Robust Bayesian inference of causal effects via randomization distributions*

Easton Huch†
Department of Statistics, University of Michigan, Ann Arbor, Michigan, USA
Walter Dempsey‡
Department of Biostatistics, University of Michigan, Ann Arbor, Michigan, USA
Fred Feinberg†
Ross School of Business, University of Michigan, Ann Arbor, Michigan, USA

June 8, 2024

This is a working paper. The most recent version can be downloaded from the first author's website: eastonhuch.com.

Abstract

We present a general framework for Bayesian inference of causal effects that delivers provably robust inferences based principally on the physical act of randomization. The framework involves fixing the observed potential outcomes and forming a likelihood based on the randomization distribution of a model-based discrepancy variable, a summary of the (imputed) complete data. We show posterior consistency of the method and derive theoretical connections to common estimators in causal inference. We evaluate the method's performance in several simulation studies and an applied data analysis.

Keywords: Bayesian methods, causal inference, design-based inference, randomization, robustness

1 Introduction

Randomization-based causal inference methods offer the promise of valid statistical inferences based solely on the physical act of randomization. Randomization-based methods encompass (a) Fisherian randomization tests (FRTs) and (b) Neymanian inferences grounded in repeated-sample evaluations. Both sets of randomization-based methods have been used

*Work in progress.
†Corresponding author. Email: ekhuch@umich.edu.
‡Equal contributions as doctoral advisors.

extensively across a variety of scientific domains. In addition to reducing the assumptions required for inference, randomization-based methods are also appealing in that they position the assignment mechanism as the conceptual focal point of the analysis, facilitating discussion of covariate balance and the risk of hidden confounders—two central issues in applied causal analysis.

Most Bayesian causal inference methods, in contrast, rely principally on correct specification of outcome models with the assignment mechanism (typically independent, subject-specific 'propensity scores') playing a subtler role. In fact, under prior independence of the parameters for the assignment mechanism and the outcome model, the assignment mechanism drops out of the likelihood—a phenomenon that has generated considerable debate in the literature. The 'ignorability' of the assignment mechanism in these cases has important implications for the robustness of Bayesian causal inference methods. In particular, Bayesian methods tend to be more sensitive to correct specification of outcomes models than their Frequentist counterparts because the propensity score does not (in general) balance subject characteristics between treatment and control groups—its primary purpose in most Frequentist methods.

Although Bayesian statisticians largely agree that the assignment mechanism is an important component of a causal analysis, a recent review of Bayesian causal inference concluded that "there is no consensus on how to proceed" (Li et al., 2023). Existing strategies include (a) treating the propensity score as a covariate in the outcome model, (b) specifying priors with dependence between the propensity score and outcome model parameters, and (c) computing frequency-based point estimators over samples from Bayesian posterior predictive distributions. However, these strategies are not universally applicable and raise challenging questions regarding trade-offs among competing analytical priorities, such as robustness to model misspecification, valid uncertainty quantification, and philosophical coherence.

Setting this challenge aside, Li et al. (2023) argues that the Bayesian approach offers several compelling advantages for causal inference, which we summarize below. First, because the Bayesian approach imputes all unobserved potential outcomes, it can be applied to any causal estimand, even those that are only partially identified, such as individual treatment effects. Second, Bayesian inferences are automatic in the sense that the inferences—including uncertainty quantification—flow directly from the probabilistic assumptions; the resulting posterior distributions can then be leveraged in decision-theoretic analysis to optimize decision making under arbitrary loss/utility functions. Third, Bayesian inferences offer a simple, straightforward solution for incorporating prior information and pooling inferences across multiple sources of information. Fourth, Bayesian methods are highly extensible and modular. In most cases, the generalization from simple parametric models to flexible non-parametric methods (Gaussian processes, mixture models, etc.) is conceptually straightforward.

By placing the assignment mechanism at the center of a Bayesian causal analysis, our proposed framework inherits both the robustness of randomization-based methods *and* the benefits of the Bayesian paradigm listed above. We name the resulting framework *Bayesian Randomization Inference* (BRI) to emphasize the combination of these complementary strengths. The key idea underlying BRI is to condition on the values of the *observed* potential outcomes. We then form a discrepancy variable that involves model-based imputations of counterfactuals, and we use its randomization distribution as a likelihood function.

Because BRI combines Fisherian, Neymanian, and Bayesian ideas, it represents somewhat of a compromise among historically distinct (often opposing) statistical paradigms. Consequently, in addition to its methodological contribution, BRI also occupies a position of historical and philosophical interest in the field of statistics, especially in the analysis of experiments. To place BRI in appropriate context, we briefly review the related historical developments and discussions in Section 2. Section 3 introduces the basic framework for BRI. Section 4 provides methodological details for nonlinear discrepancy variables, discrete treatments, and how to implement BRI with existing Bayesian software. Section 5 provides results regarding the frequentist properties of BRI, including posterior consistency, asymptotic equivalence of certain maximum *a posteriori* (MAP) estimators to common Frequentist estimators, and asymptotic posterior variance calculations. Section 6 details straightforward extensions of BRI. Section 7 concludes with a discussion of the main results, limitations, and promising future directions.

2 Historical relevance

In this section, we briefly review early controversies surrounding randomization-based inference; then we turn our attention to recently proposed Bayesian procedures that can be viewed as precursors to BRI.

2.1 Early controversies

Neyman introduced the concept of potential outcomes in his Master's thesis in 1923; however, his work in this area only became well-known in the statistics community following its posthumous publication in 1990 (Neyman et al., 1990). Fisher developed FRTs during nearly the same time period, leading to their 1935 publication in his book, *The Design of Experiments* (Fisher, 1960).

