Exploring the Role of Randomization in Causal Inference

Exploring the Role of Randomization in Causal Inference

A dissertation presented

by

Peng Ding

to

The Department of Statistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Statistics

Harvard University

Cambridge, Massachusetts

May 2015

Exploring the Role of Randomization in Causal Inference

Abstract

This manuscript includes three topics in causal inference, all of which are under the randomization inference framework (Neyman, 1923; Fisher, 1935a; Rubin, 1978). This manuscript contains three self-contained chapters.

Chapter 1. Under the potential outcomes framework, causal effects are defined as comparisons between potential outcomes under treatment and control. To infer causal effects from randomized experiments, Neyman proposed to test the null hypothesis of zero average causal effect (Neyman's null), and Fisher proposed to test the null hypothesis of zero individual causal effect (Fisher's null). Although the subtle difference between Neyman's null and Fisher's null has caused lots of controversies and confusions for both theoretical and practical statisticians, a careful comparison between the two approaches has been lacking in the literature for more than eighty years. I fill in this historical gap by making a theoretical comparison between them and highlighting an intriguing paradox that has not been recognized by previous researchers. Logically, Fisher's null implies Neyman's null. It is therefore surprising that, in actual completely randomized experiments, rejection of Neyman's null does not imply rejection of Fisher's null for many realistic situations, including the case with constant causal effect. Furthermore, I show that this paradox also exists in other commonly-used experiments, such as stratified experiments, matched-pair experiments, and factorial experiments. Asymptotic analyses, numerical examples, and

real data examples all support this surprising phenomenon. Besides its historical and theoretical importance, this paradox also leads to useful practical implications for modern researchers.

Chapter 2. Causal inference in completely randomized treatment-control studies with binary outcomes is discussed from Fisherian, Neymanian and Bayesian perspectives, using the potential outcomes framework. A randomization-based justification of Fisher's exact test is provided. Arguing that the crucial assumption of constant causal effect is often unrealistic, and holds only for extreme cases, some new asymptotic and Bayesian inferential procedures are proposed. The proposed procedures exploit the intrinsic non-additivity of unit-level causal effects, can be applied to linear and non-linear estimands, and dominate the existing methods, as verified theoretically and also through simulation studies.

Chapter 3. Recent literature has underscored the critical role of treatment effect variation in estimating and understanding causal effects. This approach, however, is in contrast to much of the foundational research on causal inference; Neyman, for example, avoided such variation through his focus on the average treatment effect and his definition of the confidence interval. In this chapter, I extend the Neymanian framework to explicitly allow both for treatment effect variation explained by covariates, known as the systematic component, and for unexplained treatment effect variation, known as the idiosyncratic component. This perspective enables estimation and testing of impact variation without imposing a model on the marginal distributions of potential outcomes, with the workhorse approach of regression with interaction terms being a special case. My approach leads to two practical results.

Abstract

First, I combine estimates of systematic impact variation with sharp bounds on overall treatment variation to obtain bounds on the proportion of total impact variation explained by a given model—this is essentially an R^2 for treatment effect variation. Second, by using covariates to partially account for the correlation of potential outcomes problem, I exploit this perspective to sharpen the bounds on the variance of the average treatment effect estimate itself. As long as the treatment effect varies across observed covariates, the resulting bounds are sharper than the current sharp bounds in the literature. I apply these ideas to a large randomized evaluation in educational research, showing that these results are meaningful in practice.

Contents

Citations to Previously Published Work

Chapter 1 is based on a paper submitted to Arxiv. Professors Luke Miratrix and Tirthankar Dasgupta helped edit early versions of this chapter.

> Ding, P. (2014) A paradox from randomization-based causal inference. Arxiv: `http://arxiv.org/abs/1402.0142`

Chapter 2 is based on a paper that will appear in the *Journal of the American Statistical Association*. This is a collaborative work with Professor Tirthankar Dasgupta.

> Ding, P. and Dasgupta, T. D. (2015) A potential tale of two by two tables from completely randomized experiments. *Journal of the American Statistical Association*, in press.

Chapter 3 extends my previous paper that will appear in the *Journal of the Royal Statistical Society, Series B*. The previous paper and the current chapter are collaborative works with Professor Luke Miratrix and fellow Ph.D. student Avi Feller. Avi Feller helped analyze the Head Start Impact Study data and edit early versions of this chapter.

> Ding, P., Feller, A., and Miratrix, L. W. (2015) Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, in press.

Acknowledgments

I want to thank all the professors in the Department of Statistics at Harvard University for teaching me statistics. Professor Donald Rubin's statistical insight motivates most parts of my thesis, and his comments have greatly improved the quality of all my papers and thesis. I could not finish writing my papers and thesis without the support from Professors Luke Miratrix and Tirthankar Dasgupta.

I want to thank fellow Ph.D. students, Avi Feller, Arman Sabbaghi, David Waston, Lo-Hua Yuan, Robin Gong, Jiannan Lu, and Xinran Li, for many helpful discussions and collaborations.

I want to thank Betsey Cogswell, Steven Finch, James Matejek, Alice Moses, and Madeleine Straubel, for their help in everyday life. They make the Harvard Statistics Department a lovely place to work and study.

To my parents.

Chapter 1

A Paradox from Randomization-Based Causal Inference

1.1 Introduction

Ever since Neyman's seminal work, the potential outcomes framework (Neyman, 1923; Rubin, 1974) has been widely used for causal inference in randomized experiments (Neyman and Iwaszkiewicz, 1935; Hinkelmann and Kempthorne, 2007). The potential outcomes framework permits us to make inference about a finite population of interest, with all potential outcomes fixed and randomness coming solely from the physical randomization of the treatment assignments. Historically, Neyman (1923) was interested in obtaining an unbiased estimator with a repeated sampling evaluation of the average causal effect, which also corresponded to a test for the null

hypothesis of zero average causal effect. On the other hand, Fisher (1935a) focused on testing the sharp null hypothesis of zero individual causal effect, and proposed the famous Fisher Randomization Test (FRT). Both Neymanian and Fisherian approaches are randomization-based inference, relying on the physical randomization of the experiments. Neyman's null and Fisher's null are closely related to each other: the latter implies the former, and they are equivalent under the constant causal effect assumption. Both approaches have existed for many decades and are widely used in current statistical practice. They are now introduced at the beginning of many causal inference courses (e.g., Rubin, 2004; Imbens and Rubin, 2015). Unfortunately, however, a detailed comparison between them has not been made in the literature.

In the past, several researchers (e.g., Rosenbaum, 2002, page 40) believed that "in most cases, their disagreement is entirely without technical consequence: the same procedures are used, and the same conclusions are reached." However, we show, via both numerical examples and theoretical investigations, that Neyman's method tends to reject the null more often than Fisher's method in many realistic randomized experiments. In fact, Neyman's method is always more powerful if there is a nonzero constant causal effect, the very alternative most often used for Fisher-style inference. This finding immediately causes a seeming paradox: logically, Fisher's null implies Neyman's null, so how can we fail to reject the former while rejecting the latter?

We demonstrate that this surprising paradox is not unique to completely randomized experiments, because it also exists in other commonly-used experiments such as stratified experiments, matched-pair experiments, and factorial experiments. The result for factorial experiments helps to explain the surprising empirical evidence

in Dasgupta et al. (2015) that interval estimators for factorial effects obtained by inverting a sequence of FRTs are often wider than Neymanian confidence intervals.

The paper proceeds as follows. We review Neymanian and Fisherian randomization-based causal inference in Section 1.2 under the potential outcomes framework. In Section 1.3, we use both numerical examples and asymptotic analyses to demonstrate the paradox from randomization-based inference in completely randomized experiments. Section 1.4 shows that a similar paradox also exists in other commonly-used experiments. Section 1.5 extends the scope of the paper to improved variance estimators and comments on the choices of test statistics. Section 1.6 illustrates the asymptotic theory of this paper with some finite sample real-life examples. We conclude with a discussion in Section 1.7, and relegate all the technical details to Appendix A.

1.2 Completely Randomized Experiments and Randomization Inference

We first introduce notation for causal inference in completely randomized experiments, and then review the Neymanian and Fisherian perspectives for causal inference.

1.2.1 Completely Randomized Experiments and Potential Outcomes

Consider N units in a completely randomized experiment. Throughout our discussion, we make the Stable Unit Treatment Value Assumption (SUTVA; Rubin,

1980), i.e., there is only one version of the treatment, and interference between subjects is absent. SUTVA allows us to define the potential outcome of unit i under treatment t as $Y_i(t)$, with $t = 1$ for treatment and $t = 0$ for control. The individual causal effect is defined as a comparison between two potential outcomes, for example, $\tau_i = Y_i(1) - Y_i(0)$. However, for each subject i, we can observe only one of $Y_i(1)$ and $Y_i(0)$ with the other one missing, and the individual causal effect τ_i is not observable. The observed outcome is a function of the treatment assignment T_i and the potential outcomes, namely, $Y_i^{obs} = T_i Y_i(1) + (1 - T_i)Y_i(0)$. Let $\boldsymbol{Y}^{obs} = (Y_1^{obs}, \ldots, Y_N^{obs})'$ be the observed outcome vector. Let $\boldsymbol{T} = (T_1, \ldots, T_N)'$ denote the treatment assignment vector, and $\boldsymbol{t} = (t_1, \ldots, t_N)' \in \{0, 1\}^N$ be its realization. Completely randomized experiments satisfy $\text{pr}(\boldsymbol{T} = \boldsymbol{t}) = N_1!N_0!/N!$, if $\sum_{i=1}^N t_i = N_1$ and $N_0 = N - N_1$. Note that in Neyman (1923)'s potential outcomes framework, all the potential outcomes are fixed numbers, and only the treatment assignment vector is random. In general, we can view this framework with fixed potential outcomes as conditional inference given the values of the potential outcomes.

1.2.2 Neymanian Inference for the Average Causal Effect

Neyman (1923) was interested in estimating the finite population average causal effect:

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i = \frac{1}{N}\sum_{i=1}^{N} \{Y_i(1) - Y_i(0)\} = \bar{Y}_1 - \bar{Y}_0,$$

where $\bar{Y}_t = \sum_{i=1}^{N} Y_i(t)/N$ is the finite population average of the potential outcomes $\{Y_i(t) : i = 1, \ldots, N\}$. He proposed an unbiased estimator

$$\widehat{\tau} \;=\; \bar{Y}_1^{obs} - \bar{Y}_0^{obs} \tag{1.1}$$

for τ, where $\bar{Y}_t^{obs} = \sum_{\{i:T_i=t\}} Y_i^{obs}/N_t$ is the sample mean of the observed outcomes under treatment t. The sampling variance of $\widehat{\tau}$ over all possible randomizations of treatment assignment is

$$\text{var}(\widehat{\tau}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}, \tag{1.2}$$

depending on $S_t^2 = \sum_{i=1}^{N}\{Y_i(t) - \bar{Y}_t\}^2/(N-1)$, the finite population variance of the potential outcomes $\{Y_i(t) : i = 1, \ldots, N\}$, and $S_\tau^2 = \sum_{i=1}^{N}(\tau_i - \tau)^2/(N-1)$, the finite population variance of the individual causal effects $\{\tau_i : i = 1, \ldots, N\}$. Since we can never jointly observe the pair of potential outcomes for each unit, the variance of individual causal effects, S_τ^2, is not identifiable from the observed data. Recognizing this difficulty, Neyman (1923) suggested using

$$\widehat{V}(\text{Neyman}) = \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}, \tag{1.3}$$

as an estimator for $\text{var}(\widehat{\tau})$, where $s_t^2 = \sum_{\{i:T_i=t\}}(Y_i^{obs} - \bar{Y}_t^{obs})^2/(N_t - 1)$ is the sample variance of the observed outcomes under treatment t. However, Neyman's variance estimator $\widehat{V}(\text{Neyman})$ overestimates the true variance $\text{var}(\widehat{\tau})$, in the sense that

$$E\{\widehat{V}(\text{Neyman})\} \geq \text{var}(\widehat{\tau}),$$

with equality holding if and only if the individual causal effects are constant: $\tau_i = \tau$ or $S_\tau^2 = 0$. The randomization distribution of $\widehat{\tau}$ enables us to test the following Neyman's null hypothesis:

$$H_0(\text{Neyman}) : \tau = 0.$$

Under H_0(Neyman) and based on the Normal approximation in Section 1.3.3, the p-value from Neyman's approach can be approximated by

$$p(\text{Neyman}) \approx 2\Phi\left\{-\frac{|\widehat{\tau}^{obs}|}{\sqrt{\widehat{V}(\text{Neyman})}}\right\}, \tag{1.4}$$

where $\widehat{\tau}^{obs}$ is the realized value of $\widehat{\tau}$, and $\Phi(\cdot)$ is the cumulative distribution function of the standard Normal distribution. When we have non-constant individual causal effects, Neyman's test for the null hypothesis of zero average causal effect tends to be "conservative," in the sense that it rejects less often than the nominal significance level when the null is true.

1.2.3 Fisherian Randomization Test for the Sharp Null

Fisher (1935a) was interested in testing the following sharp null hypothesis:

$$H_0(\text{Fisher}) : Y_i(1) = Y_i(0), \quad \forall i = 1, \ldots, N.$$

This null hypothesis is sharp because all missing potential outcomes can be uniquely imputed under H_0(Fisher). The sharp null hypothesis implies that $Y_i(1) = Y_i(0) = Y_i^{obs}$ are all fixed constants, so that the observed outcome for subject i is Y_i^{obs} under

any treatment assignment. Although we can perform randomization tests using any test statistics capturing the deviation from the null, we will first focus on the randomization test using $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) = \widehat{\tau}$ as the test statistic, in order to make a direct comparison to Neyman's method. We will comment on other choices of test statistics in the later part of this paper. Again, the randomness of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ comes solely from the randomization of the treatment assignment $\boldsymbol{T}$, since $\boldsymbol{Y}^{obs}$ is a set of constants under the sharp null. The p-value for the two-sided test under the sharp null is

$$p(\text{Fisher}) = \text{pr}\left\{|\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})| \geq |\widehat{\tau}^{obs}| \;\middle|\; H_0(\text{Fisher})\right\},$$

measuring the extremeness of $\widehat{\tau}^{obs}$ with respect to the null distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ over all possible randomizations. In practice, we can approximate the exact distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ by Monte Carlo. We draw, repeatedly and independently, completely randomized treatment assignment vectors $\{\boldsymbol{T}^1, \ldots, \boldsymbol{T}^M\}$, and with large M the p-value can be well approximated by

$$p(\text{Fisher}) \approx \frac{1}{M}\sum_{m=1}^{M} I\left\{|\widehat{\tau}(\boldsymbol{T}^m, \boldsymbol{Y}^{obs})| \geq |\widehat{\tau}^{obs}|\right\}.$$

Rubin (1980) first used the name "sharp null," and viewed the FRT as a "stochastic proof by contradiction" (Rubin, 2004). More discussion about randomization tests can also be found in Rosenbaum (2002).

1.3 A Paradox from Neymanian and Fisherian Inference

Neymanian and Fisherian approaches reviewed in Section 1.2 share some common properties but also differ fundamentally. They both rely on the distribution induced by the physical randomization, but they test two different null hypotheses and evolve from different statistical philosophies. In this section, we first compare Neymanian and Fisherian approaches using simple numerical examples, and highlight a surprising paradox. We then explain the paradox via asymptotic analysis.

1.3.1 Initial Numerical Comparisons

We compare Neymanian and Fisherian approaches using numerical examples with both balanced and unbalanced experiments. In our simulations, the potential outcomes are fixed, and the simulations are carried out over randomization distributions induced by the treatment assignments. The significance level is 0.05, and M is 10^5 for the FRT.

Example 1 (Balanced Experiments with $N_1 = N_0$)**.** The potential outcomes are independently generated from Normal distributions $Y_i(1) \sim N(1/10, 1/16)$ and $Y_i(0) \sim N(0, 1/16)$, for $i = 1, \ldots, 100$. Further, once drawn from the Normal distributions above, they are fixed. We repeatedly generate 1000 completely randomized treatment assignments with $N = 100$ and $N_1 = N_0 = 50$. For each treatment assignment, we obtain the observed outcomes and implement two tests for Neyman's null and Fisher's null. As shown in Table 1.1(a), it never happens that we reject Fisher's null but fail

to reject Neyman's null. However, we reject Neyman's null but fail to reject Fisher's null in 15 instances.

Example 2 (Unbalanced Experiments with $N_1 \neq N_0$)**.** The potential outcomes are independently generated from Normal distributions $Y_i(1) \sim N(1/10, 1/4)$ and $Y_i(0) \sim N(0, 1/16)$, for $i = 1, \ldots, 100$. They are kept as fixed throughout the simulations. The unequal variances are designed on purpose, and we will reveal the reason for choosing them later in Example 3 of Section 1.3.4. We repeatedly generate 1000 completely randomized treatment assignments with $N = 100, N_1 = 70$, and $N_0 = 30$. After obtaining each observed data set, we perform two hypothesis testing procedures, and summarize the results in Table 1.1(b). The pattern in Table 1.1(b) is more striking than in Table 1.1(a), since it happens 62 times in Table 1.1(b) that we reject Neyman's null but fail to reject Fisher's null. For this particular set of potential outcomes, Neyman's testing procedure has a power $62/1000 = 0.062$, slightly larger than 0.05, but Fisher's testing procedure has a power $8/1000 = 0.008$, much smaller than 0.05 even though the sharp null is not true. We will explain in Section 1.3.4 the reason why the FRT could have a power even smaller than the significance level under some alternative hypotheses.

1.3.2 Statistical Inference, Logic, and Paradox

Logically, Fisher's null implies Neyman's null. Therefore, Fisher's null should be rejected if Neyman's null is rejected. However, this is not always true from the results of statistical inference in completely randomized experiments. We observed in our

Table 1.1: Numerical Examples.

(a) Balanced experiments

	not reject H_0(Fisher)	reject H_0(Fisher)
not reject H_0(Neyman)	488	0
reject H_0(Neyman)	15	497

power(Fisher)=0.497, power(Neyman)=0.512

(b) Unbalanced experiments

	not reject H_0(Fisher)	reject H_0(Fisher)
not reject H_0(Neyman)	930	0
reject H_0(Neyman)	62	8

power(Fisher)=0.008, power(Neyman)=0.070

numerical examples above that it can be the case that

$$p(\text{Neyman}) < \alpha_0 < p(\text{Fisher}), \tag{1.5}$$

in which case we should reject Neyman's null, but not Fisher's null, if we choose the significance level to be α_0 (e.g., $\alpha_0 = 0.05$). When (1.5) holds, an awkward logical problem appears as illustrated in Figure 1.1. In the remaining part of this section, we will theoretically explain the empirical findings in Section 1.3.1 and the logical problem in Figure 1.1.

Logic:	not reject H_0(Fisher)	$\Longrightarrow$	not reject H_0(Neyman)
Logic:	reject H_0(Fisher)	$\Longleftarrow$	reject H_0(Neyman)
Statistical inference:	not reject H_0(Fisher)	but	reject H_0(Neyman)

Figure 1.1: A paradox from randomization-based causal inference.

1.3.3 Asymptotic Evaluations

While Neyman's testing procedure has an explicit form, the FRT is typically approximated by Monte Carlo. In order to compare them, we first discuss the asymptotic Normalities of $\widehat{\tau}$ and the randomization test statistic $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$. We provide a simplified way of doing variance calculation and a short proof for asymptotic Normalities of both $\widehat{\tau}$ and $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$, based on the finite population Central Limit Theorem (CLT; Hájek, 1960; Lehmann, 1998; Freedman, 2008). Before the formal asymptotic results, it is worth mentioning the exact meaning of "asymptotics" in the context of finite population causal inference. We need to embed the finite population of interest into a hypothetical infinite sequence of finite populations with increasing sizes, and also require the proportions of the treatment units to converge to a fixed value. Essentially, all the population quantities (e.g., τ, S_1^2, etc.) should have the index N, and all the sample quantities (e.g., $\widehat{\tau}, s_1^2$, etc.) should have double indices N and N_1. However, for the purpose of notational simplicity, we sacrifice a little bit of mathematical precision and drop all the indices in our discussion.

Theorem 1. As $N \to \infty$, the sampling distribution of $\widehat{\tau}$ satisfies

$$\frac{\widehat{\tau} - \tau}{\sqrt{\text{var}(\widehat{\tau})}} \xrightarrow{d} \mathcal{N}(0, 1).$$

In practice, the true variance $\text{var}(\widehat{\tau})$ is replaced by its "conservative" estimator $\widehat{V}(\text{Neyman})$, and the resulting test rejects less often than the nominal significance level on average. While the asymptotics for the Neymanian unbiased estimator $\widehat{\tau}$ does not depend on the null hypothesis, the following asymptotic Normality for $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ is

true only under the sharp null hypothesis.

Theorem 2. Under H_0(Fisher) and as $N \to \infty$, the null distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ satisfies

$$\frac{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})}{\sqrt{\widehat{V}(\text{Fisher})}} \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\widehat{V}(\text{Fisher}) = Ns^2/(N_1 N_0), \quad s^2 = \sum_{i=1}^{N}(Y_i^{obs} - \bar{Y}^{obs})^2/(N-1), \quad \bar{Y}^{obs} = \sum_{i=1}^{N} Y_i^{obs}/N.$$

Therefore, the p-value under H_0(Fisher) can be approximated by

$$p(\text{Fisher}) \approx 2\Phi\left\{-\frac{|\widehat{\tau}^{obs}|}{\sqrt{\widehat{V}(\text{Fisher})}}\right\}. \tag{1.6}$$

From (1.4) and (1.6), the asymptotic p-values obtained from Neymanian and Fisherian approaches differ only due to the difference between the variances $\widehat{V}$(Neyman) and $\widehat{V}$(Fisher). Therefore, a comparison of the variances will explain the different behaviors of the corresponding approaches. In the following, we use the conventional notation $R_N = o_p(N^{-1})$ for a random quantity satisfying $N \cdot R_N \to 0$ in probability as $N \to \infty$ (Lehmann, 1998).

Theorem 3. Asymptotically, the difference between the two variance estimators is

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = (N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) + N^{-1}(\bar{Y}_1 - \bar{Y}_0)^2 + o_p(N^{-1}). \tag{1.7}$$

The difference between the variance estimators depends on the ratio of the treat-

ment and control sample sizes, and differences between the means and variances of the treatment and control potential outcomes.

In order the verify the asymptotic theory above, we go back to compare the variances in the previous numerical examples.

Example 3 (Continuations of Examples 1 and 2)**.** We plot in Figure 1.2 the variances $\widehat{V}$(Neyman) and $\widehat{V}$(Fisher) obtained from the numerical examples in Section 1.3.1. In both the left and the right panels, $\widehat{V}$(Fisher) tends to be larger than $\widehat{V}$(Neyman). This pattern is more striking on the right panel with unbalanced experiments designed to satisfy $(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) > 0$. It is thus not very surprising that the FRT is much less powerful than Neyman's test, and it rejects even less often than nominal 0.05 level as shown in Table 1.1(b).

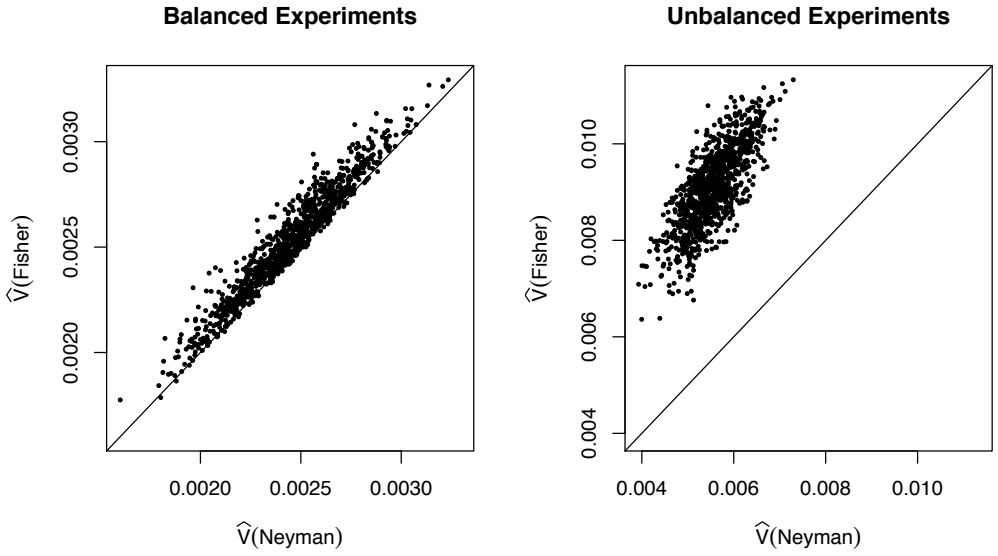


Figure 1.2: Variance estimators in balanced and unbalanced experiments

1.3.4 Theoretical Comparison

Although quite straightforward, Theorem 3 has several helpful implications to explain the paradoxical results in Section 1.3.1.

Under H_0(Fisher), $\bar{Y}_1 = \bar{Y}_0, S_1^2 = S_0^2$, and the difference between the two variances is of higher order, namely, $\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = o_p(N^{-1})$. Therefore, Neymanian and Fisherian methods coincide with each other asymptotically under the sharp null. This is the basic requirement, since both testing procedures generate correct type one errors under this circumstance.

For the case with constant causal effect, we have $\tau_i = \tau$ and $S_1^2 = S_0^2$. The difference between the two variance estimators reduces to

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = \tau^2/N + o_p(N^{-1}). \tag{1.8}$$

Under H_0(Neyman), $\bar{Y}_1 = \bar{Y}_0$, and the difference between the two variances is of higher order, and two tests have the same asymptotic performance. However, under the alternative hypothesis, $\tau = \bar{Y}_1 - \bar{Y}_0 \neq 0$, and the difference above is positive and of order $1/N$, and Neyman's test will reject more often than Fisher's test. With larger effect size $|\tau|$, the powers differ more.

For balanced experiments with $N_1 = N_0$, the difference between the two variance estimators reduces to the same formula as (1.8), and the conclusions are the same as above.

For unbalanced experiments, the difference between two variances can be either positive or negative. In practice, if we have prior knowledge $S_1^2 > S_0^2$, unbalanced

experiments with $N_1 > N_0$ are preferable to improve estimation precision. In this case, we have $(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) > 0$ and $\widehat{V}(\text{Fisher}) > \widehat{V}(\text{Neyman})$ for large N. Surprisingly, we are more likely to reject Neyman's null than Fisher's null, although Neyman's test itself is conservative with nonconstant causal effect implied by $S_1^2 > S_0^2$.

From the above cases, we can see that Neymanian and Fisherian approaches generally have different performances, unless the sharp null hypothesis holds. Fisher's sharp null imposes more restrictions on the potential outcomes, and the variance of the randomization distribution of $\widehat{\tau}$ pools the within and between group variances across treatment and control arms. Consequently, the resulting randomization distribution of $\widehat{\tau}$ has larger variance than its repeated sampling variance in many realistic cases. Paradoxically, in many situations, we tend to reject Neyman's null more often than Fisher's null, which contradicts the logical fact that Fisher's null implies Neyman's null.

Finally, we consider the performance of the FRT under Neyman's null with $\bar{Y}_1 = \bar{Y}_0$, which is often of more interest in social sciences. If $S_1^2 > S_0^2$ and $N_1 > N_0$, the rejection rate of Fisher's test is smaller than Neyman's test, even though $H_0(\text{Neyman})$ holds but $H_0(\text{Fisher})$ does not. Consequently, the difference-in-means statistic $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ has no power against the sharp null, and the resulting FRT rejects even less often than the nominal significance level. However, if $S_1^2 > S_0^2$ and $N_1 < N_0$, the FRT may not be more "conservative" than Neyman's test. Unfortunately, the FRT may reject more often than the nominal level, yielding an invalid test for Neyman's null. Gail et al. (1996) found this phenomenon through simulation studies, and here we provide a theoretical explanation. In the following, we use an example to illustrate

this possibility.

Example 4. If $Y_i(1) = aY_i(0) + (1-a)\bar{Y}(0)$ for all i, then $\tau = 0$ but $\tau_i = (a-1)Y_i(0) + (1-a)\bar{Y}(0)$. We focus on the case with $a > 1$ and $r = N_1/N < 1/2$. The FRT yields invalid type one error when $\widehat{V}(\text{Fisher})$ is asymptotically smaller than the true sampling distribution of $\widehat{\tau}$. We show in Appendix A that certain combinations of (a, r) lead to invalid FRT under Neyman's null, and we illustrate the invalid region of the FRT in Figure 1.3.

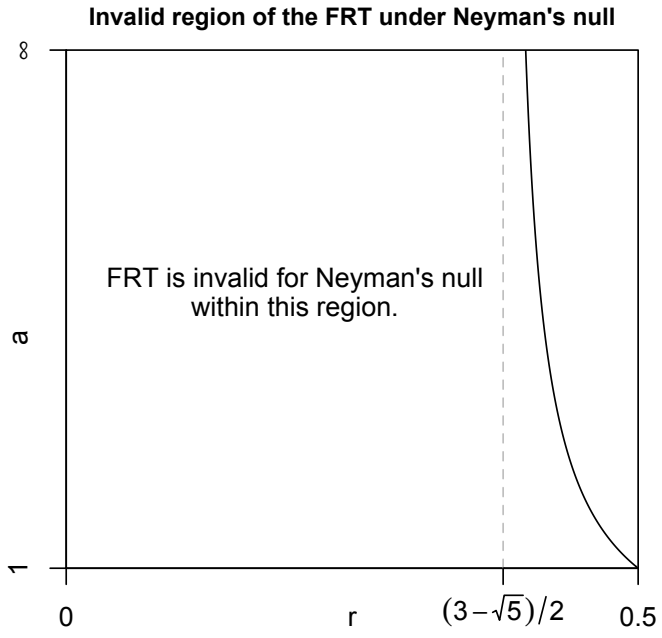


Figure 1.3: FRT under Neyman's null

1.3.5 Binary Outcomes

We close this section by investigating the special case with binary outcomes, for which more explicit results are available. Let $p_t = \bar{Y}(t)$ be the potential proportion

and $\widehat{p}_t = \bar{Y}_t^{obs}$ be the sample proportion of one under treatment t. Define $\widehat{p} = \bar{Y}^{obs}$ as the proportion of one in all the observed outcomes. The results in the following corollary are special cases of Theorems 1 to 3.

Corollary 1. Neyman's test is asymptotically equivalent to the "unpooled" test

$$\frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}_1(1-\widehat{p}_1)/N_1 + \widehat{p}_0(1-\widehat{p}_0)/N_0}} \xrightarrow{d} \mathcal{N}(0,1) \tag{1.9}$$

under H_0(Neyman); and Fisher's test is asymptotically equivalent to the "pooled" test

$$\frac{\widehat{p}_1 - \widehat{p}_0}{\sqrt{\widehat{p}(1-\widehat{p})(N_1^{-1} + N_0^{-1})}} \xrightarrow{d} \mathcal{N}(0,1) \tag{1.10}$$

under H_0(Fisher). The asymptotic difference between the two tests is due to

$$\begin{aligned} & \widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) \\ = \; & (N_0^{-1} - N_1^{-1})\{p_1(1-p_1) - p_0(1-p_0)\} + N^{-1}(p_1 - p_0)^2 + o_p(N^{-1}). \end{aligned} \tag{1.11}$$

For the case with binary outcomes, we can draw analogous but slightly different conclusions to the above. Under Neyman's null, $p_1 = p_0$ and the two tests are asymptotically equivalent. Therefore, the situation that the FRT is invalid under Neyman's null will never happen for binary outcomes. In balanced experiments, Neyman's test is always more powerful than Fisher's test under the alternative with $p_1 \neq p_0$. For unbalanced experiments, the answer is not definite, but Equation (1.11) allows us to determine the region of (p_1, p_0) that favors Neyman's test for a given

level of the ratio $r = N_1/N$. When $r > 1/2$, Figure 1.4 shows the regions in which Neyman's test is asymptotically more powerful than Fisher's test according to the value of r. When $r < 1/2$, the region has the same shape by symmetry. We provide more details about Figure 1.4 in Appendix A.

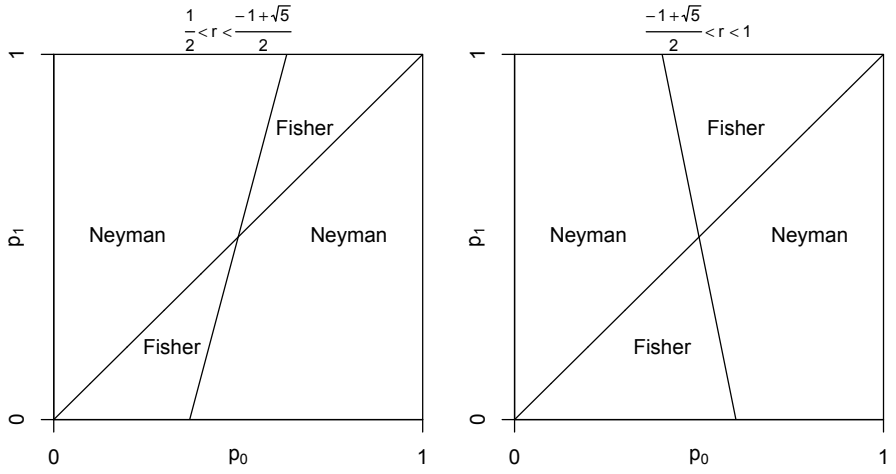


Figure 1.4: Binary Outcome with Different Proportions $r = N_1/N$. Neyman's test is more powerful in the regions marked by "Neyman."

