (32) to a multivariate case. Concretely, define the multivariate linear permutational statistics:

$$\Gamma = (\Gamma_1, \ldots, \Gamma_H)^\top, \quad \Gamma_h = \sum_{i=1}^N M_{N,h}(i,\pi(i)), \tag{34}$$

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 [96, Appendix A.1.] presented many basic facts about (34), including its mean and covariance calculation and its standardization. Fraser [94] extended Hoeffding [91] to the multi-dimensional setting by applying the Cramér–Wold device to establish a multivariate CLT. More concretely, define the centered version of $\tilde{M}_{N,h}$ in the same way as (31). Fraser [94] proposed the following condition for CLT as an extension to (32):

$$\lim_{N\to\infty} \frac{N^{-1}\sum_{i,j=1}^N \tilde{M}_{N,h}(i,j)^r}{\{N^{-1}\sum_{i,j=1}^N \tilde{M}_{N,h}(i,j)^2\}^{r/2}} = 0, \quad \text{for all integers } r > 2 \text{ and } h \in [H]. \tag{35}$$

Similarly, Fraser [94] also provided a sufficient condition for (35):

$$\lim_{N\to\infty} \frac{\max_{i,j\in[N]} \tilde{M}_{N,h}(i,j)^2}{N^{-1}\sum_{i,j=1}^{N} \tilde{M}_{N,h}(i,j)^2} = 0, \quad \text{for } h \in [H]. \tag{36}$$

The condition in (36) is further utilized by Li and Ding [59] to build asymptotic normality results for analyzing treatment effects in multi-level CREs.

## 3.5 Application of permutational CLT in randomization-based inference

In this subsection, we collect several theoretical arguments in randomization-based inference that apply permutational CLTs to deliver technical justification for studying ATEs.

### 3.5.1 Wald-type inference in CREs

Consider analyzing a multi-level CRE discussed in Section 2.1.2 and adopt the notation introduced there. Li and Ding [59] proved the following result to justify the asymptotic validity of the confidence region (15) under several regularity conditions.

**Proposition 4.** (Theorem 5 and Proposition 3 of Li and Ding [59]) *Let $Q$ be fixed and $N$ go to infinity. If the covariance matrix $S$ has 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 semi-positive definite limiting value $V_\infty$, and*

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

*where $\tau$ and $\hat{\tau}$ are the ATE and the corresponding estimator in* (9) *and* (10).

(2) *Variance estimation. The sample variance $\hat{S}(q,q)$ in* (14) *is consistent for $S(q,q)$ in* (12).

(3) *Wald-type inference. If the limit of $NF^\top \mathrm{Diag}\{N_q^{-1}S(q,q)\}F$ is nonsingular, then $\hat{V}$ in* (13) *is nonsingular with probability converging to one, and the Wald-type confidence region* (15) *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, the estimate $\hat{\tau}$ follows the same distribution as (34) with matrices $M_{N,h}$ defined as follows: for $i,j \in [N]$,

$$M_{N,h}(i,j) = N_q^{-1}F(q,h)Y_i(q), \quad N_{q-1}+1 \le j \le N_q,$$

where $N_0 = 0$ and $F(q,h)$ is the $(q,h)$th element of the contrast matrix $F$. Now, applying Condition (36), we can obtain the regularity conditions on potential outcomes in Proposition 4 and justify the multivariate asymptotic normality of the estimator $\hat{\tau}$. 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 (15). As a side remark, the matrix $V_\infty$ is not required to be invertible, because the multivariate combinatorial CLT proved by Fraser [94] does not require an invertible limit for the covariance matrix. However, to justify the validity of the Wald-type confidence intervals in part (iii), invertibility is required. 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 true asymptotic variance of $\hat{\tau}$. These then justify the asymptotic validity of the level-$(1-\alpha)$ confidence interval in (5).

### 3.5.2 Analyzing covariate adjustment

For covariate adjustment, we present a result by Zhao and Ding [60]. Let $\hat{Y}_* \in \mathbb{R}^Q$ (* = N, F, L) be the estimators for the averaged potential outcomes across all $Q$ treatment levels from three estimation strategies: $\hat{Y}_N$ from Neyman's approach, $\hat{Y}_F$ from Fisher's ANCOVA, and $\hat{Y}_L$ from Lin's regression. These three estimators correspond to the coefficients $\hat{\gamma}_*$ in front of the treatment indicators from the regressions introduced in Section 2.2.3, i.e., (21), (22), and (23). We slightly modified the notation in order to better present the results. Also, define $\hat{V}_*$ to be the corresponding EHW robust covariance estimator from the three regressions. The following proposition from Zhao and Ding [60] established the asymptotic properties of these point and variance estimators.

**Proposition 5.** (Lemma 1 of Zhao and Ding [60]) *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, where the covariates have been centered so that $N^{-1}\sum_{i=1}^N X_i = 0$. 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*:

(1) *Asymptotic normality.* $\sqrt{N}(\hat{Y}_* - \overline{Y}) \rightsquigarrow \mathcal{N}(0, V_*)$ *for some* $V_* \geq 0$, * = N, F, L.

(2) *Conservative variance estimation.* $\operatorname{plim}_{N\to\infty} N\hat{V}_{*,\mathrm{EHW}} \geq V_*$, * = N, F, L.

(3) *Efficiency comparison.* $V_L \leq V_N$ *and* $V_L \leq V_F$.

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 these estimators. Proposition 5 (iii) indicates that Lin's estimator guarantees at least as much asymptotic efficiency as the difference-in-means estimator and Fisher's ANCOVA.

In terms of the technical derivation, Proposition 5 (i) utilized Proposition 4, which is a result motivated by Hoeffding and Fraser's permutational CLT (more specifically, Conditions (33) and (36)) and can accommodate vector outcomes and multi-level randomized experiments. In particular, if we study the pseudo-potential outcome vector $(Y_i(q), X_i^\top)^\top$, for $q \in [Q]$, we can apply similar tricks as in Section 3.5.1 to establish a multivariate CLT for the arm-wise sample means $(\hat{Y}_q, \hat{X}_q^\top)^\top$ for $q \in [Q]$ based on Condition (36). The covariate-adjusted estimator $\hat{\gamma}_L$ in (23) can be formulated as linear combinations of these sample means, where the combination coefficients are consistent for certain constant coefficients in the sense that their difference is of order $o_{\mathbb{P}}(1)$. Then, a CLT can be derived after filling in the details [60]. Proposition 5 (ii) utilized the Chebyshev inequality under the bounded moment conditions. Proposition 5 (iii) involves some delicate analysis of the limiting variance structure $V_*$ for * = N, F, L, which has closed-form expressions [60].

