International Statistical Review

International Statistical Review (2024) doi: 10.1111/insr.12598

Fisher Meets Bayes: The Value of Randomisation for Bayesian Inference of Causal Effects

Thomas Leavitt

Marxe School of Public and International Affairs, Baruch College, City University of New York (CUNY), New York, NY, USA
Corresponding to: Thomas Leavitt, Marxe School of Public and International Affairs, Baruch College, City University of New York (CUNY), New York, NY, USA. Email: thomas.leavitt@baruch.cuny.edu

Summary

For a Bayesian agent with beliefs about the relationship between covariates and potential outcomes, deterministically selecting an assignment that yields optimal covariate balance rationally dominates randomisation. However, randomisation—by enabling control over the probabilities of erroneous causal conclusions due to unknown covariate imbalances—offers insurance against the possibility that an agent's beliefs may be misleading. For the most part, such rational justifications for optimum assignment have presupposed the framework of Bayesian inference, while such epistemic justifications for randomisation have presupposed the framework of significance testing. In this paper, I build on a conception of balance that seems inextricable from the significance testing framework, Fisherian balance, to show that it implies an analogous epistemic justification for randomisation within the framework of Bayesian inference. Consequently, for the choice between optimum and random assignment, this paper shows that epistemic justifications need not be wedded to significance testing nor must Bayesian inference be wedded to rational justifications.

Key words: covariate balance; exact test; philosophy of science; potential outcomes.

1 Introduction

For a single Bayesian agent in isolation, deterministic selection of an assignment that balances covariates rationally dominates randomisation given that agent's prior beliefs about the relationship between covariates and potential outcomes (Bertsimas *et al.*, 2015; Fedorov, 1972; Harville, 1975; Kallus, 2018; Kasy, 2016; Kiefer, 1959). On the other hand, a range of scholars have argued that, while randomisation is rationally suboptimal, it offers insurance against the possibility that one's prior beliefs may be misleading, that is, that unobserved covariates (potential outcomes in particular) may be imbalanced between treatment and control groups (Bai, 2023; Efron, 1971; Harshaw *et al.*, 2024; Nordin & Schultzberg, 2022; Kapelner *et al.*, 2021; Kapelner *et al.*, 2022; Wu, 1981). This long-standing tension between optimum and random assignment is known as the balance-robustness tradeoff.

As illustrated by recent debates on the balance-robustness tradeoff (Harshaw *et al.*, 2024; Johansson *et al.*, 2021; Kallus, 2018; 2020; 2021; Kapelner *et al.*, 2021), randomisation's

robustness justification can be derived in large part from R. A. Fisher's randomisation test (FRT) established via the Lady tasting tea experiment (Fisher, 1925; 1926; 1935). Building on both Hall (2007) and Senn (2013), Martinez & Teira (2024) point to the Fisherian antecedents of this robustness justification for randomisation by referring to it as 'Fisherian balance' (not to be confused with alternative conceptions of balance justifying optimum assignment). As Martinez & Teira (2024) explain, Fisherian balance, achieved through randomisation, is not about obtaining balance, per se, but rather about controlling the probability of erroneous causal conclusions due to unknown covariate imbalances.

At their core, arguments for deterministic selection of an optimal assignment based on covariate balance stand on *rational* grounds. By *rational*, I mean an act that, given an agent's prior beliefs, yields the best consequence in expectation, that is, minimises the expected distance between the unknown causal target and the expected data generated by that act.[1] By contrast, arguments for randomisation in terms of controlling the probabilities of erroneous conclusions due to unknown covariate imbalances stand on *epistemic* grounds. By *epistemic*, I mean an act (e.g., optimum or random assignment) that would lead one to recover the truth in expectation and with probability that limits to 1 as the size of an experimental population increases indefinitely.

For the most part, each of these justifications (on epistemic and rational grounds, respectively) presupposes a particular framework for drawing statistical conclusions about causal effects. Epistemic justifications for the choice between optimum and random assignment usually presuppose the framework of significance testing. Rational justifications, by contrast, usually presuppose Bayesian inference. The contribution of this paper is to show that epistemic justifications in the name of Fisherian balance need not be wedded to significance testing nor must Bayesian inference be wedded to rational justifications.