On their face, the approaches seem largely compatible in that both condition on potential outcomes. In Neyman's case, the approach requires conditioning on the full set of potential outcomes, $\mathcal{F} := \{y_{01}, y_{11}, \ldots, y_{0n}, y_{1n}\}$ (where n is the sample size), and testing the *weak* null hypothesis of no effect on average ($\bar{y}_0 = \bar{y}_1$) via a conservative variance estimator. FRTs, on the other hand, are designed to test the *sharp* null hypothesis of no causal effect for any unit; i.e., $y_{0i} = y_{1i}$ for all i. Because the sharp null implies values for the counterfactuals, it suffices to condition on only those potential outcomes that have been observed. Inference is then performed by comparing an observed statistic to simulated values from its randomization distribution under the strong null.

Despite the apparent similarity of their respective methods, Fisher and Neyman were lifelong opponents of each other's statistical philosophy. The controversy stemmed largely from differing views on model specification and *inductive inference* (Fisher) vs. *inductive behavior* (Neyman). In short, Fisher advocated for significance testing in the process of model specification without reference to an alternative hypothesis, and he viewed the resulting p-values as measures of evidence that provided information even for a single experiment. In contrast, Neyman's later work with Egon Pearson emphasized prespecified models, competing hypotheses, and long-run error frequencies for decision rules. For more complete accounts of

the controversy, we refer interested readers to Lehmann (1993) and Lenhard (2006).

Sabbaghi and Rubin (2014) argue that the controversy caused Fisher to neglect the framework of potential outcomes, which may partially explain its delayed emergence into mainstream statistics more than 50 years later. Despite Neyman's and Fisher's early disagreements, their methods for the analysis of experiments (and other areas of statistics) have since been combined into what Hubbard and Bayarri (2003) call "an anonymous hybrid of [their] competing and frequently contradictory approaches." In a similar vein, Ding (2017) points out that FRTs and Neymanian inference are often introduced in proximity to each other early in causal inference textooks and courses with limited attention paid to the differences between their respective null hypotheses. Thus, the Fisherian and Neymanian perspectives—despite their early disagreements—have largely been unified in modern statistical methodology under the heading of finite-population causal inference methods.

In contrast, Bayesian methods for causal inference have predominantly focused on superpopulation models for potential outcomes according to the prescription given in Rubin (1978). Historically, many Bayesians have been reticent to adopt FRTs and other non-likelihood-based methods (e.g., Basu, 1980). Others Bayesians, including Rubin, have argued that FRTs are logically coherent but limited to the "rare situation" of assessing point hypotheses that are credible *a priori* (Rubin, 1980). For their part, Neyman and Fisher both opposed Bayesianism, viewing it as unnecessarily subjective (Berger, 2003, p. 3); though, Fisher's emphatic advocacy of 'fiducial probability' demonstrates that he recognized the value of solving the 'inverse problem' (Zabell, 1992).

2.2 Recent developments

More recently, recognizing the benefits of 'design-based' methods, statisticians have begun developing Bayesian randomization-based procedures in specific settings. Most of these procedures require bounded outcomes (Humphreys and Jacobs, 2015; Keele and Quinn, 2017; Chiba, 2018; Ding and Miratrix, 2019). To our knowledge, the only exception is the approach of Leavitt (2023), which we discovered in preparing our manuscript. Leavitt's approach is a special case of our method with a binary treatment, a constant treatment effect model, and the difference-in-means statistic. One notable difference is that Leavitt's approach is based on a Gaussian 'working model' with a robust plug-in variance estimate, but ours involves a purely model-based randomization distribution. In large samples, our approach typically results in an approximately Gaussian randomization distribution, but the variance could be 'wrong' in the Frequentist sense if the treatment effect model is misspecified. Although we do not treat Leavitt's proposed plug-in solution in generality, this strategy could also be applied within our framework. We do not pursue Leavitt's approach because BRI offers two fully Bayesian alternatives.

The first alternative is to perform model checking and generalize the assumed causal model when we detect deviations from its assumptions (Rubin, 1984; Meng, 1994; Gelman et al., 1996). Specifically, this strategy involves performing an FRT for each posterior sample, effectively generating a 'posterior predictive p-value.' Ding and Li (2018) detail this procedure under a standard superpopulation model for potential outcomes. This procedure is particularly natural for BRI because the inference procedures already require computation of the randomization distribution of a discrepancy variable.

The second alternative is to construct specially designed discrepancy variables that provide asymptotically valid inference (in the Frequentist sense) regardless of misspecification. In Section 5.4, we show that certain studentized discrepancies have this property, a Bayesian analog to a recent line of work in Frequentist causal inference showing that certain studentized statistics can produce FRTs that are exact under sharp nulls and asymptotically valid under weak nulls (Ding, 2017; Loh et al., 2017; Ding and Dasgupta, 2018; Wu and Ding, 2021; Fogarty, 2020).

3 Proposed framework

This section introduces the general framework for BRI, including the problem setup, the use of discrepancy variables, and the structure of the probability models.

3.1 Problem setup and notation

Throughout we use lowercase unbolded characters for scalars (a, θ), lowercase bold characters for vectors $(\boldsymbol{a}, \boldsymbol{\theta})$, and uppercase bold characters for matrices $(\mathbf{A}, \boldsymbol{\Theta})$. Because all quantities are potentially random in the Bayesian approach, we do not distinguish between random and fixed (i.e., in the conditioning set) quantities in the notation, but we clarify this distinction as needed.