Note that Fisher's test is equivalent to Fisher's exact test, and (1.10) is essentially the Normal approximation of the hypergeometric distribution (Ding and Dasgupta, 2015). The two tests in (1.9) and (1.10) are based purely on randomization inference, which have the same mathematical forms as the classical "unpooled" and "pooled" tests for equal proportions under two independent Binomial models. Our conclusion is coherent with Robbins (1977) and Eberhardt and Fligner (1977) that the "unpooled" test is more powerful than the "pooled" one with equal sample size. For hypothesis testings in two by two tables, Greenland (1991) observed similar theoretical results as Corollary 1 but gave a different interpretation.

1.4 Ubiquity of the Paradox in Other Experiments

The paradox discussed in Section 1.3 is not unique to completely randomized experiments. As a direct generalization of the previous results, the paradox will appear in each stratum of stratified experiments. We will also show its existence in two other widely-used experiments: matched-pair designs and factorial designs. In order to minimize the confusion about the notation, each of the following two subsections are self-contained.

1.4.1 Matched-Pair Experiments

Consider a matched-pair experiment with $2N$ units and N pairs matched according to their observed characteristics. Within each matched pair, we randomly select one unit to receive treatment and the other to receive control. Let T_i be iid Bernoulli(1/2) for $i = 1, \ldots, N$, indicating treatment assignments for the matched pairs. For pair i, the first unit receives treatment and the second unit receives control if $T_i = 1$; and otherwise if $T_i = 0$. Under the SUTVA, we define $(Y_{ij}(1), Y_{ij}(0))$ as the potential outcomes of the jth unit in the ith pair under treatment and control, and the observed outcomes within pair i are $Y_{i1}^{obs} = T_i Y_{i1}(1) + (1 - T_i) Y_{i1}(0)$ and $Y_{i2}^{obs} = T_i Y_{i2}(0) + (1 - T_i) Y_{i2}(1)$. Let $\boldsymbol{T} = (T_1, \ldots, T_N)'$ and $\boldsymbol{Y}^{obs} = \{Y_{ij}^{obs} : i = 1, \ldots, N; j = 1, 2\}$ denote the $N \times 1$ treatment assignment vector and the $N \times 2$ observed outcome matrix, respectively. Within pair i,

$$\widehat{\tau}_i = T_i(Y_{i1}^{obs} - Y_{i2}^{obs}) + (1 - T_i)(Y_{i2}^{obs} - Y_{i1}^{obs})$$

is unbiased for the within-pair average causal effect

$$\tau_i = \{Y_{i1}(1) + Y_{i2}(1) - Y_{i1}(0) - Y_{i2}(0)\}/2.$$

Immediately, we can use

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\tau}_i$$

as an unbiased estimator for the finite population average causal effect

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\tau_i = \frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{2}\{Y_{ij}(1) - Y_{ij}(0)\}.$$

Imai (2008) discussed Neymanian inference for τ and identified the variance of $\widehat{\tau}$ with the corresponding variance estimator. To be more specific, he found that

$$\text{var}(\widehat{\tau}) = \frac{1}{4N^2}\sum_{i=1}^{N}\{Y_{i1}(1) + Y_{i1}(0) - Y_{i2}(1) - Y_{i2}(0)\}^2,$$

which can be "conservatively" estimated by

$$\widehat{V}(\text{Neyman}) = \frac{1}{N(N-1)}\sum_{i=1}^{N}(\widehat{\tau}_i - \widehat{\tau})^2.$$

The repeated sampling evaluation above allows us to test Neyman's null hypothesis of zero average causal effect:

$$H_0(\text{Neyman}) : \tau = 0.$$

On the other hand, Rosenbaum (2002) discussed intensively the FRT in matched-

pair experiments under the sharp null hypothesis:

$$H_0(\text{Fisher}) : Y_{ij}(1) = Y_{ij}(0), \forall i = 1, \ldots, N; \forall j = 1, 2,$$

which is, again, much stronger than Neyman's null. For the purpose of comparison, we choose the test statistic with the same form as $\widehat{\tau}$, denoted as $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$. In practice, the null distribution of this test statistic can be calculated exactly by enumerating all the 2^N randomizations or approximated by Monte Carlo. For our theoretical investigation, we have the following results.

Theorem 4. Under the sharp null hypothesis, we have

$$E\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = 0$$

and

$$\widehat{V}(\text{Fisher}) \equiv \text{var}\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = \frac{1}{N^2} \sum_{i=1}^{N} \widehat{\tau}_i^2.$$

Therefore, for matched-pair experiments, the difference in the variances is

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = \tau^2/N + o_p(N^{-1}).$$

The asymptotic Normality of the two test statistics holds because of the Lindberg–Feller CLT for independent random variables, and therefore the different power behaviors of Neyman and Fisher's tests is again due to the above difference in the variances. Under $H_0(\text{Neyman})$, the difference is a higher order term, leading to asymptotically equivalent behaviors of Neymanian and Fisherian inferences. However, under the al-

ternative hypothesis with nonzero τ, the same paradox appears again in matched-pair experiments: we tend to reject with Neyman's test more often than with Fisher's test.

For matched-pair experiments with binary outcomes, we let $m^{obs}_{y_1 y_0}$ be the number of pairs with treatment outcome y_1 and control outcome y_0, where $y_1, y_0 \in \{0, 1\}$. Consequently, we can summarize the observed data by a two by two table with cell counts $(m^{obs}_{11}, m^{obs}_{10}, m^{obs}_{01}, m^{obs}_{00})$. Theorem 4 can then be further simplified as follows.

Corollary 2. In matched-pair experiments with binary outcomes, Neyman's test is asymptotically equivalent to

$$\frac{m^{obs}_{10} - m^{obs}_{01}}{\sqrt{m^{obs}_{10} + m^{obs}_{01} - (m^{obs}_{10} - m^{obs}_{01})^2/N}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{1.12}$$

under H_0(Neyman), and Fisher's test is asymptotically equivalent to

$$\frac{m^{obs}_{10} - m^{obs}_{01}}{\sqrt{m^{obs}_{10} + m^{obs}_{01}}} \xrightarrow{d} \mathcal{N}(0, 1) \tag{1.13}$$

under H_0(Fisher). And the asymptotic difference between the two tests is due to

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = (m^{obs}_{10} - m^{obs}_{01})^2/N^3 + o_p(N^{-1}).$$

Note that the number of discordant pairs, $m^{obs}_{10} + m^{obs}_{01}$, is fixed over all randomizations under the sharp null hypothesis, and therefore Fisher's test is equivalent to the exact test based on $m^{obs}_{10} \sim \text{Binomial}(m^{obs}_{10} + m^{obs}_{01}, 1/2)$. Its asymptotic form (1.13) is the same as the McNemar test under a super population model (Agresti and Min, 2004).

1.4.2 Factorial Experiments

Fisher (1935a) and Yates (1937) developed the classical factorial experiments in the context of agricultural experiments, and Wu and Hamada (2009) provided a comprehensive modern discussion of design and analysis of factorial experiments. Although rooted in randomization theory (Kempthorne, 1955; Hinkelmann and Kempthorne, 2007), the analysis of factorial experiments is dominated by linear and generalized linear models, with factorial effects often defined as model parameters. Realizing the inherent drawbacks of the predominant approaches, Dasgupta et al. (2015) discussed causal inference from 2^K factorial experiments using the potential outcomes framework, which allows for defining the causal estimands based on potential outcomes instead of model parameters.

We first briefly review the notation for factorial experiments adopted by Dasgupta et al. (2015). Assume that we have K factors with levels $+1$ and -1. Let $\boldsymbol{z} = (z_1, \ldots, z_K)' \in \mathcal{F}_K = \{+1, -1\}^K$, a K-dimensional vector, denote a particular treatment combination. The number of possible values of $\boldsymbol{z}$ is $J = 2^K$, for each of which we can define $Y_i(\boldsymbol{z})$ as the corresponding potential outcome for unit i under the SUTVA. We use a J-dimensional vector $\boldsymbol{Y}_i$ to denote all potential outcomes for unit i, where $i = 1, \ldots, N = r \times 2^K$ with an integer r representing the number of replications of each treatment combination. Without loss of generality, we will discuss the inference of the main factorial effect of factor 1, and analogous discussion also holds for general factorial effects due to symmetry. The main factorial effect of factor 1 can be characterized by a vector $\boldsymbol{g}_1$ of dimension J, with one half of its elements being $+1$ and the other half being -1. Specifically, the element of $\boldsymbol{g}_1$ is $+1$ if the

corresponding z_1 is $+1$, and -1 otherwise. For example, in 2^2 experiments, we have $\boldsymbol{Y}_i = (Y_i(+1,+1), Y_i(+1,-1), Y_i(-1,+1), Y_i(-1,-1))'$ and $\boldsymbol{g}_1 = (+1,+1,-1,-1)'$. We define $\tau_{i1} = 2^{-(K-1)}\boldsymbol{g}_1'\boldsymbol{Y}_i$ as the main factorial effect of factor 1 for unit i, and

$$\tau_1 = \frac{1}{N}\sum_{i=1}^{N}\tau_{i1} = 2^{-(K-1)}\boldsymbol{g}_1'\bar{\boldsymbol{Y}}$$

as the average main factorial effect of the factor 1, where $\bar{\boldsymbol{Y}} = \sum_{i=1}^{N}\boldsymbol{Y}_i/N$.

For factorial experiments, we define the treatment assignment as $W_i(\boldsymbol{z})$, with $W_i(\boldsymbol{z}) = 1$ if the ith unit is assigned to $\boldsymbol{z}$, and 0 otherwise. Therefore, we use $\boldsymbol{W}_i = \{W_i(\boldsymbol{z}) : \boldsymbol{z} \in \mathcal{F}_K\}$ as the treatment assignment vector for unit i, and let $\boldsymbol{W}$ be the collection of all the unit-level treatment assignments. The observed outcomes are deterministic functions of the potential outcomes and the treatment assignment, namely, $Y_i^{obs} = \sum_{\boldsymbol{z}\in\mathcal{F}_K} W_i(\boldsymbol{z})Y_i(\boldsymbol{z})$ for unit i, and $\boldsymbol{Y}^{obs} = (Y_1^{obs}, \ldots, Y_N^{obs})'$ for all the observed outcomes. Since

$$\bar{Y}^{obs}(\boldsymbol{z}) = \frac{1}{r}\sum_{\{i:W_i(\boldsymbol{z})=1\}} Y_i^{obs} = \frac{1}{r}\sum_{i=1}^{N} W_i(\boldsymbol{z})Y_i(\boldsymbol{z})$$

is unbiased for $\bar{Y}(\boldsymbol{z})$, we can unbiasedly estimate τ_1 by

$$\widehat{\tau}_1 = 2^{-(K-1)}\boldsymbol{g}_1'\bar{\boldsymbol{Y}}^{obs},$$

where $\bar{\boldsymbol{Y}}^{obs}$ is the J-dimensional vector for the average observed outcomes. Dasgupta

et al. (2015) showed that the sampling variance of $\widehat{\tau}_1$ is

$$\text{var}(\widehat{\tau}_1) = \frac{1}{2^{2(K-1)}r} \sum_{\boldsymbol{z} \in \mathcal{F}_K} S^2(\boldsymbol{z}) - \frac{1}{N} S_1^2, \tag{1.14}$$

where $S^2(\boldsymbol{z}) = \sum_{i=1}^N \{Y_i(\boldsymbol{z}) - \bar{Y}(\boldsymbol{z})\}^2/(N-1)$ is the finite population variance of the potential outcomes under treatment combination $\boldsymbol{z}$, and $S_1^2 = \sum_{i=1}^N (\tau_{i1} - \tau_1)^2/(N-1)$ is the finite population variance of the unit level factorial effects $\{\tau_{i1} : i = 1, \ldots, N\}$. Similar to the discussion in completely randomized experiments, the last term S_1^2 in (1.14) cannot be identified, and consequently the variance in (1.14) can only be "conservatively" estimated by the following Neyman-style variance estimator:

$$\widehat{V}_1(\text{Neyman}) = \frac{1}{2^{2(K-1)}r} \sum_{\boldsymbol{z} \in \mathcal{F}_K} s^2(\boldsymbol{z}),$$

where the sample variance of outcomes under treatment combination $\boldsymbol{z}$,

$$s^2(\boldsymbol{z}) = \sum_{\{i: W_i(\boldsymbol{z}) = 1\}} \{Y_i^{obs} - \bar{Y}^{obs}(\boldsymbol{z})\}^2/(r-1)$$

is unbiased for $S^2(\boldsymbol{z})$. The discussion above allows us to construct a Wald-type test for Neyman's null of zero average factorial effect for factor 1:

$$H_0^1(\text{Neyman}) : \tau_1 = 0.$$

On the other hand, based on the physical act of randomization in factorial exper-

iments, the FRT allows us to test the following sharp null hypothesis:

$$H_0(\text{Fisher}) : Y_i(\boldsymbol{z}) = Y_i^{obs}, \forall \boldsymbol{z} \in \mathcal{F}_K, \forall i = 1, \ldots, N. \tag{1.15}$$

This sharp null restricts all factorial effects for all the individuals to be zero, which is much stronger than H_0^1(Neyman). For a fair comparison, we use the same test statistic as $\widehat{\tau}_1$ in our randomization test, and denote $\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs})$ as a function of the treatment assignment and observed outcomes. Under the sharp null (1.15), the randomness of $\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs})$ is induced by randomization, and the following theorem gives us its mean and variance.

Theorem 5. We have $E\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = 0$ and

$$\widehat{V}_1(\text{Fisher}) \equiv \text{var}\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = \frac{1}{2^{2(K-1)}r} J s^2,$$

where

$$\bar{Y}^{obs} = \sum_{i=1}^{N} Y_i^{obs}/N, \quad s^2 = \sum_{i=1}^{N} (Y_i^{obs} - \bar{Y}^{obs})^2/(N-1)$$

are the sample mean and variance of all the observed outcomes.

Based on Normal approximations, comparison of the p-values reduces to the difference between $\widehat{V}_1$(Neyman) and $\widehat{V}_1$(Fisher), as shown in the theorem below.

Theorem 6. With large r, the difference between $\widehat{V}_1$(Neyman) and $\widehat{V}_1$(Fisher) is

$$\widehat{V}_1(\text{Fisher}) - \widehat{V}_1(\text{Neyman}) = \frac{1}{2^{3K-1}r} \sum_{\boldsymbol{z} \in \mathcal{F}_K} \sum_{\boldsymbol{z}' \in \mathcal{F}_K} \{\bar{Y}(\boldsymbol{z}) - \bar{Y}(\boldsymbol{z}')\}^2 + o_p(r^{-1}). \tag{1.16}$$

Formula (1.8) is a special case of formula (1.16) with $K = 1$ and $r = N_1 = N_0 = N/2$, since complete randomized experiments are special cases of factorial experiments with a single factor. Therefore, in factorial experiments with the same replicates r at each level, the paradox always exists under alternative hypothesis with nonzero τ_1, just as in balanced completely randomized experiments.

1.5 Improvements and Extensions

We have shown that Neyman's test is more powerful than Fisher's test in many realistic situations. In fact, the original form of Neyman's test is suboptimal. We discuss improved Neymanian variance estimators below, which lead to even more powerful tests. Moreover, the previous sections restrict the discussion on the difference-in-means statistic. We will further comment on the importance of this choice, and other possible alternative test statistics.

1.5.1 Improvements of the Neymanian Variance Estimators

For completely randomized experiments, Neyman (1923) used $S_\tau^2 \geq 0$ as a lower bound, which is not the sharp bound. Recently, for general outcomes Aronow et al. (2014) derived the sharp bound of S_τ^2 based on the marginal distributions of the treatment and control potential outcomes using the Frechét–Hoeffding bounds. In particular, when the outcome is binary, the sharp bound for the variance of $\widehat{\tau}$ results in the following simple variance estimator (Robins, 1988; Ding and Dasgupta, 2015):

$$\widehat{V}^c(\text{Neyman}) = \widehat{V}(\text{Neyman}) - |\widehat{\tau}|(1 - |\widehat{\tau}|)/(N-1). \tag{1.17}$$

Note that the adjustment term $|\widehat{\tau}|(1-|\widehat{\tau}|)/(N-1)$ is always non-negative, resulting in smaller variance estimators.

For matched-pair experiments, Imai (2008) improved the Neymanian variance estimator by using the Cauchy–Schwarz inquality, which may not be sharp. We are currently working on deriving sharp bounds for the variance of estimated factorial effects.

In summary, Neyman's test is even more powerful with improved variance estimators, which further bolsters the paradoxical situation wherein we reject Neyman's null but fail to reject Fisher's sharp null.

1.5.2 Choice of the Test Statistic

Our discussion is restricted to tests using the difference-in-means statistics, which plays an important role in practice. First, as hinted by Ding and Dasgupta (2015), for randomized experiments with binary outcomes, all test statistics are equivalent to the difference-in-means statistic. We formally state this conclusion in the following theorem.

Theorem 7. For completely randomized experiments, matched-pair experiments, and 2^K factorial experiments, if the outcomes are binary, then all test statistics are equivalent to the difference-in-means statistic.

Therefore, for binary data, the choice of test statistic is not a problem.

Second, for continuous outcomes, the difference-in-means statistic is important, because it not only serves as a candidate test statistic for the sharp null hypothesis but also an unbiased estimator for the average causal effect. In the illustrating example in

Section 1.6.3, practitioners are interested in finding the combination of several factors that achieves an optimal mean response.

For continuous outcomes we have more options of test statistics. For instance, the Kolmogorov–Smirnov and Wilcoxon–Mann–Whitney statistics are also useful candidates for the FRT. However, the Neymanian analogues of these two statistics are not established in the literature, and direct comparisons of the Fisherian and Neymanian using these two statistics are not possible at this moment. In Appendix A, we illustrate by numerical examples that the conservative nature of the FRT is likely to be true for these two statistics, because we find that the randomization distributions under the sharp null hypothesis is more disperse than those under weaker null hypotheses. Please see the Appendix A for more details, and it is our future research topic to pursue the theoretical results.

1.6 Illustrations

In this section, we will use real-life examples to illustrate the theory in the previous sections. The first two examples have binary outcomes, and therefore there is no concern about the choice of test statistic. The goal of the third example, a 2^4 full factorial experiment, is to find the optimal combination of the factors, and therefore the difference-in-means statistic is again a natural choice for a test statistic.

1.6.1 A Completely Randomized Experiment

Consider a hypothetical completely randomized experiment with binary outcome (Rosenbaum, 2002, pp.191). Among the 32 treated units, 18 of them have outcome

being 1, and among the 21 control units, 5 of them have outcome being 1. The Neymanian p-value based on the improved variance estimator in (1.17) is 0.004. The Fisherian p-value based on the FRT or equivalently Fisher's exact test is 0.026, and the Fisherian p-value based on Normal approximation in (1.10) is 0.020. The Neymanian p-value is smaller, and if we choose significance level at 0.01 then the paradox will appear in this example.

1.6.2 A Matched-Pair Experiment

The observed data of the matched-pair experiment in Agresti and Min (2004) can be summarized by the two by two table with cell counts $(m_{11}^{obs}, m_{10}^{obs}, m_{01}^{obs}, m_{00}^{obs}) = (53, 8, 16, 9)$. The Neymanian one-sided p-value based on (1.12) is 0.049. The Fisherian p-value based on the FRT is 0.076, and the Fisherian p-value based on Normal approximation in (1.13) is 0.051. Again, Neyman's test is more powerful than Fisher's test.

1.6.3 A 2^4 Full Factorial Experiment

In the "Design of Experiments" course in Fall 2014, a group of Harvard undergraduate students followed Box (1992)'s famous paper helicopter example for factorial experiments, and tried to identify the optimal combination of the four factors: paper type (construction paper, printer paper), paperclip type (small paperclip, large paperclip), wing length (2.5 inches, 2.25 inches), and fold length (0.5 inch, 1.0 inch), with the first level coded as -1 and the second level coded as $+1$. For more details, please see Box (1992). For each combination of the factors, they recorded two replicates of

the flying times of the helicopters. We display the data, as well as some summary statistics, in Table 1.2.

We show the Neymanian and Fisherian results in the upper and lower panel of Figure 1.5, respectively. Figure 1.5(a) shows both Neymanian point estimates and p-values for the 15 factorial effects. Seven of them, $F_1, F_2, F_4, F_1F_2, F_1F_3, F_1F_4$ and $F_1F_2F_4$, are significant at level 0.05, and after the Bonferroni correction, three of them, $F_1, F_2, F_1F_2F_4$, are still significant. Figure 1.5(b) shows the randomization distribution of the factorial effects under the sharp null hypothesis by a grey histogram. Note that all factorial effects have the same randomization distribution, because all of them are essentially a comparison of a random half versus the other half of the observed outcomes. Even though the sample size 32 is not huge, the randomization distribution is well approximated by the Normal distribution with mean zero and variance $\widehat{V}_1(\text{Fisher})$. Strikingly, only two factorial effects, F_1 and F_2, are significant, and after the Bonferroni correction only F_2 is significant. We further calculate the variance estimates: $\widehat{V}_1(\text{Neyman}) = 0.025$ and $\widehat{V}_1(\text{Fisher}) = 0.034$. The empirical findings in this particular example with finite sample are coherent with our asymptotic theory developed in Section 1.4.2. In this example, the Neymanian method can help detect more significant factors for achieving optimal flying time, while the more conservative Fisherian method may miss important factors.

Table 1.2: A 2^4 Factorial Design, Observed Outcomes and Summary Statistics

F_1	F_2	F_3	F_4	replicate 1	replicate 2	mean	standard deviation
-1	-1	-1	-1	1.60	1.55	1.58	0.04
-1	-1	-1	1	1.70	1.63	1.67	0.05
-1	-1	1	-1	1.44	1.38	1.41	0.04
-1	-1	1	1	1.56	1.61	1.58	0.04
-1	1	-1	-1	1.40	1.45	1.42	0.04
-1	1	-1	1	1.36	1.38	1.37	0.01
-1	1	1	-1	1.43	1.40	1.42	0.02
-1	1	1	1	1.32	1.27	1.29	0.04
1	-1	-1	-1	1.81	1.86	1.83	0.04
1	-1	-1	1	1.70	1.57	1.64	0.09
1	-1	1	-1	2.04	2.06	2.05	0.01
1	-1	1	1	1.68	1.61	1.65	0.05
1	1	-1	-1	1.58	1.28	1.43	0.21
1	1	-1	1	1.43	1.49	1.46	0.04
1	1	1	-1	1.51	1.54	1.52	0.02
1	1	1	1	1.53	1.38	1.46	0.11

1.7 Discussion

1.7.1 Historical Controversy and Modern Discussion

As pointed out by R. A. Fisher, "the actual and physical conduct of an experiment must govern the statistical procedure of its interpretation (Fisher, 1935a, Section II)." Neyman and Fisher both proposed statistical procedures for analysis of randomized experiments, relying on the randomization distribution itself. However, whether Neyman's null or Fisher's null makes more sense in practice goes back to the famous Neyman–Fisher controversy in a meeting of the Royal Statistical Society (Fisher, 1935b; Neyman and Iwaszkiewicz, 1935). Rosenbaum (2002, page 39) gave a very insightful philosophical discussion about the controversy, and Sabbaghi and Rubin (2014) revisited this controversy recently.

While the answer may depend on different perspectives of practical problems, we discussed only the consequent paradox of Neymanian and Fisherian testing procedures for their own null hypotheses. Both our numerical examples and asymptotic theory showed that we encounter a serious logical problem in the analysis of randomized experiments, even though both Neyman's and Fisher's tests are valid Frequentists' tests, in the sense of controlling correct type one errors under their own null hypotheses. Our numerical examples and theoretical analysis reach a conclusion different from the classical book by Rosenbaum (2002).

1.7.2 Randomization-Based and Regression-Based Inference

In current statistical practice, it is also very popular among applied researchers to use regression-based methods to analyze experimental data (Angrist and Pischke, 2008). Assume the a linear model for the observed outcomes: $Y_i^{obs} = \alpha + \beta T_i + \varepsilon_i$, where $\varepsilon_i, \ldots, \varepsilon_N$ are independently and identically distributed (iid) as $\mathcal{N}(0, \sigma^2)$. The hypothesis of zero treatment effect is thus characterized by $H_0(LM) : \beta = 0$. The usual ordinary least squares variance estimator for the regression coefficient may not correctly reflect the true variance of $\widehat{\tau}$ under randomization. Schochet (2010), Samii and Aronow (2012), Lin (2013) and Imbens and Rubin (2015) pointed out that we can solve this problem by using Huber–White heteroskedasticity-robust variance estimator (Huber, 1967; White, 1980), and the corresponding Wald test is asymptotically the same as Neyman's test. In Theorem 17 of Appendix A, we further build an equivalence relationship between Rao's score test and the FRT. For more technical details, please see Appendix A. Previous results, as well as Theorem 17, do justify the usage

of linear models in analysis of experimental data.

1.7.3 Interval Estimation

Originally, Neyman (1923) proposed an unbiased estimator for the average causal effect τ with a repeated sampling evaluation, which was later developed into the concept of the confidence interval (Neyman, 1937). In order to compare Neyman's approach with the FRT, we converted the interval estimator into a hypothesis testing procedure. As a dual, we can also invert the FRT for a sequence of null hypotheses to get an interval estimator for τ (Pitman, 1937, 1938; Rosenbaum, 2002). For example, we consider the sequence of sharp null hypotheses with constant causal effects:

$$H_0^{\delta}(\text{Fisher}) : Y_i(1) - Y_i(0) = \delta, \quad \forall i = 1, \ldots, N.$$

The interval estimator for τ with coverage rate $1 - \alpha$ of τ is

$$FI_{\alpha} = \left\{ \delta : \text{Fail to reject } H_0^{\delta}(\text{Fisher}) \text{ by the FRT at significant level } \alpha \right\}.$$

Dasgupta et al. (2015) called the interval FI a "fiducial interval" or "Fisherian interval," and found some empirical evidence in 2^K factorial designs that the "fiducial interval" is wider than the Neymanian "conservative" confidence interval. Due to the duality between hypothesis testing and interval estimation, our results about hypothesis testing can partially explain the phenomenon about interval estimation in Dasgupta et al. (2015).

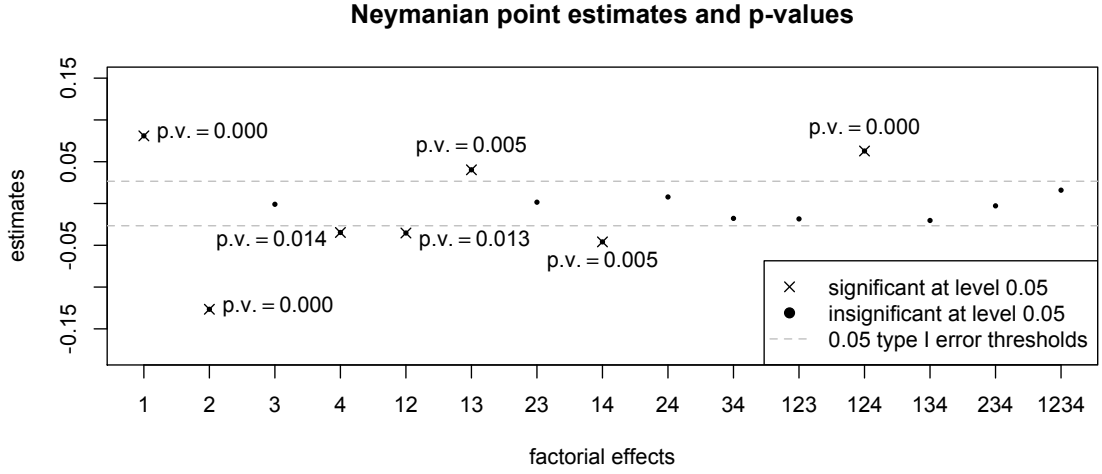
1.7.4 Practical Implications

We highlight the following practical implications of our theory developed in the above sections.

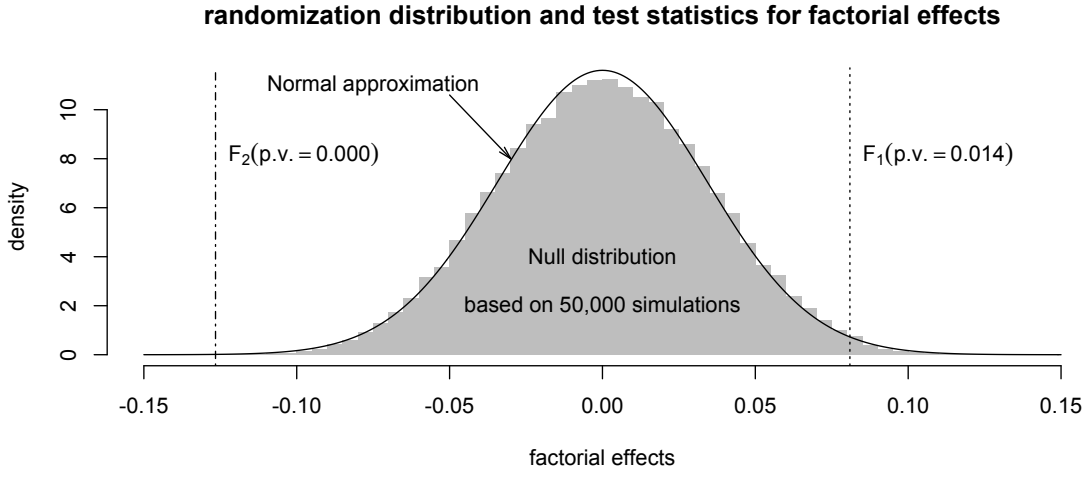
First, the FRT is usually less powerful than Neyman's test, even for the simplest case with constant causal effect. Practitioners should keep in mind that the FRT may miss important treatment factors. Our examples in Section 1.6 and the empirical evidence in Dasgupta et al. (2015) have confirmed our theoretical results.

Second, in the presence of treatment effect heterogeneity, the FRT may not be a valid test for the null hypothesis of zero average causal effect as illustrated by Example 4. Therefore, practitioners, especially those who are interested in social sciences, should always be aware of this potential danger of using the FRT, if the observed data show substantive heterogeneity in treatment and control groups. Treatment effect variation is another important issue beyond the current scope of our paper. Ding et al. (2015) investigate this problem under the randomization framework.

Third, although we have shown that the FRT is less powerful in many realistic cases, we are not concluding that Neymanian inference trumps Fisherian inference. All our comparisons are based on asymptotics under regularity conditions, and the conclusion may not be true with small sample sizes or "irregular" potential outcomes. Therefore, Fisherian inference is still useful for small sample problems and exact inference. In practice, we should always check the discrepancy between the Normal approximation and the exact randomization distribution as in Figure 1.5(b) before applying our theoretical results to applied problems.



(a) Neymanian Inference. Factorial effects $F_1, F_2, F_4, F_1F_2, F_1F_3, F_1F_4$ and $F_1F_2F_4$ are significant at level 0.05.



(b) Fisherian Inference. Factorial effects F_1 and F_2 are significant.

Figure 1.5: Randomization-Based Inference for a 2^4 Full Factorial Experiment

Chapter 2

A Potential Tale of Two by Two Tables from Completely Randomized Experiments

2.1 Introduction

The theory of causal inference from randomized treatment-control studies using the potential outcomes model has been well-developed over the past five decades and has been applied extensively to randomized experiments in the medical, behavioral and social sciences. The first formal notation for potential outcomes was introduced by Neyman (1923) in the development of randomization-based inference, and subsequently used by several researchers including Kempthorne (1952) and Cox (1958) for drawing causal inference from randomized experiments. The concept was formalized and extended by Rubin (1974, 1977, 1978) for other forms of causal inference from

randomized experiments *and* observational studies, and exposition of this transition appears in Rubin (2010). The three broad approaches to causal inference under the potential outcomes model are Fisherian, Neymanian and Bayesian.

The most common finite-population estimand in most causal inference problems is the average causal effect, defined as the finite-population average of unit-level causal effects. Since Neyman (1923)'s seminal work, additivity of unit-level treatment effects (or its lack thereof) and its influence on the inference for the average causal effect has been investigated thoroughly for continuous outcomes. In comparison, few researchers (e.g., Copas, 1973) have studied this problem for binary outcomes, in which the potential and observed outcomes can be summarized in the form of 2×2 contingency tables. In this paper, we provide a characterization of additivity based on the 2×2 table of potential outcomes, and use it to (i) justify Fisher's exact test from a randomization perspective, and (ii) propose an estimator of the variance of the average causal effect for binary outcomes that uniformly dominates the Neymanian variance estimator. As advocated by Rubin (1978), we also propose a Bayesian strategy for drawing inference about the average causal effect using the missing data perspective. Such a strategy is dependent on the assumptions related to model additivity, or more specifically, the nature and strength of the association between potential outcomes. We propose a novel sensitivity analysis which should help a practitioner understand how the analysis results might change if the assumptions are violated.