To make this contribution, I first engage with a conception of balance, standard in Bayesian critiques of randomisation (Howson & Urbach, 2006; Urbach, 1985; Worrall, 2002; 2007a; 2008; 2007b), which stipulates that treatment and control groups ought to be the same on covariates that predict potential outcomes. I establish a logical consequence of this argument: If the means of potential outcomes—themselves baseline covariates—are balanced between treatment and control groups, then the average of observed outcomes in treatment minus the average of observed outcomes in control is equal to the average treatment effect (ATE). Expressing this conception of balance directly in terms of potential outcomes is insightful. Potential outcomes are unobservable before assignment and only partially observable after; hence, whether any assignment yields sufficient balance depends on an agent's beliefs about fundamentally unobservable quantities. This dependence underscores the importance of Fisherian balance in that, no matter how many covariates one measures, the possibility of erroneous conclusions due to unknown covariate imbalances is unavoidable.

Despite the general importance of Fisherian balance, its epistemic value has been understood almost exclusively within the framework of significance testing. I use Fisher's Lady tasting tea —an experiment that seems inextricable from significance testing—to show that Fisherian balance offers an analogous epistemic justification for randomisation in the framework of Bayesian inference. In particular, I show that (1) for an agent who is initially neutral about the plausibility of different causal hypotheses, the hypothesis with the greatest posterior probability is equal to the truth, in expectation, and (2) under exceedingly mild conditions on an agent's prior distribution, so long as an experiment is sufficiently large, false causal hypotheses will receive a low posterior probability with high probability and the true causal hypothesis will receive a high posterior probability with high probability. In other words, randomisation implies that an agent who draws conclusions about causal effects via Bayes' rule will recover the true effect in a sufficiently large experiment.

1.1 Related Literature

This paper's epistemic justification for randomisation in the framework of Bayesian inference speaks directly to long-standing debates on the role of randomisation for Bayesian inference. In particular, Rubin (1976, 233) writes that randomisation is important for a Bayesian because it enables one to 'ignore the assignment mechanism when making causal inferences'. (See also Rubin, 1978; 1984.) However, one implication from Ding & Guo (2023) is that this value of randomisation can be recast as the importance of controlled experimentation (as opposed to uncontrolled observational studies). That is, randomisation implies 'ignorability' of the assignment mechanism for Bayesian causal inference; however, as Kasy (2016, Section 5.2) shows, so too would deterministic selection of an optimal assignment. 'Ignorability' fails not in the absence of randomisation, but rather in observational settings where the assignment mechanism is unknown. For this reason, Ding & Guo (2023) develop a method incorporating uncertainty over the assignment process into causal inferences (viz., posterior predictive p-values for tests of Fisher's sharp null hypothesis of no effects).

Existing arguments that do speak directly to the specific choice between random and optimum assignment allude to the epistemic value of randomisation for Bayesian inference. For example, Senn (1994, 218) writes that 'randomisation in clinical trials is not an issue which need divide Bayesians and classical statisticians, though of course they will have different views regarding analysis'. Despite such claims, the connections between Fisherian balance and Bayesian inference remain underexplored. Martinez & Teira (2024) argue that Fisherian balance justifies the decision to randomise, but within the framework of significance testing. In the framework of Bayesian inference, Martinez & Teira (2024) provide an alternative conception of balance, which provides a rational—as opposed to epistemic—justification for randomisation. This alternative conception of balance shifts from the setting of a single Bayesian agent in isolation to a strategic setting against either 'nature' (e.g., Wu, 1981) or an adversarial audience (Banerjee *et al.*, 2017; Banerjee *et al.*, 2020; Basu, 1980; Kadane & Seidenfeld, 1990; Lindley, 1982; Savage, 1954; 1962a; 1962b; Stone, 1969; Suppes, 1982). By contrast, I build on the conception of Fisherian balance from Martinez & Teira (2024) to show how it offers an epistemic—as opposed to rational—justification for randomisation within the framework of Bayesian inference.

2 Lady Tasting Tea Experiment

The FRT, which emerged from the Lady tasting tea (Fisher, 1935), is an essential ingredient of robustness justifications for random assignment (Johansson *et al.*, 2021; Kallus, 2021; Kapelner *et al.*, 2021). Hence, the Lady tasting tea is a sensible jumping off point for engaging with notions of Fisherian balance and its role in robustness justifications for random assignment. In his book, *The Design of Experiments* (1935), Fisher describes the Lady tasting tea experiment as follows:

> A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. … Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgement in a random order. The subject has been told in advance of what the test will consist, namely, that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task

is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received (Fisher, 1935, 13–14).