We denote the treatment assignment by $a \in \mathcal{A} \subseteq \mathbb{R}$. We initially consider continuous-valued treatments to clarify the framework, but we return to the common setting of discrete treatments in Section 4.2. We assume the existence of potential outcomes, $\{y_{ai}\}_{a\in\mathcal{A}}$ for all $i \in [n] := \{1, 2, \ldots, n\}$. Due to the *fundamental problem of causal inference*, we observe only a single potential outcome, y_{ai}, for each observation (Holland, 1986). We use $\boldsymbol{y_a}$ to denote the full vector of observed outcomes. At times, we employ similar notation for all potential outcomes under a single treatment assignment (e.g., $\boldsymbol{y}_0$). We employ the following standard causal assumptions:

Assumption 1. (*Consistency*) The observed outcomes, $\boldsymbol{y}$, are equal to the potential outcomes under the observed treatment assignment; i.e., $\boldsymbol{y} = \boldsymbol{y_a}$.

Assumption 2. (*Known Assignment Mechanism*) The treatments, $\boldsymbol{a}$, are randomly assigned according to a known stochastic assignment mechanism.

Assumption 3. (*Unconfoundedness*) The treatment assignments are independent of the potential outcomes: $\boldsymbol{a} \perp\!\!\!\perp \{\boldsymbol{y}_{a'}\}_{a'\in\mathcal{A}}$.

Later in the paper, we generalize Assumption 3 to *conditional unconfoundedness* given a matrix of pre-treatment covariates, $\mathbf{X} \in \mathbb{R}^{n\times q}$. BRI also requires a model for the causal effect of treatment. To ease the initial exposition, we make the following simplifying assumption:

Assumption 4. (*Deterministic Treatment Effects*) Given the parameter vector, $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{y_a}$ is a deterministic function of $\boldsymbol{y_{a'}}$ for all $\boldsymbol{a}, \boldsymbol{a'} \in \mathcal{A}$.

The extension beyond Assumption 4 is straightforward and discussed in detail in Section 6.1. A simple example of Assumption 4 is the deterministic linear model $\boldsymbol{y_a} = \boldsymbol{y}_0 + \boldsymbol{a}\theta$.

3.2 Discrepancy Variables

A *discrepancy variable* (or simply *discrepancy*) generalizes the definition of a statistic to allow dependence on parameters in addition to data. This generalization is quite natural in the Bayesian paradigm because both data and parameters are viewed as random variables. In Bayesian model checking, discrepancies typically measure deviations from modeling assumptions, as illustrated in the following example.

Example 1. *Suppose we conduct a Bayesian analysis for outcomes, z_i, under the assumption of independent and identically distributed (iid) Gaussian random variables:*

$$z_i \overset{iid}{\sim} Normal(\mu, 1). \tag{1}$$

Under this analysis, a natural choice for a discrepancy variable is

$$d(\mu) = \left| \frac{1}{n} \sum_{i=1}^{n} (z_i - \mu)^3 \right|,$$

which should be small under modeling assumption (1) *because the Gaussian distribution is symmetric. Consequently, values of $d(\mu)$ much larger than values simulated from the posterior predictive distribution (averaging over the posterior distribution of μ) provide evidence against* (1) *and suggest that the analyst should replace it with an asymmetric likelihood. We use the notation $d(\mu)$ to emphasize that the discrepancy is a function of μ.*

To see how we apply discrepancies within BRI, consider that Assumption 4 allows us to impute counterfactuals conditional on $\boldsymbol{\theta}$. Thus, within a posterior sampling algorithm, we can obtain samples (a mixture of pure observables and imputations) for all elements in $\boldsymbol{y}_{a'}$ for any fixed $a' \in \mathcal{A}$. In turn, Assumption 3 implies these values must be independent of $\boldsymbol{a}$. Consequently, if we specify a discrepancy, $\boldsymbol{d}(\boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}^q$, that measures statistical dependence between $\boldsymbol{y}_{a'}$ and $\boldsymbol{a}$, Assumptions 3 and 4 imply we should observe a value of $\boldsymbol{d}(\boldsymbol{\theta})$ that indicates little dependence relative to the known distribution of $\boldsymbol{a}$ (Assumption 2). Within the Frequentist paradigm, similar reasoning justifies the application of an FRT under a sharp null hypothesis, $H_{\boldsymbol{\theta}}$. Before exploring the solution suggested by BRI, we first explore the FRT analog in the example below.

Example 2. *Consider a randomized clinical trial (RCT) designed to estimate the effect of a medication dosage, a_i, on some continuous outcome, y_{ai}. For simplicity, we assume that $a_i \overset{iid}{\sim} Uniform(0, 1)$ and $\boldsymbol{y}_{\boldsymbol{a}} = \boldsymbol{y}_0 + \boldsymbol{a}\theta$. Now consider the null hypothesis $H_0 : \theta = 1$. Under H_0, we can impute $\boldsymbol{y}_0$ as $\boldsymbol{y}_0(\theta) = \boldsymbol{y}_a - \boldsymbol{a}\theta$ with $\theta = 1$. Assumption 3 then implies that $\boldsymbol{y}_0(\theta)$ is independent of $\boldsymbol{a}$, so conditional on $\boldsymbol{y}_{\boldsymbol{a}}$, the following statistic has an expectation approximately equal to zero:*

$$s(\theta) = \widehat{\text{Cov}}\{\boldsymbol{y}_0(\theta), \boldsymbol{a}\} / \widehat{\text{Var}}(\boldsymbol{a}).$$

Note that $s(\theta)$ is the OLS regression coefficient of $\boldsymbol{y}_0(\theta)$ on $\boldsymbol{a}$. Conceptually, $s(\theta)$ should be small for the true value of θ because the modeling assumptions imply that $\boldsymbol{a}$ contains no information that can be used to predict $\boldsymbol{y}_0$. The FRT then proceeds by simulating (or