Apart from the average causal effect, other popular estimands for binary outcomes are the log of the causal risk ratio and the log of the causal odds ratio. Although of great practical interest, to the best of our knowledge, estimators of these causal

measures have not been studied carefully from the Neymanian perspective, because unlike the average causal effect, non-linearity of these estimands and their estimators make exact variance calculations intractable. We circumvent this problem by taking an asymptotic perspective. By deriving asymptotic expressions for variances of these estimators, we explore the adequacy of the widespread practice of drawing statistical inference for such causal estimands on the basis of independent Binomial models, and propose improved methods that are justified by randomization. We conduct simulation studies under different settings to demonstrate the effectiveness of the proposed methods and also illustrate their application to a recent randomized controlled trial.

The paper is organized as follows. In the following section, we define the potential outcomes, the finite population estimands, the assignment mechanism and the observed outcomes. In Section 2.3, we discuss the Fisherian and Neymanian forms of inference for 2×2 tables. In Section 2.4, we propose a Bayesian framework for causal inference, explore its frequentists' properties and also propose a methodology for sensitivity analysis to assess the effect of violation of assumptions regarding additivity (or its lack thereof) on the inference. Causal inference for non-linear estimands is discussed in Section 2.5. A detailed simulation study is conducted in Section 2.6 to compare the different methods of inference and to demonstrate the superiority of the proposed methodology. Application of the proposed methodology to randomized experiments with binary outcomes is demonstrated with a real-life example in Section 2.7. Some concluding remarks are presented in Section 2.8. Some technical details, the proofs, additional simulation studies, and more details of the application are in Appendix B.

2.2 Potential Outcomes, Estimands, and the Observed Data

The evolution of the potential outcomes framework was motivated by the need for a clear separation between the object of interest (often referred to as the "Science") and what researchers do to learn about the Science (e.g., randomly assign treatments to units). We assume a finite population of N experimental units that are exposed to a binary treatment W and yield a binary response Y. Under the Stable Unit Treatment Value Assumption (Cox, 1958; Rubin, 1980), we define $Y_i(t)$ as the potential outcome for individual i when exposed to treatment t ($t = 1$ and $t = 0$ often refer to treatment and control, respectively). The $N \times 2$ matrix of the potential outcomes $\{(Y_i(1), Y_i(0)) : i = 1, \ldots, N\}$ is typically referred to as the Science (Rubin, 2005). Because the response Y is binary, the information contained in the Science can be condensed into a 2×2 table as shown in Table 2.1.

Table 2.1: "Science Table" of the Potential Outcomes

	$Y(0) = 1$	$Y(0) = 0$	row sum
$Y(1) = 1$	N_{11}	N_{10}	N_{1+}
$Y(1) = 0$	N_{01}	N_{00}	N_{0+}
column sum	N_{+1}	N_{+0}	N

2.2.1 Finite-Population Causal Estimands and Uniformity of Unit-Level Causal Effects

Having defined the so-called Science, we now proceed to the definition of causal estimands. A unit-level causal effect is defined as a contrast between the potential

outcomes under the treatment and the control, for example, $\tau_i = Y_i(1) - Y_i(0)$. We define the finite population average causal effect as

$$\tau = \frac{1}{N}\sum_{i=1}^{N} \tau_i = p_1 - p_0,$$

where $p_t = \sum_{i=1}^{N} Y_i(t)/N$ is the finite population average of $Y(t)$ for $t = 0, 1$. For binary outcomes, the average causal effect is also called the causal risk difference (CRD). From Table 2.1, it follows that

$$\tau = N_{1+}/N - N_{+1}/N = (N_{10} - N_{01})/N.$$

A measure of uniformity (or its lack, thereof) of the unit-level causal effects is the finite population variance of the individual causal effect τ_i, given by

$$S_\tau^2 = \frac{1}{N-1}\sum_{i=1}^{N} (\tau_i - \tau)^2 = \frac{1}{N(N-1)}\left\{(N_{10} + N_{01})(N_{11} + N_{00}) + 4N_{10}N_{01}\right\}. \quad (2.1)$$

Note that S_τ^2 can also be represented as

$$S_\tau^2 = S_1^2 + S_0^2 - 2S_{10}, \quad (2.2)$$

where $S_t^2 = \sum_{i=1}^{N}\{Y_i(t) - p_t\}^2/(N-1)$ is the finite population variance of the potential outcome $Y_i(t)$, and

$$S_{10} = \sum_{i=1}^{N}\{Y_i(1) - p_1\}\{Y_i(0) - p_0\}/(N-1) = (N_{11}N_{00} - N_{10}N_{01})/\{N(N-1)\}$$

is the finite population covariance between $Y_i(1)$ and $Y_i(0)$.

Note that constant causal effect or additivity of unit-level causal effects implies that $S_1^2 = S_0^2, S_{10} = S_1 S_0$, and the uniformity measure $S_\tau^2 = 0$. Copas (1973) considered a representation of the potential outcomes similar to that in Table 2.1, and defined parameters $\alpha = \tau$ as the treatment effect and $\beta = (N_{10} + N_{01})/N$ as a measure of "the differential effect." However, we feel that β, which essentially equals $(\sum_{i=1}^{N} \tau_i^2)/N$, is not an adequate representation of the differential effect, because it does not reduce to zero when all unit-level causal effects are equal to 1 or -1. To discuss this aspect further, we consider the case of *strict additivity* of treatment effects, where $\tau_i = \tau$ for all $i = 1, \ldots, N$, and summarize its impact on the Science and its summary measures τ, S_τ^2 and β in Table 2.2.

Table 2.2: Effect of additivity on the Science

$\tau(=\tau_i)$	Entries of Table 2.1	$\tau = \alpha$	S_τ^2	β
1	$N_{11} = N_{01} = N_{00} = 0$, $N_{10} = N$	1	0	1
-1	$N_{11} = N_{10} = N_{00} = 0$, $N_{01} = N$	-1	0	1
0	$N_{10} = N_{01} = 0$, $N_{00} + N_{11} = N$	0	0	0

Note that the last row of Table 2.2 represents a special case of additivity, with zero treatment effect for each unit. Such a hypothesis about the Science case is referred to as Fisher's sharp null hypothesis of no treatment effect, and forms the basis of the "Fisherian" inference described in Section 2.3.1.

To sum up our discussion on the degree of uniformity of treatment effects, we define another condition referred to as *monotonicity* (Angrist et al., 1996).

Definition 1. The Science table is said to satisfy the monotonicity condition if either of the following two conditions hold: (i) $Y_i(1) \geq Y_i(0)$ for all i (or equivalently

$N_{01} = 0$), (ii) $Y_i(1) \leq Y_i(0)$ for all i (or equivalently $N_{10} = 0$).

Under monotonicity, we have $\tau = \alpha = \beta = N_{10}/N, \quad S_\tau^2 = N_{10}N_{11}/\{N(N-1)\}$ if (i) holds, and $\tau = \alpha = -N_{01}/N, \quad \beta = N_{01}/N, \quad S_\tau^2 = N_{01}N_{11}/\{N(N-1)\}$ if (ii) holds. We also note that any one of the three additivity conditions as described in Table 2.2 implies at least one of the monotonicity assumptions. We shall discuss the impact of strict additivity and monotonicity on the inference for τ in Section 2.3.2.

2.2.2 Treatment Assignment and the Observed Data

We consider a completely randomized treatment assignment in which N_1 and N_0 units receive treatments 1 and 0 respectively. Let $\boldsymbol{W} = (W_1, \ldots, W_N)$ be the vector of treatment assignments and let $\boldsymbol{w} = (w_1, \ldots, w_N)$ be a realization of $\boldsymbol{W}$. Then, a completely randomized experiment satisfies $P(\boldsymbol{W} = \boldsymbol{w}) = N_1!N_0!/N!$ if $\sum_{i=1}^N w_i = N_1$ and $P(\boldsymbol{W} = \boldsymbol{w})$ otherwise. The observed outcomes are deterministic functions of both the treatment and the potential outcomes, since $Y_i^{\text{obs}} = W_iY_i(1) + (1 - W_i)Y_i(0)$. Let $\boldsymbol{Y}^{\text{obs}} = (Y_1^{\text{obs}}, \cdots, Y_N^{\text{obs}})$ be the vector of the observed outcomes. Since the treatment and the outcome are both binary, the observed data form a 2×2 contingency table as shown in Table 2.3. The row sums in Table 2.3, (N_1, N_0), are the numbers of individuals receiving treatment and control, and the column sums, (n_{+1}, n_{+0}), are the number of individuals with outcomes 1 and 0, respectively.

Table 2.3: Summary of the Observed Data

	$Y^{\text{obs}} = 1$	$Y^{\text{obs}} = 0$	row sum
$W = 1$	n_{11}	n_{10}	N_1
$W = 0$	n_{01}	n_{00}	N_0
column sum	n_{+1}	n_{+0}	N

We conclude this section by emphasizing that the fundamental problem of causal inference is the missingness of one element of each pair $(Y_i(1), Y_i(0))$. Consequently, the key idea is to infer about the entries of Table 2.1 (and the estimands that are functions of these unknown entries) using those of Table 2.3 and the distribution of these entries under randomization.

2.3 Fisherian and Neymanian Approaches to Inference

In this section, the potential outcomes of the finite population are assumed to be fixed numbers, and the randomness in the observed outcomes comes only from randomization of the treatment assignment (Neyman, 1923; Rubin, 1990). We discuss two forms of finite-population inference — Fisherian and Neymanian — under this set-up. Fisher's form of randomization-based inference focuses on assessing the sharp null hypothesis of no treatment effect using the randomization distribution of a test statistic, which is obtained by imputing the missing outcomes under the sharp null. Neyman's form of randomization-based inference can be viewed as drawing inferences by evaluating the expectations of statistics over the distribution induced by the assignment mechanism in order to calculate a confidence interval for the typical causal effect. Using asymptotic results is one way of achieving this. In the following subsection (Section 2.3.1), we briefly discuss the Fisher randomization test and establish its connection to Fisher's exact test. In Section 2.3.2, we discuss Neymanian inference and propose an improvement over the traditional Neymanian estimator.

2.3.1 Fisherian Randomization Test and Its Connection to Fisher's Exact Test

According to (Fisher, 1935a), randomization yields "a reasoned basis for inference," and it allows for testing the sharp null hypothesis of zero individual causal effect, i.e., $Y_i(1) = Y_i(0)$ for $i = 1, \ldots, N$, characterized by the last row of Table 2.2. Such a null hypothesis permits imputation of all the missing potential outcomes. Although in principle any test statistic can be used, the most natural one is $\widehat{\tau} = \widehat{p}_1 - \widehat{p}_0$, where $\widehat{p}_1 = \sum_{i=1}^{N} W_i Y_i^{\text{obs}}/N_1 = n_{11}/N_1$, and $\widehat{p}_0 = \sum_{i=1}^{N}(1 - W_i)Y_i^{\text{obs}}/N_0 = n_{01}/N_0$. The test statistic $\widehat{\tau}$ is a function of both the treatment assignment and the observed outcomes. Under the sharp null hypothesis, the randomness of $\widehat{\tau}$ comes only from the randomization of the treatment assignment $\boldsymbol{W}$. The p-value under the sharp null is a measure of the extremeness of the observed value of the test statistic with respect to its randomization distribution under the sharp null. For a two-sided test, the p-value is typically defined as the proportion of values of $|\widehat{\tau}|$ generated under all possible randomizations that exceed its observed value $|\widehat{\tau}^{\text{obs}}|$. In general, the null distribution of $\widehat{\tau}$ and the p-value can either be calculated exactly, or approximated by Monte Carlo.

However, we can obtain the "exact" distribution of the randomization test statistic for a binary outcome. In Table 2.3, the margins N_1 and N_0 are fixed by design. Under the sharp null hypothesis, the margins n_{+1} and n_{+0} represent the number of observations with potential outcomes $Y_i(1) = Y_i(0) = 1$ and $Y_i(1) = Y_i(0) = 0$, respectively, and are equal to N_{11} and N_{00} in Table 2.1. It follows that

$$\widehat{\tau} = \frac{n_{11}}{N_1} - \frac{n_{01}}{N_0} = \frac{n_{11}}{N_1} - \frac{n_{+1} - n_{11}}{N_0} = \frac{N}{N_1 N_0} n_{11} - \frac{N_{11}}{N_0}, \tag{2.3}$$

i.e., the test statistic $\widehat{\tau}$ is a monotone function of n_{11}. Therefore, the rejection region based on $\widehat{\tau}$ is equivalent to the rejection region based on n_{11}, which has the usual Hypergeometric null distribution of the exact test for a two by two contingency table.

More interestingly, any randomization test using statistics other than $\widehat{\tau}$ is also equivalent to the test based on n_{11}, since any test statistic is a function of n_{11} under Fisher's sharp null hypothesis. Numerically, the test has exactly the same form as Fisher's exact test, although the two tests were originally derived based on completely different statistical reasonings. In observational studies under Multinomial or independent Binomial sampling, Fisher (1935c) justified his exact test for association as a conditional test, by arguing that the marginal counts are nearly ancillary. However, it turns out that the marginal counts contain some information about the association (Chernoff, 2004), and they are not ancillary. Here, we give a justification of the validity for Fisher's exact test based on randomization, if the data truly come from a completely randomized experiment. For more discussion about the hypothesis testing issue, see Berkson (1978), Yates (1984) and Chernoff (2004) for observational studies, and Ding (2014) for randomized experiments.

2.3.2 Neymanian Inference for the Average Causal Effect

Neyman (1923) showed that $\widehat{\tau} = \widehat{p}_1 - \widehat{p}_0$ is unbiased for τ, with the sampling variance

$$\text{var}(\widehat{\tau}) = \frac{N_0}{N_1 N} S_1^2 + \frac{N_1}{N_0 N} S_0^2 + \frac{2}{N} S_{10} = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}, \tag{2.4}$$

where S_τ^2, S_1^2 and S_0^2 are defined in Section 2.2.1. The proof can be found in Neyman

(1923) or directly from Lemma 5 in Appendix B. Since the third term in (2.4), S_τ^2/N, depends on the joint distribution of the potential outcomes, it is not identifiable from the observed data without further assumptions. Because of this difficulty, Neyman (1923) proposed a "conservative" estimator for $\text{var}(\widehat{\tau})$, defined as

$$\widehat{V}_{Neyman} = \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0}, \tag{2.5}$$

where $s_1^2 = \sum_{W_i=1}(Y_i^{\text{obs}} - \widehat{p}_1)^2/(N_1 - 1)$ and $s_0^2 = \sum_{W_i=0}(Y_i^{\text{obs}} - \widehat{p}_0)^2/(N_0 - 1)$ are the sample variances of the observed outcomes under the treatment and the control, respectively. For binary outcomes, the variance estimator can be simplified as

$$\begin{aligned} \widehat{V}_{Neyman} &= \frac{1}{N_1(N_1-1)}\left(n_{11} - N_1\frac{n_{11}^2}{N_1^2}\right) + \frac{1}{N_0(N_0-1)}\left(n_{01} - N_0\frac{n_{01}^2}{N_0^2}\right) \\ &= \frac{\widehat{p}_1(1-\widehat{p}_1)}{N_1-1} + \frac{\widehat{p}_0(1-\widehat{p}_0)}{N_0-1}. \end{aligned} \tag{2.6}$$

As we will discuss later in Section 2.5.2, (2.6) is very close to the standard formula for the variance of the difference of sample proportions, except for the fact that the coefficient denominator in the latter are N_w instead of $N_w - 1$ for $w = 0, 1$.

The variance estimator $\widehat{V}_{Neyman}$ is "conservative" in the sense that it only unbiasedly estimates the first two terms of (2.4), $S_1^2/N_1 + S_0^2/N_0$, and therefore $E(\widehat{V}_{Neyman}) \geq \text{var}(\widehat{\tau})$, a fact pointed out by several authors, e.g., Gadbury (2001), who provided an expression for the bias of the estimator. The variance estimator $\widehat{V}_{Neyman}$ is unbiased for the true variance if and only if the individual causal effects are constant ($\tau_i = \tau$) or, equivalently, the conditions in Table 2.2 are satisfied. Neyman (1923)'s constant causal effect assumption is equivalent to using 0 as a lower bound for S_τ^2, which is

not sharp for binary outcomes. Consequently, Neyman's "conservative" variance estimator can be improved for binary outcomes, even if the potential outcomes are not strictly additive. The following result gives the sharp lower bound for S_τ^2/N in terms of τ.

Theorem 8. A lower bound for S_τ^2/N is

$$\frac{S_\tau^2}{N} \geq \frac{|\tau|(1-|\tau|)}{N-1}, \tag{2.7}$$

and equality holds if and only if the potential outcomes satisfy the monotonicity condition as stated in Definition 1.

Theorem 8 implies that the monotonicity assumptions are the most "conservative" cases for variance estimation. As shown in the proof of Theorem 8, the lower bound (2.7) for S_τ^2 is obtained via an optimization approach, which minimizes S_τ^2 under the constraints of the marginal distributions p_1 and p_0. Therefore, the lower bound in (2.7) is "sharp", in the sense that it cannot be uniformly improved without further assumptions.

The lower bound for S_τ^2/N allows us to define the following estimator, which is an improvement over Neyman's variance estimator given by (2.6):

$$\widehat{V}^c_{Neyman} = \frac{\widehat{p}_1(1-\widehat{p}_1)}{N_1-1} + \frac{\widehat{p}_0(1-\widehat{p}_0)}{N_0-1} - \frac{|\widehat{\tau}|(1-|\widehat{\tau}|)}{N-1}. \tag{2.8}$$

This estimator cannot be larger than $\widehat{V}_{Neyman}$, and is also an improvement of the variance estimator given in Robins (1988). If $\tau \in \{1, -1, 0\}$, i.e., if $S_\tau^2 = 0$, we have $|\tau|(1-|\tau|) = 0$, and with large sample size, the adjusting term $|\widehat{\tau}|(1-|\widehat{\tau}|)/(N-1)$ is of

higher order relative to the two leading terms of the variance estimator (2.8). Therefore, if $S_\tau^2 = 0$, then asymptotically, the adjusting term does not hurt, and $\widehat{V}^c_{Neyman}$ and $\widehat{V}_{Neyman}$ are equivalent. However, for small samples, we may under-estimate the true sampling variance due to the positive adjusting term $|\tau|(1-|\tau|)/(N-1)$. We will investigate this finite sample issue further in the simulation studies. If the true average causal effect is not $-1, 0$, or 1, i.e., $S_\tau^2 \neq 0$ the correction term $|\widehat{\tau}|(1-|\widehat{\tau}|)/(N-1)$ in the variance estimator cannot be asymptotically neglected, and the "adjusted" variance estimator will improve Neyman (1923)'s original variance estimator. For example, if we observed a two by two table with cell counts $(n_{11}^{\text{obs}}, n_{10}^{\text{obs}}, n_{01}^{\text{obs}}, n_{00}^{\text{obs}}) = (15, 5, 5, 15)$, then we have $\widehat{p}_1 = 0.75, \widehat{p}_0 = 0.25, \widehat{V}_{Neyman} = 0.020$, and $\widehat{V}^c_{Neyman} = 0.013$, with the latter variance estimator 32.48% smaller than the former one.

2.4 Bayesian Causal Inference for Binary Outcomes

In this section, we adopt the Bayesian causal inference framework advocated by Rubin (1978). We assume that all the potential outcomes are drawn from a hypothetical super-population, while we are still interested in making inference on the finite population average causal effect τ. Similar to Neymanian randomization inference, the association between the potential outcomes is also crucial for our Bayesian causal inference. We first propose a Bayesian procedure based on a simple model with independent potential outcomes, and discuss its frequentists' repeated sampling property Rubin (1984). We then propose a sensitivity analysis procedure to investigate the impact of departures from the independence assumption on Bayesian inference.

2.4.1 Independent Potential Outcomes

Assume that we have the following model with independent potential outcomes:

$$Y_i(1) \sim \text{Bern}(\pi_{1+}), \quad Y_i(0) \sim \text{Bern}(\pi_{+1}), \quad Y_i(1) \perp\!\!\!\perp Y_i(0), \quad i = 1, \cdots, N,$$

where "$\perp\!\!\!\perp$" denotes independence. The notation (π_{1+}, π_{+1}) is chosen to be coherent with the marginal probabilities in Table 2.4 to be discussed later. We will relax the independence assumption in Section 2.4.3. We postulate the following priors $\pi_{1+} \sim \text{Beta}(\alpha_1, \beta_1), \pi_{+1} \sim \text{Beta}(\alpha_0, \beta_0)$, and assume that they are independent *a priori*.

Since the treatment assignment mechanism is ignorable (Rubin, 1978) in completely randomized experiments, the joint posterior distribution of π_{1+} and π_{+1} is

$$\begin{aligned} f(\pi_{1+}, \pi_{+1} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) \quad \propto \quad & \pi_{1+}^{\alpha_1 - 1}(1-\pi_{1+})^{\beta_1 - 1}\pi_{+1}^{\alpha_0 - 1}(1-\pi_{+1})^{\beta_0 - 1} \\ & \cdot \pi_{1+}^{n_{11}}(1-\pi_{1+})^{n_{10}}\pi_{+1}^{n_{01}}(1-\pi_{+1})^{n_{00}}, \end{aligned} \tag{2.9}$$

or equivalently,

$$\pi_{1+}|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}} \sim \text{Beta}(n_{11} + \alpha_1, n_{10} + \beta_1), \pi_{+1}|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}} \sim \text{Beta}(n_{01} + \alpha_0, n_{00} + \beta_0),$$

and they are independent *a posteriori*. After obtaining the posterior distribution of (π_{1+}, π_{+1}), we can impute all the missing potential outcomes, conditioning on (π_{1+}, π_{+1}). If $W_i = 1$, we impute $Y_i(0)|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{+1} \sim \text{Bern}(\pi_{+1})$; and if $W_i = 0$, we impute $Y_i(1)|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+} \sim \text{Bern}(\pi_{1+})$. Therefore, the posterior distribution of

τ conditioning on π_{1+} and π_{+1} is

$$\tau|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1} \sim \frac{n_{11} + B_0 - n_{01} - B_1}{N}, \tag{2.10}$$

where $B_1 \sim \text{Binomial}(N_1, \pi_{+1}), B_0 \sim \text{Binomial}(N_0, \pi_{1+})$, and they are independent. The description above also illustrates a Monte Carlo strategy for simulating the posterior distribution of τ. For theoretical comparison with Neymanian inference, we can also obtain the posterior mean and variance of τ as follows. We give the exact formulae for posterior mean and variance in Appendix B, and here for simplicity we give approximate formulae.

Theorem 9. Assume that the prior pseudo counts $(\alpha_0, \beta_0, \alpha_1, \beta_1)$ are small compared to n_{ij}'s. The posterior mean of τ is

$$E(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) \approx \widehat{\tau},$$

and the posterior variance of τ is

$$\text{var}(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) \approx \frac{N_0}{N}\frac{\widehat{p}_1(1-\widehat{p}_1)}{N_1 - 1} + \frac{N_1}{N}\frac{\widehat{p}_0(1-\widehat{p}_0)}{N_0 - 1}. \tag{2.11}$$

From Theorem 9, we can see that the posterior variance of τ is smaller than Neyman's variance estimator. These variances are different because Neyman (1923) assumed perfect correlation between the potential outcomes, while the Bayesian model assumes independence between the potential outcomes. As shown in (2.4), the assumption that $S_\tau^2 = 0$ is the worst case for the variance of $\text{var}(\widehat{\tau})$, and Neyman (1923)

adopted this as the most "conservative" estimator for the true variance.

2.4.2 Frequency Evaluation of the Bayesian Procedure Under Independence

Going back to the finite population perspective, the sampling distribution of $\widehat{\tau}$ depends on the finite population covariance between $Y_i(1)$ and $Y_i(0)$, as shown in (2.4). Assuming independence between $Y_i(1)$ and $Y_i(0)$, we have $S_{10} = 0$, and (2.4) becomes

$$\text{var}(\widehat{\tau}) = \frac{N_0}{N_1 N} S_1^2 + \frac{N_1}{N_0 N} S_0^2.$$

The variance of $\widehat{\tau}$ can be unbiasedly estimated by

$$\widehat{V}_{ind} = \frac{N_0}{N_1 N} s_1^2 + \frac{N_1}{N_0 N} s_0^2 = \frac{N_0}{N} \frac{\widehat{p}_1(1-\widehat{p}_1)}{N_1 - 1} + \frac{N_1}{N} \frac{\widehat{p}_0(1-\widehat{p}_0)}{N_0 - 1}. \tag{2.12}$$

The estimator of the sampling variance of $\widehat{\tau}$ in (2.12) and the approximated posterior variance of τ in (2.11) under independence are the same.

Therefore, the Bayesian credible interval under independence will have a correct asymptotic coverage property, if the finite population covariance of the potential outcomes is zero. However, if the finite population covariance between $Y_i(1)$ and $Y_i(0)$, S_{10}, is negative, we have

$$\text{var}(\widehat{\tau}) < \frac{N_0}{N_1 N} S_1^2 + \frac{N_1}{N_0 N} S_0^2$$

according to (2.4), which implies that the Bayesian credible interval will over-cover

the truth over repeated sampling. If the finite population covariance between $Y_i(1)$ and $Y_i(0)$, S_{10}, is positive, the Bayesian credible interval may not have a correct frequentists' coverage property.

2.4.3 Bayesian Sensitivity Analysis

The independence between potential outcomes may not be plausible even conditionally on observed covariates. In particular, if the potential outcomes are positively correlated, the Bayesian credible interval may not have a correct frequentists' coverage property. However, the observed data provide no information about the association between the two potential outcomes, since they are never jointly observed. Therefore, we propose a sensitivity analysis approach for the Bayesian model discussed above.

Table 2.4: Model of the Potential Outcomes

	$Y(0)=1$	$Y(0)=0$	row sum
$Y(1)=1$	π_{11}	π_{10}	π_{1+}
$Y(1)=0$	π_{01}	π_{00}	$1-\pi_{1+}$
column sum	π_{+1}	$1-\pi_{+1}$	1

The joint distribution of $(Y_i(1), Y_i(0))$ follows a Multinomial distribution with parameters $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ as shown in Table 2.4, which can be equivalently characterized by the marginal distributions (π_{1+}, π_{+1}) and an association parameter. We propose a new characterization of association between the potential outcomes in terms of the sensitivity parameter:

$$\gamma = \frac{P\{Y(1)=1 \mid Y(0)=1\}}{P\{Y(1)=1 \mid Y(0)=0\}} = \frac{\pi_{11}}{\pi_{10}} \frac{1-\pi_{+1}}{\pi_{+1}} \in (0, \infty).$$

When the potential outcomes are independent, we have $\gamma = 1$; when $\pi_{11} \to 0$, we have

$\gamma \to 0$; when $\pi_{10} \to 0$, we have $\gamma \to \infty$. In practice, we propose varying our sensitivity parameter γ over a wide range of values, and performing Bayesian inference at each fixed value of γ.

There is a one-to-one mapping between $(\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00})$ and $(\pi_{1+}, \pi_{+1}, \gamma)$, and thus the cell probabilities π_{jk}'s can be expressed as

$$\pi_{11} = \frac{\gamma \pi_{1+} \pi_{+1}}{1 - \pi_{+1} + \gamma \pi_{+1}}, \qquad \pi_{10} = \frac{\pi_{1+}(1 - \pi_{+1})}{1 - \pi_{+1} + \gamma \pi_{+1}}, \tag{2.13}$$

$$\pi_{01} = \pi_{+1} - \frac{\gamma \pi_{1+} \pi_{+1}}{1 - \pi_{+1} + \gamma \pi_{+1}}, \qquad \pi_{00} = 1 - \pi_{+1} - \pi_{1+} + \pi_{11}. \tag{2.14}$$

Since all the cell probabilities are within the interval $[0, 1]$, the equations in (2.13) and (2.14) impose the following restrictions on $(\pi_{1+}, \pi_{+1}, \gamma)$:

$$\gamma(\pi_{1+} - \pi_{+1}) \leq 1 - \pi_{+1}, \quad \gamma \pi_{+1} > \pi_{1+} + \pi_{+1} - 1. \tag{2.15}$$

The posterior distributions of (π_{1+}, π_{+1}) are the same as (2.9). However, the imputations of the missing potential outcomes are different from Section 2.4.1. For $W_i = 1$, we impute

$$\begin{aligned} Y_i(0)|Y_i(1) = 1 &\sim \text{Bern}\left(\frac{\pi_{11}}{\pi_{1+}} = \frac{\gamma \pi_{+1}}{1 - \pi_{+1} + \gamma \pi_{+1}}\right), \\ Y_i(0)|Y_i(1) = 0 &\sim \text{Bern}\left(\frac{\pi_{01}}{1 - \pi_{1+}} = \frac{\pi_{+1}}{1 - \pi_{1+}} - \frac{\gamma \pi_{1+} \pi_{+1}}{(1 - \pi_{+1} + \gamma \pi_{+1})(1 - \pi_{1+})}\right). \end{aligned}$$

For $W_i = 0$, we impute

$$\begin{aligned} Y_i(1)|Y_i(0) = 1 &\sim \text{Bern}\left(\frac{\pi_{11}}{\pi_{+1}} = \frac{\gamma\pi_{1+}}{1-\pi_{+1}+\gamma\pi_{+1}}\right), \\ Y_i(1)|Y_i(0) = 0 &\sim \text{Bern}\left(\frac{\pi_{10}}{1-\pi_{+1}} = \frac{\pi_{1+}}{1-\pi_{+1}+\gamma\pi_{+1}}\right). \end{aligned}$$

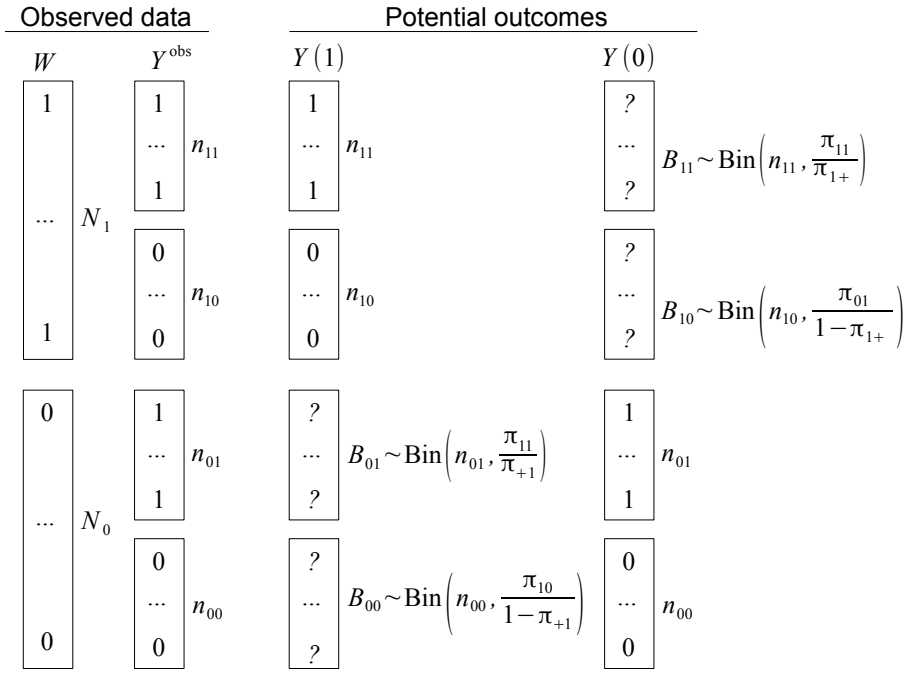


Figure 2.1: Imputation of the Missing Potential Outcomes

We illustrate the strategy for imputing missing potential outcomes in Figure 2.1, in which we have

$$\tau|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1} \sim \frac{n_{11} + B_{01} + B_{00} - B_{11} - B_{10} - n_{01}}{N},$$

where $B_{11} \sim \text{Binomial}(n_{11}, \pi_{11}/\pi_{1+})$, $B_{10} \sim \text{Binomial}\{n_{10}, \pi_{01}/(1-\pi_{1+})\}$, $B_{01} \sim \text{Binomial}(n_{01}, \pi_{11}/\pi_{+1})$, $B_{00} \sim \text{Binomial}\{n_{00}, \pi_{10}/(1-\pi_{+1})\}$, and $\{B_{11}, B_{10}, B_{01}, B_{00}\}$

are independent. Note that although the posterior distribution of (π_{1+}, π_{+1}) does not depend on the association parameter γ, the posterior distribution of τ does. While there is no explicit form of the posterior distribution of τ, we can approximate it via Monte Carlo. We will apply the proposed sensitivity analysis in Section 2.7.

2.5 General Causal Measures

Up to now, we have considered the most commonly used causal estimand, the average causal effect (or CRD). However, researchers and practitioners are also often interested in the log of the causal risk ratio (relative risk)

$$\log(\text{CRR}) = \log(p_1) - \log(p_0) = \log\left(\frac{N_{11} + N_{10}}{N_{11} + N_{01}}\right), \tag{2.16}$$

and the log of the causal odds ratio

$$\log(\text{COR}) = \text{logit}(p_1) - \text{logit}(p_0) = \log\left(\frac{N_{11} + N_{10}}{N_{01} + N_{00}}\right) - \log\left(\frac{N_{11} + N_{01}}{N_{10} + N_{00}}\right), \tag{2.17}$$

where $\text{logit}(x) = \log(x) - \log(1-x)$. One attractive feature of CRD is that it is linear in the individual causal effects. On the contrary, log(CRR) and log(COR) are finite population level causal estimands, which are not simple averages of individual causal effects. The linearity of the average causal effect permitted Neyman (1923) to obtain an unbiased estimator with exact variance. However, the elegant mathematics of Neyman (1923)'s randomization inference for CRD is not directly applicable to non-linear causal measures. We will fill in the gap by obtaining asymptotic randomization

inference for log(CRR) and log(COR).