### 3.5.3 Analyzing rerandomization

For rerandomization using the Mahalanobis distance discussed in Section 2.3, we adopt the notation introduced there and present the following result by Li et al. [77]. Define $V$ as the variance of $\sqrt{N}\hat{\tau}$ under the CRE, and $R^2$ as the squared multiple correlation between the difference-in-means of outcome and covariates (see Proposition 1 of Li et al. [77] for its explicit expression). Let $\varepsilon_0 \sim \mathcal{N}(0, 1)$, $L_{K,a} \sim D_1 | D'D \leq a$ with $D = (D_1, \ldots, D_K)^\top \sim \mathcal{N}(0, I_K)$, and $\varepsilon_0 \perp\!\!\!\perp L_{K,a}$.

**Proposition 6.** (Li et al. [77], Theorems 1 and 2 and Appendix A4.2) *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) $N^{-1}\max_{1\leq i\leq N}|Y_i(z) - \overline{Y}(z)|^2 \to 0$ and $N^{-1}\max_{1\leq i\leq N}||X_i - \overline{X}||_2^2 \to 0$.*

(i) *Asymptotic distribution.* $\sqrt{N}(\hat{\tau} - \tau)|M \leq a \rightsquigarrow V^{1/2}(\sqrt{1-R^2}\cdot\varepsilon_0 + \sqrt{R^2}\cdot L_{K,a})$.

(ii) *Conservative inference. We can construct estimators* $\hat{V}$ *and* $\hat{R}^2$ *such that, as* $N \to \infty$, $\text{plim}_{N\to\infty}(\hat{V} - V) \geq 0$ *and* $\text{plim}_{N\to\infty}(\hat{V}\hat{R}^2 - VR^2) = 0$.

(iii) *Efficiency comparison. The asymptotic distribution under ReM has a smaller (or equal) variance and narrower (or equal) symmetric quantile ranges than that under the CRE. Here, the symmetric quantile range means the interval formulated by the lower and upper* $\alpha/2$ *th quantile of the asymptotic distribution of* $\sqrt{N}(\hat{\tau} - \tau)$, *for* $\alpha \in (0, 1)$.

Proposition 6 (i) means that the distribution of the difference-in-means estimator under rerandomization converges in distribution to the same limit as $V^{1/2}(\sqrt{1 - R^2} \cdot \varepsilon_0 + \sqrt{R^2} \cdot L_{K,a})$, a convolution of a Gaussian and a constrained Gaussian random variable, where the coefficient depends crucially on $R^2$ that represents an $R^2$-type measure for the association between potential outcomes and covariates (see Li et al. [77] 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 [77]. 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. In Proposition 6(ii), we omit the explicit expressions of the estimators for conciseness. As a side remark, Li et al. [77] used an estimator for $V$ that is less conservative than or asymptotically equivalent to references (4) by utilizing the covariate information; we refer interesting readers to references [77,97] for details. Importantly, Proposition 6(ii) shows that we can consistently estimate the coefficient of $L_{K,a}$ in the asymptotic distribution, while only conservatively estimating the coefficient of $\varepsilon_0$. 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 [77]. Branson et al. [98] recently extended the comparison to non-symmetric quantile ranges.

We now discuss the technical aspects of Proposition 6. A key for its derivation is to note 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 Li et al. [77]). Proposition 6(iii) involves careful analysis of the non-Gaussian distribution.

# 4 Permutational BEBs

## 4.1 Several univariate and multivariate permutational BEBs

Recently, permutational BEBs (also called combinatorial BEBs) start to raise attention in randomization-based inference for experiments. BEBs depict the distance between the sampling distribution of a statistic and a target, often normal, distribution. Theoretically speaking, it measures the convergence rate of CLTs. 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, BEBs consider $\mathcal{H}$ 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:

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

In the multivariate case, there are many choices of sets for different purposes, such as Euclidean balls [99], rectangular sets [100], and measurable convex sets [101,102].

Below, we review some theoretical progresses of permutational BEBs and their important application for analyzing randomized experiments in finite populations.

### 4.1.1 Univariate case

We consider the univariate linear permutational statistics in (30) and adopt the notation from Section 3.3. We will summarize BEB results for $\Gamma$ upon standardization. The standardized version of $\Gamma$ can be expressed as

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

where

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

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

**Condition 1.** (Normalizing $\Gamma$) $\Gamma$ in Definition (30) is defined with $M_N$ satisfying the following normalizing condition:

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

Von Bahr [103] and Ho and Chen [104] established some early results. Bolthausen [105] applied one version of Stein's method [106] to establish the following result requiring only conditions concerning the third moment of the matrix $M_N$.

**Proposition 7.** (Main theorem of Bolthausen [105]) *Assume Condition* 1. *There exists some universal constant* $C > 0$, *such that*

$$\sup_{t\in\mathbb{R}}|\mathbb{P}\{\Gamma \le t\} - \Phi(t)| \le CN^{-1}\sum_{i,j\in[N]}|M_N(i,j)|^3. \tag{37}$$

The upper bound on the right-hand-side of (37) achieves the rate of $O(N^{-1/2})$ if

$$N^{-1/2}\sum_{i,j\in[N]}|M_N(i,j)|^3 = O(1). \tag{38}$$

von Bahr [103] imposed the following boundedness condition, which is sufficient for (38):

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

As a side note, the aforementioned boundedness condition is also sufficient for the BEB in the study of Ho and Chen [104] to achieve the $O(N^{-1/2})$ convergence rate. Chen et al. [107, Chapter 6.1] presented a thorough discussion about the univariate permutational BEB. Proposition 7 is very helpful for analyzing the finite-

sample quality of normal approximation for linear estimators. In Section 4.2, we provide an example of using Proposition 7 to analyze CREs with possibly varying group sizes and diverging treatment levels.

### 4.1.2 Multivariate case

We now consider the multivariate linear permutational statistics (34) in Section 3.4. For the ease of presentation, we focus on results upon standardization. Specifically, Shi and Ding [96, Lemma S2] proved that the standardized version of (34) 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:

$$\check{M}_{N,h}(i,+) = \check{M}_{N,h}(+,j) = \check{M}_{N,h}(+,+) = 0, \qquad \text{for all } i,j \in [N] \text{ and } h \in [H]; \tag{39}$$

$$\sum_{i,j\in[N]}\check{M}_{N,h}(i,j)^2 = N-1, \qquad \text{for all } h \in [H]; \tag{40}$$

$$\sum_{i,j\in[N]}\check{M}_{N,h}(i,j)\check{M}_{N,h'}(i,j) = 0, \qquad \text{for all } h \neq h' \in [H]. \tag{41}$$

$\check{M}_{N,h}$s can be constructed from $M_{N,h}$s by performing the centering step as in (31) and then applying a linear combination using the matrix $\mathrm{Var}\{\Gamma\}^{-1/2}$. Therefore, without loss of generality, we assume the following condition.