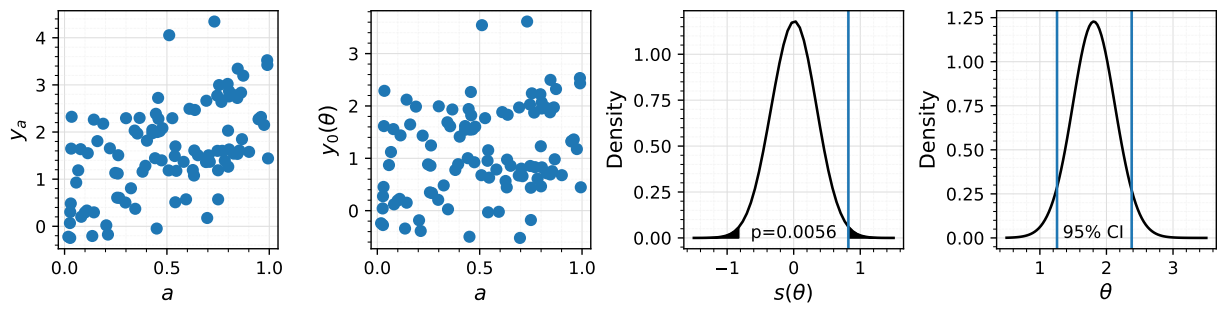

Figure 1: (Far left) Scatter plot showing linear relationship between $\boldsymbol{a}$ and $\boldsymbol{y_a}$ in Example 2. (Middle left) Scatter plot showing relationship between $\boldsymbol{a}$ and $\boldsymbol{y_0}(\theta)$ at $\theta = 1$. (Middle right) Randomization distribution at $\theta = 1$; the low p-value indicates that the FRT rejects H_0. (Far right) The density of the discrepancy (imputed statistic) in black with the 95% Frequentist confidence interval (from test inversion) in blue.

analytically approximating) the randomization distribution of $s(\theta)$ and calculating a p-value that quantifies the extremeness of $s(\theta)$, such as its absolute value. Figure 1 summarizes the analysis. Setting $\theta = 1$ and imputing $\boldsymbol{y}_0(\theta)$ removes some of the correlation between $\boldsymbol{y_a}$ and $\boldsymbol{a}$ in the first two panels. However, the third panel shows that the observed statistic value is still extreme relative to the randomization distribution (p=0.0056), resulting in a rejection of H_0. The fourth panel compares the 95% confidence interval from inverting the FRT to the density of $s(\theta)$ as a function of θ.

The Frequentist paradigm views $s(\theta)$ as a statistic because θ is fixed under H_0. Within BRI, however, θ is an unknown parameter with a plausible range of values, so the term 'discrepancy' is more appropriate. When conducting the FRT, a value of θ is deemed plausible if it implies a statistic that is not 'too unlikely,' as quantified by the p-value. The reasoning underlying BRI is similar with one crucial exception: the plausibility of θ is quantified in terms of a *density*—not a p-value. The latter two panels of Figure 1 hint that we can, in fact, interpret the randomization distribution for $s(\theta)$ as a likelihood and obtain valid Bayesian inferences *without* specifying a likelihood for the potential outcomes; the next section provides the setup.

3.3 Model structure

The goal of BRI is to calculate the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{y_a}$ and $\boldsymbol{d}(\boldsymbol{\theta})$. The probability models take the following form:

$$p\{\boldsymbol{\theta}|\boldsymbol{y_a}, \boldsymbol{d}(\boldsymbol{\theta})\} \propto p(\boldsymbol{\theta}|\boldsymbol{y_a})p\{\boldsymbol{d}(\boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y_a}\}j(\boldsymbol{\theta}), \tag{2}$$

where $p(\boldsymbol{\theta}|\boldsymbol{y_a})$ plays the role of prior and $p(\boldsymbol{d}(\boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y_a})$, of likelihood. Both of these densities condition on $\boldsymbol{y_a}$. The final term, $j(\boldsymbol{\theta})$, is a Jacobian area adjustment term defined in Section 4.1; when $\boldsymbol{d}(\boldsymbol{\theta})$ is linear, $j(\boldsymbol{\theta})$ is constant so it drops out of (2).

The observed potential outcomes, $\boldsymbol{y_a}$, on their own provide limited information about the treatment effect without knowledge of the assignment vector, $\boldsymbol{a}$. Under the justifiable belief that $\boldsymbol{y_a}$ alone provides no information about θ—that is, $p(\boldsymbol{y_a}|\boldsymbol{\theta}) \propto 1$—the prior employed in BRI corresponds with the analyst's prior before observing $\boldsymbol{y_a}$ because

$$p(\boldsymbol{\theta}|\boldsymbol{y_a}) \propto p(\boldsymbol{\theta})p(\boldsymbol{y_a}|\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}).$$

The likelihood term, $p\{\boldsymbol{d}(\boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y_a}\}$, is precisely the randomization distribution employed in an FRT. In terms of Example 2, it corresponds with the middle-right panel of Figure 1. It represents the likelihood of the observed discrepancy, $\boldsymbol{d}(\boldsymbol{\theta})$, within its randomization distribution, holding the *full set* of potential outcomes fixed. The reason we hold the full set fixed as opposed to only $\boldsymbol{y_a}$ is that, conditional on $\boldsymbol{\theta}$, Assumption 4 implies values for the counterfactuals. In practice, we often do not have access to a tractable closed-form representation of $p\{\boldsymbol{d}(\boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y_a}\}$. In these situations, we recommend approximating $p\{\boldsymbol{d}(\boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y_a}\}$ using either (a) its limiting distribution (if available) or (b) Monte Carlo methods. The latter involve drawing independent samples of $\boldsymbol{a}$, computing the implied values of $\boldsymbol{d}(\boldsymbol{\theta})$, and performing density estimation at the observed discrepancy value.