We can obtain estimators for the log of the causal risk ratio and odds ratio by substituting estimators of p_1 and p_0 in (2.16) and (2.17), i.e., $\log(\widehat{\text{CRR}}) = \log(\widehat{p}_1) - \log(\widehat{p}_0)$ and $\log(\widehat{\text{COR}}) = \text{logit}(\widehat{p}_1) - \text{logit}(\widehat{p}_0)$. As mentioned earlier, general nonlinear causal measures have not been studied carefully from the Neymanian perspective, because the absence of linearity makes exact variance calculations intractable for such measures. Instead, we take an asymptotic perspective in this section. In the following subsections, we will (i) propose asymptotic randomization inference for $\log(\widehat{\text{CRR}})$ and $\log(\widehat{\text{COR}})$; (ii) compare them with the results under traditional independent Binomial models; (iii) discuss Bayesian inference for the general causal measures.

2.5.1 Neymanian Asymptotic Randomization Inference

Unfortunately, the unbiasedness is not preserved by plugging $\widehat{p}_1$ and $\widehat{p}_0$ into the nonlinear functions (2.16) and (2.17). Furthermore, the plug-in estimators do not have finite means or variances, since $\widehat{p}_1$ and $\widehat{p}_0$ can equal to 0 or 1 with positive probabilities. In spite of these limitations, when p_1 and p_0 are both bounded away from 0 and 1, the estimators $\log(\widehat{\text{CRR}})$ and $\log(\widehat{\text{COR}})$ have regular asymptotic distributions, summarized in the following two theorems.

Theorem 10. If $0 < p_0, p_1 < 1$, as $N \to \infty$, $\log(\widehat{\text{CRR}})$ is consistent for log(CRR) and asymptotically Normal with asymptotic variance

$$\frac{N_1 p_1 + N_0 p_0}{N p_1 p_0} \left(\frac{S_1^2}{N_1 p_1} + \frac{S_0^2}{N_0 p_0} - \frac{S_\tau^2}{N_1 p_1 + N_0 p_0} \right). \tag{2.18}$$

Assuming $S_\tau^2 = 0$ as in Neyman (1923), we can estimate the asymptotic variance by

$$\widehat{V}_{\text{CRR}} = \frac{n_{10}}{n_{11}N_1}\frac{(n_{11}+n_{01})N_0}{n_{01}N} + \frac{n_{00}}{n_{01}N_0}\frac{(n_{11}+n_{01})N_1}{n_{11}N}. \tag{2.19}$$

Theorem 11. If $0 < p_0, p_1 < 1$, as $N \to \infty$, $\log(\widehat{\text{COR}})$ is consistent for $\log(\text{COR})$ and asymptotically Normal with asymptotic variance

$$\frac{N_1p_1(1-p_1)+N_0p_0(1-p_0)}{Np_1(1-p_1)p_0(1-p_0)}\left\{\frac{S_1^2}{N_1p_1(1-p_1)}+\frac{S_0^2}{N_0p_0(1-p_0)} - \frac{S_\tau^2}{N_1p_1(1-p_1)+N_0p_0(1-p_0)}\right\}. \tag{2.20}$$

Assuming $S_\tau^2 = 0$ as in Neyman (1923), we can estimate the asymptotic variance by

$$\widehat{V}_{\text{COR}} = \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}. \tag{2.21}$$

The variance formulae (2.18) and (2.20) for $\log(\widehat{\text{CRR}})$ and $\log(\widehat{\text{COR}})$ are similar to the variance formula (2.4) for $\widehat{\tau}$, depending on the finite population variances of the potential outcomes S_1^2 and S_0^2, and the unidentifiable finite population variance of the individual causal effect S_τ^2.

Furthermore, borrowing the idea of bias-correction for ratio estimators (Cochran, 1977), we can obtain bias-corrected estimators for log(CRR) and log(COR), which have lower order asymptotic biases than the naïve moment estimators. Similar to Neyman's variance estimator for $\text{var}(\widehat{\tau})$, the variance estimators in (2.19) and (2.21) are conservative unless the constant causal effects assumption holds. Analogous to the result in (2.8) for CRD, using the lower bound for S_τ^2 in Theorem 8, we can

improve the variance estimators (2.19) and (2.21) for the bias corrected estimators for log(CRR) and log(COR). These bias-corrected point estimators and improved variance estimators improve the moment-based Neymanian inference asymptotically, and we call them improved Neymanian inference hereinafter. We provide technical details about bias and variance reduction with proofs in Appendix B.

2.5.2 Independent Binomial Models Versus Neymanian Inference

In current clinical practice, the following independent Binomial models are widely used:

$$n_{11} \sim \text{Binomial}(N_1, p_1), \quad n_{01} \sim \text{Binomial}(N_0, p_0), \quad n_{11} \perp\!\!\!\perp n_{01}. \tag{2.22}$$

In the model above, n_{11} and n_{01} are assumed to be Binomial random variables. Such an assumption cannot, however, be justified by randomization using the potential outcomes model.

The maximum likelihood estimators for $p_1 - p_0, \log(p_1) - \log(p_0), \text{logit}(p_1) - \text{logit}(p_0)$ are the same as $\widehat{\tau}, \log(\widehat{\text{CRR}}), \log(\widehat{\text{COR}})$, and their asymptotic variances (Woolf, 1955; Rothman et al., 2008) can be estimated by

$$\widehat{V}^{\text{Bin}}_{\text{CRD}} = \frac{\widehat{p}_1(1-\widehat{p}_1)}{N_1} + \frac{\widehat{p}_0(1-\widehat{p}_0)}{N_0}, \tag{2.23}$$

$$\widehat{V}^{\text{Bin}}_{\text{CRR}} = \frac{1}{n_{11}} - \frac{1}{N_1} + \frac{1}{n_{01}} - \frac{1}{N_0} = \frac{n_{10}}{n_{11}N_1} + \frac{n_{00}}{n_{01}N_0}, \tag{2.24}$$

$$\widehat{V}^{\text{Bin}}_{\text{COR}} = \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}. \tag{2.25}$$

Here, the superscript "Bin" is for "Binomial" models. For CRD and log(COR), the estimated variances under independent Binomial models are the same as Neymanian inference assuming constant causal effects. However, this does not hold for log(CRR). One sufficient condition for the equivalence of the variances from Neymanian inference and independent Binomial models is

$$\frac{N_1}{N_0} = \frac{n_{11}}{n_{01}}, \text{ or equivalently, } \widehat{p}_1 = \widehat{p}_0,$$

which essentially assumes the null hypothesis of zero average causal effect.

However, all the conclusions here are based on the constant causal effects assumption which may not be realistic in applications with binary outcomes. Without assuming constant causal effects and by using the new sharp bound for S_τ^2 in (2.7), we obtain different results from independent Binomial models, as shown in Appendix A. One surprising property of the log odds ratio is that the variance estimator under independent Binomial models (2.25) is symmetric with respect to treatment and outcome, which coincides with the randomization-based variance estimator (2.21) assuming $S_\tau^2 = 0$. However, the true variance of $\log(\widehat{\text{COR}})$ over all possible randomizations, (2.20), and the improved variance estimator in Appendix A, (B.4), do not have this symmetry.

2.5.3 Bayesian Inference for General Causal Measures

As shown above, Neymanian randomization inference for nonlinear measures of causal effects involves tedious algebra, and relies on asymptotics under regularity conditions. In contrast, the Bayesian inference for log(CRR) and log(COR) is quite

natural, once we impute all the missing potential outcomes based on their posterior predictive distributions.

For example, under the independent potential outcomes model, we have

$$\begin{aligned}
\log(\text{CRR})|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1} &\sim \log\left(\frac{n_{11}+B_0}{n_{01}+B_1}\right), \\
\log(\text{COR})|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1} &\sim \log\left(\frac{n_{11}+B_0}{N-n_{11}-B_0}\right) - \log\left(\frac{n_{01}+B_1}{N-n_{01}-B_1}\right).
\end{aligned}$$

Also, we can apply the Bayesian sensitivity analysis technique, similar to Section 2.4.3, and obtain

$$\begin{aligned}
\log(\text{CRR})|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1} &\sim \log\left(\frac{n_{11}+B_{01}+B_{00}}{n_{01}+B_{11}+B_{10}}\right), \\
\log(\text{COR})|\boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1} &\sim \log\left(\frac{n_{11}+B_{01}+B_{00}}{N-n_{11}-B_{01}-B_{00}}\right) \\
&\quad - \log\left(\frac{n_{01}+B_{11}-B_{10}}{N-n_{01}-B_{11}-B_{10}}\right).
\end{aligned}$$

The posterior distributions of these causal measures can then be approximated by Monte Carlo.

2.6 Simulation Studies

In order to compare the finite sample properties of Neyman's original method, the modified Neyman's method, and the Bayesian method assuming independent potential outcomes, we conduct two sets of simulation studies with independent and positively associated potential outcomes. The first set listed as Cases 1–5 in Table 2.5 represent independent potential outcomes, while those listed as Cases 6–12 represent

positively associated potential outcomes. Table 2.5 also shows the marginal variances, correlations of the potential outcomes, and causal measures for each set of potential outcomes. To save space in the main text, we present only the results for CRD and log(COR). The results for log(CRR) and the simulation studies for negatively associated potential outcomes are discussed in Appendix B.

Table 2.5: "Science table" for the simulation studies

Case	N_{11}	N_{10}	N_{01}	N_{00}	S_1^2	S_0^2	S_{10}	S_τ^2	τ	log(CRR)	log(COR)
1	50	50	50	50	0.251	0.251	0.000	0.503	0.000	0.000	0.000
2	30	70	30	70	0.251	0.211	0.000	0.462	0.200	0.511	0.847
3	30	90	20	60	0.241	0.188	0.000	0.430	0.350	0.875	1.504
4	80	20	80	20	0.251	0.161	0.000	0.412	-0.300	-0.470	-1.386
5	60	20	90	30	0.241	0.188	0.000	0.430	-0.350	-0.629	-1.504
6	60	40	40	60	0.251	0.251	0.050	0.402	0.000	0.000	0.000
7	50	50	30	70	0.251	0.241	0.050	0.392	0.100	0.223	0.405
8	50	70	30	50	0.241	0.241	0.010	0.462	0.200	0.405	0.811
9	40	110	10	40	0.188	0.188	0.013	0.352	0.500	1.099	2.197
10	70	30	50	50	0.251	0.241	0.050	0.392	-0.100	-0.182	-0.405
11	50	30	70	50	0.241	0.241	0.010	0.462	-0.200	-0.405	-0.811
12	30	10	110	50	0.161	0.211	0.010	0.352	-0.500	-1.253	-2.234

For given potential outcomes, we draw, repeatedly and independently, the treatment assignment vectors 5000 times, obtain the observed outcomes, and then apply three methods: Neymanian inference assuming constant treatment effects, improved Neymanian inference, and Bayesian inference assuming independent potential outcomes. The improved Neymanian inference means using the improved variance estimator (2.8) for CRD, and bias-corrected estimators (B.1) and (B.3) and improved variance estimators (B.2) and (B.4) for nonlinear causal measures log(CRR) and log(COR). Comparison of these methods are summarized in Figure 2.2, with average biases, average lengths of the 95% confidence/credible intervals, and the coverage probabilities.

First, the bias-corrected estimators for nonlinear causal measure log(COR) do have smaller biases than the original Neymanian estimators and Bayes estimators in most cases. Second, the confidence intervals from the modified Neyman's method are narrower than Neyman's original method, while still maintaining correct coverage properties. They indeed improve Neyman's original method. Third, the average widths of the Bayesian credible intervals are much narrower than the original and modified Neyman's method. Moreover, when the potential outcomes are independent, the Bayesian credible intervals have correct frequentists' coverage property. When the potential outcomes are positively associated and the average causal effect is small, the Bayesian credible intervals slightly under-cover the truth. The results in Appendix B show that when the potential outcomes are negatively associated, even the narrowest Bayesian credible intervals over-cover the true causal measures, and the Neymanian intervals and their modification are too "conservative."

It is also interesting to investigate the frequency coverage property of the improved variance estimator (2.8) for CRD under the sharp null. Table 2.6 compares the frequency properties of Neymanian variance estimator (2.6) and its improved version (2.8), with moderate sample size $N = 30$ and different choices of N_{11} and N_{00} such that $N_1 = N_0 = 15$. Except for the case with $N_{11} = N_{00} = 15$, the improved variance estimators have shorter confidence intervals but preserve the same coverage rates. Under the sharp null, with sample size $N = 30$, the sampling variance of $\widehat{\tau}$ is maximized at $N_{11} = N_{00} = 15$, and in this case the adjusting term $|\widehat{\tau}|(1-|\widehat{\tau}|)/(N-1)$ has the wildest behavior. Therefore, in practice, if we have small sample sizes and observe that $\widehat{p}_1 \approx \widehat{p}_0 \approx 0.5$, the improved variance estimator may hurt our inference.

In all other situations, we suggest using the improved variance estimator.

Table 2.6: Neymanian and its improved variance estimators for CRD, (2.6) and (2.8), under the sharp null hypothesis with $N = 30$ and $N_{10} = N_{01} = 0$

N_{11}	N_{00}	Length (Using Neyman's estimator)	Coverage (Using Neyman's estimator)	Length^c (Using corrected estimator)	Coverage^c (Using corrected estimator)
20	10	0.686	0.951	0.644	0.951
25	5	0.542	0.959	0.491	0.959
15	15	0.728	0.971	0.683	0.858
12	18	0.713	0.942	0.672	0.942
8	22	0.633	0.96	0.601	0.96

2.7 Application to a Randomized Controlled Trial

This example is taken from Bissler et al. (2013), where the authors compare the rate of adverse events in the treatment group versus the control group. The adverse event naspharyngitis occurred in 19 out of 79 subjects in the treatment group with everolimus, and it occurred in 12 out of 39 subjects in the control group. Therefore, the 2×2 table representing the observed data has cell counts $(n_{11}, n_{10}, n_{01}, n_{00}) = (19, 60, 12, 27)$. In Figure 2.3(a), we show the results for three causal measures using Neymanian inference, modified Neymanian inference, and Bayesian posterior inference assuming independent potential outcomes. The results match with those in our simulation studies in the sense that the bias-corrected estimators are slightly different from the original estimators, and the Bayes posterior credible intervals are much narrower than the confidence intervals from Neymanian inference. However, in this particular example, all intervals cover zero.

Since the independence assumption between potential outcomes has a strong impact on the Bayesian inference for the finite population causal measures, we conduct

a sensitivity analysis as proposed in Section 2.4.3 by varying $\log(\gamma)$ within $[-2, 4]$, and obtain Bayesian credible intervals for the causal measures at each $\log(\gamma)$. Figure 2.3(b) shows the sensitivity analysis for CRD and log(COR), with similar patterns log(CRR) as shown in the Supplementary Materials. Finally, the widths of the credible interval depend on $\log(\gamma)$; however, in the example, even the widest credible intervals are narrower than the "conservative" Neymanian confidence intervals.

2.8 Discussion

In this paper, we have discussed causal inference of completely randomized treatment-control studies with binary outcomes under the potential outcomes model. We first made a connection between the Fisher randomization test (Fisher, 1935a) and Fisher's exact test (Fisher, 1935b) for binary outcomes, and proposed a procedure which uniformly dominates Neyman (1923)'s method. Although widely used in clinical practice, statistical inference for general nonlinear causal measures are based on the assumption of independent Binomial models, which is not justified by randomization. Based on randomization, our asymptotic analysis shows that the widely used variance estimators are either incorrect or inefficient, unless the null hypothesis of zero average causal effect is true.

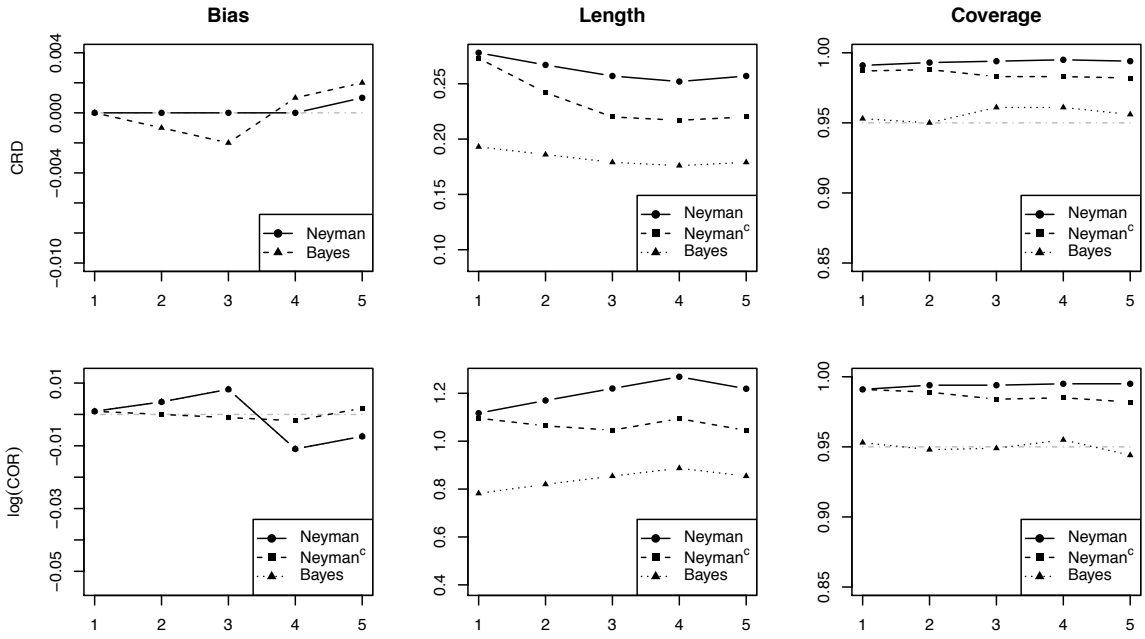
Ding (2014) shows that the Neyman's test for zero average causal effect tends to be more powerful than Fisher's test for zero individual causal effect for many realistic cases including balanced designs. Our variance estimator (2.8) further improves Neyman's test. Our new result is not contradictory to the classical result that Fisher's exact test is the uniformly most powerful unbiased test for equal probability of two

independent Binomials (Lehmann and Romano, 2006). Both the Fisherian and Neymanian approaches are derived under the potential outcomes model, but the classical result for Fisher's exact test is derived under the independent Binomial models.

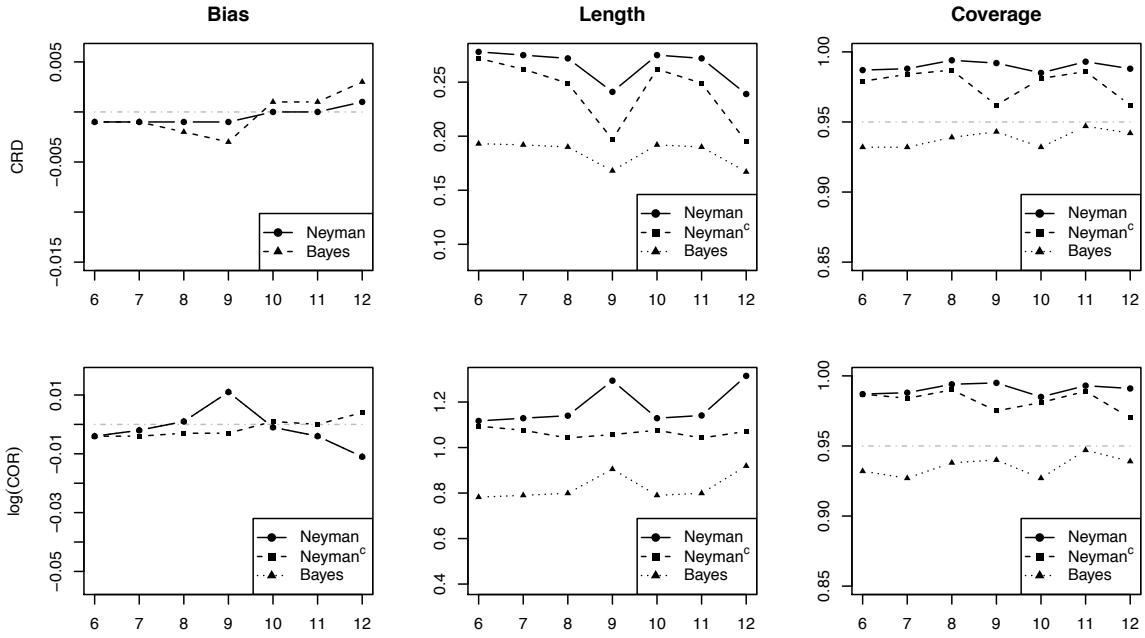
Traditionally, the variance formulae in (2.21) and (2.25) for log(COR) have been used in both experimental and observational studies (including both prospective and retrospective observational studies). Due to the symmetry of the variance formulae in (2.21) and (2.25) with respect to the treatment and outcome, researchers found that statistical inference of the log odds ratio measure is invariant to the sampling scheme (experimental study, prospective or retrospective observational studies), which was regarded as a celebrated and also mysterious feature of the log of the odds ratio. As a pioneer in epidemiology and biostatistics, Cornfield (1959) said that "there is a distinction seems undeniable, but its exact nature is elusive," when he was discussing experimental and observational studies. However, randomized experiments are fundamentally different from observational studies, and especially different from retrospective observational studies. Under the potential outcomes model with the potential outcomes treated as fixed quantities, the randomness of the observed outcomes comes only from the physical randomization in experiments. Therefore, the treatment and the observed outcome are asymmetric unless the sharp null is true, and the variance in (2.20) for log(COR) and its estimator in (B.4) reflect the asymmetric nature explicitly. In a recent comment on Cornfield (1959), Rubin (2012) suggested revealing the hidden nature of different studies using potential outcomes. Indeed, our results verify Rubin (2012)'s conjecture.

In order to reveal the importance of the correlation between potential outcomes

and the intrinsic lack of additivity for binary outcomes, we focus our discussion on two by two tables from completely randomized experiments. The same idea can be also applied to observational studies as long as the ignorability assumption holds. We can either stratify on the observed covariates or propensity scores (Rosenbaum and Rubin, 1983), and then within each strata the data can be approximately viewed as generated from randomized experiments (Rosenbaum, 2002). The findings of this paper can be generalized in many ways. For instance, we can discuss Neymanian randomization inference for full factorial or fractional factorial designs with binary outcomes, since the current discussion Dasgupta et al. (2015) is restricted to continuous outcomes. It will also be interesting to discuss causal inference under the potential outcomes model for general outcomes (categorical data, counts, survival times, etc.). These topics are our ongoing or future research projects.

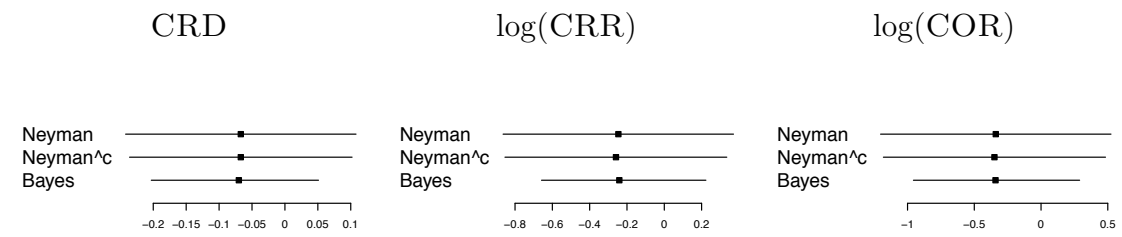


(a) Independent Potential Outcomes: Cases 1 to 5 with the x-axis denoting the case numbers

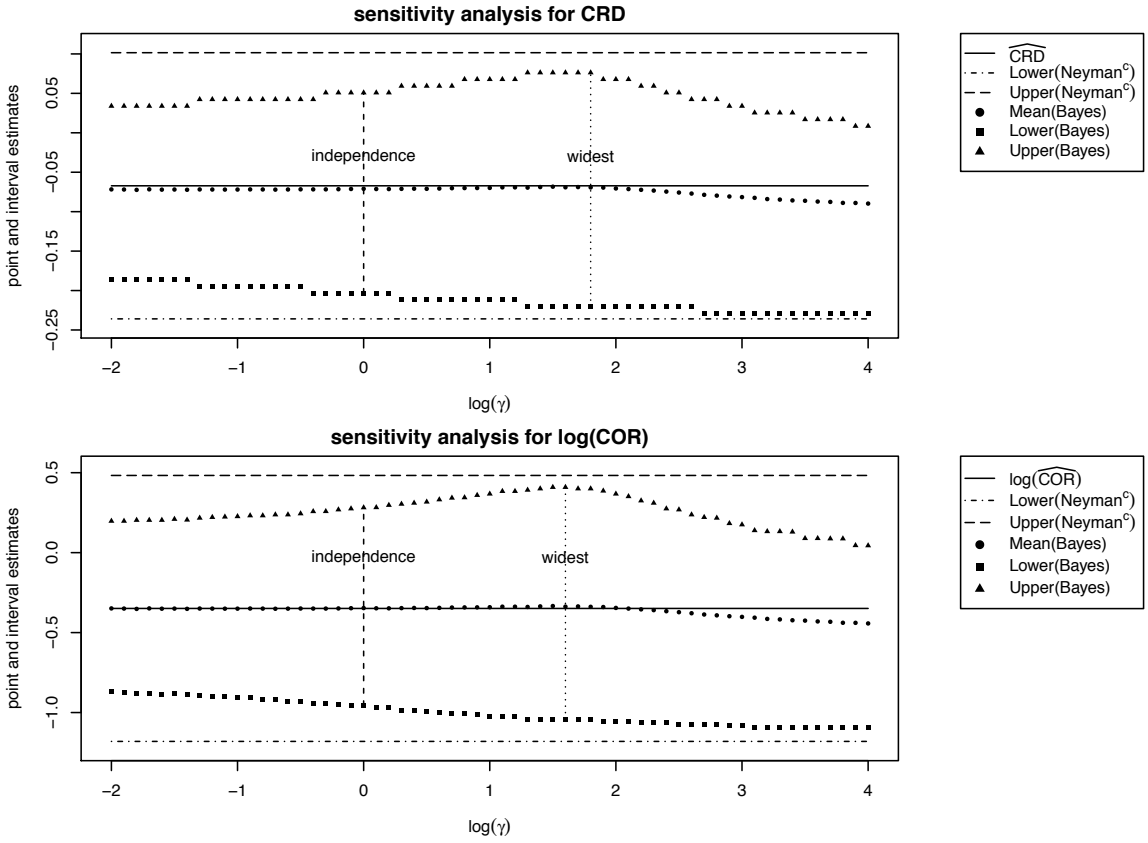


(b) Positively Associated Potential Outcomes: Cases 6 to 13 with the x-axis denoting the case number

Figure 2.2: Simulation Studies. Each subfigure is a 2×3 matrix summarizing results for 2 parameters and 3 properties. Note that "Neyman" and "Bayes" are indistinguishable for biases of log(COR).



(a) Inference for the Causal Measures. We apply Neymanian, improved Neymanian (Neyman^c above) and Bayesian approaches with the segments representing the 95% confidence/credible intervals and centers illustrating the point estimators.



(b) Bayesian Sensitivity Analysis for CRD and log(COR). The intervals named "independence" are the 95% posterior credible intervals under independence of the potential outcomes, and the intervals named "widest" are the widest 95% credible intervals over the ranges of the sensitivity parameters.

Figure 2.3: A Randomized Experiments with Observed Data $(n_{11}, n_{10}, n_{01}, n_{00}) = (19, 60, 12, 27)$.

Chapter 3

Treatment Effect Heterogeneity in Randomized Experiments

3.1 Introduction

Researchers and practitioners are increasingly interested in whether and how treatment effects vary in randomized experiments. For example, we might be interested in assessing the effect of scaling up a promising intervention evaluated on a limited subpopulation (O'Muircheartaigh and Hedges, 2014). If we only use observed characteristics to predict the program's effectiveness on the new population, we might wonder if we are missing critical unexplained variation, which could undermine our generalization. Similarly, we might want to determine whether different theoretical models are sufficiently rich to explain observed behavior in a randomized experiment. For instance, is a simple model of constant treatment effects within subgroups sufficient to explain observed labor supply behavior in welfare reform experiments? Or is

there meaningful unexplained variation, as predicted by labor supply theory (Bitler et al., 2010)?

Unfortunately, assessing such variation is difficult. In general, researchers investigating specific types of idiosyncratic variation must therefore rely on strong modeling assumptions to draw meaningful conclusions from the data (Cox, 1984; Heckman et al., 1997). Instead, we propose a randomization-based inferential framework for treatment effect variation that does not strongly rely on the modeling assumptions. The statistical procedures are justified by the physical randomization itself (Neyman, 1923; Fisher, 1935a; Rosenbaum, 2002).

Of course, all treatment effects vary in practice, especially in the social sciences. The key question is whether the unexplained variation is sufficiently large to be of substantive importance. We first offer a randomization-based test for the presence of idiosyncratic treatment effect variation that cannot be explained by the observed covariates, and then propose a measure of the fraction of the systematic treatment effect variation that can be explained by the observed covariates. Applying our method to the Head Start Impact Study, a large-scale randomized evaluation of a Federal preschool program, gives meaningful practical conclusions.

3.2 Treatment Effect Decomposition

Assume that we have n units in a completely randomized experiment with n_1 units receiving treatment and n_0 units receiving control. For unit i, let T_i denote the treatment indicator with 1 for treatment and 0 for control. We use the potential outcomes framework (Neyman, 1923) to define causal effects. Under the Stable Unit

Treatment Value Assumption (Rubin, 1980) that there is only one version of the treatment and no interference among units, we define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes under treatment and control. The observed outcome, $Y_i^{\text{obs}} = T_i Y_i(1) + (1 - T_i)Y_i(0)$, is a deterministic function of the treatment and potential outcomes. Finally, let $X_i \in \mathbb{R}^K$ denote the vector of pretreatment covariates, with the constant one as its first component. Under the potential outcomes framework, $\{Y_i(1), Y_i(0), X_i\}_{i=1}^n$ are all fixed numbers; the randomness comes solely from T_i, the physical randomization itself.

We define the individual treatment effect as $\tau_i = Y_i(1) - Y_i(0)$, which has the following decomposition:

$$\tau_i = Y_i(1) - Y_i(0) = X_i'\beta + \varepsilon_i = \delta_i + \varepsilon_i \quad (i = 1, \ldots, n). \tag{3.1}$$

If we could observe τ_i for each individual, β would be the usual linear regression coefficients of τ_i on X_i. We can then define $\delta_i = X_i'\beta$ as the systematic treatment effect variation explained by the observed covariates X_i, and ε_i as the idiosyncratic treatment effect variation unexplained by X_i (Djebbari and Smith, 2008).

Even though we cannot observe τ_i, we can still leverage these regression concepts. Define

$$S_{xx} = \sum_{i=1}^n X_i X_i'/n, \quad S_{x\varepsilon} = \sum_{i=1}^n X_i \varepsilon_i/n, \quad S_{x\tau} = \sum_{i=1}^n X_i \tau_i/n.$$

First, we assume that $\det(S_{xx}) > 0$, which is analogous to the usual full rank assumption in any linear model. Second, we assume that $S_{x\varepsilon} = 0$, i.e., that ε_i and X_i have covariance zero. This assumption will hold automatically, if we re-define (β, ε_i) to

be $(\beta + S_{xx}^{-1}S_{x\varepsilon}, \varepsilon_i - X_i'S_{xx}^{-1}S_{x\varepsilon})$. Therefore, following the agnostic regression framework (Lin, 2013), we view the systematic component, $\delta_i = X_i'\beta$, as a linear projection of τ_i onto the linear space spanned by X_i; the idiosyncratic treatment effect ε_i is the corresponding residual.

3.3 Statistical Inference of Treatment Effect Variation

3.3.1 Randomization Inference

We now derive the randomization inference-based estimator of β. Define

$$\widehat{S}_{x1} = \sum_{i=1}^{n} T_i X_i Y_i^{\text{obs}}/n_1, \quad \widehat{S}_{x0} = \sum_{i=1}^{n}(1-T_i)X_i Y_i^{\text{obs}}/n_0$$

as the sample covariances between X_i and Y_i^{obs} under treatment and control, and let $\mathcal{S}(V) = \sum_{i=1}^{n}(V_i - \bar{V})(V_i - \bar{V})'/(n-1)$ be the covariance operator, where $\bar{V} = \sum_{i=1}^{n} V_i/n$. The physical randomization of T_i's justifies the following theorem.

Theorem 12. Under model (3.1), an unbiased estimator for β is

$$\widehat{\beta}_{\text{RI}} = S_{xx}^{-1}(\widehat{S}_{x1} - \widehat{S}_{x0}),$$

with covariance over all possible randomizations as $\text{cov}(\widehat{\beta}_{\text{RI}}) = S_{xx}^{-1}\text{cov}(\widehat{S}_{x1} - \widehat{S}_{x0})S_{xx}^{-1}$,

where

$$\text{cov}(\widehat{S}_{x1} - \widehat{S}_{x0}) = \frac{\mathcal{S}\{XY(1)\}}{n_1} + \frac{\mathcal{S}\{XY(0)\}}{n_0} - \frac{\mathcal{S}(X\tau)}{n}. \tag{3.2}$$

Therefore, $\widehat{\beta}_{\text{RI}}$ is an unbiased estimator of the systematic treatment effect variation over X_i. Moreover, the covariance formula (3.2) generalizes Neyman (1923)'s classical result for the average treatment effect, reducing to Neyman's formula if $X_i = 1$ for all units. As with Neyman's original formula, we can assume that $\varepsilon_i = 0$ for all units in order to obtain a "conservative" estimate of $\text{cov}(\widehat{S}_{x1} - \widehat{S}_{x0})$. Note that S_{xx} is known for the population, rather than estimated.