**Condition 2.** (Normalizing $\Gamma$ in the multivariate case) $\Gamma$ in (34) is defined with $M_{N,h}$'s satisfying the normalizing conditions in (39), (40), and (41), which guarantees that $\mathbb{E}\{\Gamma\} = 0_H$ and $\mathrm{Var}\{\Gamma\} = I_H$.

Bolthausen and Gotze [108] extended the univariate result in Proposition 7 to the multivariate, possibly nonlinear setting. In particular, for the multivariate linear case, Bolthausen and Gotze [108, Theorem 1] established the following BEB.

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

$$\sup_{A\in\mathcal{A}}|\mathbb{P}\{\Gamma \in A\} - \mathbb{P}\{\Gamma_Z \in A\}| \le \frac{C_H}{N}\sum_{i,j\in[N]}\left(\sum_{h=1}^{H}M_{N,h}(i,j)^2\right)^{3/2}. \tag{42}$$

The BEB in (42) covers the univariate case in Proposition 7 as a special case with $H = 1$. However, Bolthausen and Gotze [108] did not give a closed form expression for $C_H$, whose dependence on the dimension $H$ is unknown. Raic [109] conjectured the following result:

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

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 [110] made one step forward to reveal the dimensional dependence using Stein's method with multivariate exchangeable pairs. In Chatterjee and Meckes [110, 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 in the following a special case of Chatterjee and Meckes [110]'s result.

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

$$\sup_{i,j\in[N],h\in[H]} |M_{N,h}(i,j)| = O(N^{-1/2}), \tag{43}$$

*we have*

$$\sup_{g\in C^2(\mathbb{R}^H)} |\mathbb{E}\{g(\Gamma)\} - \mathbb{E}\{g(\Gamma_Z)\}| = O\left(\frac{H^3}{N^{1/2}}\right). \tag{44}$$

Nevertheless, (44) does not translate directly into a BEB under the Kolmogorov metric because the indicator functions are not members of $C^2(\mathbb{R}^H)$. Shi and Ding [96, Theorem S2] made use of one key result established by Fang and Röllin [111] 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* (43), *we have*

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

Proposition 10 is also useful for analyzing the finite-sample performance of many non-linear permutational statistics. Shi and Ding [96] used Proposition 10 to obtain a BEB for quadratic forms of a multi-dimensional estimator for causal effects in CRE, which builds up the ground for Wald-type inference (see Appendix E of Shi and Ding [96] for more discussion).

## 4.2 Application of permutational BEBs to randomization-based inference

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

### 4.2.1 CREs with possibly varying group sizes and diverging treatment levels

Many classical experiments only involve a small number of treatment levels. For example, classical factorial experiments typically include a small number of factors (like $K \le 5$) [52]. However, many modern experiments involve a much larger number of treatment levels and units due to the need for analyzing more complex relations as well as the development of experimentation technologies. For example, in political science, powered by the development of computers and web-based technology, conjoint survey experiments [53,112,113] (as a special type of factorial experiments) are very popular for analyzing the effects of many factors together and answering complex causal questions. Zhirkov [113] investigated an experiment examining the impact of six different attributes of immigrants on public support for their admission to the United States. Caughey et al. [112] studied the impact of 12 ($K = 12$) personal traits on citizens' preference for U.S. presidential candidates. A large number of treatment levels pose new challenges to the analysis of randomized experiments and call for new methodological and theoretical developments. Shi and Ding [96] and Shi et al. [114] 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 presents several possible regimes that are of interest both technically and practically.

**Table 3:** Theoretical results for multi-level experiments under the randomization model, originally from Table 1 of the study of Shi and Ding [96]

| 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 |

The columns "$Q$" 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.

Most of the regimes in Table 3 are less visited by literature and lack scientific justification. Shi and Ding [96] 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 on the usage of permutational BEBs with a canonical example in factorial experiments from Shi and Ding [96].

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 $\tau = F^\top \overline{Y}(\cdot)$ may consist of a subset of factorial effects, where $F$ is a contrast matrix with orthogonal columns and entries of $\pm(Q/2)^{-1}$ (see Dasgupta et al. [35] 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 $\underline{c} \leq \bar{c}$ such that $N_q = c_q N_0$ with $\underline{c} \leq c_q \leq \bar{c}$, for all $q = 1, \ldots, Q$. Such a setup can cover many cases in regimes (R1)–(R4) in Table 3. Shi and Ding [96] established the following result for the plug-in estimator $\hat{\tau} = F^\top \hat{Y}(\cdot)$:

**Proposition 11.** (Shi and Ding [96], 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 interactions. Recall the definition of $S(q, q)$ from* (12)*. Under some mild regularity conditions, we have*

$$\sup_{b \in \mathbb{R}^H, \|b\|_2 = 1} \sup_{t \in \mathbb{R}} |\mathbb{P}\{b^\top \tilde{\tau} \leq t\} - \Phi(t)| \leq C\sigma_F \frac{\max_{q \in [Q], i \in [N]} |Y_i(q) - \overline{Y}(q)|}{\{\min_{q \in [Q]} S(q, q)\}^{1/2}} \sqrt{\frac{K^2}{N}}, \tag{45}$$

*where $C > 0$ is an absolute constant, and $\sigma_F > 0$ is a certain constant related to the matrix $F$.*

Proposition 11 is established based on the permutational BEB from Bolthausen [105] (presented in Proposition 7 in Section 4.1). In particular, one can formulate $b^\top \tilde{\tau}$ as a linear permutational statistic with a carefully defined matrix $M_N$ and apply Proposition 7 to obtain a raw BEB. After taking supreme over all unit-norm vector $b$, the BEB can be simplified to the presented form (45). More technical details are provided in Appendix A of Shi and Ding [96]. Also, the BEB in (45) is uniform in the linear coefficient vector $b \in \mathbb{R}^H$ with $\|b\|_2 = 1$. This uniformity results in the additional dependence in $K^2$ (or the dimension $H$). Intuitively with higher dimension $H$, the uniform bound becomes larger. From Proposition 11, we can obtain a sufficient condition for the upper bound (45) to converge to 0, which implies a CLT for any one-dimensional linear transformation of $\tilde{\tau}$. In addition, Proposition 11 requires mainly the total sample size $N$ to be large enough, and therefore allows either a fixed number of treatment levels $Q$ and diverging replications $N_0$, or a diverging $Q$ with limited replications $N_0$. Shi and Ding [96] also established design-based properties of Wald-type inference under general CREs, which utilizes multivarite permutational BEBs such as Proposition 10.