Remark 1. Because $p\{\boldsymbol{d}(\boldsymbol{\theta})|\boldsymbol{\theta}, \boldsymbol{y_a}\}$ is the same randomization distribution as that employed in an FRT, it may be tempting to replace the likelihood with a p-value. However, this approach generally produces anti-conservative inference. The reason is that p-values quantify the likelihood of an event *at least as extreme* as the one observed; hence, these events include outcomes much *more* extreme than the observed outcome.

In contrast to most Bayesian methods, which condition on purely observable quantities, BRI conditions on the random function $\boldsymbol{d}(\boldsymbol{\theta})$—a *conditionally* observable quantity. In doing so, we effectively discard information in $\boldsymbol{a}$ that we believe is unrelated to the treatment effects. In this respect, BRI is similar to limited-information Bayesian methods (Kwan, 1999; Kim, 2002; Greco et al., 2008). More formally, we define the event of observing $\boldsymbol{d}(\boldsymbol{\theta})$ as

$$\{\boldsymbol{d}(\boldsymbol{\theta})\} = (\boldsymbol{\theta}, \boldsymbol{a}) \in \underset{\boldsymbol{\theta'} \in \text{supp}(\boldsymbol{\theta})}{\cup} \{(\boldsymbol{\theta'}, \boldsymbol{a}) : \boldsymbol{f}(\boldsymbol{\theta'}, \boldsymbol{a}; \boldsymbol{y_a}) = \boldsymbol{d}(\boldsymbol{\theta'})\},$$

where $\boldsymbol{f}$ is the function of observables and $\boldsymbol{\theta}$ defining the discrepancy, $\boldsymbol{d}(\boldsymbol{\theta})$. Crucially, we treat $\boldsymbol{\theta}$ as random in the event $\{\boldsymbol{d}(\boldsymbol{\theta})\}$; otherwise, the likelihood would involve computing the density of $\boldsymbol{d}(\boldsymbol{\theta})$ *for all* $\boldsymbol{\theta}$. Treating $\boldsymbol{\theta}$ as random means that conditioning on $\boldsymbol{\theta}$ in the likelihood reveals a single plausible value of $\boldsymbol{d}(\boldsymbol{\theta})$ which, in turn, reveals a plausible set of values for $\boldsymbol{a}$—in particular, $\boldsymbol{a} \in \boldsymbol{g}_{\boldsymbol{\theta}}^{-1}\{\boldsymbol{d}(\boldsymbol{\theta}); \boldsymbol{y_a}\}$, where $\boldsymbol{g}_{\boldsymbol{\theta}}(\boldsymbol{a}; \boldsymbol{y_a}) = \boldsymbol{f}(\boldsymbol{\theta}, \boldsymbol{a}; \boldsymbol{y_a})$. In some cases, we may choose a discrepancy that is purely observable—a statistic in the classical sense—in which case the value of the discrepancy is constant over $\boldsymbol{\theta}$.

We now return briefly to the setting of Example 2. Example 3 below discusses a BRI analysis of the same data.

Example 3. *Within the BRI framework, $s(\theta)$ is properly interpreted as a discrepancy variable because its value depends on θ. In fact, straightforward algebra reveals that*

$$s(\theta) = \widehat{\text{Cov}}\{\boldsymbol{y_a}, \boldsymbol{a}\} / \widehat{\text{Var}}(\boldsymbol{a}) - \theta.$$

Because $s(\theta)$ is a linear function, $j(\boldsymbol{\theta})$ is constant and can be omitted in the estimation process (in fact, $j(\boldsymbol{\theta}) = 1$ in this case). This representation also reveals that we would obtain identical

inferences if we based the BRI analysis on the fully observable statistic $\widehat{\text{Cov}}\{\boldsymbol{y_a}, \boldsymbol{a}\}/\widehat{\text{Var}}(\boldsymbol{a})$ *because this statistic is simply a translated version of* $s(\theta)$*.*

The likelihood function is precisely the function plotted in the far right panel of Figure 1. Were we to assume a uniform prior on θ*, this function would also be the BRI posterior distribution. Because the model includes only a single, scalar parameter, we could easily estimate it by evaluating the posterior on a fine grid as we did for Figure 1. The posterior could then be summarized by the usual quantities, such as its mean and quantiles.*

3.4 Comparison to Related Methods

In contrast to randomization tests, which produce p-values for point hypotheses, BRI produces a full posterior distribution over counterfactuals, enabling automatic Bayesian inference of any sample-based estimand. An important difference between BRI and Neymanian methods is that BRI conditions on only those potential outcomes that have been observed. In contrast, Neymanian methods also condition on (unobserved) counterfactuals, a strategy that would lack coherence in the Bayesian paradigm because, to a Bayesian, conditioning implies knowledge—there would be nothing left to infer.

The primary difference between BRI and standard, superpopulation-based Bayesian causal inference methods is that BRI does not require researchers to specify models for the marginal distributions of potential outcomes. Instead, the key ingredient is a model for the individual treatment effects, which typically requires substantially fewer parameters than a full joint model. Thus, in addition to added robustness, in many cases BRI also produces simpler models with lower computational requirements relative to standard Bayesian approaches.

4 Model Specification and Estimation

Having explained the basic structure of BRI models, we now explore several details involved in specifying and estimating these models.

4.1 Jacobians

When $\boldsymbol{d}(\boldsymbol{\theta})$ is a smooth nonlinear function, BRI applies Bayes rule on a curved manifold, which requires us to introduce the Jacobian term $j(\boldsymbol{\theta})$ into (2). This section defines $j(\boldsymbol{\theta})$ and provides an illustrative example showing why the Jacobian is needed. For simplicity, we assume that the prior distribution, $p(\boldsymbol{\theta}|\boldsymbol{y_a})$, is defined with respect to $\mathcal{L}^p$, Lebesgue measure on $\mathbb{R}^p$. Further, we introduce a regularity assumption on $\boldsymbol{d}(\boldsymbol{\theta})$:

Assumption 5. The discrepancy $\boldsymbol{d}(\boldsymbol{\theta}) : \mathbb{R}^p \to \mathbb{R}^q$ is Lipschitz.