To formalise the setting of the Lady tasting tea experiment, let the index $i = 1, \ldots, N$ run over the individual cups with $N = 8$. Each cup can be assigned to either the tea-first (control) condition, $z_i = 0$, or the milk-first (treatment) condition, $z_i = 1$. The vector $\boldsymbol{z} = [z_1 \quad z_2 \quad \ldots \quad z_N]^\top$, where the superscript $\top$ denotes matrix transposition, is the collection of N individual treatment indicator variables. The set of possible ways a researcher could assign these N units to treatment or control is $\{0,1\}^N$ of which there are 2^N possibilities. However, under the assignment process Fisher (1935) describes, there are $\binom{8}{4} = \frac{8!}{4!(8-4)!} = 70$ possible ways in which the experimenter could order the 8 cups such that 4 cups are in the milk-first (treatment) condition and the remaining 4 cups are in the tea-first (control) condition. I denote this set of allowable assignments by $\Omega \subseteq \{0,1\}^N$.

The potential responses of 'the lady' are a mapping from $\{0,1\}^N$ to an N-dimensional vector of real numbers, $\mathbb{R}^N$. With 2^N assignments, there are 2^N corresponding vectors of potential outcomes. However, under the Stable Unit Treatment Value Assumption (SUTVA) (Cox, 1958; Rubin, 1980; 1986), let $y_i(1)$ (the treated potential outcome) denote the outcome value of the ith cup for all $\boldsymbol{z} \in \{0,1\}^N$ with $z_i = 1$. Likewise, let $y_i(0)$ (the control potential outcome) denote the outcome value of the ith cup for all $\boldsymbol{z} \in \{0,1\}^N$ with $z_i = 0$. Under SUTVA, the collection of all cups' treated potential outcomes is equal to the vector of outcomes if all cups had been assigned to treatment, $\boldsymbol{y}(\boldsymbol{1})$, and the collection of all cups' control potential outcomes is equal to the vector of outcomes if all cups had been assigned to control, $\boldsymbol{y}(\boldsymbol{0})$. Both $\boldsymbol{y}(\boldsymbol{1})$ and $\boldsymbol{y}(\boldsymbol{0})$ are baseline covariates in that they are fixed quantities, not changing depending on how the random assignment process turns out. Observable outcomes, by contrast, can vary depending on which assignment happens to be realised; I denote them by $\boldsymbol{y}(\boldsymbol{z})$ for all $\boldsymbol{z} \in \Omega$.

Under SUTVA, the collection of N individual effects, $\boldsymbol{\tau}$, is defined as $\boldsymbol{y}(\boldsymbol{1}) - \boldsymbol{y}(\boldsymbol{0})$. The average of these N individual effects (the ATE) is denoted by τ. When effects are homogeneous, the causal target is the 1-dimensional, constant effect. When effects are heterogeneous, this causal target can be interpreted as the 1-dimensional effect that best approximates $\boldsymbol{\tau}$, which, under a standard Euclidean distance measure, will be the ATE, τ.

The outcome for each cup can be either 0 or 1, denoting whether the 'lady' identifies cup i as either tea-first (0) or milk-first (1). According to the historical record (Box, 1978, 131 – 135), the 'lady' in question—fellow scientist at the Rothamsted Experimental Station, Muriel Bristol—correctly identified all of the 8 cups. In accordance with this historical record, suppose (without loss of generality) that $\boldsymbol{z}_1 = [1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]^\top$ was randomly selected from Ω and that the realised data from the experiment were as in Table 1.

Table 1. *Results of R. A. Fisher's Lady tasting tea experiment.*

Unit	$\boldsymbol{z}$	$\boldsymbol{y}(\boldsymbol{z})$
1	1	1
2	1	1
3	1	1
4	1	1
5	0	0
6	0	0
7	0	0
8	0	0

For simplicity and following Rosenbaum (2002), I consider two rival causal hypotheses. Each causal hypothesis assigns some probability to the observed data, summarised by a test statistic, based on only the potential outcomes schedule implied by that hypothesis and the known random assignment mechanism. The first hypothesis is Fisher's sharp causal hypothesis of no effects, hereafter referred to as *Fisher's null*. The second rival causal hypothesis is that of a positive effect for each of the 8 cups, that is, the sharp causal hypothesis of Perfect discrimination, hereafter referred to as *Perfect discrimination*. Both of these rival causal hypotheses satisfy SUTVA, which implies that we can write the two potential outcome schedules compactly as in Table 2.