#### 4.2.2 Rerandomization with diminishing covariate imbalance and diverging number of covariates

Li et al. [77] studied the asymptotic theory of rerandomization with a fixed covariate imbalance threshold that does not vary with the sample size, as discussed in Sections 2.3 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 asymptotic 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 asymptotic analysis.

In the context of simple random sampling, Wang and Li [115] derived a multivariate BEB for the sample average using Hájek's [86] coupling and the BEB for sums of independent random vectors [116] with explicit dependence on the dimension. The bound, although weaker than that implied by the conjecture in Raic [116], is sufficient for studying rerandomization with diminishing covariate imbalance threshold (or equivalently acceptance probability) and diverging number of covariates. With the derived BEBs, Wang and Li [115] presented 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 $N$. 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 average 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.** (Wang and Li [115], Theorems 1 and 3) *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*

$$\sup_{c\in\mathbb{R}} |\Pr\{\mathrm{Var}\{\hat{\tau}\}^{-1/2}(\hat{\tau}-\tau) \le c | M \le a_N\} - \Pr(\sqrt{1-R^2}\,\varepsilon_0 + \sqrt{R^2}L_{K_N,a_N} \le c)| \to 0,$$

*where* $\tau$ *is the true ATE.*

Wang and Li [115] further studied additional conditions such that the constrained Gaussian random variable $L_{K_N,a_N}$ becomes ignorable as $N \to \infty$, under which $\mathrm{Var}\{\hat{\tau}\}^{-1/2}(\hat{\tau}-\tau)$ 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 [115] further proposed large-sample valid confidence intervals for the ATE under rerandomization.

## 5 Extensions

Neyman [1] 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 Neyman [1].

## 5.1 Other randomized experiments

In this section, we discuss several other widely used and studied randomized experiments, beyond Neyman's [1] focus on the CRE.

### 5.1.1 Stratified (block) randomized experiments

Stratified randomized experiments (SRE) have been used widely in many fields, including agricultural study [117], biomedical study [118], and social science [119]. An 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 the remaining $N_{[k]0} = N_{[k]} - N_{[k]1}$ 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

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

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

$$\hat{\tau}_{\mathrm{S}} = \sum_{k=1}^{K} \pi_{[k]} \hat{\tau}_{[k]}, \tag{46}$$

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

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

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

Under certain regularity conditions, the point estimator (46) is asymptotically normal and Wald-type inference is proved to be asymptotically valid (see, for example, previous studies [120,121]). The random assignment mechanism requires studying a convolution of independent permutational distributions, which motivates new theoretical tools. When the total number of strata $K$ is small and the sizes of the strata are large, the permutational CLTs 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, previous studies [120,121]). Moreover, Liu et al. [122] and Wang et al. [80] further investigated covariate adjustment and rerandomization in SREs.

### 5.1.2 Matched-pair experiments (MPEs)

The MPE is another popular experimental design in practice [19,123,124]. 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 notations for the 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 set so that we can calculate the stratum-specific sample variances. Imai [124] 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

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

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 in the study by Fogarty [37].

In general stratified experiments with possibly one treated or one control unit in some strata, Fogarty [125] and Pashley and Miratrix [126] 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 and $M$ 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 $j = 1, \ldots, n_i$. The experimenter randomly assigns $M_1$ clusters to receive the treatment and $M_0$ clusters to receive the control, where $M_1$ and $M_0$ are fixed positive integers satisfying $M_1 + M_0 = M$. 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 ATE 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 $\tau$, including individual-level estimators and cluster-level estimators [127], both enjoying desirable asymptotic properties implied by permutational CLTs. We refer interested readers to a collection of works on analyzing cluster randomized experiments [127–132].

## 5.2 Some technical aspects for permutations

Permutation is a core element in the design and analysis of randomized experiments, and its development, such as CLTs and BEBs, has involved many technical tools including moment matching [91], coupling [93,133], Stein's method [105–107,110]. In this section, we discuss some technical aspects for analyzing permutation-related problems.

### 5.2.1 Hajek's coupling

Hajek's coupling is one technique developed in the study by Hájek [86] for proving CLTs in simple random sampling. The idea is based on constructing a coupling between simple random sampling and Bernoulli random sampling so that CLTs for independent and identically distributed (i.i.d.) sampling can be applied. Wang and Li [115] utilized Hajek's coupling together with a multivariate BEB for sum of independent random vectors [116] to study rerandomization with diminishing covariate imbalance. The techniques are useful to establish theories for a wide range of permutational statistics (see, e.g., previous studies [122,133]).

### 5.2.2 Double and multiple permutations

Nowadays, there are many new variants of permutations in designing randomized experiments. For example, Fredrickson and Chen [134] and Chen and Friedman [135] discussed permutation and randomization tests for analyzing network data. D'Amour and Airoldi [136] and Deng et al. [137] studied randomized experiments with dyadic outcomes, i.e., outcomes that measure the relationship between pairs of units. When randomization occurs at the unit level, the dyadic outcomes are in turn randomized with double permutations. *Doubly indexed permutation statistics* (DIPS) is useful in these settings because the dyadic potential outcomes are functions of pairs of treatments, and the statistics for studying causal effects in these problems are generally represented as DIPS. Bajari et al. [138,139] 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. [107], Zhao et al. [140], Reinert and Röllin [141], among others, used Stein's method [106] to study the asymptotic properties of DIPS.

### 5.2.3 Concentration inequalities

Another technical tool that has been recognized by the literature is permutational/combinatorial concentration inequalities. Bloniarz et al. [66] and Lei and Ding [64] used permutational concentration inequalities to analyze regression adjustment in CREs when the dimension of the covariates is diverging. It will be interesting to explore related potential research questions that involve delicate analysis of finite sample properties of permutational statistics and inspire the use of concentration inequalities.

## 6 Conclusion

In this review, we revisited the fundamental contributions of Neyman's [1] seminal work regarding the introduction of potential outcomes, the promotion of physical randomization, and the emphasis of repeated sampling properties of statistics over the randomization distribution. These contributions lay down the foundation for the design and analysis of randomized experiments. We also reviewed permutational CLTs and BEBs in great detail, and listed applications of these technical results in randomization-based inference.

Beyond what we have covered in the review, many research works are closely related to Neyman [1]. From a technical point of view, many theoretical tools are not fully covered in the discussion. For example, when analyzing SREs, we need CLTs and BEBs that combine the independent permutational distributions [121,122]. This is also closely related to Rosenbaum's sensitivity analysis for matched observational studies with biased permutations in each matched set [20,142,143]. As another example, for the design and analysis of adaptive experiments, a general martingale structure typically exists, which requires a martingale CLT or Berry–Esseen result [144,145].