Rademacher's theorem then implies that $\boldsymbol{d}(\boldsymbol{\theta})$ is also differentiable $\mathcal{L}^p$-a.e. We now define the Jacobian matrix $\mathbf{J}(\boldsymbol{\theta}) := \partial \boldsymbol{d}(\boldsymbol{\theta})^\intercal / \partial \boldsymbol{\theta}$ and the related function

$$h(\boldsymbol{\theta}) = \begin{cases} \det\{\mathbf{J}(\boldsymbol{\theta})^\intercal \mathbf{J}(\boldsymbol{\theta})\}, & \text{if } p \leq q \\ \det\{\mathbf{J}(\boldsymbol{\theta})\mathbf{J}(\boldsymbol{\theta})^\intercal\}, & \text{otherwise.} \end{cases}$$

The following assumption imposes additional regularity, eliminating pathological cases, such as overparametrized models:

Assumption 6. If $\mathbf{J}(\boldsymbol{\theta}) \neq \mathbf{0}$, then $h(\boldsymbol{\theta}) \in (0, \infty)$, $\mathcal{L}^p$-a.e.

After observing $\boldsymbol{d}(\boldsymbol{\theta})$, we know that the pair $(\boldsymbol{\theta}, \boldsymbol{d}(\boldsymbol{\theta}))$ is confined to a manifold, $\mathcal{M} \subset \mathbb{R}^{p+q}$. In turn, Assumption 6 implies that the Hausforff dimension of $\mathcal{M}$ is p. We are now in a position to define the Jacobian area adjustment factor:

$$j(\boldsymbol{\theta}) = \begin{cases} 1, & \text{if } \mathbf{J}(\boldsymbol{\theta}) = \mathbf{0} \\ \sqrt{h(\boldsymbol{\theta})}, & \text{otherwise.} \end{cases} \tag{3}$$

This factor follows from standard results in geometric measure theory. The case $p \leq q$ can be derived from the 'area formula,' a generalization of the standard change-of-variables formula in introductory calculus. The case $p > q$ (and $p = q$) can similarly be derived according to the 'coarea formula.' When $p = q$, $\sqrt{\det\{\mathbf{J}(\boldsymbol{\theta})^\intercal \mathbf{J}(\boldsymbol{\theta})\}} = \sqrt{\det\{\mathbf{J}(\boldsymbol{\theta})\}^2} = \det\{\mathbf{J}(\boldsymbol{\theta})\}$, which shows that the formulas given above are equivalent to the standard change-of-variables formula and that $h(\boldsymbol{\theta})$ can equivalently be defined with the strict inequality $p < q$ for the first case. The following modification of Example 3 demonstrates why $j(\boldsymbol{\theta})$ is needed.

Example 4. *As in Examples 2, we assume linear treatment effects:* $\boldsymbol{y}_{\boldsymbol{a}} = \boldsymbol{y}_0 + \boldsymbol{a}\theta$. *To easily derive closed-form densities, we assume* $a_i \overset{iid}{\sim}$ *Normal*$(0, 1)$. *We consider two transformations of the following discrepancy variable:*

$$d(\theta) = \frac{1}{n} \sum_{i=1}^{n} y_{0i}(\theta) a_i = \frac{1}{n} \left(\boldsymbol{y}_{\boldsymbol{a}}{}^\intercal \boldsymbol{a} - \theta \boldsymbol{a}^\intercal \boldsymbol{a} \right).$$

The two transformations are $d_1(\theta) = |d(\theta)|$ *and* $d_2(\theta) = d(\theta)^2$. *These two transformed discrepancies should yield the same inferences because they contain the same information. Standard computations reveal that*

$$d_1(\theta) | \theta, \boldsymbol{y}_{\boldsymbol{a}} \sim \textit{Half-Normal}\left\{ \boldsymbol{y}_0(\theta)^\intercal \boldsymbol{y}_0(\theta) / n^2 \right\} \tag{4}$$

$$d_2(\theta) | \theta, \boldsymbol{y}_{\boldsymbol{a}} \sim \textit{Gamma}\left\{ 0.5,\ 2\boldsymbol{y}_0(\theta)^\intercal \boldsymbol{y}_0(\theta) / n^2 \right\}. \tag{5}$$

However, these densities differ by a ratio of $2/|\boldsymbol{y}_0(\theta)^\intercal \boldsymbol{a}|$, *resulting in an apparent paradox in which the same information yields different inferences. Including the Jacobian factors of* $j_1(\theta) = \boldsymbol{a}^\intercal \boldsymbol{a}$ *and* $j_2(\theta) = 2\,\boldsymbol{a}^\intercal \boldsymbol{a} \cdot |\boldsymbol{y}_0(\theta)^\intercal \boldsymbol{a}|$, *respectively, resolves the paradox. Figure 2 graphically depicts* $d_1(\theta)$, $d_2(\theta)$, *and their likelihoods with and without the Jacobian adjustment. The final panel reveals that an analysis neglecting the Jacobian would yield highly anticonservative inferences in this case.*

When closed-form likelihoods are available, as in Example 4, the choice of transformation has little or no impact on the analysis. In the more common case where a closed-form likelihood is *not* available, however, the choice of transformation can inform the selection of an appropriate method for approximating the likelihood. Were we to approximate the likelihoods in Example in 4, a Monte Carlo half-Gaussian approximation for $d_1(\theta)$ would