Both potential outcome schedules imply a 1-dimensional, constant effect for all experimental units, which is equal to the average of the N individual effects, 0 (under Fisher's null) and 1 (under Perfect discrimination). One can, of course, test more complex, N-dimensional effects in which individual effects differ across units. However, as alluded to above, tests of 1-dimensional effects can be interpreted as the single value that best approximates the true N-dimensional vector of individual effects, $\boldsymbol{\tau}$ (Rosenbaum, 2010, Section 2.4.4, 44–46).

3 Fisherian Balance

Arguments over the role of balance in experiments typically focus on differences between treated and control groups in the means of covariates that predict potential outcomes. However, as Senn (2013, 1447) writes, 'what really matters is differences in outcome. Differences in covariates are only relevant to the extent that they help us predict outcomes we would have seen between groups in the absence of treatment'. Under SUTVA, treated and control potential outcomes, $\boldsymbol{y}(\boldsymbol{1})$ and $\boldsymbol{y}(\boldsymbol{0})$, are baseline covariates, just like any other quantities that are fixed over allowable assignments. These two covariates, $\boldsymbol{y}(\boldsymbol{1})$ and $\boldsymbol{y}(\boldsymbol{0})$, have a special status in that balance in their means suffices for the average difference in observed outcomes between treated and control groups to be equal to the causal target of interest. Balance in other covariates is meaningful only insofar as such balance suggests that treated and control potential outcomes are balanced between treated and control groups.

To more precisely explicate this point, suppose that the observable data for all N units is summarised by a 1-dimensional statistic, the canonical Difference-in-Means. The Difference-in-Means is

$$\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})) \quad = \left(\frac{1}{\boldsymbol{z}^\top \boldsymbol{z}}\right)\boldsymbol{z}^\top \boldsymbol{y}(\boldsymbol{z}) \; - \; \left(\frac{1}{(\boldsymbol{1} - \boldsymbol{z})^\top (\boldsymbol{1} - \boldsymbol{z})}\right)(\boldsymbol{1} - \boldsymbol{z})^\top \boldsymbol{y}(\boldsymbol{z}), \tag{1}$$

Table 2. *Potential outcome schedules implied by Fisher's null and Perfect discrimination.*

	Fisher's null			*Perfect discrimination*		
Unit	$\boldsymbol{y}(\boldsymbol{0})$	$\boldsymbol{y}(\boldsymbol{1})$	$\boldsymbol{\tau}$	$\boldsymbol{y}(\boldsymbol{0})$	$\boldsymbol{y}(\boldsymbol{1})$	$\boldsymbol{\tau}$
1	1	1	0	0	1	1
2	1	1	0	0	1	1
3	1	1	0	0	1	1
4	1	1	0	0	1	1
5	0	0	0	0	1	1
6	0	0	0	0	1	1
7	0	0	0	0	1	1
8	0	0	0	0	1	1

where the 'hat' operator denotes functions of observed quantities. Proposition 1 formally establishes that balance in the means of treated and control potential outcomes implies that the Difference-in-Means is equal to the ATE (or, equivalently, the 1-dimensional effect that best approximates $\boldsymbol{\tau}$).

Proposition 1. *Suppose SUTVA and, without loss of generality, a set of possible assignments with a fixed number of $n_1 \geq 1$ treated units and a fixed number of $N - n_1 = n_0 \geq 1$ control units. For any assignment in this set that yields balance in the means of both treated and control potential outcomes, that is,*

$$\left(\frac{1}{n_1}\right)\boldsymbol{z}^\top \boldsymbol{y}(\boldsymbol{1}) = \left(\frac{1}{n_0}\right)(\boldsymbol{1} - \boldsymbol{z})^\top \boldsymbol{y}(\boldsymbol{1}) \text{ and} \tag{2}$$

$$\left(\frac{1}{n_1}\right)\boldsymbol{z}^\top \boldsymbol{y}(\boldsymbol{0}) = \left(\frac{1}{n_0}\right)(\boldsymbol{1} - \boldsymbol{z})^\top \boldsymbol{y}(\boldsymbol{0}), \tag{3}$$

the Difference-in-Means is equal to the ATE, that is,

$$\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})) = \tau. \tag{4}$$

The proof of Proposition 1 is in Appendix A. For a simple illustration of this proposition, consider the Lady tasting tea experiment under two different settings (both of which would be unbeknownst to the researcher): when Fisher's null is true and when Perfect discrimination is true. In both settings, a Difference-in-Means far from the true effect implies large covariate imbalances in potential outcomes.