From a practical point of view, many real-world examples can motivate the study of new designs, outcomes, assumptions, and causal estimands under the finite population framework. For example, interference among units is a common phenomenon in many experimental and observational studies. The study of interference and peer influence has motivated a lot of new designs and methods, such as designing and analyzing bipartite experiments [8,146], multiple randomization [138], randomized experiments with network interference [147], group formulation design [148,149], etc. Another example is randomization with missing observations or covariates. Zhao and Ding [150] discussed several strategies for randomization-based inference with missing covariates, and Zhao et al. [151] further studied covariate adjustment in randomized experiments with both missing outcomes and covariates. Censored survival outcomes are another type of missingness that occurs frequently in clinical trials. In these settings, to test the null hypothesis of no treatment effect for any unit, Rosenbaum [20] proposed randomization tests for censored outcomes using a partial ordering, and Zhang and Rosenberger [152] established asymptotic normality of the randomization distribution of the log-rank statistic. Both approaches require the assumption of identical potential censoring times under treatment and control. Recently, Li and Small [153] relaxed this assumption and proved that, under a Bernoulli randomized experiment, with non-informative i.i.d. censoring, the log-rank test is asymptotically valid for testing Fisher's null hypothesis of no treatment effect on any unit.

At the same time, there have been extensive progresses for analyzing treatment effects from a super-population perspective, and many of them share similar spirit as the randomization-based inference [154]. For example, Yang and Tsiatis [155] have suggested linear covariate adjustment with treatment–covariate interaction under a semiparametric model (see also previous studies [156,157]). In the presence of censoring, there have been many works studying semiparametric estimation of treatment effect [158–160], as well as covariate adjustment to improve inference efficiency [161–163].

**Acknowledgements:** We thank the reviewers for carefully reading the manuscript and providing many constructive suggestions for improving the manuscript.

**Funding information**: X. L. was partly supported by the National Science Foundation under Grant DMS-2400961.

**Author contributions**: All authors have accepted responsibility for the entire content of this manuscript and consented to its submission to the journal, reviewed all the results, and approved the final version of the manuscript. LS and XL worked together to frame the structure of the review, collect related literature, and organize the presentation of the methodology and theory for the design and analysis of randomized experiments.

**Conflict of interest:** Authors state no conflict of interest.

## References

[1] Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science. 1923/1990. pp. 465–72.
[2] Pitman EJ. Significance tests which may be applied to samples from any populations. Suppl J R Stat Soc. 1937;4(1):119–30.
[3] Welch BL. On the z-test in randomized blocks and Latin squares. Biometrika. 1937;29(1/2):21–52.
[4] Kempthorne O. The design and analysis of experiments. New York: Wiley; 1952.
[5] Kempthorne O. The randomization theory of experimental inference. J Amer Stat Assoc. 1955;50(271):946–67.
[6] Hudgens MG, Halloran ME. Toward causal inference with interference. J Amer Stat Assoc. 2008;103(482):832–42.
[7] Tchetgen EJT, VanderWeele TJ. On causal inference in the presence of interference. Stat Methods Med Res. 2012;21(1):55–75.
[8] Zigler CM, Papadogeorgou G. Bipartite causal inference with interference. Stat Sci Rev J Inst Math Stat. 2021;36(1):109.
[9] Liu L, Wang Y, Xu Y. A practical guide to counterfactual estimators for causal inference with time-series cross-sectional data. Amer J Politic Sci. 2022;68:160–76.

[10] Sjölander A, Frisell T, Kuja-Halkola R, Öberg S, Zetterqvist J. Carryover effects in sibling comparison designs. Epidemiology. 2016;27(6):852–8.
[11] Imai K, Kim IS, Wang EH. Matching methods for causal inference with time-series cross-sectional data. Amer J Politic Sci. 2023;67(3):587–605.
[12] Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Amer Stat Assoc. 1996;91(434):444–55.
[13] VanderWeele T. Explanation in causal inference: methods for mediation and interaction. New York, NY: Oxford University Press; 2015.
[14] VanderWeele TJ. Mediation analysis: a practitioner's guide. Ann Rev Public Health. 2016;37:17–32.
[15] Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688.
[16] Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci. 1990;5(4):472–80.
[17] Student. On testing varieties of cereals. Biometrika. 1923;15:271–93.
[18] Fisher RA, Mackenzie WA. Studies in crop variation. II. The manurial response of different potato varieties. J Agricult Sci. 1923;13(3):311–20.
[19] Fisher RA. The design of experiments. 1st ed. Edinburgh, London: Oliver and Boyd; 1935.
[20] Rosenbaum PR. Observational studies. New York, NY: Springer-Verlag; 2002.
[21] Hinkelmann K, Kempthorne O. Design and analysis of experiments, Volume 1: Introduction to experimental design. vol. 1. Hoboken, NJ: John Wiley & Sons; 2007.
[22] Imbens GW, Rubin DB. Causal inference in statistics, social, and biomedical sciences. New York: Cambridge University Press; 2015.
[23] Splawa-Neyman J. Contributions to the theory of small samples drawn from a finite population. Biometrika. 1925;17:472–9.
[24] Neyman J. On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. J R Stat Soc Ser A: Stat Soc. 1934;97(4):558–606.
[25] Neyman J, Iwaszkiewicz K. Statistical problems in agricultural experimentation. Suppl J R Stat Soc. 1935;2(2):107–80.
[26] Fienberg SE, Tanur JM. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation and sampling. Int Stat Rev/Revue Int de Stat. 1996;64:237–53.
[27] Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. J Amer Stat Assoc. 2005;100(469):322–31.
[28] Scheffé H. The analysis of variance. New York: John Wiley & Sons; 1959.
[29] Copas JB. Randomization models for the matched and unmatched $2 \times 2$ tables. Biometrika. 1973;60(3):467–76.
[30] Robins JM. Confidence intervals for causal parameters. Stat Med. 1988;7(7):773–85.
[31] Hinkelmann K, Kempthorne O. Design and analysis of experiments, introduction to experimental design. vol. 1. New York: John Wiley & Sons; 2007.
[32] Freedman DA. On regression adjustments to experimental data. Adv Appl Math. 2008;40(2):180–93.
[33] Freedman DA. On regression adjustments in experiments with several treatments. Ann Appl Stat. 2008;2:176–96.
[34] Lin W. Agnostic notes on regression adjustments to experimental data: Reexamining Freedmanas critique. Ann Appl Stat. 2013;7(1):295–318. doi: https://doi.org/10.1214/12-AOAS583.
[35] Dasgupta T, Pillai NS, Rubin DB. Causal inference from $2^K$ factorial designs by using potential outcomes. J R Stat Soc Ser B. 2015;77:727–53.
[36] Athey S, Imbens GW. The econometrics of randomized experiments. In: Banerjee A, Duflo E, editors. Handbook of economic field experiments. vol. 1. North-Holland, Amsterdam; 2017. p. 73–140.
[37] Fogarty CB. Regression assisted inference for the average treatment effect in paired experiments. Biometrika. 2018;105:994–1000.
[38] Guo K, Basse G. The generalized Oaxaca-Blinder estimator. J Amer Stat Assoc. 2021;118:1–13.
[39] Wu J, Ding P. Randomization tests for weak null hypotheses in randomized experiments. J Amer Stat Assoc. 2021;116(536):1898–913.
[40] Rubin DB. Randomization analysis of experimental data: The Fisher randomization test comment. J Amer Stat Assoc. 1980;75(371):591–3.
[41] Ding P. A paradox from randomization-based causal inference. Stat Sci. 2017;32:331–45.
[42] Ding P, Dasgupta T. A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. Biometrika. 2018;105(1):45–56.
[43] Zhao A, Ding P. Covariate-adjusted Fisher randomization tests for the average treatment effect. J Econ. 2021;225(2):278–94.
[44] Cohen PL, Fogarty CB. Gaussian prepivoting for finite population causal inference. J R Stat Soc Ser B Stat Meth. 2022;84(2):295–320.
[45] Rosenbaum PR. Effects attributable to treatment: inference in experiments and observational studies within a discrete pivot. Biometrika. 2001;88:219–31.
[46] Rigdon J, Hudgens MG. Exact confidence intervals in the presence of interference. Stat Probab Lett. 2015;105:130–5.
[47] Li X, Ding P. Exact confidence intervals for the average causal effect on a binary outcome. Stat Med. 2016;35:957–60.
[48] Caughey D, Dafoe A, Li X, Miratrix L. Randomization inference beyond the sharp null: bounded null hypotheses and quantiles of individual treatment effects. J R Stat Soc Ser B (Stat Meth). 2023;85:1471–91.
[49] Su Y, Li X. Treatment effect quantiles in stratified randomized experiments and matched observational studies. Biometrika. 2023;111:235–54.