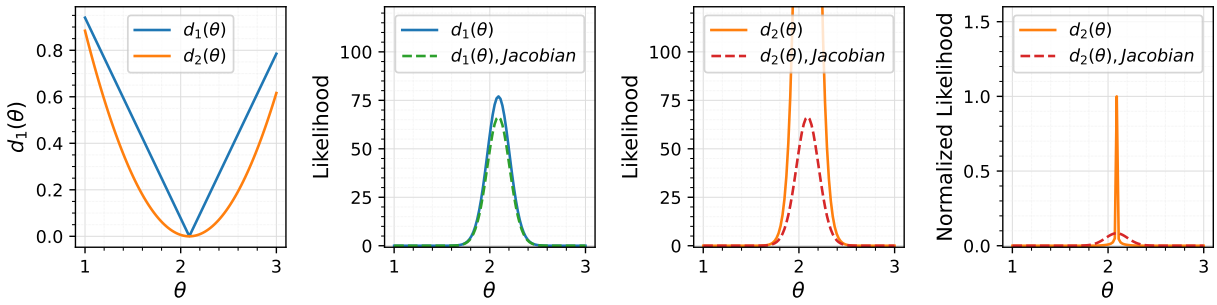

Figure 2: (Far left) Shapes of $d_1(\theta)$ and $d_2(\theta)$ from Example 4. (Middle left) Likelihood function for $d_1(\theta)$ with and without the Jacobian adjustment; the likelihoods are proportional because $d_1(\theta)$ is linear $\mathcal{L}^1$-a.e. (Middle right) Likelihood function for $d_2(\theta)$ with and without the Jacobian adjustment; with the Jacobian adjustment, the likelihoods for $d_1(\theta)$ and $d_2(\theta)$ are equal. (Far right) Same as previous, except with likelihoods normalized; the normalization reveals that the Jacobian must be included to obtain valid inferences.

perform well, but it would fail for $d_2(\theta)$ because $p(d_2(\theta)|\theta, \boldsymbol{y_a})$ increases without bound as $d_2(\theta) \to 0$; consequently, a Gamma approximation would perform much better. Similarly, although a KDE would likely perform adequately for the untransformed discrepancy, $d(\theta)$, its performance would be poor for $d_1(\theta)$ and $d_2(\theta)$ because KDEs smooth over boundary points (zero, in this case). A better nonparametric alternative would be the local regression density estimators described in Cattaneo et al. (2020, 2021).

4.2 Continuous outcomes and discrete treatments

[To be added later. Two solutions are (a) condition on neighborhoods of $\boldsymbol{d}(\boldsymbol{\theta})$ scaled by $j(\boldsymbol{\theta})$ or (b) use a continuous latent variable representation. The theoretical results show that (a) results in desirable theoretical properties.]

4.3 Estimation Algorithms

In principle, we can apply any standard Bayesian computational method to estimate BRI models. The primary challenge compared to standard Bayesian models is accurately approximating the likelihood function using density estimation techniques, which may be challenging (or even impossible) with some existing software packages.

In our testing, we found that the NumPyro package is especially well suited to fit BRI models due to its flexible interface and the availability of modern MCMC algorithms, such as the No U-turn Sampler (NUTS; Hoffman et al., 2014; Phan et al., 2019; Bingham et al., 2019). We also found that implementation in the Engine for Likelihood-free Inference (ELFI) is particularly straightforward because ELFI's paradigm of simulation-based likelihoods is a natural fit for BRI's randomization-based likelihood function (Lintusaari et al., 2018). A key benefit of NumPyro and other automatic differentiation packages is that $j(\boldsymbol{\theta})$ can be computed automatically; in contrast, ELFI and other packages would require the analyst to

manually derive and program the computation of $j(\boldsymbol{\theta})$. Appendix A provides example code for fitting the model of example 3 via NumPyro.

5 Theoretical results

[To be added later. I have derived the main results already and am actively working on including them here. The informal takeaway is that BRI estimates the parameter value that places $\boldsymbol{d}(\boldsymbol{\theta})$ in the center of it randomization distribution (similar to a Hodges–Lehmann estimator).]

5.1 Assumptions

5.2 Posterior consistency

5.3 Connections to common estimators

5.4 Asymptotic variances

6 Extensions

[To be added later. I have tested both of these ideas in simulation and am actively working on adding the details here.]

6.1 Beyond deterministic models

6.2 Residualization techniques

7 Discussion

References

Basu, D. (1980) Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, **75**, 575–582.

Berger, J. O. (2003) Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science*, **18**, 1–32.

Bingham, E., Chen, J. P., Jankowiak, M., Obermeyer, F., Pradhan, N., Karaletsos, T., Singh, R., Szerlip, P. A., Horsfall, P. and Goodman, N. D. (2019) Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, **20**, 28:1–28:6. URL: `http://jmlr.org/papers/v20/18-403.html`.

Cattaneo, M. D., Jansson, M. and Ma, X. (2020) Simple local polynomial density estimators. *Journal of the American Statistical Association*, **115**, 1449–1455.

— (2021) Local regression distribution estimators. *Journal of Econometrics*, 105074.

Chiba, Y. (2018) Bayesian inference of causal effects for an ordinal outcome in randomized trials. *Journal of Causal Inference*, **6**, 20170019.

Ding, P. (2017) A paradox from randomization-based causal inference. *Statistical Science*, 331–345.

Ding, P. and Dasgupta, T. (2018) A randomization-based perspective on analysis of variance: a test statistic robust to treatment effect heterogeneity. *Biometrika*, **105**, 45–56.

Ding, P. and Li, F. (2018) Causal inference: A missing data perspective. *Statistical Science*, **33**, 214–237.

Ding, P. and Miratrix, L. W. (2019) Model-free causal inference of binary experimental data. *Scandinavian Journal of Statistics*, **46**, 200–214.

Fisher, R. A. (1960) *The design of experiments.* No. 7th Ed. Oliver and Boyd. London and Edinburgh.