As Figure 1 shows, when Perfect discrimination is true, every allowable assignment is perfectly balanced in the means of both treated and control potential outcomes. Hence, every

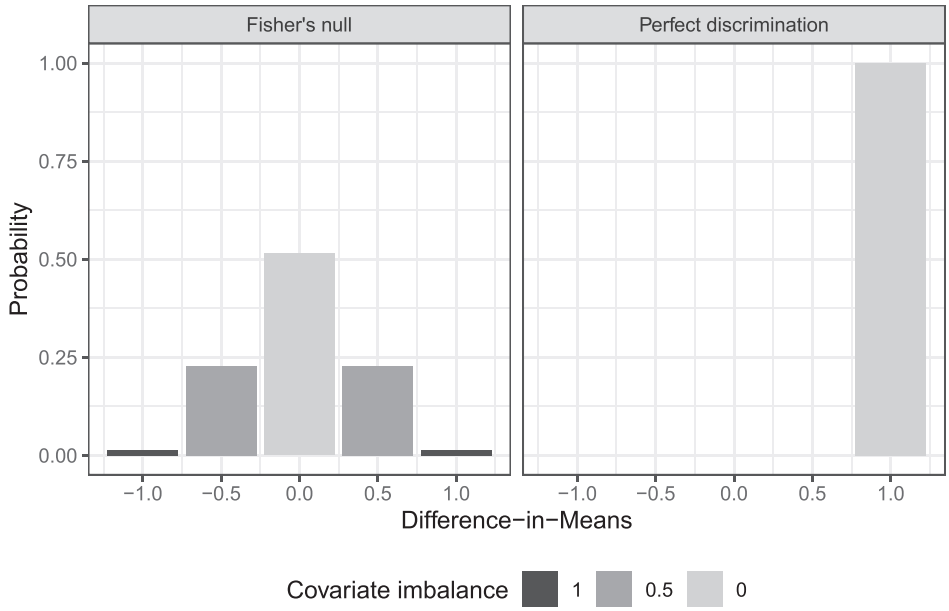

FIGURE 1. *Relationship between Difference-in-Means and covariate balance under Fisher's null and Perfect discrimination, respectively.*

one of the allowable assignments yields a Difference-in-Means exactly equal to the true 1-dimensional causal effect. By contrast, when Fisher's null is true, any positive distance between the Difference-in-Means and the true 1-dimensional effect implies covariate imbalance.[2]

Proposition 1 relates to the Bayesian conception of balance in Howson & Urbach (2006), Urbach (1985), (Worrall, 2002; 2007a; 2007b; 2008) and others, whereby treatment and control groups ought to be the same on prognostic covariates. Proposition 1 recasts this conception of balance in terms of potential outcomes themselves. In so doing, Proposition 1 bears on an important argument made by Fisher (1935, 21): 'It would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable'. Proposition 1 implies that the relevant covariates to be balanced are *not* 'always strictly innumerable'. In fact, under SUTVA, there are only two such covariates.

The importance of Fisherian balance is not because the prognostic covariates are 'always strictly innumerable'. Rather, Fisherian balance matters because it is impossible to directly calculate covariate balance in (2) and (3) since both equations contain fundamentally unobservable quantities. Before assignment, no potential outcomes can be observed and, after assignment, treated potential outcomes in the control group and control potential outcomes in the treated group are unobservable. Consequently, if, for example, one were to observe a difference in average outcomes between treated and control groups equal to 1, one would be unable to definitively distinguish between two possibilities: the existence of a causal effect (Perfect discrimination) or the absence of a causal effect (Fisher's null) and an imbalance in potential outcomes.