[50] Chen Z, Li X, Zhang B. The role of randomization inference in unraveling individual treatment effects in clinical trials: Application to HIV vaccine trials. 2023. arXiv: http://arXiv.org/abs/arXiv:231014399.
[51] Ding P. A first course in causal inference. 2023. arXiv: http://arXiv.org/abs/arXiv:230518793.
[52] Wu CJ, Hamada MS. Experiments: planning, analysis, and optimization. Hoboken, NJ: John Wiley & Sons; 2011.
[53] Hainmueller J, Hopkins DJ, Yamamoto T. Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. Politic Anal. 2014;22(1):1–30.
[54] Hainmueller J, Hopkins DJ. The hidden American immigration consensus: A conjoint analysis of attitudes toward immigrants. Amer J Politic Sci. 2015;59(3):529–48.
[55] Bauer DJ, Sterba SK, Hallfors DD. Evaluating group-based interventions when control participants are ungrouped. Multivariate Behav Res. 2008;43(2):210–36.
[56] Hallfors D, Cho H, Sanchez V, Khatapoush S, Kim HM, Bauer D. Efficacy vs effectiveness trial results of an indicated "model" substance abuse program: implications for public health. Amer J Public Health. 2006;96(12):2254–9.
[57] Branson Z, Dasgupta T. Sampling-based randomised designs for causal inference under the potential outcomes framework. Int Stat Rev. 2020;88:101–21.
[58] Yang Z, Qu T, Li X. Rejective sampling, rerandomization, and regression adjustment in survey experiments. J Amer Stat Assoc. 2021;118:1207–21.
[59] Li X, Ding P. General forms of finite population central limit theorems with applications to causal inference. J Amer Stat Assoc. 2017;112(520):1759–69.
[60] Zhao A, Ding P. Covariate adjustment in multiarmed, possibly factorial experiments. J R Stat Soc Ser B Stat Methodol. 2023;85(1):1–23.
[61] Fisher RA. Statistical methods for research workers. 1st ed. Edinburgh: Oliver and Boyd; 1925.
[62] Cochran WG. Sampling techniques. Hoboken, NJ: John Wiley & Sons; 1977.
[63] Lu J. Covariate adjustment in randomization-based causal inference for $2^K$ factorial designs. Stat Probabil Lett. 2016;119:11–20.
[64] Lei L, Ding P. Regression adjustment in completely randomized experiments with a diverging number of covariates. Biometrika. 2020 Dec;108(4):815–28. doi: https://doi.org/10.1093.
[65] Lu X, Yang F, Wang Y. Debiased regression adjustment in completely randomized experiments with moderately high-dimensional covariates. 2023. arXiv: http://arXiv.org/abs/arXiv:230902073.
[66] Bloniarz A, Liu H, Zhang CH, Sekhon JS, Yu B. Lasso adjustments of treatment effect estimates in randomized experiments. Proc Nat Acad Sci. 2016;113(27):7383–90.
[67] Cohen PL, Fogarty CB. No-harm calibration for generalized oaxaca-blinder estimators. 2020. arXiv: http://arXiv.org/abs/arXiv:201209246.
[68] Morgan KL, Rubin DB. Rerandomization to improve covariate balance in experiments. Ann Stat. 2012;40(2):1263–82.
[69] Sprott D, Farewell V. Randomization in experimental science. Stat Papers. 1993;34:89–94.
[70] Rubin DB. Comment: The design and analysis of gold standard randomized experiments. J Amer Stat Assoc. 2008;103(484):1350–3.
[71] Worrall J. Evidence: philosophy of science meets medicine. J Evaluat Clin Practice. 2010;16(2):356–62.
[72] Cox D. Randomization in the design of experiments. Int Stat Rev. 2009;77(3):415–29.
[73] Bruhn M, McKenzie D. In pursuit of balance: Randomization in practice in development field experiments. Amer Econ J Appl Econ. 2009;1(4):200–32.
[74] Maclure M, Nguyen A, Carney G, Dormuth C, Roelants H, Ho K, et al. Measuring prescribing improvements in pragmatic trials of educational tools for general practitioners. Basic Clin Pharm Toxicol. 2006;98(3):243–52.
[75] Bruhn M, McKenzie D. In pursuit of balance: randomization in practice in development field experiments. Amer Econ J Appl Econ. 2009;1:200–32.
[76] Lee JN, Morduch J, Ravindran S, Shonchoy A, Zaman H. Poverty and migration in the digital age: experimental evidence on mobile banking in Bangladesh. Amer Econ J Appl Econ. 2021;13:38–71.
[77] Li X, Ding P, Rubin DB. Asymptotic theory of rerandomization in treatment-control experiments. Proc Nat Acad Sci. 2018;115(37):9157–62.
[78] Branson Z, Dasgupta T, Rubin DB. Improving covariate balance in 2 K factorial designs via rerandomization with an application to a New York city department of education high school study. Ann Appl Stat. 2016:1958–76.
[79] Li X, Ding P, Rubin D. Rerandomization in $2^K$ factorial experiments. Ann Stat. 2020;48(1):43–63.
[80] Wang X, Wang T, Liu H. Rerandomization in stratified randomized experiments. J Amer Stat Assoc. 2023;118(542):1295–304.
[81] Johansson P, Schultzberg M. Rerandomization: A complement or substitute for stratification in randomized experiments? J Stat Plan Inference. 2022;218:43–58.
[82] Li X, Ding P. Rerandomization and regression adjustment. J R Stat Soc Ser B Stat Meth. 2020;82(1):241–68.
[83] Zhao A, Ding P. No star is good news: A unified look at rerandomization based on $p$-values from covariate balance tests. J Econ. 2024;241(1):105724.
[84] Wang Y, Li X. Asymptotic theory of the best-choice rerandomization using the Mahalanobis distance. 2023. arXiv: http://arXiv.org/abs/arXiv:231202513.
[85] Paul E, Rényi A. On the central limit theorem for samples from a finite population. Publ Math Inst Hungarian Acad Sci. 1959;4:49–61.