Thus far, the role of covariates has been to model the treatment effect alone; $\widehat{\beta}_{\text{RI}}$ is unbiased for β regardless of the marginal distributions of potential outcomes. In general, we also want to use covariates to reduce sampling variability. Let $W_i \in \mathbb{R}^J$ denote a vector of pretreatment covariates, with the constant vector as its first component. Since X_i and W_i have different roles in estimation they may also contain different sets of covariates, though, in practice, X is likely to be a subset of W.

Following the covariate adjustment approach in survey sampling (Cochran, 1977), we can therefore obtain a model-assisted estimator for β that uses W to reduce sampling variability. To see this, we need several definitions. Define $\bar{W}$ and S_{ww} as the population mean and covariance of W, with $\det(S_{ww}) > 0$; define $\bar{W}_1$ and $\bar{W}_0$ as the sample means under treatment and control; define $\widehat{B}_t$ as the regression coefficient of $Y^{\text{obs}}X$ on W for treatment arm t; and define $e_i(t)$ be the residual of the regression of $Y_i(t)X_i$ on W_i, with $\Delta_i = e_i(1) - e_i(0)$. The model-assisted estimator for S_{tx} is

then

$$\widehat{S}^w_{tx} = \widehat{S}_{tx} + \widehat{B}'_t(\bar{W} - \bar{W}_t)$$

for treatment t. More generally, we have the following theorem.

Theorem 13. The model-assisted estimator

$$\widehat{\beta}^w_{\text{RI}} = S^{-1}_{xx}(\widehat{S}^w_{1x} - \widehat{S}^w_{0x})$$

has asymptotic covariance matrix $S^{-1}_{xx}\text{cov}(\widehat{S}^w_{1x} - \widehat{S}^w_{0x})S^{-1}_{xx}$, where

$$\text{cov}(\widehat{S}^w_{1x} - \widehat{S}^w_{0x}) = \frac{\mathcal{S}\{e(1)\}}{n_1} + \frac{\mathcal{S}\{e(0)\}}{n_0} - \frac{\mathcal{S}(\Delta)}{n}.$$

The resulting estimator, $\widehat{\beta}^w_{\text{RI}}$, therefore uses covariates both to estimate treatment effect variation and to reduce sampling variability. Asymptotically, as long as W is predictive of the marginal potential outcomes, then the model-assisted estimator will improve precision over the unassisted estimator. Finally, when X_i and W_i are matrices of dummy variables generated by the same categorical covariates, this estimator reduces to the post-stratification estimator (Miratrix et al., 2013).

3.3.2 Regression with Treatment-Covariate Interactions

We now use the results from the randomization inference to better understand the familiar case of linear regression with treatment-covariate interactions (Berrington de

González and Cox, 2007; Crump et al., 2008):

$$Y_i^{\text{obs}} = X_i'\gamma + T_i X_i'\beta + \varepsilon_i \qquad (i = 1, \ldots, n), \tag{3.3}$$

where ε_i is implicitly assumed to induce the randomness. Written in the usual matrix form, it is difficult to compare the regression-based estimator, $\widehat{\beta}_{\text{OLS}}$, to the randomization inference-based estimator, $\widehat{\beta}_{\text{RI}}$. We therefore re-write $\widehat{\beta}_{\text{OLS}}$ using the following theorem.

Theorem 14. The ordinary least squares estimator for β in equation (3.1) can be written as:

$$\widehat{\beta}_{\text{OLS}} = \widehat{S}_{xx,1}^{-1}\widehat{S}_{x1} - \widehat{S}_{xx,0}^{-1}\widehat{S}_{x0},$$

where $\widehat{S}_{xx,t}$ is the sample covariance matrix of X_i under treatment arm t. $\widehat{\beta}_{\text{OLS}}$ is a consistent estimator for β.

The differences betwen $\widehat{\beta}_{\text{OLS}}$ and $\widehat{\beta}_{\text{RI}}^{w}$ are minor. First, while the point estimates differ slightly—$\widehat{\beta}_{\text{RI}}$ is unbiased while $\widehat{\beta}_{\text{OLS}}$ is consistent—they are asymptotically equivalent. Second, unlike the variance of $\widehat{\beta}_{\text{RI}}^{w}$, the variance of $\widehat{\beta}_{\text{OLS}}$ has a complex form in finite samples. The usual OLS standard errors are not appropriate in this case, and we must instead use Huber–White standard errors (Lin, 2013). Even so, the resulting variance estimates for $\widehat{\beta}_{\text{RI}}^{w}$ and $\widehat{\beta}_{\text{OLS}}$ are asymptotically equivalent.

Therefore, even though $\widehat{\beta}_{\text{RI}}^{w}$ and $\widehat{\beta}_{\text{OLS}}$ are not identical, the close connections between the two suggest that, in practice, $\widehat{\beta}_{\text{OLS}}$ is justified by the randomization, just as Fisher (1935a) suggested nearly a century ago.

3.4 Testing and Decomposing Treatment Effect Variation

3.4.1 Testing Systematic Treatment Effect Variation

Under the assumption of no idiosyncratic treatment effect variation, i.e., $\varepsilon_i = 0$, the covariance of $\widehat{\beta}_{\mathrm{RI}}$ reduces to $\mathrm{cov}(\widehat{\beta}_{RI}) = S_{xx}^{-1}\ \mathrm{cov}(\widehat{S}_{1x} - \widehat{S}_{0x})\ S_{xx}^{-1}$, where

$$\mathrm{cov}(\widehat{S}_{1x} - \widehat{S}_{0x}) = \frac{\mathcal{S}\{Y_i(1)X_i\}}{n_1} + \frac{\mathcal{S}\{Y_i(0)X_i\}}{n_0} - \frac{\mathcal{S}\{X_iX_i'\beta\}}{n}.$$

We can estimate $\mathcal{S}\{Y_i(t)X_i\}$ by the sample covariance of $\{Y_i^{\mathrm{obs}}X_i : T_i = t\}$, and $\mathcal{S}\{X_iX_i'\beta\}$ by the sample covariance of $\{X_iX_i'\widehat{\beta}_{\mathrm{RI}} : i = 1, \ldots, n\}$. Therefore, we can estimate the sampling covariance of $\widehat{\beta}_{\mathrm{RI}}$, and then construct a confidence region for β.

The null hypothesis of no treatment effect variation explained by the observed covariates can be characterized by the null hypothesis

$$H_0(X) : \beta_1 = 0,$$

where β_1 contains all the components of β except the first one corresponding to the intercept. We can use Wald-type test for the null hypothesis $H_0(X)$, because we already have a point estimate and confidence region for β.

3.4.2 Testing Idiosyncratic Treatment Effect Variation

In many practical problems, we are interested in whether there is any meaningful idiosyncratic treatment effect variation beyond that can be explained away by the observed covariates X (Ding et al., 2015). Statistically, this yields the following hypothesis test:

$$H_0(S) : \tau_i = Y_i(1) - Y_i(0) = X_i'\beta \text{ for some } \beta \quad (i = 1, \ldots, n).$$

Intuitively, we can use the shifted Kolmogorov–Smirnov statistic

$$t_{SKS} = \sup_y |\widehat{F}_1(y) - \widehat{F}_0(y)|$$

to capture the deviation from the null hypothesis $H_0(S)$, where $\widehat{F}_1$ and $\widehat{F}_0$ are empirical cumulative distributions of $\{Y_i^{\text{obs}} - X_i'\widehat{\beta}_{\text{RI}} : T_i = 1\}$ and $\{Y_i^{\text{obs}} : T_i = 0\}$, respectively, although many other test statistics are possible. Unfortunately, the presence of the nuisance parameter β in $H_0(S)$ complicates calculations of the null distribution of the test statistic. This nuisance parameter problem can be bypassed by maximizing the Fisher randomization test p-values over a confidence region of β with a small adjustment. To be more specific, with a known β we can impute all missing potential outcomes based on the known individual treatment effect $\tau_i = X_i'\beta$, and then obtain the p-value $p(\beta)$ from the Fisher randomization test. We can then calculate the p-value against the null hypothesis $H_0(S)$ by

$$p = \sup_{\beta \in \text{CR}_\gamma} p(\beta) + \gamma,$$

where CR_γ is a confidence region for β with coverage rate $(1-\gamma)$, and γ is typically very small compared to α. The resulting p-value is valid in the sense of being stochastically dominated by a uniform random variable (Berger and Boos, 1994). Ding et al. (2015) give extensive discussion on this approach.

3.4.3 Variance on the Average Treatment Effect Estimate

Decomposing treatment effect variation into systematic and idiosyncratic components is important even when we are only interested in the Average Treatment Effect

$$\bar{\tau} = \frac{1}{n}\sum_{i=1}^{n} \tau_i.$$

To see this, we begin with Neyman (1923), who proposed the difference-in-means statistic, $\widehat{\tau} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}}$, which is an unbiased estimator $\bar{\tau}$. Its sampling variance is

$$\text{var}(\widehat{\tau}) = \frac{S_{11}}{n_1} + \frac{S_{00}}{n_0} - \frac{S_{\tau\tau}}{n},$$

which depends on S_{11}, S_{00}, and $S_{\tau\tau}$, the finite population variances of $Y_i(1), Y_i(0)$ and τ_i, respectively. While S_{11} and S_{00} are estimable quantities, $S_{\tau\tau}$ depends on the correlation of potential outcomes and is unidentified.

There are a range of variance estimators that circumvent this unidentifiability. In a classic result, Neyman (1923) proposed a lower bound for $\text{var}(\widehat{\tau})$ under the assumption of a constant treatment effect, $S_{\tau\tau} = 0$. More recently, Aronow et al. (2014) build on an idea from Heckman et al. (1997), proposing to bound $S_{\tau\tau}$ rather than to assume $S_{\tau\tau} = 0$. The authors use Fréchet–Hoeffding bounds (Hoeffding, 1941; Fréchet, 1951),

which bound a joint distribution via its marginal distributions, and show that these bounds are sharp for $S_{\tau\tau}$ and, therefore, for $\text{var}(\widehat{\tau})$.

We now use the results from Section 3.3.1 to derive lower-variance bounds than those in Aronow et al. (2014). First, define

$$S_{\delta\delta} = \sum_{i=1}^{n}(\delta_i - \bar{\tau})^2/n, \quad S_{\varepsilon\varepsilon} = \sum_{i=1}^{n}\varepsilon_i^2/n,$$

with δ_i and ε_i as in equation (3.1). Then $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$. For $t = 0$ and 1, we further define $Y_i(t) - X_i'\gamma_t$ as the residual of linear projection of the potential outcomes $Y_i(t)$ onto the linear space spanned by X_i, where γ_t is the regression coefficient. Let $\widetilde{F}_1(y)$ and $\widetilde{F}_0(y)$ be the empirical cumulative distribution functions of $\{Y_i(1) - X_i'\gamma_1 : i = 1, \ldots, n\}$ and $\{Y_i(0) - X_i'\gamma_0 : 1, \ldots, n\}$, respectively.

We have the following theorem based on Fréchet–Hoeffding bounds (Hoeffding, 1941; Fréchet, 1951).

Theorem 15. *$S_{\tau\tau}$ has the following sharp bounds:*

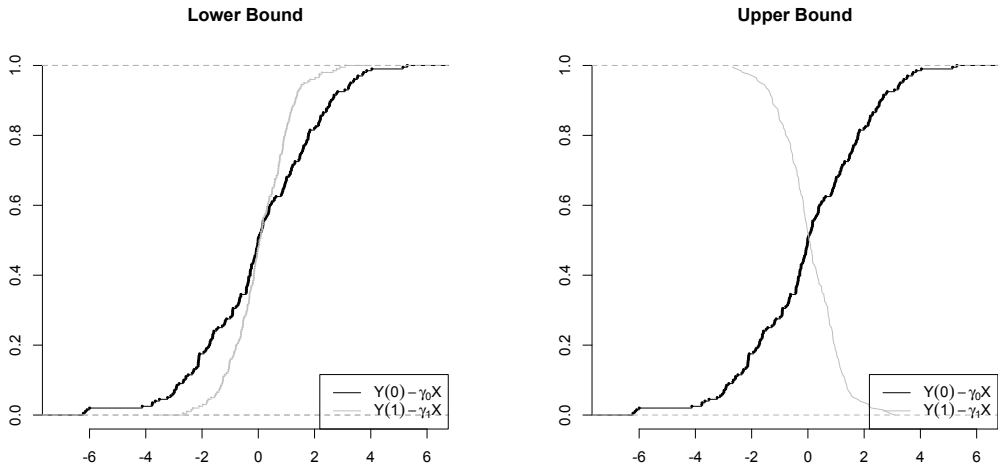
$$S_{\delta\delta} + \underline{S}_{\varepsilon\varepsilon} \leq S_{\tau\tau} \leq S_{\delta\delta} + \overline{S}_{\varepsilon\varepsilon},$$

where

$$\underline{S}_{\varepsilon\varepsilon} = \int_0^1 \{\widetilde{F}_1^{-1}(u) - \widetilde{F}_0^{-1}(u)\}^2 du, \quad \overline{S}_{\varepsilon\varepsilon} = \int_0^1 \{\widetilde{F}_1^{-1}(u) - \widetilde{F}_0^{-1}(1-u)\}^2 du.$$

From the proof of Theorem 15, we can see that when $\{Y_i(1) - X_i'\gamma_1 : i = 1, \ldots, n\}$ and $\{Y_i(0) - X_i'\gamma_0 : 1, \ldots, n\}$ have the same ranks as in Figure 3.1(a), $\underline{S}_{\varepsilon\varepsilon}$ attains its

lower bound; when they have opposite ranks as in Figure 3.1(b), $\underline{S}_{\varepsilon\varepsilon}$ attains its upper bound.



(a) lower bound (b) upper bound

Figure 3.1: Frechét–Hoeffding Bounds.

We can obtain consistent estimators for each quantity; $S_{\delta\delta}$ can be estimated by the sample variance of $X_i'\widehat{\beta}$, and $\widetilde{F}_{e1}(y)$ and $\widetilde{F}_{e0}(y)$ can be estimated by $\widehat{F}_1$ and $\widehat{F}_0$, the empirical cumulative distribution functions of $\{Y_i^{\text{obs}} - X_i'\widehat{\gamma}_1 : T_i = 1\}$ and $\{Y_i^{\text{obs}} - X_i'\widehat{\gamma}_0 : T_i = 0\}$ based on the residuals from least squares, respectively. Overall, so long as $S_{\delta\delta} > 0$, this yields strictly tighter bounds on $\text{var}(\widehat{\tau})$ than the corresponding bounds that do not incorporate covariate information.

3.4.4 Bounding the Fraction of Treatment Effect Variation Explained by Covariates

A natural question in practice is the relative size of $S_{\delta\delta}$ and $S_{\varepsilon\varepsilon}$. Continuing the regression analogy, we desire an R^2-like measure for the proportion of total treatment effect variation explained by the systematic component:

$$R^2_\tau = \frac{S_{\delta\delta}}{S_{\tau\tau}} = \frac{S_{\delta\delta}}{S_{\delta\delta} + S_{\varepsilon\varepsilon}},$$

which is the ratio between the variances of δ_i and τ_i. As above, this measure is not identifiable since ε_i depends on the joint distribution of the potential outcomes. However, applying Theorem 15, we obtain the following bounds on R^2_τ:

$$\frac{S_{\delta\delta}}{S_{\delta\delta} + \overline{S}_{\varepsilon\varepsilon}} \leq R^2_\tau \leq \frac{S_{\delta\delta}}{S_{\delta\delta} + \underline{S}_{\varepsilon\varepsilon}}.$$

3.5 The Head Start Impact Study

Head Start is the largest Federal preschool program today, serving around $900,000$ children each year at a cost of roughly $8 billion. The Head Start Impact Study (HSIS) is the first major randomized evaluation of the program. The published report found that, on average, providing children and their families with the opportunity to enroll in Head Start improved childrens key cognitive and social-emotional outcomes. The report also included average treatment effect estimates for a variety of subgroups of interest, though there is only significant impact variation across a small number of the reported, pre-treatment covariates. After these findings were released, many

researchers argued that the reported topline results masked critical variation in program impacts. All of these approaches, however, estimate treatment effect variation by relying on a specific set of models, such as quantile or hierarchical regression. We investigate this question by focusing on the Peabody Picture Vocabulary Test (PPVT), a widely used measure of cognitive ability in early childhood. We also utilize a rich set of pre-treatment covariates, including pre-test score, child's age, child's race, mother's education level, and mother's marital status. For the sake of exposition, we restrict our analysis to a complete-case subset of HSIS, with $N_1 = 2,238$ in the treatment group and $N_0 = 1,348$ in the control group. Ding et al. (2015) perform randomization tests to detect treatment effect variation that cannot be explained away by the observed covariates. Here we focus on measuring the fraction of the systematic treatment effect variation.

As shown in Table 3.1, we find that among the top three covariates that are most predictive of treatment effect variation, Dual Language Learner status has the largest upper bound of the R^2 measure. By itself, Dual Language Learner status alone can explain away up to 32.4% of the overall treatment effect variation, and the top three covariates in total can explain away up to 45.3% of the overall treatment effect variation.

Table 3.1: R^2 for important covariates.

covariates	Lower R^2_τ	Upper R^2_τ
dual language learner	0.001	0.324
academic skills	0.000	0.096
age	0.000	0.017
all above	0.002	0.453

Intuitively, Dual Language Learner status is very predictive to the control po-

tential outcome. Therefore, we have a conjecture that the treatment effect varies by the control potential outcome itself. Fortunately, this conjecture can be empirically tested. In particular, in HSIS we have the following inequality about sample variances:

$$\widehat{\text{var}}\{Y_i^{\text{obs}} - X_i'\widehat{\beta}_{\text{RI}} : Z_i = 1\} \leq \widehat{\text{var}}\{Y_i^{\text{obs}} : Z_i = 0\}.$$

Transforming this back to the potential outcomes, we have

$$\text{var}\{Y_i(1) - X_i'\beta\} \leq \text{var}\{Y_i(0)\}.$$

Simple algebra reduces the above inequality to

$$\text{cov}\{Y_i(0), \varepsilon_i\} \leq 0.$$

Therefore, the individual treatment effects are negatively associated with the control potential outcomes, implying that the treatment effects are larger for smaller values of $Y_i(0)$ even after controlling for covariates. As a result, we find that the treatment effect not only varies by important covariates, e.g., Dual Language Learner status, but also varies by the control potential outcome itself.

3.6 Generalization to Accommodate Nonlinear Treatment Effect

In our previous discussion, we model the systematic treatment effect variation as a linear combination of the observed covariates, which may reasonably approximate the true effect variation surface by incorporating polynomials of the covariates. However, the potential outcomes framework does not restrict us to the linear model. Assume that we have

$$\tau_i = Y_i(1) - Y_i(0) = f(X_i; \beta) + \varepsilon_i, \quad (i = 1, \ldots, n), \tag{3.4}$$

where $f(X_i; \beta) \equiv f_i$ is a smooth nonlinear function with first derivative $\nabla f(X_i; \beta) \equiv \nabla f_i$ with respect to the unknown parameter β, and ε_i is the idiosyncratic treatment effect variation satisfying the condition $\sum_{i=1}^{n} \varepsilon_i \nabla f_i / n = 0$. Define

$$m(T_i, Y_i^{\text{obs}}; \beta) = \frac{T_i}{n_1/n} Y_i^{\text{obs}} \nabla f_i - \frac{1 - T_i}{n_0/n} Y_i^{\text{obs}} \nabla f_i - f_i \nabla f_i.$$

Because $E\{m(T_i, Y_i^{\text{obs}}; \beta)\} = 0$ is an unbiased estimating equation for β for all i, we can solve $\widehat{\beta}$ from

$$n^{-1} \sum_{i=1}^{n} m(T_i, Y_i^{\text{obs}}; \widehat{\beta}) = 0.$$

Analogous to Theorem 12, we have the following theorem.

Theorem 16. Under model (3.4), $\widehat{\beta}$ is consistent for β and asymptotically normal

with variance

$$S_{\nabla f \nabla f}\left[\frac{\mathcal{S}\{Y(1)\nabla f\}}{n_1}+\frac{\mathcal{S}\{Y(0)\nabla f\}}{n_0}-\frac{\mathcal{S}(\tau\nabla f)}{n}\right]S_{\nabla f \nabla f},$$

where $S_{\nabla f \nabla f}=n^{-1}\sum_{i=1}^{n}\nabla f_i(\nabla f_i)'$.

As a sanity check, when $f(X_i;\beta)=X_i'\beta$, we have $\nabla f_i = X_i$ and Theorem 16 reduces to Theorem 12.

Similarly to the discussions in Sections 3.4.1 and 3.4.2, we can test the presence of systematic and idiosyncratic treatment effect variations based on the above theorem.

Appendix A

Technical Details for Chapter 1

A.1 Lemmas

Lemma 1. The completely randomized treatment assignment $\boldsymbol{T} = (T_1, \ldots, T_N)'$ satisfies

$$E(T_i) = \frac{N_1}{N}, \quad \text{var}(T_i) = \frac{N_1 N_0}{N^2}, \quad \text{cov}(T_i, T_j) = -\frac{N_1 N_0}{N^2(N-1)}.$$

If $c_1, \ldots, c_N$ are constants and $\bar{c} = \sum_{i=1}^N c_i/N$, we have

$$E\left(\sum_{i=1}^N T_i c_i\right) = N_1 \bar{c}, \quad \text{var}\left(\sum_{i=1}^N T_i c_i\right) = \frac{N_1 N_0}{N(N-1)} \sum_{i=1}^N (c_i - \bar{c})^2.$$

Proof of Lemma 1. The treatment vector $\boldsymbol{T}$ can be viewed as the inclusion indicator vector of a simple random sample of size N_1 from a finite population of size N. The conclusion follows from Cochran (1977). □

Lemma 2 (Finite Population Central Limit Theorem; Hájek, 1960; Lehmann, 1998)**.**

Suppose we have a finite population $\{x_1, \ldots, x_N\}$ with size N and mean $\bar{x} = \sum_{i=1}^{N} x_i/N$, and a simple random sample of size n with inclusion indicators $\{I_i : i = 1, \ldots, N\}$. Let $\bar{X}_n = \sum_{i=1}^{N} I_i x_i/n$ be the sample mean. As $N \to \infty$, if

$$\frac{\max_{1\le i\le N}(x_i - \bar{x})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2/N} \text{ is bounded and } \frac{n}{N} \to c \in (0,1), \tag{A.1}$$

we have that

$$\frac{\bar{X}_n - \bar{x}}{\sqrt{\text{var}(\bar{X}_n)}} \xrightarrow{d} N(0,1).$$

Lemma 3. If $\{W_i(\boldsymbol{z}) : i = 1, \ldots, N; \boldsymbol{z} \in \mathcal{F}_K\}$ is the collection of treatment indicators from a 2^K factorial experiment, then we have the following correlation structure: for $i \neq i'$ and $\boldsymbol{z} \neq \boldsymbol{z}'$,

$$\text{cov}\{W_i(\boldsymbol{z}), W_i(\boldsymbol{z})\} = \frac{r(N-r)}{N^2}, \quad \text{cov}\{W_i(\boldsymbol{z}), W_{i'}(\boldsymbol{z})\} = -\frac{r(N-r)}{N^2(N-1)},$$
$$\text{cov}\{W_i(\boldsymbol{z}), W_i(\boldsymbol{z}')\} = -\frac{r^2}{N^2}, \quad \text{cov}\{W_i(\boldsymbol{z}), W_{i'}(\boldsymbol{z}')\} = \frac{r^2}{N^2(N-1)}.$$

Proof of Lemma 3. Dasgupta et al. (2015) show the above results in their Lemmas 4 and 5. □

A.2 Proofs of the Theorems

Proof of Theorem 1. First, $\widehat{\tau}$ has the following representation

$$\begin{aligned}
\widehat{\tau} &= \frac{1}{N_1}\sum_{i=1}^{N} T_i Y_i^{obs} - \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)Y_i^{obs} \\
&= \frac{1}{N_1}\sum_{i=1}^{N} T_i Y_i(1) - \frac{1}{N_0}\sum_{i=1}^{N}(1-T_i)Y_i(0) \\
&= \sum_{i=1}^{N} T_i \left\{ \frac{Y_i(1)}{N_1} + \frac{Y_i(0)}{N_0} \right\} - \frac{1}{N_0}\sum_{i=1}^{N} Y_i(0). \qquad \text{(A.2)}
\end{aligned}$$

Since all the potential outcomes are fixed, we use Lemma 1 to obtain that the mean is

$$E(\widehat{\tau}) = \frac{N_1}{N}\sum_{i=1}^{N}\left\{ \frac{Y_i(1)}{N_1} + \frac{Y_i(0)}{N_0} \right\} - \frac{1}{N_0}\sum_{i=1}^{N} Y_i(0) = \frac{1}{N}\sum_{i=1}^{N} Y_i(1) - \frac{1}{N}\sum_{i=1}^{N} Y_i(0) = \tau,$$

and the variance is

$$\begin{aligned}
\text{var}(\widehat{\tau}) &= \frac{N_1 N_0}{N(N-1)}\sum_{i=1}^{N}\left\{ \frac{Y_i(1)}{N_1} + \frac{Y_i(0)}{N_0} - \frac{\bar{Y}_1}{N_1} - \frac{\bar{Y}_0}{N_0} \right\}^2 \\
&= \frac{N_1 N_0}{N(N-1)}\left[\frac{1}{N_1^2}\sum_{i=1}^{N}\{Y_i(1)-\bar{Y}_1\}^2 + \frac{1}{N_0^2}\sum_{i=1}^{N}\{Y_i(0)-\bar{Y}_0\}^2 \right. \\
&\quad \left. + \frac{2}{N_1 N_0}\sum_{i=1}^{N}\{Y_i(1)-\bar{Y}_1\}\{Y_i(0)-\bar{Y}_0\} \right].
\end{aligned}$$

Because of the following decomposition based on $2ab = a^2 + b^2 - (a-b)^2$:

$$2\{Y_i(1)-\bar{Y}_1\}\{Y_i(0)-\bar{Y}_0\} = \{Y_i(1)-\bar{Y}_1\}^2 + \{Y_i(0)-\bar{Y}_0\}^2 - \{Y_i(1)-Y_i(0)-\bar{Y}_1+\bar{Y}_0\}^2,$$

we have $2S_{10} = S_1^2 + S_0^2 - S_\tau^2$, and therefore

$$\text{var}(\widehat{\tau}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} - \frac{S_\tau^2}{N}.$$

Furthermore, $\sum_{i=1}^N T_i \left\{Y_i(1)/N_1 + Y_i(0)/N_0\right\}/N_1$ is the mean of a simple random sample from $\{x_i = Y_i(1)/N_1 + Y_i(0)/N_0 : i = 1, \ldots, N\}$, and the asymptotic Normality of $\widehat{\tau}$ follows from (A.2) and Lemma 2 if $x_i = Y_i(1)/N_1 + Y_i(0)/N_0$ satisfies the condition in (A.1). □

Proof of Theorem 2. Under Fisher's sharp null, all the potential outcomes are fixed constants with $Y_i(1) = Y_i(0) = Y_i^{obs}$. The randomization statistic can be represented as

$$\begin{aligned} \widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) &= \frac{1}{N_1}\sum_{i=1}^N T_i Y_i^{obs} - \frac{1}{N_0}\sum_{i=1}^N (1-T_i) Y_i^{obs} \\ &= \frac{N}{N_1 N_0}\sum_{i=1}^N T_i Y_i^{obs} - \frac{1}{N_0}\sum_{i=1}^N Y_i^{obs}. \end{aligned} \tag{A.3}$$

Using Lemma 1, we have

$$E\left\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\right\} = \frac{N}{N_1 N_0}\frac{N_1}{N}\sum_{i=1}^N Y_i^{obs} - \frac{1}{N_0}\sum_{i=1}^N Y_i^{obs} = 0,$$

and

$$\text{var}\left\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\right\} = \frac{N}{N_1 N_0 (N-1)}\sum_{i=1}^N (Y_i^{obs} - \bar{Y}^{obs})^2.$$

Since $\sum_{i=1}^N T_i Y_i^{obs}/N_1$ is the mean of a simple random sample from $\left\{x_i = Y_i^{obs} : 1, \ldots, N\right\}$, the randomization statistic $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ follows a Normal distribution asymptotically

by (A.3) and Lemma 2 if $x_i = Y_i^{obs}$ satisfies the condition in (A.1). □

Proof of Theorem 3. We have the following variance decomposition for $\boldsymbol{Y}^{obs}$:

$$\begin{aligned}
&\sum_{i=1}^{N}(Y_i^{obs} - \bar{Y}^{obs})^2 \\
= &\sum_{\{i:T_i=1\}}(Y_i^{obs} - \bar{Y}_1^{obs} + \bar{Y}_1^{obs} - \bar{Y}^{obs})^2 + \sum_{\{i:T_i=0\}}(Y_i^{obs} - \bar{Y}_0^{obs} + \bar{Y}_0^{obs} - \bar{Y}^{obs})^2 \\
= &\sum_{\{i:T_i=1\}}(Y_i^{obs} - \bar{Y}_1^{obs})^2 + N_1(\bar{Y}_1^{obs} - \bar{Y}^{obs})^2 + \sum_{\{i:T_i=0\}}(Y_i^{obs} - \bar{Y}_0^{obs})^2 + N_0(\bar{Y}_0^{obs} - \bar{Y}^{obs})^2.
\end{aligned}$$

Ignoring the difference between N and $N-1$ contributes only a higher order term $o_p(N^{-1})$ in the asymptotic analysis. Therefore, we obtain that

$$\begin{aligned}
&\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) \\
= &N_0^{-1}s_1^2 + N_1^{-1}s_0^2 + N_0^{-1}(\bar{Y}_1^{obs} - \bar{Y}^{obs})^2 + N_1^{-1}(\bar{Y}_0^{obs} - \bar{Y}^{obs})^2 - N_1^{-1}s_1^2 - N_0^{-1}s_0^2 + o_p(N^{-1}) \\
= &(N_0^{-1} - N_1^{-1})(s_1^2 - s_0^2) + N_0^{-1}(\bar{Y}_1^{obs} - \bar{Y}^{obs})^2 + N_1^{-1}(\bar{Y}_0^{obs} - \bar{Y}^{obs})^2 + o_p(N^{-1}).
\end{aligned}$$

Since $\bar{Y}^{obs} = (N_1\bar{Y}_1^{obs} + N_0\bar{Y}_0^{obs})/N$, we have

$$(\bar{Y}_1^{obs} - \bar{Y}^{obs})^2/N_0 = N_0(\bar{Y}_1^{obs} - \bar{Y}_0^{obs})^2/N^2, \quad (\bar{Y}_0^{obs} - \bar{Y}^{obs})^2/N_1 = N_1(\bar{Y}_1^{obs} - \bar{Y}_0^{obs})^2/N^2.$$

It follows that

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = (N_0^{-1} - N_1^{-1})(s_1^2 - s_0^2) + N^{-1}(\bar{Y}_1^{obs} - \bar{Y}_0^{obs})^2 + o_p(N^{-1}).$$

Replacing the sample quantities $(s_1^2, s_0^2, \bar{Y}_1^{obs}, \bar{Y}_0^{obs})$ by the population quantities $(S_1^2, S_0^2, \bar{Y}_1, \bar{Y}_0)$

adds only higher order terms $o_p(N^{-1})$, and we eventually have

$$\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) = (N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) + N^{-1}(\bar{Y}_1 - \bar{Y}_0)^2 + o_p(N^{-1}).$$

□

Proof of Corollary 1. For binary outcomes, the conclusions follow from

$$\begin{aligned} s_t^2 &= \frac{1}{N_t - 1} \sum_{\{i:T_i=t\}} (Y_i^{obs} - \bar{Y}_t^{obs})^2 = \frac{N_t}{N_t - 1}\widehat{p}_t(1 - \widehat{p}_t), \\ s^2 &= \frac{1}{N-1}\sum_{i=1}^{N}(Y_i^{obs} - \bar{Y}^{obs})^2 = \frac{N}{N-1}\widehat{p}(1-\widehat{p}). \end{aligned}$$

□

Proof of Theorem 4. Under the sharp null hypothesis, $\{|\widehat{\tau}_i| = |Y_{i1}^{obs} - Y_{i2}^{obs}| : i = 1, \ldots, N\}$ are all fixed numbers, and $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ has the same distribution as

$$\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) \sim \frac{1}{N}\sum_{i=1}^{N}(1 - 2T_i)|\widehat{\tau}_i| \sim \frac{1}{N}\sum_{i=1}^{N}\delta_i|\widehat{\tau}_i|,$$

where δ_i's are iid random signs with mean zero and variance one. Therefore, the randomization distribution of $\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs})$ has mean zero by symmetry, and variance

$$\widehat{V}(\text{Fisher}) = \text{var}\{\widehat{\tau}(\boldsymbol{T}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = \frac{1}{N^2}\sum_{i=1}^{N}\text{var}(\delta_i)|\widehat{\tau}_i|^2 = \frac{1}{N^2}\sum_{i=1}^{N}\widehat{\tau}_i^2.$$

The classical Lindberg–Feller Central Limit Theorem (Lehmann, 1998) guarantees its asymptotic normality.