Thus, randomisation's value is not its ability to yield covariate balance (although randomisation does so in expectation). Rather, randomisation is valuable because it yields Fisherian balance. In contrast to a standard conception of balance, Fisherian balance is about controlling the probability of errors due to unknown imbalances in the relevant covariates or, as Martinez and Teira (2024, 4) write 'measuring *ex post* the influence of uncontrolled covariates'.

Such control over the probability of errors due to unknown imbalances in the relevant covariates is understood largely against the background of the significance testing framework. Suppose a significance level (i.e. α-level) of 0.05 and a test of Fisher's null against the alternative of Perfect discrimination. When Fisher's null is true, Figure 1 shows that one would erroneously reject this null if and only if covariate imbalance is large (equal to 1). However, randomisation implies that the probability of such covariate imbalance is low (1/70 to be exact). When Perfect discrimination is true, Figure 1 shows that the Difference-in-Means is always equal to the true causal effect because all allowable assignments yield perfect covariate balance. Since the Difference-in-Means under Perfect discrimination lies in the rejection region of the Difference-in-Means' distribution implied by Fisher's null, one would never fail to reject Fisher's null when Perfect discrimination is true.[3]

Figure 1 shows that randomisation enables one to control the probabilities of drawing erroneous conclusions (i.e. randomisation implies Fisherian balance) within the significance testing framework. More generally, as Rosenbaum (2002, Chapter 2) shows, randomisation controls the probability of a type I error (rejecting the null when it is true) and, for a specific class of alternatives, ensures that a test's power (rejecting the null when it is false) is always weakly (and often strictly) greater than a test's type I error probability. Moreover, as the size of an experiment increases indefinitely, the type I error probability remains at least as small as a test's significance level and the power of a test limits to 1.

Consequently, within the framework of either rejection (1) or not (0) of a null hypothesis, randomisation ensures that the expected inference (i.e. rejection probability) is greater when the null hypothesis is false compared with when it is true. Likewise, as the size of an experiment increases indefinitely, randomisation ensures that the expected inference limits to 1 when the null hypothesis is false. When the null hypothesis is true, the expected inference in an experiment of any size is always less than the expected inference when the null is false. Thus, within the framework of significance testing, randomisation stands on solid epistemic ground—that is, randomisation leads one to recover the truth in expectation and with probability that limits to 1 as the size of an experimental population increases indefinitely.

4 From Significance Testing to Bayesian Inference

One of the primary concerns with significance testing is its coarse decision calculus in which one either rejects or fails to reject causal hypotheses. This framework enables one to conclude that a causal hypothesis one rejects is less supported by the evidence than a causal hypothesis one does not reject. Yet, amongst the hypotheses one rejects or amongst the hypotheses one fails to reject, the significance testing framework does *not* enable statements about which hypotheses are more plausible given experimental evidence. For example, a null hypothesis with a p-value slightly above $\alpha = 0.05$ is no less plausible than a null hypothesis with a p-value much greater than $\alpha = 0.05$; both hypotheses are simply those the researcher was unable to reject. Bayesian inference, by contrast, permits more nuanced inferences by encoding the relative plausibility of each causal hypothesis in terms of a continuous probability measure.

In Section 5, I show that Fisherian balance, implied by randomisation, leads to analogous epistemic properties of inference by a Bayesian. Usually a Bayesian agent is defined as one who behaves rationally by choosing an act that maximises the expected consequence given that agent's prior beliefs. By contrast, I embrace a different conception of a Bayesian—one with a 'human face', so to speak (Jeffrey, 1983). That is, for the purposes of this paper, a Bayesian is simply one who conducts Bayesian inference—that is, updates prior beliefs about causal hypotheses via a likelihood implied by the known assignment process.

This type of Bayesian need not embrace axiomatic principles grounded in rational decision theory (see Ramsey, 1990; 1931; Savage, 1954, and related 'Dutch book' arguments). Instead, one might conduct Bayesian inference for its pragmatic benefits—specifically, the flexible and nuanced inferences it affords about the relative plausibility of competing causal hypotheses (Howson & Urbach, 2006; McElreath, 2020; Strevens, 2012). The existence of prior beliefs about causal effects, to be updated upon observing data, does not imply that one must choose rationally between random and optimum assignment given those prior beliefs.