Fogarty, C. B. (2020) Studentized sensitivity analysis for the sample average treatment effect in paired observational studies. *Journal of the American Statistical Association*, **115**, 1518–1530.

Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 733–760.

Greco, L., Racugno, W. and Ventura, L. (2008) Robust likelihood functions in Bayesian inference. *Journal of Statistical Planning and Inference*, **138**, 1258–1270. URL: `https://www.sciencedirect.com/science/article/pii/S0378375807001899`.

Hoffman, M. D., Gelman, A. et al. (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Holland, P. W. (1986) Statistics and causal inference (with discussion). *Journal of the American Statistical Association*, **81**, 945–960.

Hubbard, R. and Bayarri, M. J. (2003) Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *The American Statistician*, **57**, 171–178.

Humphreys, M. and Jacobs, A. M. (2015) Mixing methods: A Bayesian approach. *American Political Science Review*, **109**, 653–673.

Keele, L. and Quinn, K. M. (2017) Bayesian sensitivity analysis for causal effects from 2×2 tables in the presence of unmeasured confounding with application to presidential campaign visits.

Kim, J.-Y. (2002) Limited information likelihood and Bayesian analysis. *Journal of Econometrics*, **107**, 175–193. URL: `https://www.sciencedirect.com/science/article/pii/S0304407601001191`. Information and Entropy Econometrics.

Kwan, Y. K. (1999) Asymptotic Bayesian analysis based on a limited information estimator. *Journal of Econometrics*, **88**, 99–121. URL: https://www.sciencedirect.com/science/article/pii/S0304407698000244.

Leavitt, T. (2023) Randomization-based, Bayesian inference of causal effects. *Journal of Causal Inference*, **11**, 20220025.

Lehmann, E. L. (1993) The Fisher, Neyman–Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, **88**, 1242–1249.

Lenhard, J. (2006) Models and statistical inference: The controversy between Fisher and Neyman–Pearson. *The British Journal for the Philosophy of Science*.

Li, F., Ding, P. and Mealli, F. (2023) Bayesian causal inference: A critical review. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **381**, 20220153. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2022.0153.

Lintusaari, J., Vuollekoski, H., Kangasrääsiö, A., Skytén, K., Järvenpää, M., Marttinen, P., Gutmann, M. U., Vehtari, A., Corander, J. and Kaski, S. (2018) ELFI: Engine for likelihood-free inference. *Journal of Machine Learning Research*, **19**, 1–7. URL: http://jmlr.org/papers/v19/17-374.html.

Loh, W. W., Richardson, T. S. and Robins, J. M. (2017) An apparent paradox explained. *Statistical Science*, **32**, 356–361.

Meng, X.-L. (1994) Posterior predictive p-values. *The Annals of Statistics*, **22**, 1142–1160.

Neyman, J. S., Dabrowska, D. M. and Speed, T. P. (1990) On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 465–472.

Phan, D., Pradhan, N. and Jankowiak, M. (2019) Composable effects for flexible and accelerated probabilistic programming in NumPyro. *arXiv preprint arXiv:1912.11554*.

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

— (1980) Randomization analysis of experimental data: The Fisher randomization test comment. *Journal of the American Statistical Association*, **75**, 591–593.

— (1984) Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Annals of Statistics*, 1151–1172.

Sabbaghi, A. and Rubin, D. B. (2014) Comments on the Neyman–Fisher controversy and its consequences.

Wu, J. and Ding, P. (2021) Randomization tests for weak null hypotheses in randomized experiments. *Journal of the American Statistical Association*, **116**, 1898–1913.

Zabell, S. L. (1992) RA Fisher and fiducial argument. *Statistical Science*, 369–387.

A Example NumPyro Code

The code below fits the model of Example 3. The code draws Monte Carlo samples for $\boldsymbol{a}$ to approximate the likelihood via a kernel-density estimator (KDE). Alternatively, one could apply a Gaussian approximation based on either (a) theoretical asymptotic expressions or (b) simulated moments.

```
import numpyro
import jax.numpy as jnp
from numpyro.infer import MCMC, NUTS
from jax import random
import numpy as np


# Model function for computing log posterior density up to proportionality
def model(y, a, bandwidth_adjustment=1., num_samples=2000):
    # Prior
    theta = numpyro.sample("theta", numpyro.distributions.Normal(0., 10.))

    # Compute discrepancy values
    y0_theta = y - a*theta  # Implied value of y0
    s_observed = compute_s(y0_theta, a)  # Observed discrepancy value
    s_samples = compute_s(y0_theta, num_samples=num_samples)  # Sampled
    ↪  discrepancy values

    # Approximate log likelihood up to proportionality
    bandwidth = bandwidth_adjustment * s_samples.std()
    zs = (s_samples - s_observed) / bandwidth
    kde_values = jnp.exp(-0.5 * (zs**2)) / bandwidth
    log_like = jnp.log(kde_values.mean())  # Average over samples, then log
    numpyro.factor("log_like", log_like)  # Include in posterior density

# Helper function for computing (or sampling) s
def compute_s(y0, a=None, num_samples=2000):
    y0_centered = y0 - y0.mean()  # Center y0 for cov/var calculations
    axis = 1 if a is None else 0  # For calculating means over correct axis
    if a is None:
        a = np.random.random((num_samples, y.size))
    a_centered = (a.T - a.mean(axis=axis)).T  # .T for dimension compatibility
    cov = (y0_centered * a_centered).mean(axis=axis)
    var = (a_centered**2).mean(axis=axis)
    return cov / var


# Fit model
kernel = NUTS(model)
mcmc = MCMC(kernel, num_warmup=1000, num_samples=1000)
```

```
rng_key = random.PRNGKey(0)
mcmc.run(rng_key, y=y, a=a, bandwidth_adjustment=0.2)
mcmc.print_summary()
```