The difference between the Neymanian and Fisherian variances is

$$\begin{aligned}
\widehat{V}(\text{Fisher}) - \widehat{V}(\text{Neyman}) &= \frac{1}{N^2}\sum_{i=1}^{N}\widehat{\tau}_i^2 - \frac{1}{N(N-1)}\sum_{i=1}^{N}(\widehat{\tau}_i - \widehat{\tau})^2 \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\widehat{\tau}_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}\widehat{\tau}_i^2 - N\widehat{\tau}^2\right) + o_p(N^{-1}) \\
&= \frac{\tau^2}{N} + o_p(N^{-1}),
\end{aligned}$$

where the $o_p(N^{-1})$ appears due to the difference between N and $N-1$, and $\widehat{\tau} - \tau = o_p(1)$. □

Proof of Corollary 2. For matched-pair experiments with binary outcomes, we have

$$\widehat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\widehat{\tau}_i = \frac{m_{10}^{obs} - m_{01}^{obs}}{N},$$

since only the pairs with discordant outcomes contribute to the $\widehat{\tau}_i$ terms. The Fisherian variance is

$$\widehat{V}(\text{Fisher}) = \frac{1}{N^2}\sum_{i=1}^{N}\widehat{\tau}_i^2 = \frac{m_{10}^{obs} + m_{01}^{obs}}{N^2},$$

and the Neymanian variance is

$$\widehat{V}(\text{Neyman}) = \frac{1}{N(N-1)}\left(\sum_{i=1}^{N}\widehat{\tau}_i^2 - N\widehat{\tau}^2\right) = \frac{1}{N(N-1)}\left\{m_{10}^{obs} + m_{01}^{obs} - \frac{(m_{10}^{obs} - m_{01}^{obs})^2}{N}\right\}.$$

Therefore, the Fisherian test is asymptotically equivalent to

$$\frac{\widehat{\tau}}{\sqrt{\widehat{V}(\text{Fisher})}} = \frac{m_{10}^{obs} - m_{01}^{obs}}{\sqrt{m_{10}^{obs} + m_{01}^{obs}}} \xrightarrow{d} \mathcal{N}(0,1)$$

under H_0(Fisher), and the Neymanian test is asymptotically equivalent to

$$\frac{\widehat{\tau}}{\sqrt{\widehat{V}(\text{Neyman})}} = \frac{m_{10}^{obs} - m_{01}^{obs}}{\sqrt{m_{10}^{obs} + m_{01}^{obs} - (m_{10}^{obs} - m_{01}^{obs})^2/N}} \xrightarrow{d} \mathcal{N}(0,1)$$

under H_0(Neyman). □

Proof of Theorem 5. It is direct to obtain $E\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = 0$ by symmetry. Under H_0(Fisher), $\boldsymbol{Y}^{obs} = \{Y_i^{obs} : i = 1, \ldots, N\}$ is a fixed vector. Lemma 3 implies that $\bar{Y}^{obs}(\boldsymbol{z})$ is the sample mean of a simple random sample of size r from the population $\boldsymbol{Y}^{obs}$ of size N. Therefore, we have

$$\text{var}\{\bar{Y}^{obs}(\boldsymbol{z}) \mid H_0(\text{Fisher})\} = \left(\frac{1}{r} - \frac{1}{N}\right) s^2. \tag{A.4}$$

Based on the correlation structure in Lemma 3, we obtain that

$$\begin{aligned}
& \text{cov}\{\bar{Y}^{obs}(\boldsymbol{z}_1), \bar{Y}^{obs}(\boldsymbol{z}_2) \mid H_0(\text{Fisher})\} \\
= \; & \frac{1}{r^2}\text{cov}\left\{\sum_{i=1}^N W_i(\boldsymbol{z}_1) Y_i^{obs}, \sum_{i=1}^N W_i(\boldsymbol{z}_2) Y_i^{obs} \mid H_0(\text{Fisher})\right\} \\
= \; & \frac{1}{r^2}\left[\sum_{i=1}^N \text{cov}\{W_i(\boldsymbol{z}_1), W_i(\boldsymbol{z}_2)\}(Y_i - \bar{Y}^{obs})^2 \right. \\
& \qquad \left. + \sum_{i=1}^N \sum_{i' \neq i} \text{cov}\{W_i(\boldsymbol{z}_1), W_i(\boldsymbol{z}_2)\}(Y_i - \bar{Y}^{obs})(Y_{i'} - \bar{Y}^{obs})\right] \\
= \; & -\frac{1}{N^2}\sum_{i=1}^N (Y_i - \bar{Y}^{obs})^2 + \frac{1}{N^2(N-1)}\sum_{i=1}^N \sum_{i' \neq i} (Y_i - \bar{Y}^{obs})(Y_{i'} - \bar{Y}^{obs}) \\
= \; & -\frac{1}{N^2}\sum_{i=1}^N (Y_i - \bar{Y}^{obs})^2 - \frac{1}{N^2(N-1)}\sum_{i=1}^N (Y_i - \bar{Y}^{obs})^2 \\
= \; & -\frac{1}{N} s^2.
\end{aligned} \tag{A.5}$$

Therefore, the variance of the test statistic is

$$\begin{aligned}
& \text{var}\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} \\
= \;& 2^{-2(K-1)} \boldsymbol{g}_1' \text{cov}(\bar{\boldsymbol{Y}}^{obs}) \boldsymbol{g}_1 \\
= \;& 2^{-2(K-1)} \left[\sum_{j=1}^{J} g_{1j}^2 \text{var}\{\bar{Y}^{obs}(\boldsymbol{z}_j) \mid H_0(\text{Fisher})\} \right. \\
& \qquad \left. + \sum_{j=1}^{J} \sum_{j' \neq j}^{J} g_{1j} g_{1j'} \text{cov}\{\bar{Y}^{obs}(\boldsymbol{z}_j), \bar{Y}^{obs}(\boldsymbol{z}_{j'}) \mid H_0(\text{Fisher})\} \right] \\
= \;& 2^{-2(K-1)} s^2 \left\{ \sum_{j=1}^{J} g_{1j}^2 \left(\frac{1}{r} - \frac{1}{N} \right) - \sum_{j=1}^{J} \sum_{j' \neq j}^{J} g_{1j} g_{1j'} \frac{1}{N} \right\},
\end{aligned}$$

where the last equation is due to (A.4) and (A.5). Since

$$0 = \left(\sum_{j=1}^{J} g_{1j} \right)^2 = \sum_{j=1}^{J} g_{1j}^2 + \sum_{j=1}^{J} \sum_{j' \neq j}^{J} g_{1j} g_{1j'},$$

we have

$$- \sum_{j=1}^{J} \sum_{j' \neq j}^{J} g_{1j} g_{1j'} = \sum_{j=1}^{J} g_{1j}^2 = J.$$

Therefore, we can simplify the variance as

$$\text{var}\{\widehat{\tau}_1(\boldsymbol{W}, \boldsymbol{Y}^{obs}) \mid H_0(\text{Fisher})\} = 2^{-2(K-1)} s^2 J / r.$$

□

Proof of Theorem 6. We first observe the following variance decomposition:

$$\begin{aligned}
&\sum_{i=1}^{N}(Y_i^{obs}-\bar{Y}^{obs})^2 \\
=\;& \sum_{\boldsymbol{z}\in\mathcal{F}_K}\sum_{\{i:W_i(\boldsymbol{z})=1\}}\{Y_i^{obs}-\bar{Y}^{obs}(\boldsymbol{z})+\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2 \\
=\;& \sum_{\boldsymbol{z}\in\mathcal{F}_K}\sum_{\{i:W_i(\boldsymbol{z})=1\}}\{Y_i^{obs}-\bar{Y}^{obs}(\boldsymbol{z})\}^2+r\sum_{\boldsymbol{z}\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
s^2 &= \frac{1}{N-1}\sum_{\boldsymbol{z}\in\mathcal{F}_K}\sum_{\{i:W_i(\boldsymbol{z})=1\}}\{Y_i^{obs}-\bar{Y}^{obs}(\boldsymbol{z})\}^2+\frac{r}{N-1}\sum_{\boldsymbol{z}\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2 \\
&= \frac{r-1}{N-1}\sum_{\boldsymbol{z}\in\mathcal{F}_K}s^2(\boldsymbol{z})+\frac{r}{N-1}\sum_{\boldsymbol{z}\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2 \\
&= \frac{1}{J}\sum_{\boldsymbol{z}\in\mathcal{F}_K}s^2(\boldsymbol{z})+\frac{1}{J}\sum_{\boldsymbol{z}\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2+o_p(r^{-1}),
\end{aligned}$$

where ignoring the difference between N and $N-1$ and between r and $r-1$ in the last equation contributes the higher order term. Therefore, we have

$$2^{2(K-1)}r\left\{\widehat{V}_1(\text{Fisher})-\widehat{V}_1(\text{Neyman})\right\}=Js^2-\sum_{\boldsymbol{z}\in\mathcal{F}_K}s^2(\boldsymbol{z})=\sum_{\boldsymbol{z}\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2+o_p(r^{-1}).$$

Since $\bar{Y}^{obs}=\sum_{\boldsymbol{z}\in\mathcal{F}_K}\bar{Y}^{obs}(\boldsymbol{z})/2^K$, the formula $\sum_{i=1}^n(x_i-\bar{x})^2=\sum_{i=1}^n\sum_{j=1}^n(x_i-x_j)^2/(2n)$ gives us

$$\sum_{\boldsymbol{z}\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}\}^2=\sum_{\boldsymbol{z}\in\mathcal{F}_K}\sum_{\boldsymbol{z}'\in\mathcal{F}_K}\{\bar{Y}^{obs}(\boldsymbol{z})-\bar{Y}^{obs}(\boldsymbol{z}')\}^2/2^{K+1}.$$

Consequently, we have

$$\widehat{V}_1(\text{Fisher}) - \widehat{V}_1(\text{Neyman}) = \frac{1}{2^{3K-1}r} \sum_{\boldsymbol{z}\in\mathcal{F}_K} \sum_{\boldsymbol{z}'\in\mathcal{F}_K} \{\bar{Y}^{obs}(\boldsymbol{z}) - \bar{Y}^{obs}(\boldsymbol{z}')\}^2 + o_p(r^{-1}),$$

which leads to the final conclusion since replacing $\bar{Y}^{obs}(\boldsymbol{z})$ by $\bar{Y}(\boldsymbol{z})$ contributes only $o_p(r^{-1})$. □

Proof of Theorem 7. In the following, we will prove the results for completely randomized experiments, matched-pair experiments, and factorial experiments, respectively.

For completely randomized experiments with binary outcomes, we can summarize the observed data by a two by two table with cell counts $n_{ty}^{obs} = \#\{i : T_i = t, Y_i^{obs} = y\}$, where $t, y = 0, 1$. The row sums $N_1 = n_{11}^{obs} + n_{10}^{obs}$ and $N_0 = n_{01}^{obs} + n_{00}^{obs}$ are fixed by the design of experiments, and the column sums $n_{11}^{obs}+n_{01}^{obs}$ and $n_{10}^{obs}+n_{00}^{obs}$ are also fixed under the sharp null hypothesis. Therefore, n_{11}^{obs} is the only random component in the two by two table, because other cell counts are deterministic functions of it. According to the treatment assignment mechanism, we know that n_{11}^{obs} follows the hypergeometric distribution the same as the one in Fisher's exact test. All test statistics are functions of the two by two table, and thus functions of n_{11}^{obs}. Consequently, all test statistics are equivalent to the difference-in-means statistic under the sharp null.

For matched-pair experiments with binary outcomes, we can summarize the observed data by the two by two table with cell counts $m_{y_1y_0}^{obs}$ defined in the main text. Under the sharp null hypothesis, m_{11}^{obs}, m_{00}^{obs}, and $m_{dis}^{obs} = m_{10}^{obs} + m_{01}^{obs}$ are all fixed numbers, implying that the only random component in the two by two table is m_{10}^{obs}. According to the treatment assignment mechanism, we know $m_{10}^{obs} \sim$

Binomial$(m_{dis}^{obs}, 1/2)$. All test statistics are functions of the two by two table, and thus functions of m_{10}^{obs}. Consequently, all test statistics are equivalent to the difference-in-means statistic under the sharp null.

For 2^K factorial experiments, by symmetry we only need to show the result for factorial effect 1. It has the same structure as completely randomized experiments, and therefore, the conclusion follows. □

A.3 Connections with Regression-Based Inference

Assume the following linear model for the observed outcomes:

$$Y_i^{obs} = \alpha + \beta T_i + \varepsilon_i, \tag{A.6}$$

where $\varepsilon_i, \ldots, \varepsilon_N$ are independently and identically distributed (iid) as $\mathcal{N}(0, \sigma^2)$. The hypothesis of zero treatment effect is thus characterized by $H_0(LM) : \beta = 0$.

Hinkelmann and Kempthorne (2007) called

$$Y_i^{obs} = T_i Y_i(1) + (1 - T_i) Y_i(0) = Y_i(0) + \{Y_i(1) - Y_i(0)\} T_i = \alpha + \beta T_i + \varepsilon_i$$

the "derived linear model", assuming that $Y_i(1) - Y_i(0) = \beta$ is a constant and $Y_i(0) = \alpha + \varepsilon_i$ for all $i = 1, \ldots, N$. But the linear model for observed outcomes ignores the design of the randomized experiment, and the "iid" assumption contradicts $\text{cov}(T_i, T_j) \neq 0$ and $\text{cov}(Y_i^{obs}, Y_j^{obs}) \neq 0$ for $i \neq j$. Although linear regression has been criticized for analyzing experimental data (Freedman, 2008), the least square

estimator $\widehat{\beta}_{OLS} = \widehat{\tau}$ is unbiased for the average causal effect τ. However, the correct variance of $\widehat{\beta}_{OLS}$ requires careful discussion.

A.3.1 Wald Test and Neymanian Inference

The residual is defined as $\widehat{\varepsilon}_i = Y_i^{obs} - \bar{Y}_1$ if $T_i = 1$ and $\widehat{\varepsilon}_i = Y_i^{obs} - \bar{Y}_0$ if $T_i = 0$. Since the variance σ^2 in the linear model can be estimated by

$$\widehat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\widehat{\varepsilon}_i^2 = \frac{N_1 - 1}{N-2}s_1^2 + \frac{N_0 - 1}{N-2}s_0^2,$$

the variance of $\widehat{\beta}_{OLS}$, $\text{var}(\widehat{\beta}_{OLS}) = N\sigma^2/(N_1N_0)$, can be estimated by

$$\widehat{V}_{OLS} = \frac{N(N_1 - 1)}{(N-2)N_1N_0}s_1^2 + \frac{N(N_0 - 1)}{(N-2)N_1N_0}s_0^2 \approx \frac{s_1^2}{N_0} + \frac{s_0^2}{N_1}.$$

It is different from Neyman's variance estimator unless $N_1 = N_0$. Fortunately, we can avoid this problem by using Huber–White heteroskedasticity-robust variance estimator:

$$\widehat{V}_{HW} = \frac{\sum_{i=1}^{N}\widehat{\varepsilon}_i^2(T_i - \bar{T})^2}{\left\{\sum_{i=1}^{N}(T_i - \bar{T})^2\right\}^2} = \frac{s_1^2}{N_1}\frac{N_1 - 1}{N_1} + \frac{s_0^2}{N_0}\frac{N_0 - 1}{N_0} \approx \frac{s_1^2}{N_1} + \frac{s_0^2}{N_0},$$

which is asymptotically equivalent to the Neymanian variance estimator. Therefore, the Wald statistic using $\widehat{V}_{HW}$ for testing $H_0(LM)$ is asymptotically the same as the Neymanian test.

A.3.2 Rao's Score Test and the FRT

While the connection between the behavior of the Wald test for $H_0(LM)$ and Neyman's test has been established in previous studies, we make a similar connection between Rao's score test for $H_0(LM)$ and the FRT in the following theorem.

Theorem 17. Rao's score test for $H_0(LM)$ under model (A.6) is equivalent to

$$\frac{\widehat{\tau}}{\sqrt{\widehat{V}_S}} \xrightarrow{d} \mathcal{N}(0,1),$$

where $\widehat{V}_S = (N-1)s^2/(N_1N_0)$.

Ignoring the difference between $(N-1)$ and N when N is large, the difference between $\widehat{V}_S$ and $\widehat{V}$(Fisher) is of higher order, and Rao's score test is asymptotically equivalent to the FRT. The sharp null hypothesis imposes the equal variance assumption on potential outcomes under treatment and control, leading to the equivalence of Rao's score test under the homoskedastic model and the FRT.

Proof of Theorem 17. The log likelihood function for the linear model in is

$$l(\alpha, \beta, \sigma^2) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{\sum_{i=1}^{N}(Y_i^{obs} - \alpha - \beta T_i)^2}{2\sigma^2}.$$

Therefore, the score functions are

$$\begin{aligned}
\partial l/\partial\alpha &= \sum_{i=1}^{N}(Y_i-\alpha-\beta T_i)/\sigma^2,\\
\partial l/\partial\beta &= \sum_{i=1}^{N}(Y_i-\alpha-\beta T_i)T_i/\sigma^2,\\
\partial l/\partial\sigma^2 &= -N/(2\sigma^2)+\sum_{i=1}^{N}(Y_i-\alpha-\beta T_i)^2/\{2(\sigma^2)^2\}.
\end{aligned}$$

Plugging the MLEs under the null hypothesis with $\beta = 0$, $\widetilde{\alpha} = \bar{Y}^{obs}$ and $\widetilde{\sigma}^2 = \sum_{i=1}^{N}(Y_i^{obs}-\bar{Y}^{obs})^2/N$ into the score functions, we obtain that only the second component of the score functions is non-zero: $\sum_{i=1}^{N}(Y_i-\bar{Y})T_i/\widetilde{\sigma}^2 = N_1N_0\widehat{\tau}/(N\widetilde{\sigma}^2)$.

The second order derivatives of the log likelihood function are

$$\begin{aligned}
\partial^2 l/\partial\alpha^2 &= -N/\sigma^2,\\
\partial^2 l/\partial\beta^2 &= \sum_{i=1}^{N}T_i^2/\sigma^2 = -N_1/\sigma^2,\\
\partial^2 l/\partial(\sigma^2)^2 &= N/(2\sigma^4)-\sum_{i=1}^{N}(Y_i-\alpha-\beta T_i)^2/\sigma^6,\\
\partial^2 l/\partial\alpha\partial\beta &= -N_1/\sigma^2,\\
\partial^2 l/\partial\alpha\partial\sigma^2 &= -\sum_{i=1}^{N}(Y_i-\alpha-\beta T_i)/\sigma^4,\\
\partial^2 l/\partial\beta\partial\sigma^2 &= -\sum_{i=1}^{N}(Y_i-\alpha-\beta T_i)T_i/\sigma^4.
\end{aligned}$$

Therefore, the expected Fisher information matrix is

$$\boldsymbol{I}_N = \begin{pmatrix} N/\sigma^2 & N_1/\sigma^2 & 0 \\ N_1/\sigma^2 & N_1/\sigma^2 & 0 \\ 0 & 0 & N/(2\sigma^4) \end{pmatrix},$$

with the $(2,2)$-th element of $\boldsymbol{I}_N^{-1}$ being $N\sigma^2/(N_1N_0)$. Thus, Rao's score test for $H_0(LM)$ is

$$\left(\frac{N_1N_0\widehat{\tau}}{N\widetilde{\sigma}^2}\right)^2 \frac{N\widetilde{\sigma}^2}{N_1N_0} \xrightarrow{d} \chi^2(1),$$

or equivalently,

$$\widehat{\tau}\Big/\sqrt{\frac{N\widetilde{\sigma}^2}{N_1N_0}} = \widehat{\tau}\Big/\sqrt{\frac{(N-1)s^2}{N_1N_0}} = \frac{\widehat{\tau}}{\sqrt{\widehat{V}_S}} \xrightarrow{d} \mathcal{N}(0,1).$$

□

A.4 More Details About Figures 1.3 and 1.4

A.4.1 Figure 1.3

Asymptotically, the FRT is invalid under Neyman's null if and only if $\widehat{V}$(Fisher) is asymptotically smaller than the true sampling variance of $\widehat{\tau}$, V(Neyman), i.e.,

$$(N_0^{-1} - N_1^{-1})(S_1^2 - S_0^2) + N^{-1}S_\tau^2 < 0.$$

When $Y_i(1) = aY_i(0) + (1-a)\bar{Y}(0)$, we have $S_1^2 = a^2 S_0^2$ and $S_\tau^2 = (a-1)^2 S_0^2$. The above inequality reduces to

$$\left(\frac{1}{1-r} - \frac{1}{r}\right)(a^2 - 1) + (a-1)^2 < 0.$$

If $a > 1$, the inequality further reduces to

$$\left(\frac{1}{1-r} - \frac{1}{r}\right)(a+1) + (a-1) < 0 \Longleftrightarrow \frac{2}{1+a} > -\frac{r^2 - 3r + 1}{(1-r)r}.$$

We only consider the case with $r < 1/2$. It is straightforward to see that when $0 < r < (3-\sqrt{5})/2$, we have $r^2 - 3r + 1 > 0$ and the above inequality holds automatically. When $r > (3-\sqrt{5})/2$, then the above inequality reduces to

$$a < \frac{r^2 + r - 1}{r^2 - 3r + 1}.$$

The above discussion allows us to determine the region that the FRT rejects more often than the nominal level under Neyman's null.

A.4.2 Figure 1.4

According to Corollary 1 in the main text, the Neymanian test has larger asymptotic power than the Fisherian test if and only if

$$\left(\frac{1}{1-r} - \frac{1}{r}\right)\{p_1(1-p_1) - p_0(1-p_0)\} + (p_1 - p_0)^2 > 0.$$

After some simple algebra, we can simplify the above inequality as

$$(p_1 - p_0)(ap_1 + bp_0 + c) > 0,$$

where

$$a = \frac{1 - r - r^2}{(1-r)r}, \quad b = \frac{1 - 3r + r^2}{(1-r)r}, \quad c = \frac{2r - 1}{(1-r)r}.$$

The shape of the region depends on the signs of a and b, because the line $ap_1 + bp_0 + c = 0$ intersects with the line $p_1 - p_0 = 0$ at the point $(p_1, p_0) = (1/2, 1/2)$. It is easy to show that $a > 0$ if and only if $0 \leq r \leq \Gamma$, and $b > 0$ if and only if $0 \leq r \leq 1 - \Gamma$, where $\Gamma = (-1 + \sqrt{5})/2 \approx 0.618$ is the reciprocal of the golden ratio. Therefore, when $r > 1/2$, the region may have two shapes according the value of r compared to Γ, as shown in Figure 1.4 of the main text. By symmetry, we can also plot the region when $r < 1/2$.

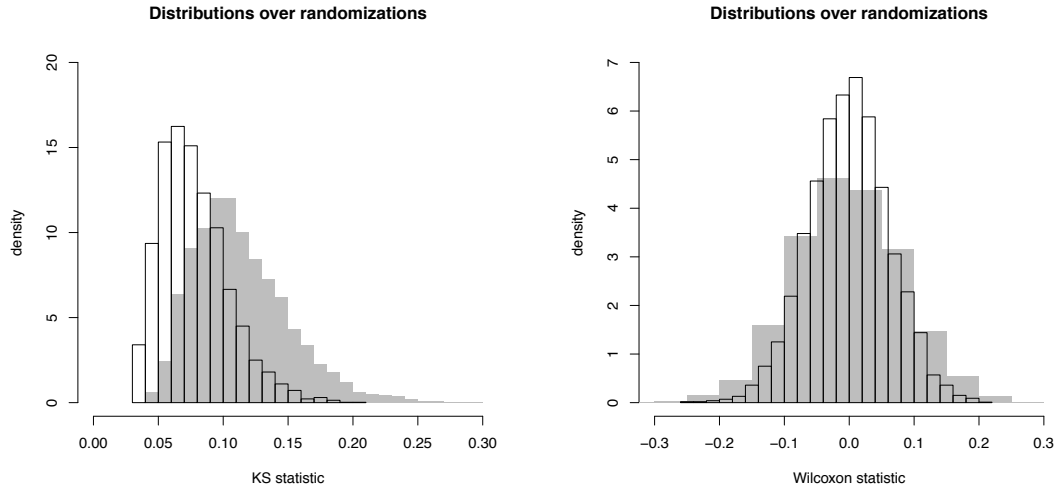
A.5 Other Test Statistics

Consider a finite population of size $N = 200$, and balanced completely randomized experiments. Under the sharp null hypothesis, we generate potential outcomes $Y_i(1) = Y_i(0)$ from $\mathcal{N}(0, 1)$; under the average null hypothesis, we generate $Y_i(1)$ from $\mathcal{N}(0, 1)$, and generate $Y_i(0)$ as the order statistics of $Y_i(1)$. Clearly, the marginal distributions are the same but the correlation of the potential outcomes are different under different null hypothesis.

The grey histogram in Figure A.1(a) is the randomization distribution of the Kolmogorov–Smirnov statistic under the sharp null hypothesis, and the white his-

togram with border is the randomization distribution under the average null hypothesis. The former is more disperse than the latter, indicating that the FRT using the Kolmogorov–Smirnov statistic tends to be conservative under the average null hypothesis.

The results for the Wilcoxon–Mann–Whitney rank sum statistic in Figure A.1(b) are the same as above.



(a) Kolmogorov–Smirnov Statistic

(b) Wilcoxon–Mann–Whitney Rank Sum Statistic

Figure A.1: Randomization Distributions of Different Test Statistics Under the Sharp Null (grey histograms) and Average Null (white histograms with borders).

Appendix B

Technical Details for Chapter 2

B.1 Bias and Variance Reduction for Nonlinear Causal Measures

Result for log(**CRR**)**.** A bias-corrected estimator for log(CRR) is

$$\log(\widehat{\mathrm{CRR}})^c = \log(\widehat{\mathrm{CRR}}) + \frac{N_0}{2\widehat{p}_1^2 N_1 N} s_1^2 - \frac{N_1}{2\widehat{p}_0^2 N_0 N} s_0^2, \tag{B.1}$$

with improved variance estimator

$$\widehat{V}_{\mathrm{CRR}}^c = \widehat{V}_{\mathrm{CRR}} - \frac{|\widehat{\tau}|(1-|\widehat{\tau}|)}{\widehat{p}_1\widehat{p}_0(N-1)}, \tag{B.2}$$

where $\widehat{V}_{\mathrm{CRR}}$ is defined in (2.19).

Result for log(**COR**)**.** A bias-corrected estimator for log(COR) is

$$\log(\widehat{\mathrm{COR}})^c = \log(\widehat{\mathrm{COR}}) + \frac{1-2\widehat{p}_1}{2\widehat{p}_1^2(1-\widehat{p}_1)^2}\frac{N_0}{N_1 N}s_1^2 - \frac{1-2\widehat{p}_0}{2\widehat{p}_0^2(1-\widehat{p}_0)^2}\frac{N_1}{N_0 N}s_0^2, \quad \text{(B.3)}$$

with improved variance estimator

$$\widehat{V}_{\mathrm{COR}}^c = \widehat{V}_{\mathrm{COR}} - \frac{|\widehat{\tau}|(1-|\widehat{\tau}|)}{\widehat{p}_1(1-\widehat{p}_1)\widehat{p}_0(1-\widehat{p}_0)(N-1)}, \quad \text{(B.4)}$$

where $\widehat{V}_{\mathrm{COR}}$ is defined in (2.21).

B.2 Lemmas and Their Proofs

Lemma 4. The completely randomized treatment assignment $\boldsymbol{W}$ satisfies $E(W_i) = N_1/N$, $\mathrm{var}(W_i) = N_1N_0/N^2$, and $\mathrm{cov}(W_i, W_j) = -N_1N_0/\{N^2(N-1)\}$. If $(c_1, \cdots, c_N)$ and $(d_1, \cdots, d_N)$ are constants with $\overline{c} = \sum_{i=1}^N c_i/N$ and $\overline{d} = \sum_{i=1}^N d_i/N$, we have

$$E\left(\sum_{i=1}^N W_i c_i\right) = N_1\overline{c}, \quad \mathrm{cov}\left(\sum_{i=1}^N W_i c_i, \sum_{i=1}^N W_i d_i\right) = \frac{N_1N_0}{N(N-1)}\sum_{i=1}^N (c_i-\overline{c})(d_i-\overline{d}).$$

Proof of Lemma 4. The observed outcomes in the treatment and control can be viewed as two sets of simple random samples from the finite population of $\{Y_i(1) : i = 1, \cdots, N\}$ and $\{Y_i(0) : i = 1, \cdots, N\}$, respectively. Therefore, the conclusion follows from classic survey sampling textbooks such as Cochran (1977). □

Lemma 5. The estimators, $\widehat{p}_1$ and $\widehat{p}_0$, are unbiased for p_1 and p_0, with variances and

covariance:

$$\text{var}(\widehat{p}_1) = \frac{N_0}{N_1 N} S_1^2, \quad \text{var}(\widehat{p}_0) = \frac{N_1}{N_0 N} S_0^2, \quad \text{cov}(\widehat{p}_1, \widehat{p}_0) = -\frac{1}{N} S_{10} = -\frac{1}{2N}(S_1^2 + S_0^2 - S_\tau^2).$$

Proof of Lemma 5. The unbiasedness and variances of $\widehat{p}_1$ and $\widehat{p}_0$ follow directly from Lemma 4. The covariance between $\widehat{p}_1$ and $\widehat{p}_0$ is

$$\text{cov}(\widehat{p}_1, \widehat{p}_0) = -\frac{1}{N_1 N_0} \frac{N_1 N_0}{N} S_{10} = -\frac{1}{N} S_{10}.$$

Summing from $i = 1$ to N over the following decomposition

$$2\{Y_i(1) - p_1\}\{Y_i(0) - p_0\} = \{Y_i(1) - p_1\}^2 + \{Y_i(0) - p_0\}^2 - (\tau_i - \tau)^2,$$

we have $2S_{10} = S_1^2 + S_0^2 - S_\tau^2$, and therefore the covariance can also be expressed as

$$\text{cov}(\widehat{p}_1, \widehat{p}_0) = -\frac{1}{2N}(S_1^2 + S_0^2 - S_\tau^2).$$

□

B.3 Proofs of the Theorems

Proof of Theorem 8. Define the proportions $p_{jk} = N_{jk}/N$. We first rewrite S_τ^2/N as

$$\begin{aligned}\frac{S_\tau^2}{N} &= \frac{1}{N(N-1)}\sum_{i=1}^{N}(\tau_i - \tau)^2 \\ &= \frac{1}{N(N-1)}\left\{(N_{10}+N_{01}) - \frac{(N_{10}-N_{01})^2}{N}\right\} \\ &= \frac{1}{N-1}(\tau + 2p_{01} - \tau^2).\end{aligned}$$

In order to find the lower bound of S_τ^2/N, we only need to find the lower bound for p_{01}. This reduces to the following linear programming problem:

$$\begin{cases} \min\limits_{p_{11},p_{10},p_{01},p_{00}} & p_{01} \\ s.t. & p_{11}+p_{10} = p_1, \\ & p_{11}+p_{01} = p_0, \\ & p_{11}+p_{10}+p_{01}+p_{00} = 1, \\ & p_{ij} \geq 0, i,j = 0,1. \end{cases}$$

Since $p_{01} = p_0 - p_1 + p_{10} \geq -\tau$ and $p_{01} \geq 0$, the lower bound of p_{01} is

$$p_{01} \geq \max(-\tau, 0),$$

and therefore, the lower bound of S_τ^2/N is

$$\frac{S_\tau^2}{N} \geq \frac{1}{N-1}\{\tau + 2\max(-\tau,0) - \tau^2\} = \frac{1}{N-1}\{\max(-\tau,\tau) - \tau^2\} = \frac{|\tau|(1-|\tau|)}{N-1}.$$

From the derivation above, the bound is sharp, and is attained if and only if $p_{10} = 0$ or $p_{01} = 0$. Or, equivalently, S_τ^2 attains its minimum at either of the two vertices within the feasible region of the linear programming problem above: $(p_{11}, p_{10}, p_{01}, p_{00}) = (p_1, 0, -\tau, 1 - p_0)$ if $\tau \leq 0$, and $(p_{11}, p_{10}, p_{01}, p_{00}) = (p_0, \tau, 0, 1 - p_1)$ if $\tau \geq 0$. □

Proof of Theorem 9. Define $N'_w = N_w + \alpha_w + \beta_w$ and $\widehat{p}'_w = (n_{w1} + \alpha_w)/N'_w$ as the sample sizes and proportions adjusted by the pseudo counts of the prior distributions. The posterior means of π_{1+} and π_{+1} are

$$E(\pi_{1+} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) = \widehat{p}'_1, \quad \text{and} \quad E(\pi_{+1} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) = \widehat{p}'_0.$$

The posterior variances of π_{1+} and π_{+1} are

$$\text{var}(\pi_{1+} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) = \frac{\widehat{p}'_1(1 - \widehat{p}'_1)}{N'_1 + 1}, \quad \text{and} \quad \text{var}(\pi_{+1} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) = \frac{\widehat{p}'_0(1 - \widehat{p}'_0)}{N'_0 + 1}.$$