[86] Hájek J. Limiting distributions in simple random sampling from a finite population. Publ Math Inst Hungarian Acad Sci. 1960;5:361–74.
[87] Madow WG. On the limiting distributions of estimates based on samples from finite universes. Ann Math Stat. 1948;19:535–45.
[88] David F. Limiting distributions connected with certain methods of sampling human populations. Stat Res Mem. 1938;2:69–90.
[89] Wald A, Wolfowitz J. Statistical tests based on permutations of the observations. Ann Math Stat. 1944;15(4):358–72.
[90] Noether GE. On a theorem by Wald and Wolfowitz. Ann Math Stat. 1949;20(3):455–8.
[91] Hoeffding W. A combinatorial central limit theorem. Ann Math Stat. 1951;22:558–66.
[92] Motoo M. On the Hoeffding's combinatorial central limit theorem. Ann Inst Stat Math. 1956;8:145–54.
[93] Hájek J. Some extensions of the Wald-Wolfowitz-Noether theorem. Ann Math Stat. 1961;32:506–23.
[94] Fraser D. A vector form of the Wald-Wolfowitz-Hoeffding theorem. Ann Math Stat. 1956;27:540–3.
[95] DiCiccio CJ, Romano JP. Robust permutation tests for correlation and regression coefficients. J Amer Stat Assoc. 2017;112(519):1211–20.
[96] Shi L, Ding P. Berry–Esseen bounds for design-based causal inference with possibly diverging treatment levels and varying group sizes. 2022. arXiv: http://arXiv.org/abs/arXiv:220912345.
[97] Ding P, Feller A, Miratrix L. Decomposing treatment effect variation. J Amer Stat Assoc. 2019;114:304–17.
[98] Branson Z, Li X, Ding P. Power and sample size calculations for rerandomization. Biometrika. 2023;111:355–63.
[99] Bentkus V. On the dependence of the Berry–Esseen bound on dimension. J Stat Plan Infer. 2003;113(2):385–402.
[100] Chernozhukov V, Chetverikov D, Kato K. Central limit theorems and bootstrap in high dimensions. Ann Probability. 2017;45(4):2309.
[101] Bentkus V. A Lyapunov-type bound in $\mathbb{R}^d$. Theory Probabil Appl. 2005;49(2):311–23.
[102] Bhattacharya RN, Rao RR. Normal approximation and asymptotic expansions. Philadelphia, PA: SIAM; 2010.
[103] von Bahr B. Remainder term estimate in a combinatorial limit theorem. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete. 1976;35(2):131–9.
[104] Ho ST, Chen LH. An $L_p$ bound for the remainder in a combinatorial central limit theorem. Ann Probability. 1978;6(2):231–49.
[105] Bolthausen E. An estimate of the remainder in a combinatorial central limit theorem. Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete. 1984;66(3):379–86.
[106] Stein C. 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. vol. 6. University of California Press; 1972. p. 583–603.
[107] Chen LH, Goldstein L, Shao QM. Normal approximation by Steinas method. vol. 2. New York, NY: Springer; 2011.
[108] Bolthausen E, Gotze F. The rate of convergence for multivariate sampling statistics. Ann Stat. 1993;21:1692–710.
[109] Raic M. Multivariate normal approximation: Permutation statistics, local dependence and beyond; 2015.
[110] Chatterjee S, Meckes E. Multivariate normal approximation using exchangeable pairs. 2007. arXiv: http://arXiv.org/abs/math/0701464v1.
[111] Fang X, Röllin A. Rates of convergence for multivariate normal approximation with applications to dense graphs and doubly indexed permutation statistics. Bernoulli. 2015;21:2157–89.
[112] Caughey D, Katsumata H, Yamamoto T. Item response theory for conjoint survey experiments. Working Paper; 2019.
[113] Zhirkov K. Estimating and using individual marginal component effects from conjoint experiments. Politic Anal. 2022;30(2):236–49.
[114] Shi L, Wang J, Ding P. Forward screening and post-screening inference in factorial designs. 2023. arXiv: http://arXiv.org/abs/arXiv:230112045.
[115] Wang Y, Li X. Rerandomization with diminishing covariate imbalance and diverging number of covariates. Ann Stat. 2022;50(6):3439–65.
[116] Raiccc M. A multivariate Berry–Esseen theorem with explicit constants. Bernoulli. 2019;25(4A):2824–53.
[117] Petersen RG. Agricultural field experiments: design and analysis. Boca Raton, FL: CRC Press; 1994.
[118] Goldner MG, Knatterud GL, Prout TE. Effects of hypoglycemic agents on vascular complications in patients with adult-onset diabetes: III. Clinical implications of UGDP results. JAMA. 1971;218(9):1400–10.
[119] Chong A, Cohen I, Field E, Nakasone E, Torero M. Iron deficiency and schooling attainment in Peru. Amer Econ J Appl Econ. 2016;8(4):222–55.
[120] Bickel PJ, Freedman DA. Asymptotic normality and the bootstrap in stratified sampling. Ann Stat. 1984;12:470–82.
[121] Liu H, Yang Y. Regression-adjusted average treatment effect estimates in stratified randomized experiments. Biometrika. 2020;107(4):935–48.
[122] Liu H, Ren J, Yang Y. Randomization-based joint central limit theorem and efficient covariate adjustment in randomized block $2^K$ factorial experiments. J Amer Stat Assoc. 2022;119:1–15.
[123] Ball S, et al. Reading with television: an evaluation of the electric company. A report to the children's television workshop. Volumes 1 and 2. 1973.
[124] Imai K. Variance identification and efficiency analysis in randomized experiments under the matched-pair design. Stat Med. 2008;27(24):4857–73.
[125] Fogarty CB. On mitigating the analytical limitations of finely stratified experiments. J R Stat Soc Ser B Stat Methodol. 2018;80(5):1035–56.