Immediately, we have

$$\begin{aligned} E\{\pi_{1+}(1 - \pi_{1+}) \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}\} &= \widehat{p}'_1(1 - \widehat{p}'_1) - \frac{\widehat{p}'_1(1 - \widehat{p}'_1)}{N'_1 + 1} = \frac{N'_1}{N'_1 + 1}\widehat{p}'_1(1 - \widehat{p}'_1), \\ E\{\pi_{+1}(1 - \pi_{+1}) \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}\} &= \widehat{p}'_0(1 - \widehat{p}'_0) - \frac{\widehat{p}'_0(1 - \widehat{p}'_0)}{N'_0 + 1} = \frac{N'_0}{N'_0 + 1}\widehat{p}'_0(1 - \widehat{p}'_0). \end{aligned}$$

Applying the laws of conditional expectation and variance to (2.10) in the main text,

we obtain the posterior mean

$$\begin{aligned}
& E(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) \\
= \;& E\left\{E(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1})\right\} \\
= \;& E\left(\frac{n_{11} + N_0\pi_{1+} - n_{01} - N_1\pi_{+1}}{N} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}\right) \\
= \;& \frac{n_{11} + N_0\widehat{p}_1' - n_{01} - N_1\widehat{p}_0'}{N} \\
= \;& \frac{N_1' + N_0}{N}\widehat{p}_1' - \frac{N_0' + N_1}{N}\widehat{p}_0' - \frac{\alpha_1 - \alpha_0}{N},
\end{aligned}$$

and posterior variance

$$\begin{aligned}
& \text{var}(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) \\
= \;& E\left\{\text{var}(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1})\right\} + \text{var}\left\{E(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}, \pi_{1+}, \pi_{+1})\right\} \\
= \;& E\left\{\frac{N_0}{N^2}\pi_{1+}(1-\pi_{1+}) + \frac{N_1}{N^2}\pi_{+1}(1-\pi_{+1}) \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}\right\} \\
& + \text{var}\left\{\frac{N_0}{N}\pi_{1+} - \frac{N_1}{N}\pi_{+1} \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}\right\} \\
= \;& \frac{N_0 N_1'}{N^2(N_1'+1)}\widehat{p}_1'(1-\widehat{p}_1') + \frac{N_1 N_0'}{N^2(N_0'+1)}\widehat{p}_0'(1-\widehat{p}_0') \\
& + \frac{N_0^2}{N^2(N_1'+1)}\widehat{p}_1'(1-\widehat{p}_1') + \frac{N_1^2}{N^2(N_0'+1)}\widehat{p}_0'(1-\widehat{p}_0') \\
= \;& \frac{N_0(N_1'+N_0)}{N^2}\frac{\widehat{p}_1'(1-\widehat{p}_1')}{N_1'+1} + \frac{N_1(N_1+N_0')}{N^2}\frac{\widehat{p}_0'(1-\widehat{p}_0')}{N_0'+1}.
\end{aligned}$$

When we have large sample size, the prior pseudo counts are overwhelmed by the observed counts n_{jk}'s, and the posterior mean and variance of τ can be approximately

by

$$\begin{aligned} E(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) &\approx \widehat{\tau}, \\ \text{var}(\tau \mid \boldsymbol{W}, \boldsymbol{Y}^{\text{obs}}) &\approx \frac{N_0}{N}\frac{\widehat{p}_1(1-\widehat{p}_1)}{N_1-1} + \frac{N_1}{N}\frac{\widehat{p}_0(1-\widehat{p}_0)}{N_0-1}. \end{aligned}$$

□

Proof of Theorem 10. Applying Taylor expansion, we have

$$\log(\widehat{\text{CRR}}) - \log(\text{CRR}) = \frac{1}{p_1}(\widehat{p}_1 - p_1) - \frac{1}{p_0}(\widehat{p}_0 - p_0) + o_p\left(\frac{1}{N^{1/2}}\right).$$

According to Lemma 5, the asymptotic variance of $\log(\widehat{\text{CRR}})$ is

$$\begin{aligned} & \frac{1}{p_1^2}\text{var}(\widehat{p}_1) + \frac{1}{p_0^2}\text{var}(\widehat{p}_0) - \frac{2}{p_1 p_0}\text{cov}(\widehat{p}_1, \widehat{p}_0) \\ = & \frac{1}{p_1^2}\frac{N_0}{N_1 N}S_1^2 + \frac{1}{p_0^2}\frac{N_1}{N_0 N}S_0^2 + \frac{1}{p_1 p_0 N}(S_1^2 + S_0^2 - S_\tau^2) \\ = & \frac{N_1 p_1 + N_0 p_0}{p_1^2 p_0 N_1 N}S_1^2 + \frac{N_1 p_1 + N_0 p_0}{p_1 p_0^2 N_0 N}S_0^2 - \frac{1}{p_1 p_0 N}S_\tau^2 \\ = & \frac{N_1 p_1 + N_0 p_0}{p_1 p_0 N}\left(\frac{S_1^2}{N_1 p_1} + \frac{S_0^2}{N_0 p_0} - \frac{S_\tau^2}{N_1 p_1 + N_0 p_0}\right). \end{aligned}$$

Assume $S_\tau^2 = 0$, and we can estimate the asymptotic variance by

$$\begin{aligned} \widehat{V}_{\text{CRR}} &= \frac{N_1\widehat{p}_1 + N_0\widehat{p}_0}{\widehat{p}_1^2\widehat{p}_0 N_1 N}s_1^2 + \frac{N_1\widehat{p}_1 + N_0\widehat{p}_0}{\widehat{p}_1\widehat{p}_0^2 N_0 N}s_0^2 \\ &= \frac{N_1\widehat{p}_1 + N_0\widehat{p}_0}{\widehat{p}_1^2\widehat{p}_0 N_1 N}\frac{N_1}{N_1-1}\widehat{p}_1(1-\widehat{p}_1) + \frac{N_1\widehat{p}_1 + N_0\widehat{p}_0}{\widehat{p}_1\widehat{p}_0^2 N_0 N}\frac{N_0}{N_0-1}\widehat{p}_0(1-\widehat{p}_0) \\ &= \frac{(N_1\widehat{p}_1 + N_0\widehat{p}_0)(1-\widehat{p}_1)}{\widehat{p}_1\widehat{p}_0(N_1-1)N} + \frac{(N_1\widehat{p}_1 + N_0\widehat{p}_0)(1-\widehat{p}_0)}{\widehat{p}_1\widehat{p}_0(N_0-1)N} \\ &= \frac{n_{10}}{n_{11}(N_1-1)}\frac{(n_{11}+n_{01})N_0}{n_{01}N} + \frac{n_{00}}{n_{01}(N_0-1)}\frac{(n_{11}+n_{01})N_1}{n_{11}N}. \end{aligned} \tag{B.5}$$

Ignoring the difference between N_w and $(N_w - 1)$ $(w = 0, 1)$ in asymptotic analysis, we obtain the formula in Theorem 3. □

Proof of Theorem 11. Applying Taylor expansion, we have

$$\log(\widehat{\mathrm{COR}}) - \log(\mathrm{COR}) = \frac{1}{p_1(1-p_1)}(\widehat{p}_1 - p_1) - \frac{1}{p_0(1-p_0)}(\widehat{p}_0 - p_0) + o_p\left(\frac{1}{N^{1/2}}\right).$$

According to Lemma 5, the asymptotic variance of $\log(\widehat{\mathrm{COR}})$ is

$$\begin{aligned}
& \frac{1}{p_1^2(1-p_1)^2}\mathrm{var}(\widehat{p}_1) + \frac{1}{p_0^2(1-p_0)^2}\mathrm{var}(\widehat{p}_0) - \frac{2}{p_1(1-p_1)p_0(1-p_0)}\mathrm{cov}(\widehat{p}_1, \widehat{p}_0) \\
= & \frac{1}{p_1^2(1-p_1)^2}\frac{N_0}{N_1 N}S_1^2 + \frac{1}{p_0^2(1-p_0)^2}\frac{N_1}{N_0 N}S_0^2 + \frac{1}{p_1(1-p_1)p_0(1-p_0)N}(S_1^2 + S_0^2 - S_\tau^2) \\
= & \frac{N_1 p_1(1-p_1) + N_0 p_0(1-p_0)}{p_1^2(1-p_1)^2 p_0(1-p_0)NN_1}S_1^2 + \frac{N_1 p_1(1-p_1) + N_0 p_0(1-p_0)}{p_1(1-p_1)p_0^2(1-p_0)^2 NN_0}S_0^2 \\
& - \frac{1}{Np_1(1-p_1)p_0(1-p_0)}S_\tau^2 \\
= & \frac{N_1 p_1(1-p_1) + N_0 p_0(1-p_0)}{Np_1(1-p_1)p_0(1-p_0)}\left\{\frac{S_1^2}{N_1 p_1(1-p_1)} + \frac{S_0^2}{N_0 p_0(1-p_0)}\right. \\
& \left. - \frac{S_\tau^2}{N_1 p_1(1-p_1) + N_0 p_0(1-p_0)}\right\}.
\end{aligned}$$

The Neyman-type "conservative" variance estimator for the asymptotic variance is

$$\begin{aligned}
\widehat{V}_{\mathrm{COR}} &= \frac{N_1\widehat{p}_1(1-\widehat{p}_1) + N_0\widehat{p}_0(1-\widehat{p}_0)}{\widehat{p}_1^2(1-\widehat{p}_1)^2\widehat{p}_0(1-\widehat{p}_0)NN_1}s_1^2 + \frac{N_1\widehat{p}_1(1-\widehat{p}_1) + N_0\widehat{p}_0(1-\widehat{p}_0)}{\widehat{p}_1(1-\widehat{p}_1)\widehat{p}_0^2(1-\widehat{p}_0)^2 NN_0}s_0^2 \\
&= \frac{N_1\widehat{p}_1(1-\widehat{p}_1) + N_0\widehat{p}_0(1-\widehat{p}_0)}{\widehat{p}_1(1-\widehat{p}_1)\widehat{p}_0(1-\widehat{p}_0)N(N_1-1)} + \frac{N_1\widehat{p}_1(1-\widehat{p}_1) + N_0\widehat{p}_0(1-\widehat{p}_0)}{\widehat{p}_1(1-\widehat{p}_1)\widehat{p}_0(1-\widehat{p}_0)N(N_0-1)} \\
&\approx \frac{N_1\widehat{p}_1(1-\widehat{p}_1) + N_0\widehat{p}_0(1-\widehat{p}_0)}{\widehat{p}_1(1-\widehat{p}_1)\widehat{p}_0(1-\widehat{p}_0)}\left(\frac{1}{NN_1} + \frac{1}{NN_0}\right) \\
&= \frac{(n_{01}+n_{00})n_{11}n_{10} + (n_{11}+n_{10})n_{01}n_{00}}{n_{11}n_{10}n_{01}n_{00}} \\
&= \frac{1}{n_{11}} + \frac{1}{n_{10}} + \frac{1}{n_{01}} + \frac{1}{n_{00}}, \qquad \text{(B.6)}
\end{aligned}$$

where the approximation is due to the difference between $N_w - 1$ and N_w for $w = 0, 1$. □

B.4 Proofs for the Results in Appendix B.1 about Bias and Variance Reduction for Nonlinear Causal Measures

Proof of the Result for log(CRR). Applying Taylor expansion, we have

$$\begin{aligned}\log(\widehat{\mathrm{CRR}}) - \log(\mathrm{CRR}) &= \frac{1}{p_1}(\widehat{p}_1 - p_1) - \frac{1}{2p_1^2}(\widehat{p}_1 - p_1)^2 \\ &\quad -\frac{1}{p_0}(\widehat{p}_0 - p_0) + \frac{1}{2p_0^2}(\widehat{p}_0 - p_0)^2 + o_p\left(\frac{1}{N}\right).\end{aligned}$$

Therefore, the asymptotic bias of $\log(\widehat{\mathrm{CRR}})$ is

$$-\frac{1}{2p_1^2}\mathrm{var}(\widehat{p}_1) + \frac{1}{2p_0^2}\mathrm{var}(\widehat{p}_0) = -\frac{N_0}{2p_1^2 N_1 N}S_1^2 + \frac{N_1}{2p_0^2 N_0 N}S_0^2,$$

and the bias-corrected estimator for log(CRR) in Appendix B.1 can be obtained by subtracting the estimated asymptotic bias from $\log(\widehat{\mathrm{CRR}})$. □

Proof of the Result for $\log(\mathrm{COR})$. Applying Taylor expansion, we have

$$\begin{aligned} & \log(\widehat{\mathrm{COR}}) - \log(\mathrm{COR}) \\ = \; & \frac{1}{p_1(1-p_1)}(\widehat{p}_1 - p_1) - \frac{1-2p_1}{2p_1^2(1-p_1)^2}(\widehat{p}_1 - p_1)^2 \\ & -\frac{1}{p_0(1-p_0)}(\widehat{p}_0 - p_0) + \frac{1-2p_0}{2p_0^2(1-p_0)^2}(\widehat{p}_0 - p_0)^2 + o_p\left(\frac{1}{N}\right). \end{aligned}$$

Therefore, the asymptotic bias of $\log(\widehat{\mathrm{COR}})$ is

$$\begin{aligned} & -\frac{1-2p_1}{2p_1^2(1-p_1)^2}\mathrm{var}(\widehat{p}_1) + \frac{1-2p_0}{2p_0^2(1-p_0)^2}\mathrm{var}(\widehat{p}_0) \\ = & -\frac{1-2p_1}{2p_1^2(1-p_1)^2}\frac{N_0}{N_1 N}S_1^2 + \frac{1-2p_0}{2p_0^2(1-p_0)^2}\frac{N_1}{N_0 N}S_0^2, \end{aligned}$$

and the bias-corrected estimator for $\log(\mathrm{COR})$ in Appendix B.1 can be obtained by subtracting the estimated asymptotic bias from $\log(\widehat{\mathrm{COR}})$. □

B.5 More Simulation Studies

In order to compare the finite sample properties of Neyman's original method, the modified Neyman's method, and the Bayesian method, we conduct the following set of simulation studies. In the main text, we choose the following two sets of potential outcomes: the first set of potential outcomes $(N_{11}, N_{10}, N_{01}, N_{00})$ are independent: $(50, 50, 50, 50)$, $(30, 70, 30, 70)$, $(30, 90, 20, 60)$, $(80, 20, 80, 20)$, $(60, 20, 90, 30)$; the second set of potential outcomes are positively associated: $(60, 40, 40, 60)$, $(50, 50, 30, 70)$, $(50, 70, 30, 50)$, $(40, 110, 10, 40)$, $(70, 30, 50, 50)$, $(50, 30, 70, 50)$, $(30, 10, 110, 50)$. In addition, in this Supplementary Materials, we also choose negatively associated poten-

tial outcomes: $(40, 60, 60, 40)$, $(30, 70, 50, 50)$, $(40, 80, 40, 40)$, $(30, 120, 20, 30)$, $(50, 50, 70, 30)$, $(40, 40, 80, 40)$, $(20, 20, 120, 40)$. We summarize the "Science" in Table B.1, where Cases 1–5 represent independent potential outcomes, Cases 6–12 represent positively associated potential outcomes, and Cases 13–19 represent negatively associated potential outcomes.

Table B.1: "Science table" for the simulation studies

Case	N_{11}	N_{10}	N_{01}	N_{00}	S_1^2	S_0^2	S_{10}	S_τ^2	τ	log(CRR)	log(COR)
1	50	50	50	50	0.251	0.251	0.000	0.503	0.000	0.000	0.000
2	30	70	30	70	0.251	0.211	0.000	0.462	0.200	0.511	0.847
3	30	90	20	60	0.241	0.188	0.000	0.430	0.350	0.875	1.504
4	80	20	80	20	0.251	0.161	0.000	0.412	-0.300	-0.470	-1.386
5	60	20	90	30	0.241	0.188	0.000	0.430	-0.350	-0.629	-1.504
6	60	40	40	60	0.251	0.251	0.050	0.402	0.000	0.000	0.000
7	50	50	30	70	0.251	0.241	0.050	0.392	0.100	0.223	0.405
8	50	70	30	50	0.241	0.241	0.010	0.462	0.200	0.405	0.811
9	40	110	10	40	0.188	0.188	0.013	0.352	0.500	1.099	2.197
10	70	30	50	50	0.251	0.241	0.050	0.392	-0.100	-0.182	-0.405
11	50	30	70	50	0.241	0.241	0.010	0.462	-0.200	-0.405	-0.811
12	30	10	110	50	0.161	0.211	0.010	0.352	-0.500	-1.253	-2.234
13	40	60	60	40	0.251	0.251	-0.050	0.603	0.000	0.000	0.000
14	30	70	50	50	0.251	0.241	-0.050	0.593	0.100	0.223	0.405
15	40	80	40	40	0.241	0.241	-0.040	0.563	0.200	0.405	0.811
16	30	120	20	30	0.188	0.188	-0.038	0.452	0.500	1.099	2.197
17	50	50	70	30	0.251	0.241	-0.050	0.593	-0.100	-0.182	-0.405
18	40	40	80	40	0.241	0.241	-0.040	0.563	-0.200	-0.405	-0.811
19	20	20	120	40	0.161	0.211	-0.040	0.452	-0.500	-1.253	-2.234

For given potential outcomes, we draw, repeatedly and independently, the treatment assignment vectors 5000 times, and apply the three methods after obtaining the observed outcomes. We compare three methods: Neymanian inference assuming constant treatment effects, improved Neymanian inference, and Bayesian inference assuming independent potential outcomes.

The results are summarized in Figures B.1, B.2 and B.3, with average biases,

average lengths of the 95% confidence/credible intervals, and the coverage probabilities. The main text only reports the results for CRD and log(COR), here we report the results for all causal measures. When the potential outcomes are independent or positively associated, the results for log(CRR) are similar to those for log(COR) as discussed in the main text. When the potential outcomes are negatively associated, all the interval estimates over cover the true causal measures, while the Bayesian credible intervals are the narrowest.

B.6 More Details about the Application

As in the main text, the example is taken from Bissler et al. (2013), and they compare the rate of adverse events in the treatment group versus the control group. The adverse event naspharyngitis occurred in 19 among 79 subjects in the treatment group with everolimus, and it occurred in 12 among 39 subjects in the control group. Therefore, the 2×2 table representing the observed data has cell counts $(n_{11}, n_{10}, n_{01}, n_{00}) = (19, 60, 12, 27)$. Figure B.4 shows the sensitivity analysis for $\text{CRD}, \log(\text{CRR})$ and $\log(\text{COR})$, with similar patterns for all of them.

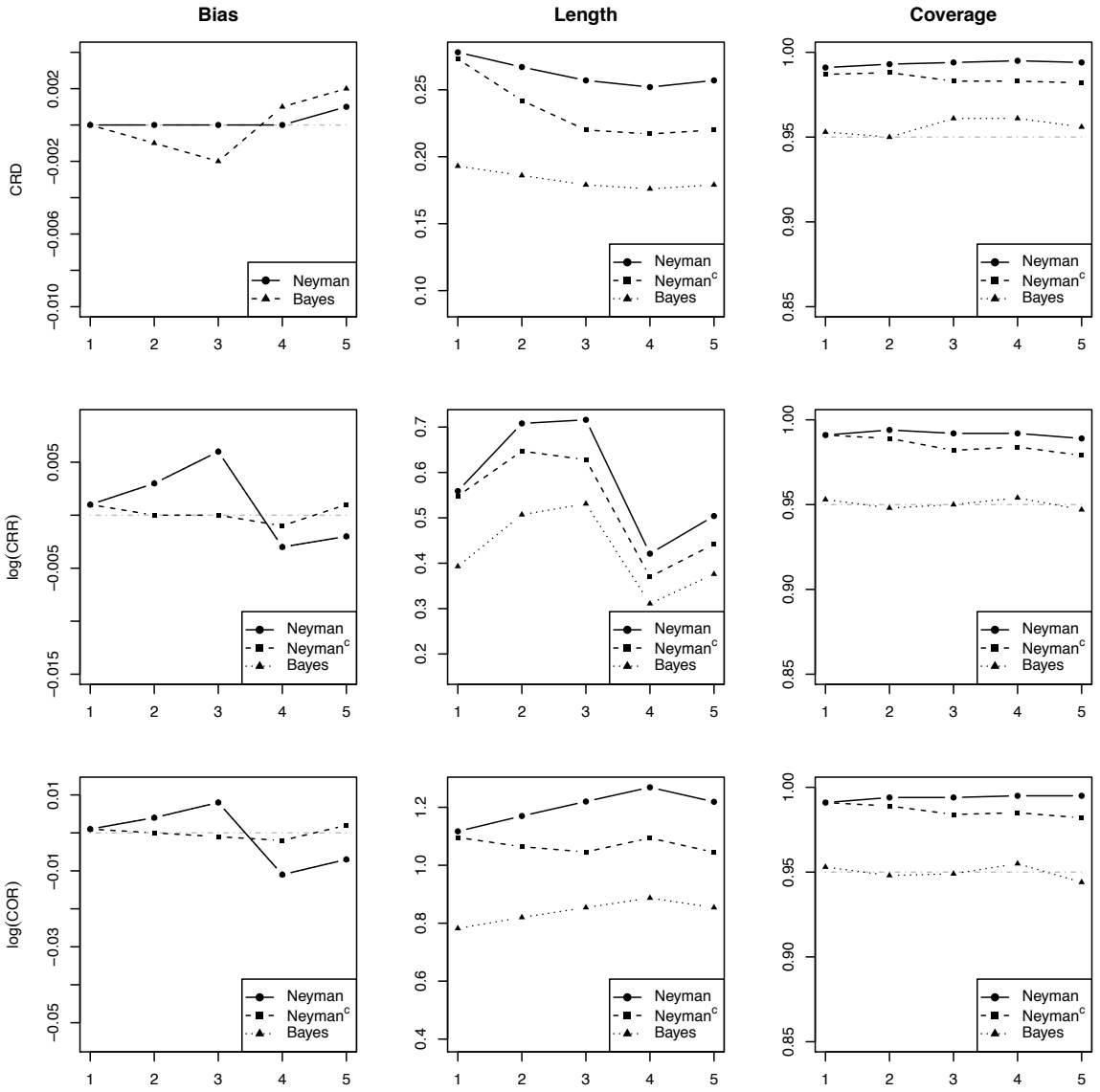


Figure B.1: Simulation Results for Independent Potential Outcomes. Each subfigure is a 2×3 matrix summarizing 3 repeated sampling properties (average bias, average length, and coverage of interval estimates) for 2 causal measures. Note that "Neyman" and "Bayes" are indistinguishable for biases of log(CRR) and log(COR).

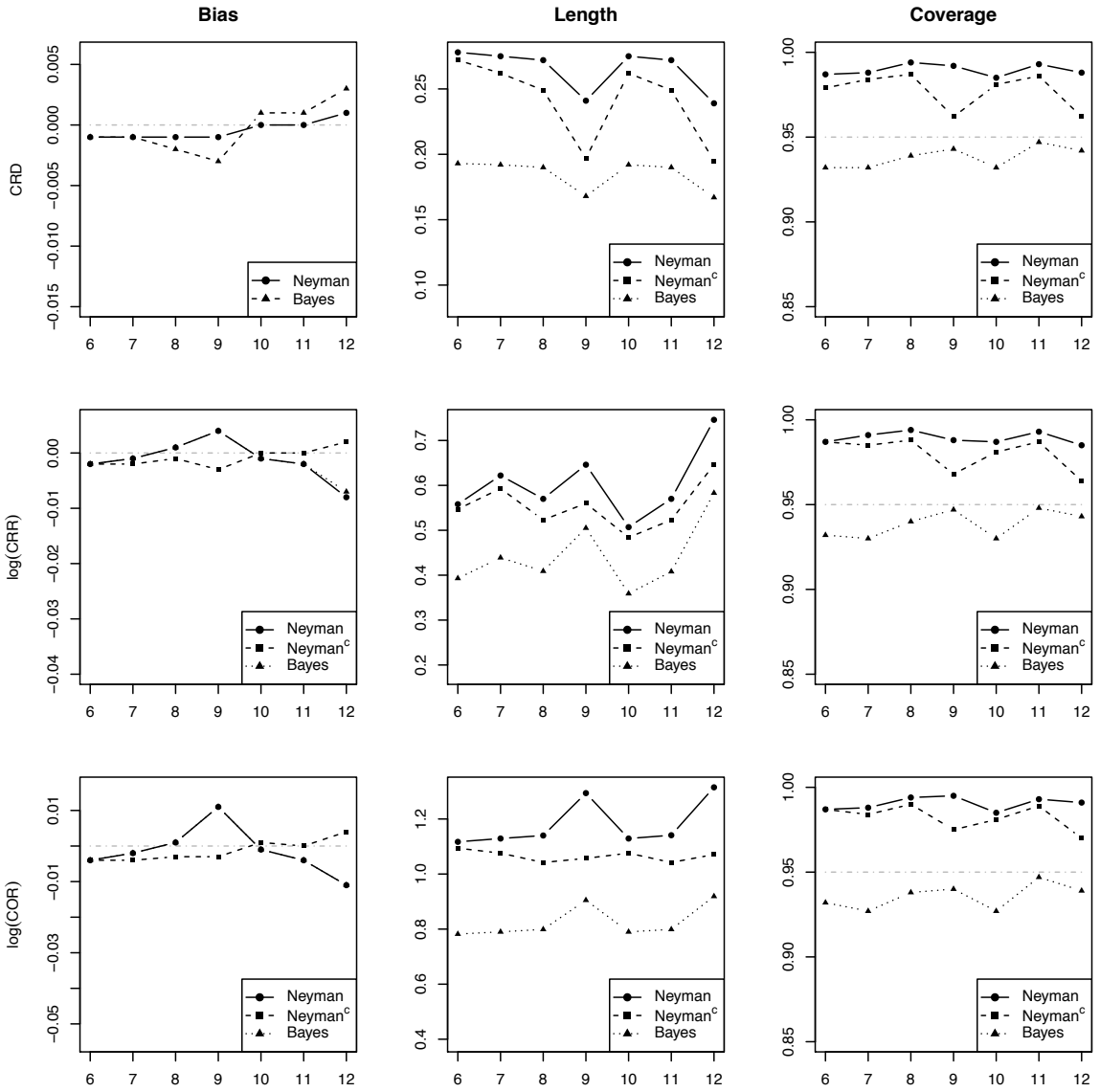


Figure B.2: Simulation Results for Positively Associated Potential Outcomes. Each subfigure is a 2×3 matrix summarizing 3 repeated sampling properties (average bias, average length, and coverage of interval estimates) for 2 causal measures. Note that "Neyman" and "Bayes" are indistinguishable for biases of log(CRR) and log(COR).

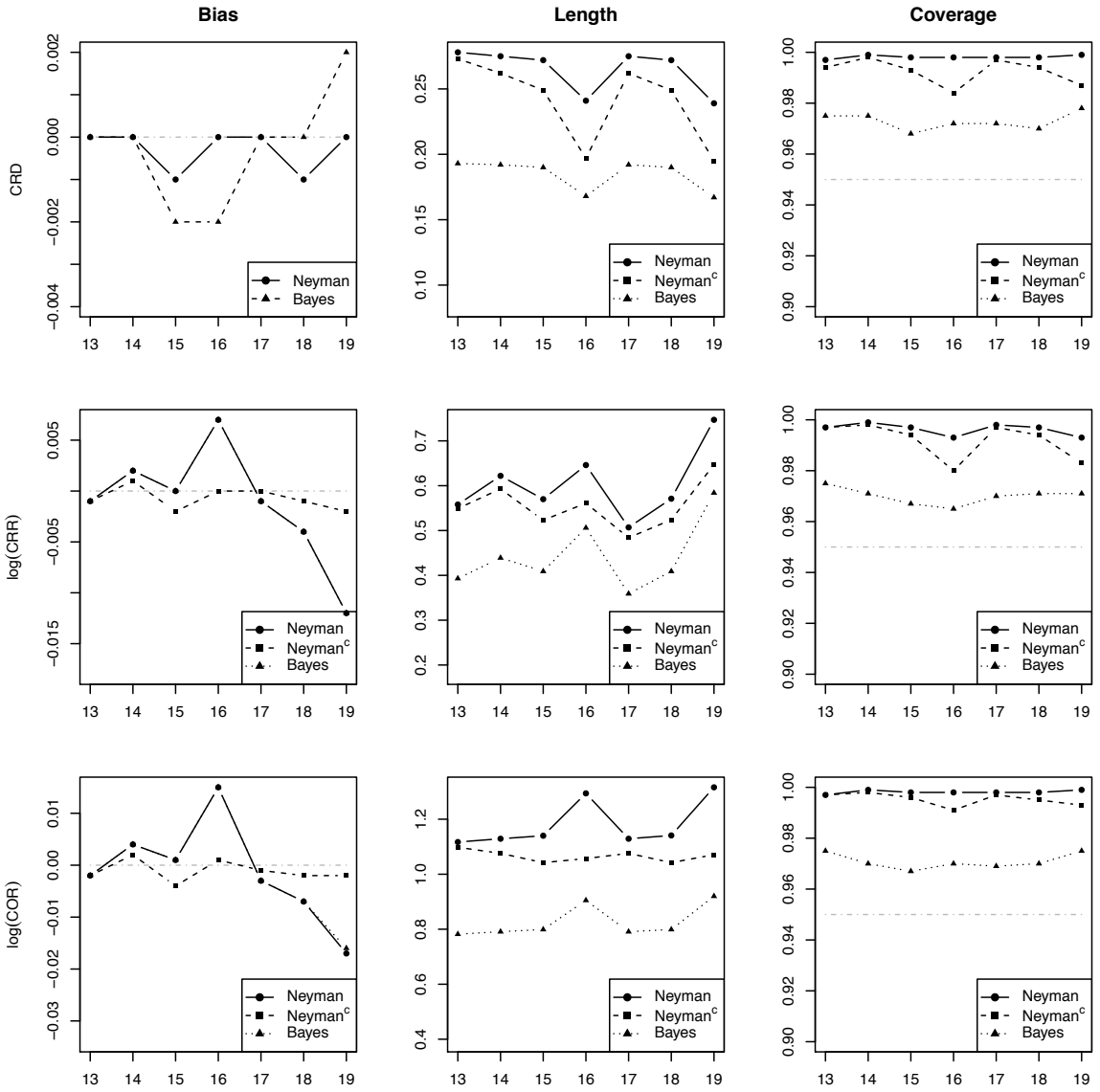


Figure B.3: Simulation Results for Negatively Associated Potential Outcomes. Each subfigure is a 2×3 matrix summarizing 3 repeated sampling properties (average bias, average length, and coverage of interval estimates) for 2 causal measures. Note that "Neyman" and "Bayes" are indistinguishable for biases of log(CRR) and log(COR).

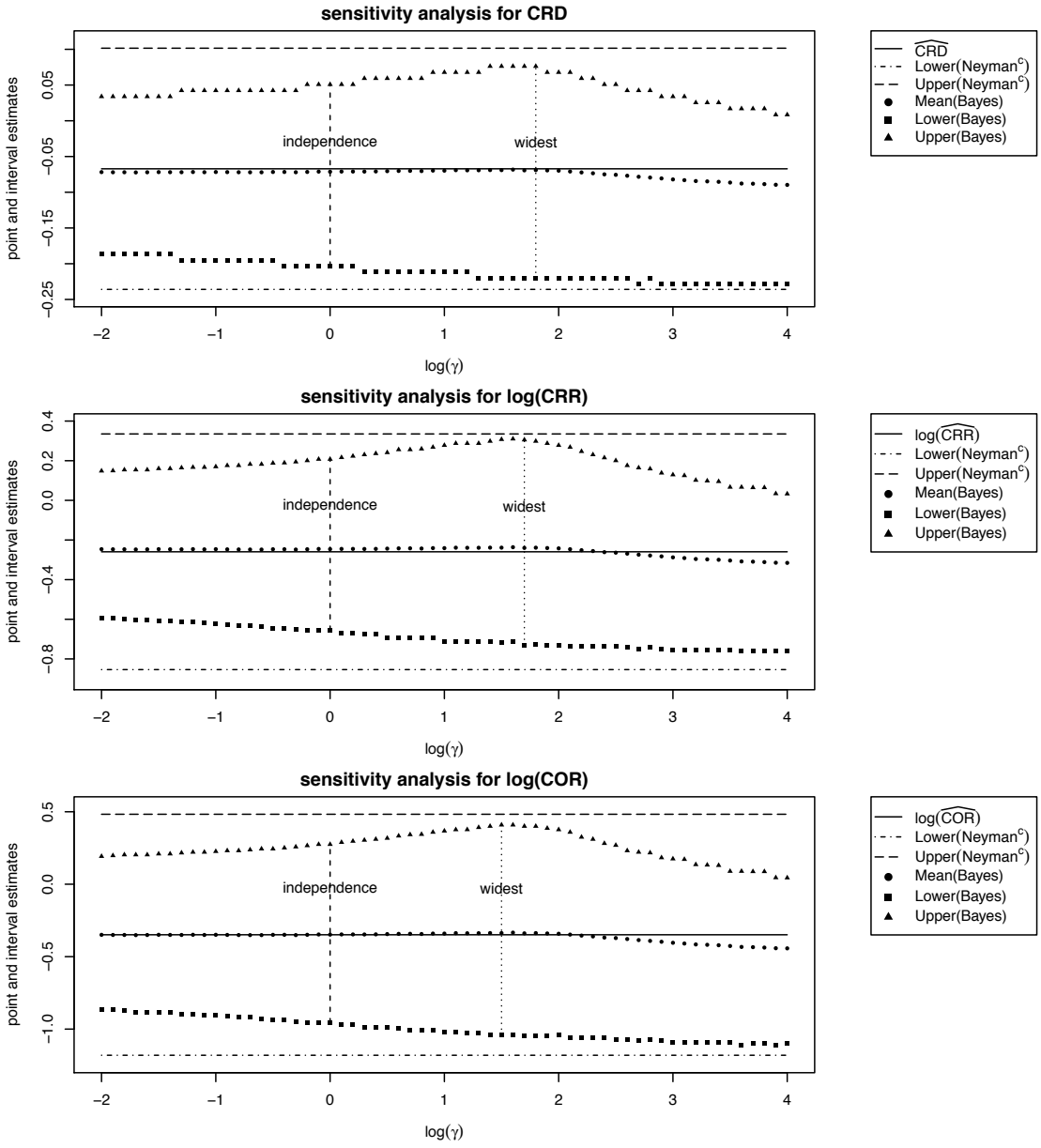


Figure B.4: Bayesian Sensitivity Analysis of the Trial with $(n_{11}, n_{10}, n_{01}, n_{00}) = (19, 60, 12, 27)$. Three panels are for CRD, log(CRR), and log(COR), respectively. The intervals named "independence" are the 95% posterior credible intervals under independence of the potential outcomes, and the intervals named "widest" are the widest 95% credible intervals over the ranges of the sensitivity parameters.

Appendix C

Technical Details for Chapter 3

C.1 Lemmas

In order to prove the theorems, we need the following Lemmas 6 to 8, which are also of independent interest in other contexts. Let $1_n = (1, \ldots, 1)'$ and $0_n = (0, \ldots, 0)'$ be column vectors of length n, and I_n be the $n \times n$ identity matrix. Then $S_n = I_n - n^{-1} 1_n 1_n'$ is the projection matrix orthogonal to 1_n with $S_n 1_n = 0_n$.

Lemma 6. [Covariance of the treatment assignment vector] The completely randomized treatment assignment $T = (T_1, \ldots, T_n)'$ has mean and covariance matrix:

$$E(T) = \frac{n_1}{n} 1_n, \quad \text{cov}(T) = \frac{n_1 n_0}{n(n-1)} S_n.$$

Proof of Lemma 6. The conclusions follow from the facts that $E(T_i) = n_1/n, \text{var}(T_i) = n_1 n_0 / n^2$, and $\text{cov}(T_i, T_j) = -n_1 n_0 / \{n^2(n-1)\}$ for $i \neq j$. □

Lemma 7. [S_n as a covariance operator] If U_i and V_i are column vectors of length

p, define $\mathcal{U} = [U_1, U_2, \ldots, U_n]$ and $\mathcal{V} = [V_1, V_2, \ldots, V_n]$ as two matrices of dimension $p \times n$, and we have

$$\mathcal{U} S_n \mathcal{V}' = \sum_{i=1}^{n} (U_i - \bar{U})(V_i - \bar{V})',$$

where $\bar{U} = \sum_{i=1}^{n} U_i/n$ and $\bar{V} = \sum_{i=1}^{n} V_i/n$. In particular, when $U_i = V_i$, the project matrix S_n reduces to our covariance operator defined in the main text, since

$$\mathcal{V} S_n \mathcal{V}' = (n-1)\mathcal{S}(V).$$

Proof of Lemma 7. The left hand side is equal to

$$\mathcal{U} S_n \mathcal{V}' = \mathcal{U}\mathcal{V}' - n^{-1} (\mathcal{U} 1_n) (\mathcal{V} 1_n)' = \sum_{i=1}^{n} U_i V_i' - n^{-1} (n\bar{U})(n\bar{V}) = \sum_{i=1}^{n} U_i V_i' - n\bar{U}\bar{V},$$

which is the same as the right hand side. □

Lemma 8. [Neymanian randomization inference for vector outcomes] In completely randomized experiments with a vector outcome Z, the Neymanian unbiased estimator for the finite population average treatment effect on Z,

$$\tau_Z = n^{-1} \sum_{i=1}^{n} \{Z_i(1) - Z_i(0)\},$$

is

$$\widehat{\tau}_Z = \bar{Z}_1^{\text{obs}} - \bar{Z}_0^{\text{obs}},$$

where $\bar{Z}_t^{\text{obs}} = \sum_{T_i = t} Z_i/n_t$ is the sample mean of the observed outcomes under treat-

ment arm t. The covariance of $\widehat{\tau}_Z$ is

$$\text{cov}(\widehat{\tau}_Z) = \frac{\mathcal{S}\{Z(1)\}}{n_1} + \frac{\mathcal{S}\{Z(0)\}}{n_0} - \frac{\mathcal{S}\{Z(1) - Z(0)\}}{n}.$$

Proof of Lemma 8. The Neymanian unbiased estimator has the following representation:

$$\begin{aligned}
\widehat{\tau}_Z &= \bar{Z}_1^{\text{obs}} - \bar{Z}_0^{\text{obs}} \\
&= \frac{1}{n_1}\sum_{i=1}^{n} T_i Z_i(1) - \frac{1}{n_0}\sum_{i=1}^{n}(1 - T_i)Z_i(0) \\
&= \sum_{i=1}^{n} T_i \left\{ \frac{Z_i(1)}{n_1} + \frac{Z_i(0)}{n_0} \right\} - \frac{1}{n_0}\sum_{i=1}^{n} Z_i(0).
\end{aligned}$$

The unbiasedness of $\widehat{\tau}_Z$ follows from the linearity of the expectation and Lemma 6. Define $\mathcal{Z}_1 = [Z_1(1), \ldots, Z_n(1)]$ and $\mathcal{Z}_0 = [Z_1(0), \ldots, Z_n(0)]$ as the matrices of the potential outcomes. The estimator $\widehat{\tau}_Z$ can be represented as

$$\widehat{\tau}_Z = \left(\frac{\mathcal{Z}_1}{n_1} + \frac{\mathcal{Z}_0}{n_0} \right) T - \frac{1}{n_0}\sum_{i=1}^{n} Z_i(0).$$

Applying Lemmas 6 and 7, we can obtain the covariance matrix of $\widehat{\tau}_Z$ as:

$$\begin{aligned}
& \operatorname{cov}(\widehat{\tau}_Z) \\
= & \left(\frac{\mathcal{Z}_1}{n_1}+\frac{\mathcal{Z}_0}{n_0}\right)\operatorname{cov}(T)\left(\frac{\mathcal{Z}_1}{n_1}+\frac{\mathcal{Z}_0}{n_0}\right)' \\
= & \frac{n_1 n_0}{n(n-1)}\left(\frac{\mathcal{Z}_1}{n_1}+\frac{\mathcal{Z}_0}{n_0}\right)S_n\left(\frac{\mathcal{Z}_1}{n_1}+\frac{\mathcal{Z}_0}{n_0}\right)' \\
= & \frac{n_1 n_0}{n(n-1)}\left(\frac{1}{n_1^2}\mathcal{Z}_1 S_n \mathcal{Z}_1' + \frac{1}{n_0^2}\mathcal{Z}_0 S_n \mathcal{Z}_0' + \frac{1}{n_1 n_0}\mathcal{Z}_0 S_n \mathcal{Z}_1' + \frac{1}{n_1 n_0}\mathcal{Z}_1 S_n \mathcal{Z}_0'\right) \\
= & \frac{n_0}{n n_1}\mathcal{S}\{Z(1)\} + \frac{n_1}{n n_0}\mathcal{S}\{Z(0)\} + \frac{1}{n(n-1)}(\mathcal{Z}_0 S_n \mathcal{Z}_1' + \mathcal{Z}_1 S_n \mathcal{Z}_0').
\end{aligned}$$

Using the fact $ab' + ba' = aa' + bb' - (a-b)(a-b)'$ for two column vectors a and b, we have

$$\begin{aligned}
& \{Z_i(1)-\bar{Z}(1)\}\{Z_i(0)-\bar{Z}(0)\}' + \{Z_i(0)-\bar{Z}(0)\}\{Z_i(1)-\bar{Z}(1)\}' \\
= & \{Z_i(1)-\bar{Z}(1)\}\{Z_i(1)-\bar{Z}(1)\}' + \{Z_i(1)-\bar{Z}(1)\}\{Z_i(1)-\bar{Z}(1)\}' \\
& -\{Z_i(1)-Z_i(0)-\bar{Z}(1)+\bar{Z}(0)\}\{Z_i(1)-Z_i(0)-\bar{Z}(1)+\bar{Z}(0)\}'.
\end{aligned}$$

Summing over $i = 1, \ldots, n$ and applying Lemma 7, we have

$$\frac{\mathcal{Z}_0 S_n \mathcal{Z}_1'}{n-1} + \frac{\mathcal{Z}_1 S_n \mathcal{Z}_0'}{n-1} = \mathcal{S}\{Z(1)\} + \mathcal{S}\{Z(0)\} - \mathcal{S}\{Z(1)-Z(0)\}.$$

Therefore, the covariance of $\widehat{\tau}_Z$ can be simplified as:

$$\begin{aligned} & \operatorname{cov}(\widehat{\tau}_Z) \\ = \; & \frac{n_0}{nn_1}\mathcal{S}\{Z(1)\} + \frac{n_1}{nn_0}\mathcal{S}\{Z(0)\} + \frac{1}{n}\left[\mathcal{S}\{Z(1)\} + \mathcal{S}\{Z(0)\} - \mathcal{S}\{Z(1) - Z(0)\}\right] \\ = \; & \frac{\mathcal{S}\{Z(1)\}}{n_1} + \frac{\mathcal{S}\{Z(0)\}}{n_0} - \frac{\mathcal{S}\{Z(1) - Z(0)\}}{n}. \end{aligned}$$

□

C.2 Proof of the Theorems

Proof of Theorem 12. Multiplying both sides of $\tau_i = X_i'\beta$ by X_i, we obtain:

$$X_i\tau_i = X_iX_i'\beta.$$

Summing over $i = 1, \ldots, n$, we have

$$\sum_{i=1}^{n} X_i\tau_i = \sum_{i=1}^{n} X_iX_i'\beta,$$

or equivalently,

$$n^{-1}\sum_{i=1}^{n} X_iY_i(1) - n^{-1}\sum_{i=1}^{n} X_iY_i(0) = n^{-1}\sum_{i=1}^{n} X_iX_i'\beta.$$

And therefore, $S_{x1} - S_{x0} = S_{xx}\beta$, implying that

$$\beta = S_{xx}^{-1}(S_{x1} - S_{x0}).$$

The above equations are deterministic under model (3.1), with S_{xx} known directly from the observed data. Since the sample means for $\{X_i Y_i^{\text{obs}} : T_i = t\} = \{X_i Y_i(t) : T_i = t\}$, $\widehat{S}_{xt}$, is unbiased for the population mean S_{xt}, the estimator $\widehat{\beta}_{\text{RI}}$ is also unbiased for β. Its sampling covariance over all possible randomizations is

$$\text{cov}(\widehat{\beta}_{\text{RI}}) = S_{xx}^{-1} \text{cov}(\widehat{S}_{x1} - \widehat{S}_{x0}) S_{xx}^{-1}.$$

Therefore, we only need to determine the covariance of $\widehat{S}_{x1} - \widehat{S}_{x0}$. We can view

$$\widehat{S}_{x1} - \widehat{S}_{x0} = \frac{1}{n_1} \sum_{i=1}^{n} T_i X_i Y_i^{\text{obs}} - \frac{1}{n_0} \sum_{i=1}^{n} (1 - T_i) X_i Y_i^{\text{obs}}$$

as the difference between the sample means of $\{X_i Y_i(1) : i = 1, \ldots, n\}$ and $\{X_i Y_i(0) : i = 1, \ldots, N\}$ under treatment and control. Viewing $X_i Y_i^{\text{obs}}$ as a vector outcome of the completely randomized experiment, we can apply Lemma 8 to obtain the following result:

$$\text{cov}(\widehat{S}_{x1} - \widehat{S}_{x0}) = \frac{\mathcal{S}\{XY(1)\}}{n_1} + \frac{\mathcal{S}\{XY(0)\}}{n_0} - \frac{\mathcal{S}(X\tau)}{n},$$

which completes the proof. □

Proof of Theorem 13. First, we have the population-level ordinary least squares regression matrix of $Y(t)X$ onto W:

$$B_t = S_{ww}^{-1} \left\{ n^{-1} \sum_{i=1}^{n} Y_i(t) W_i X_i' \right\},$$

which is a $J \times K$ matrix and minimizes $\sum_{i=1}^{n} ||Y_i(t)X_i - B_t' W_i||_2^2$ with $||\cdot||_2^2$ being the

L_2-norm. Define $\widetilde{S}_{tx} = \widehat{S}_{tx} + B_t'(\bar{W} - \bar{W}_t)$ and $\widetilde{\beta}^w_{\mathrm{RI}} = S_{xx}^{-1}(\widetilde{S}_{1x} - \widetilde{S}_{0x})$. We first observe that

$$\begin{aligned}
\widehat{\beta}^w_{\mathrm{RI}} - \widetilde{\beta}^w_{\mathrm{RI}} &= S_{xx}^{-1}\left\{(\widehat{B}_1 - B_1)'(\bar{W} - \bar{W}_1) + (\widehat{B}_0 - B_0)'(\bar{W} - \bar{W}_0)\right\} \\
&= S_{xx}^{-1}\left\{O_P(n^{-1/2})O_P(n^{-1/2}) + O_P(n^{-1/2})O_P(n^{-1/2})\right\} \\
&= O_P(n^{-1}),
\end{aligned}$$

based on the same rationale of regression estimator in surveys (Cochran, 1977). Therefore, $\widehat{\beta}_{\mathrm{RI}}$ and $\widetilde{\beta}^w_{\mathrm{RI}}$ have the same asymptotic covariance, and in the following we only need to discuss the covariance of $\widetilde{\beta}^w_{\mathrm{RI}}$. Since

$$\begin{aligned}
\widetilde{S}_{1x} - \widetilde{S}_{0x} &= n_1^{-1}\sum_{i=1}^{n} T_i\left\{Y_i(1)X_i + B_1'(\bar{W} - W_i)\right\} \\
&\quad -n_0^{-1}\sum_{i=1}^{n}(1 - T_i)\left\{Y_i(0)X_i + B_0'(\bar{W} - W_i)\right\}
\end{aligned}$$

can be represented as the difference between the sample means of $e_i(1)$ and $e_i(0)$, applying Lemma 8 we can obtain its variance:

$$\mathrm{cov}\left(\widetilde{S}_{1x} - \widetilde{S}_{0x}\right) = \frac{\mathcal{S}\{e(1)\}}{n_1} + \frac{\mathcal{S}\{e(0)\}}{n_0} - \frac{\mathcal{S}\{\Delta\}}{n},$$

which completes the proof. □

Proof of Theorem 14. Denote $p_1 = n_1/n, p_0 = n_0/n$, and

$$\widehat{S}_{xx,1} = n_1^{-1}\sum_{i=1}^{n} T_i X_i X_i', \quad \widehat{S}_{xx,0} = n_0^{-1}\sum_{i=1}^{n}(1 - T_i)X_i X_i', \quad S_{xy} = n^{-1}\sum_{i=1}^{n} X_i Y_i^{\mathrm{obs}}.$$

Therefore, we have $S_{xx} = p_1\widehat{S}_{xx,1} + p_0\widehat{S}_{xx,0}$ and $S_{xy} = p_1\widehat{S}_{x1} + p_0\widehat{S}_{x0}$. The least square estimators of the regression coefficients are

$$\begin{aligned}
\begin{pmatrix} \widehat{\gamma}_{\text{OLS}} \\ \widehat{\beta}_{\text{OLS}} \end{pmatrix} &= \left\{ n^{-1} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ T_i X_i \end{pmatrix} (X_i', T_i X_i') \right\}^{-1} \left\{ n^{-1} \sum_{i=1}^{n} \begin{pmatrix} X_i \\ T_i X_i \end{pmatrix} Y_i^{\text{obs}} \right\} \\
&= \begin{pmatrix} n^{-1}\sum_{i=1}^n X_i X_i' & n^{-1}\sum_{i=1}^n T_i X_i X_i' \\ n^{-1}\sum_{i=1}^n T_i X_i X_i' & n^{-1}\sum_{i=1}^n T_i X_i X_i' \end{pmatrix}^{-1} \begin{pmatrix} n^{-1}\sum_{i=1}^n X_i Y_i^{\text{obs}} \\ n^{-1}\sum_{i=1}^n X_i Y_i^{\text{obs}} T_i \end{pmatrix} \\
&= \begin{pmatrix} p_1\widehat{S}_{xx,1} + p_0\widehat{S}_{xx,0} & p_1\widehat{S}_{xx,1} \\ p_1\widehat{S}_{xx,1} & p_1\widehat{S}_{xx,1} \end{pmatrix}^{-1} \begin{pmatrix} p_1\widehat{S}_{x1} + p_0\widehat{S}_{x0} \\ p_1\widehat{S}_{x1} \end{pmatrix}.
\end{aligned}$$

We will use the following formula for the inverse of a block matrix

$$\begin{pmatrix} A & B \\ B & B \end{pmatrix}^{-1} = \begin{pmatrix} (A-B)^{-1} & -A^{-1}B(B - BA^{-1}B)^{-1} \\ -(A-B)^{-1} & (B - BA^{-1}B)^{-1} \end{pmatrix}.$$

Take $A = p_1\widehat{S}_{xx,1} + p_0\widehat{S}_{xx,0}$ and $B = p_1\widehat{S}_{xx,1}$, and we can simplify each of the components above as $(A-B)^{-1} = p_0^{-1}\widehat{S}_{xx,0}^{-1}$, $(B - BA^{-1}B)^{-1} = (p_0\widehat{S}_{xx,0})^{-1} + (p_1\widehat{S}_{xx,1})^{-1}$, $-A^{-1}B(B - BA^{-1}B)^{-1} = -(p_0\widehat{S}_{xx,0})^{-1}$. Therefore, the least square estimator can be rewritten as

$$\begin{aligned}
\begin{pmatrix} \widehat{\gamma}_{\text{OLS}} \\ \widehat{\beta}_{\text{OLS}} \end{pmatrix} &= \begin{pmatrix} (p_0\widehat{S}_{xx,0})^{-1} & -(p_0\widehat{S}_{xx,0})^{-1} \\ -(p_0\widehat{S}_{xx,0})^{-1} & (p_0\widehat{S}_{xx,0})^{-1} + (p_1\widehat{S}_{xx,1})^{-1} \end{pmatrix} \begin{pmatrix} p_1\widehat{S}_{x1} + p_0\widehat{S}_{x0} \\ p_1\widehat{S}_{x1} \end{pmatrix} \\
&= \begin{pmatrix} \widehat{S}_{xx,0}^{-1}\widehat{S}_{x0} \\ \widehat{S}_{xx,1}^{-1}\widehat{S}_{x1} - \widehat{S}_{xx,0}^{-1}\widehat{S}_{x0} \end{pmatrix},
\end{aligned}$$

from which we can see that $\widehat{\gamma}_{\text{OLS}}$ can be obtained by running regression of Y^{obs} onto X using the control group data, and $\widehat{\gamma}_{\text{OLS}} + \widehat{\beta}_{\text{OLS}}$ can be obtained by running regression of Y^{obs} onto X using the treatment group data. □

In order to prove Theorem 15, we need to invoke the following Fréchet–Hoeffding inequality (Hoeffding, 1941; Fréchet, 1951; Heckman et al., 1997; Aronow et al., 2014).

Lemma 9. If we only know the marginal distributions $X \sim F_X(x)$ and $Y \sim F_Y(y)$, then $E(XY)$ can be sharply bounded by

$$\int_0^1 F_X^{-1}(u)F_Y^{-1}(1-u)du \leq E(XY) \leq \int_0^1 F_X^{-1}(u)F_Y^{-1}(u)du.$$

Lemma 9 immediately implies the following bound for $\text{var}(X-Y)$ if $E(X-Y)=0$.

Lemma 10. If we know the marginal distributions $X \sim F_X(x), Y \sim F_Y(y)$ and $E(X-Y)=0$, then $\text{var}(X-Y)$ can be sharply bounded by

$$\int_0^1 \{F_X^{-1}(u) - F_Y^{-1}(u)\}^2 du \leq \text{var}(X-Y) \leq \int_0^1 \{F_X^{-1}(u) - F_Y^{-1}(1-u)\}^2 du$$

Proof of Lemma 10. The variance $\text{var}(X-Y)$ can be decomposed as $\text{var}(X-Y) = E(X-Y)^2 = E(X^2) + E(Y^2) - 2E(XY)$, depending on the following three terms:

$$\begin{aligned} E(X^2) = \quad & \textstyle\int x dF_X(x) & &= \int_0^1 F_X^{-1}(u)du, \\ E(Y^2) = \quad & \textstyle\int_0^1 F_Y^{-1}(u)du & &= \int_0^1 F_Y^{-1}(1-u)du, \\ \int_0^1 F_X^{-1}(u)F_Y^{-1}(1-u)du \leq \quad & E(XY) & &\leq \int_0^1 F_X^{-1}(u)F_Y^{-1}(u)du. \end{aligned}$$

Plugging the above expressions into the variance of $X - Y$, we can obtain the desired bounds. □

Applying Lemma 10, we can easily prove Theorem 15.

Proof of Theorem 15. Since $S_{\tau\tau} = S_{\delta\delta} + S_{\varepsilon\varepsilon}$, we only need to bound $S_{\varepsilon\varepsilon}$, which is the variance of

$$\varepsilon_i = \{Y_i(1) - X_i'\gamma_1\} - \{Y_i(0) - X_i'\gamma_0\}$$

for the finite population $i = 1, \ldots, n$. We can identify the marginal distributions of $\{Y_i(1) - X_i'\gamma_1 : i = 1, \ldots, n\}$ and $\{Y_i(0) - X_i'\gamma_0 : i = 1, \ldots, n\}$, and $n^{-1}\sum_{i=1}^n \varepsilon_i = 0$. Therefore, the bounds in Lemma 10 imply the bounds in Theorem 15. □

Proof of Theorem 16. Assume that we have

$$\tau_i = Y_i(1) - Y_i(0) = f(X_i;\beta) + \varepsilon_i, \quad (i = 1, \ldots, n),$$

where $f(X_i;\beta) \equiv f_i$ is a smooth nonlinear function with gradient $\nabla f(X_i;\beta) \equiv \nabla f_i$ and Hessian matrix $\nabla^2 f(X_i;\beta) \equiv \nabla^2 f_i$ with respect to the unknown parameter β. We can obtain $\widehat{\beta}$ by solving

$$\begin{aligned} 0 &= n^{-1}\sum_{i=1}^n m(T_i, Y_i^{\text{obs}}; \widehat{\beta}) \\ &= n_1^{-1}\sum_{i=1}^n T_i Y_i^{\text{obs}} \nabla \widehat{f}_i - n_0^{-1}\sum_{i=1}^n (1 - T_i) Y_i^{\text{obs}} \nabla \widehat{f}_i - n^{-1}\sum_{i=1}^n \widehat{f}_i \nabla \widehat{f}_i, \end{aligned}$$

where $\widehat{f}_i = f(X_i;\widehat{\beta})$ and similar definitions applied to its derivatives. Applying Taylor

expansion, we have

$$\begin{aligned}
& \widehat{\beta} - \beta \\
\approx & \left\{ n_1^{-1} \sum_{i=1}^{n} T_i Y_i^{\text{obs}} \nabla^2 f_i - n_0^{-1} \sum_{i=1}^{n} (1 - T_i) Y_i^{\text{obs}} \nabla^2 f_i \right. \\
& \left. - n^{-1} \sum_{i=1}^{n} \nabla f_i (\nabla f_i)' - n^{-1} \sum_{i=1}^{n} f_i \nabla^2 f_i \right\} \\
& \cdot \left\{ n_1^{-1} \sum_{i=1}^{n} T_i Y_i^{\text{obs}} \nabla f_i - n_0^{-1} \sum_{i=1}^{n} (1 - T_i) Y_i^{\text{obs}} \nabla f_i - n^{-1} \sum_{i=1}^{n} f_i \nabla f_i \right\} \\
\approx & N(0, CDC),
\end{aligned}$$

in distribution. In the above, the term

$$n_1^{-1} \sum_{i=1}^{n} T_i Y_i^{\text{obs}} \nabla^2 f_i - n_0^{-1} \sum_{i=1}^{n} (1 - T_i) Y_i^{\text{obs}} \nabla^2 f_i - n^{-1} \sum_{i=1}^{n} \nabla f_i (\nabla f_i)' - n^{-1} \sum_{i=1}^{n} f_i \nabla^2 f_i$$

has expectation

$$\begin{aligned}
C &= n^{-1} \sum_{i=1}^{n} \left\{ Y_i(1) \nabla^2 f_i - Y_i(0) \nabla^2 f_i - \nabla f_i (\nabla f_i)' - f_i \nabla^2 f_i \right\} \\
&= n^{-1} \sum_{i=1}^{n} \nabla f_i (\nabla f_i)' \equiv -S_{\nabla f \nabla f},
\end{aligned}$$

which is the second order moment matrix of ∇f up to an ignorable negative sign. The term

$$n_1^{-1} \sum_{i=1}^{n} T_i Y_i^{\text{obs}} \nabla f_i - n_0^{-1} \sum_{i=1}^{n} (1 - T_i) Y_i^{\text{obs}} \nabla f_i - n^{-1} \sum_{i=1}^{n} f_i \nabla f_i$$

has sampling variance over all possible randomization:

$$D = \frac{\mathcal{S}\{Y(1)\nabla f\}}{n_1} + \frac{\mathcal{S}\{Y(0)\nabla f\}}{n_0} - \frac{\mathcal{S}(\tau\nabla f)}{n},$$

which is the sampling variance of the treatment on the vector outcome $Y\nabla f$ according to Lemma 8. □

Bibliography

A. Agresti and Y. Min. Effects and non-effects of paired identical observations in comparing proportions with binary matched-pairs data. *Statistics in Medicine*, 23: 65–75, 2004.

J. D. Angrist and J. Pischke. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton: Princeton University Press, 2008.

J. D. Angrist, G. W Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91:444–455, 1996.

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

R. L. Berger and D. D. Boos. P values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association*, 89:1012–1016, 1994.

J. Berkson. Do the marginal totals of the 2×2 table contain relevant information respecting the table proportions? *Journal of Statistical Planning and Inference*, 2: 43–44, 1978.

A. Berrington de González and D. R. Cox. Interpretation of interaction: A review. *The Annals of Applied Statistics*, 1:371–385, 2007.

J. J. Bissler, J. C. Kingswood, E. Radzikowska, et al. Everolimus for angiomyolipoma associated with tuberous sclerosis complex or sporadic lymphangioleiomyomatosis (exist-2): a multicentre, randomised, double-blind, placebo-controlled trial. *The Lancet*, 381:817–824, 2013.

M. P. Bitler, J. B. Gelbach, and H. W. Hoynes. Can variation in subgroups' average treatment effects explain treatment effect heterogeneity? Evidence from a social experiment. `http://www.socsci.uci.edu/ mbitler/papers/bgh-subgroups-paper.pdf`, 2010. Working Paper.

G. E. P. Box. Teaching engineers experimental design with a paper helicopter. *Quality Engineering*, 4, 1992.

H. Chernoff. Information for testing the equality of two probabilities, from the margins of the 2× 2 table. *Journal of Statistical Planning and Inference*, 121:209–214, 2004.

W. G. Cochran. *Sampling Techniques*. New York: John Wiley & Sons, 1977.

J. B. Copas. Randomization models for the matched and unmatched 2×2 tables. *Biometrika*, 60:467–476, 1973.

J. Cornfield. Principles of research. *American Journal of Mental Deficiency*, 64: 240–252, 1959.

D. R. Cox. *Planning of Experiments.* New York: Wiley, 1958.

D. R. Cox. Interaction. *International Statistical Review*, 52:1–24, 1984.

R. K. Crump, V. J. Hotz, G. W. Imbens, and O. A. Mitnik. Nonparametric tests for treatment effect heterogeneity. *Review of Economics and Statistics*, 90:389–405, 2008.

T. Dasgupta, N. S. Pillai, and D. B. Rubin. Causal inference from 2^k factorial designs using the potential outcomes model. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, page in press, 2015.

P. Ding. A paradox from randomization-based causal inference. *arXiv preprint arXiv:1402.0142*, 2014.

P. Ding and T. Dasgupta. A potential tale of two by two tables from completely randomized experiments. *Journal of the American Statistical Association*, page in press, 2015.

P. Ding, A. Feller, and L. W. Miratrix. Randomization inference for treatment effect variation. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, page in press, 2015.

H. Djebbari and J. Smith. Heterogeneous impacts in PROGRESA. *Journal of Econometrics*, 145:64–80, 2008.

K. R. Eberhardt and M. A. Fligner. A comparison of two tests for equality of two proportions. *The American Statistician*, 31:151–155, 1977.

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

R. A. Fisher. Comment on "Statistical problems in agricultural experimentation". *Supplement to the Journal of the Royal Statistical Society*, pages 154–157, 173, 1935b.

R. A. Fisher. The logic of inductive inference. *Journal of the Royal Statistical Society*, pages 39–82, 1935c.

M. Fréchet. Sur les tableaux de corrélation dont les marges son données. *Ann Univ Lyon Sect A*, 9:53–77, 1951.

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

G. L. Gadbury. Randomization inference and bias of standard errors. *The American Statistician*, 55:310–313, 2001.

M. H. Gail, S. D. Mark, R. J. Carroll, S. B. Green, and D. Pee. On design considerations and randomization-based inference for community intervention trials. *Statistics in Medicine*, 15:1069–1092, 1996.

S. Greenland. On the logical justification of conditional tests for two-by-two contingency tables. *The American Statistician*, 45:248–251, 1991.

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

J. J. Heckman, J. Smith, and N. Clements. Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, 64:487–535, 1997.

K. Hinkelmann and O. Kempthorne. *Design and Analysis of Experiments, Volume 1, Introduction to Experimental Design, 2nd Edition*. New York: John Wiley & Sons, 2007.

W. Hoeffding. Masstabinvariante korrelationsmasse für diskontinuierliche verteilungen. *Arkiv fr matematischen Wirtschaften und Sozialforschung*, pages 49–70, 1941.

P. J. Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 221–233, 1967.

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

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

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

O. Kempthorne. The randomization theory of experimental inference. *Journal of the American Statistical Association*, 50:946–967, 1955.

E. L. Lehmann. *Elements of Large-Sample Theory*. New York: Springer, 1998.

E. L. Lehmann and J. P. Romano. *Testing Statistical Hypotheses*. New York: Springer, 2006.

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

L. W. Miratrix, J. S. Sekhon, and B. Yu. Adjusting treatment effect estimates by post-stratification in randomized experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:369–396, 2013.

J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–472, 1923.

J. Neyman. Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236:333–380, 1937.

J. Neyman and K. Iwaszkiewicz. Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, pages 107–180, 1935.

C. O'Muircheartaigh and L. V. Hedges. Generalizing from unrepresentative experiments: a stratified propensity score approach. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63:195–210, 2014.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, pages 119–130, 1937.

E. J. G. Pitman. Significance tests which may be applied to samples from any populations: III. the analysis of variance test. *Biometrika*, pages 322–335, 1938.

H. Robbins. A fundamental question of practical statistics. *The American Statistician*, 31:97, 1977.

J. M. Robins. Confidence intervals for causal parameters. *Statistics in Medicine*, 7: 773–785, 1988.

P. R. Rosenbaum. *Observational Studies*. New York: Springer, 2002.

P.R. Rosenbaum and D. B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

K. J. Rothman, S. Greenland, and T. L. Lash. *Modern Epidemiology*. Philadelphia, PA.: Lippincott Williams & Wilkins, 2008.

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

D. B. Rubin. Assignment to treatment group on the basis of a covariate. *Journal of Educational and Behavioral statistics*, 2:1–26, 1977.

D. B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, pages 34–58, 1978.

D. B. Rubin. Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *Journal of the American Statistical Association*, 75:591–593, 1980.

D. B. Rubin. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 12:1151–1172, 1984.

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

D. B. Rubin. Teaching statistical inference for causal effects in experiments and observational studies. *Journal of Educational and Behavioral Statistics*, 29:343–367, 2004.

D. B. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322–331, 2005.

D. B. Rubin. Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). *Psychological Methods*, pages 38–40, 2010.

D. B. Rubin. Potential updates to Cornfield's 1959 "Principles of Research". *Statistics in Medicine*, 31:2778–2779, 2012.

A. Sabbaghi and D. B. Rubin. Comments on the Neyman–Fisher controversy and its consequences. *Statistical Science*, 29:267–284, 2014.

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

P. Z. Schochet. Is regression adjustment supported by the Neyman model for causal inference? *Journal of Statistical Planning and Inference*, 140:246–259, 2010.

H. White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, pages 817–838, 1980.

B. Woolf. On estimating the relation between blood group and disease. *Annals of Human Genetics*, 19:251–253, 1955.

C. F. J. Wu and M. S. Hamada. *Experiments: Planning, Analysis, and Optimization*. New York: John Wiley & Sons, 2009.

F. Yates. *The design and analysis of factorial experiments*. Harpenden: Imperial Bureau of Soil Science, 1937.

F. Yates. Tests of significance for 2× 2 contingency tables. *Journal of the Royal Statistical Society. Series A (General)*, pages 426–463, 1984.