[126] Pashley NE, Miratrix LW. Insights on variance estimation for blocked and matched pairs designs. J Educat Behav Stat. 2021;46(3):271–96.
[127] Su F, Ding P. Model-assisted analyses of cluster-randomized experiments. J R Stat Soc Ser B Stat Meth. 2021;83(5):994–1015.
[128] Abadie A, Athey S, Imbens GW, Wooldridge JM. When should you adjust standard errors for clustering? Quarter J Econ. 2023;138(1):1–35.
[129] Middleton JA, Aronow PM. Unbiased estimation of the average treatment effect in cluster-randomized experiments. Stat Politic Policy. 2015;6(1–2):39–75.
[130] Lu X, Liu T, Liu H, Ding P. Design-based theory for cluster rerandomization. Biometrika. 2023;110(2):467–83.
[131] Schochet PZ, Pashley NE, Miratrix LW, Kautz T. Design-based ratio estimators and central limit theorems for clustered, blocked RCTs. J Amer Stat Assoc. 2022;117(540):2135–46.
[132] Athey S, Imbens GW. The econometrics of randomized experiments. In: Handbook of economic field experiments. vol. 1. Amsterdam: Elsevier; 2017. p. 73–140.
[133] Hájek J. Asymptotic normality of simple linear rank statistics under alternatives. Ann Math Stat. 1968;39:325–46.
[134] Fredrickson MM, Chen Y. Permutation and randomization tests for network analysis. Soc Networks. 2019;59:171–83.
[135] Chen H, Friedman JH. A new graph-based two-sample test for multivariate and object data. J Amer Stat Assoc. 2017;112(517):397–409.
[136] D'Amour A, Airoldi E. Causal inference for dyadic outcomes in social network analysis. 2016.
[137] Deng L, Li Y, Zhang J, Wang Y, Chen C. Unbiased estimation for total treatment effect under interference using aggregated dyadic data. 2024. arXiv: http://arXiv.org/abs/arXiv:240212653.
[138] Bajari P, Burdick B, Imbens GW, Masoero L, McQueen J, Richardson T, et al. Multiple randomization designs; 2021. arXiv:2112.13495.
[139] Bajari P, Burdick B, Imbens GW, Masoero L, McQueen J, Richardson TS, et al. Experimental design in marketplaces. Stat Sci. 2023;1(1):1–19.
[140] Zhao L, Bai Z, Chao CC, Liang WQ. Error bound in a central limit theorem of double-indexed permutation statistics. Ann Stat. 1997;25(5):2210–27.
[141] Reinert G, Röllin A. Multivariate normal approximation with Stein's method of exchangeable pairs under a general linearity condition. Ann Probability. 2007;37(6):2150–73.
[142] Gastwirth JL, Krieger AM, Rosenbaum PR. Asymptotic separability in sensitivity analysis. J R Stat Soc Ser B. 2000;62:545–55.
[143] Wu D, Li X. Sensitivity analysis for quantiles of hidden biases in matched observational studies. 2023. arXiv: http://arXiv.org/abs/arXiv:230906459.
[144] Hu F, Rosenberger WF. The theory of response-adaptive randomization in clinical trials. Hoboken, NJ: John Wiley & Sons; 2006.
[145] Hall P, Heyde CC. Martingale limit theory and its application. San Diego, CA: Academic Press; 2014.
[146] Harshaw C, Sävje F, Eisenstat D, Mirrokni V, Pouget-Abadie J. Design and analysis of bipartite experiments under a linear exposure-response model. Elect J Stat. 2023;17(1):464–518.
[147] Leung MP. Causal inference under approximate neighborhood interference. Econometrica. 2022;90(1):267–93.
[148] Li X, Ding P, Lin Q, Yang D, Liu JS. Randomization Inference for Peer Effects. J Amer Stat Assoc. 2019;114:1651–64.
[149] Basse G, Ding P, Feller A, Toulis P. Randomization tests for peer effects in group formation experiments. 2019. arXiv: http://arXiv.org/abs/arXiv:190402308.
[150] Zhao A, Ding P. To adjust or not to adjust? estimating the average treatment effect in randomized experiments with missing covariates. J Amer Stat Assoc. 2022;119:1–11.
[151] Zhao A, Ding P, Li F. Covariate adjustment in randomized experiments with missing outcomes and covariates. Biometrika. 2024;111:asae017.
[152] Zhang Y, Rosenberger WF. On asymptotic normality of the randomization-based logrank test. Nonparametric Stat. 2005;17(7):833–9.
[153] Li X, Small DS. Randomization-based test for censored outcomes: a new look at the Logrank test. Stat Sci. 2023;38(1):92–107.
[154] Ding P, Li X, Miratrix LW. Bridging finite and super population causal inference. J Causal Infer. 2017;5:20160027.
[155] Yang L, Tsiatis AA. Efficiency study of estimators for a treatment effect in a Pretest-Posttest trial. Amer Stat. 2001;55:314–21.
[156] Rosenblum M, van der Laan MJ. Simple, efficient estimators of treatment effects in randomized trials using generalized linear models to Leverage baseline variables. Int J Biostat. 2010;6:6.
[157] Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005;61:962–73.
[158] Laan MJ, Robins JM. Unified methods for censored longitudinal data and causality. New York, NY: Springer; 2003.
[159] Rubin D, van der Laan MJ. A doubly robust censoring unbiased transformation. Int J Biostat. 2007;3(1):4. doi: 10.2202/1557-4679.1052.
[160] Van der Laan MJ, Rose S, et al. Targeted learning: causal inference for observational and experimental data. vol. 4. New York, NY: Springer; 2011.
[161] Hernández AV, Eijkemans MJ, Steyerberg EW. Randomized controlled trials with time-to-event outcomes: how much does prespecified covariate adjustment increase power? Ann Epidemiol. 2006;16(1):41–8.
[162] Lu X, Tsiatis AA. Improving the efficiency of the log-rank test using auxiliary covariates. Biometrika. 2008;95(3):679–94.
[163] Moore KL, van der Laan MJ. Increasing power in randomized trials with right censored outcomes through covariate adjustment. J Biopharm Stat. 2009;19(6):1099–131.