5 Randomisation's Epistemic Value for Bayesian Inference

To situate the Lady tasting tea within a Bayesian framework, the first step is to provide a prior distribution on the set of two hypotheses, H_{Fisher} (Fisher's null) and H_{Perfect} (Perfect discrimination). Let the prior distribution be uniform in which $\Pr(H_{\text{Fisher}}) = 0.5$ and $\Pr(H_{\text{Perfect}}) = 0.5$. However, I will later relax this condition so that the prior can have an arbitrary distribution so long as the true causal hypothesis is in the prior distribution's support.

Under randomisation, the true probability mass function (PMF) of the Difference-in-Means is

$$f(t) = \sum_{\boldsymbol{z} \in \Omega} 1\{\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})) = t\} \Pr(\boldsymbol{Z} = \boldsymbol{z}) \tag{5}$$

for $t \in \mathbb{R}$. The PMF in (5) is unknown since one can observe outcomes under only one assignment. Under whichever assignment is realised, one can construct *reference* PMFs, analogous to reference cumulative distribution functions (CDFs) for the calculation of p-values. The reference PMFs for H_{Fisher} and H_{Perfect} implied by randomisation are

$$\hat{f}_{\text{Fisher}}(t) = \sum_{\boldsymbol{w} \in \Omega} 1\{\hat{\tau}(\boldsymbol{w}, \boldsymbol{y}(\boldsymbol{z})) = t\} \Pr(\boldsymbol{W} = \boldsymbol{w}) \tag{6}$$

$$\hat{f}_{\text{Perfect}}(t) = \sum_{\boldsymbol{w} \in \Omega} 1\{\hat{\tau}(\boldsymbol{w}, \boldsymbol{w}) = t\} \Pr(\boldsymbol{W} = \boldsymbol{w}), \tag{7}$$

where $\boldsymbol{w}$ is a relabelling of the assignments in Ω after observing data. Both (6) and (7) can be calculated because H_{Fisher} and H_{Perfect} fully specify the potential outcomes schedule. With more general hypotheses (specifically those that are not sharp in this sense), the calculation of such reference PMFs would not be trivial.

Each reference PMF in (6) and (7) is implied by the known assignment process and the supposition that a given null hypothesis is true. Both PMFs facilitates the calculation of likelihoods. Random assignment implies that the probability of observing $\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})) = 1$ if H_{Fisher} were true is 1/70. If H_{Perfect} were true, randomisation implies that the probability of observing $\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})) = 1$ is 1.

With both a prior and likelihood, the posterior distribution immediately follows from an application of Bayes' rule:

$$\hat{\Pr}(H_{\text{Fisher}} | \hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z}))) = \frac{\hat{f}_{\text{Fisher}}(\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})))\Pr(H_{\text{Fisher}})}{\hat{f}_{\text{Fisher}}(\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})))\Pr(H_{\text{Fisher}}) + \hat{f}_{\text{Perfect}}(\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})))\Pr(H_{\text{Perfect}})} \approx 0.01$$

$$\hat{\Pr}(H_{\text{Perfect}} | \hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z}))) = \frac{\hat{f}_{\text{Perfect}}(\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})))\Pr(H_{\text{Perfect}})}{\hat{f}_{\text{Fisher}}(\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})))\Pr(H_{\text{Fisher}}) + \hat{f}_{\text{Perfect}}(\hat{\tau}(\boldsymbol{z}, \boldsymbol{y}(\boldsymbol{z})))\Pr(H_{\text{Perfect}})} \approx 0.99.$$

This posterior distribution intuitively makes sense in that correctly guessing all cups yields much greater inductive support for Perfect discrimination relative to Fisher's null. But, unlike in the significance testing framework, one does not conclude that Fisher's null is categorically false.

In this Bayesian framework, Fisherian balance also plays a crucial role in avoiding erroneous causal conclusions. Figure 2 shows the true and reference PMFs when one or the other hypothesis is, in fact, true.

From Figure 2, it is straightforward to deduce the likelihoods for each causal hypothesis (when one or the other is true) over all possible assignments. Table 3 presents the distribution of likelihoods for each causal hypothesis when Fisher's null is true and when it is false.

Cross-referencing Figure 2 and Table 3 with Figure 1 illustrates the connection between Bayesian inference and Fisherian balance. For example, when Fisher's null is true, the assignment that yields a high likelihood for Perfect discrimination (and a low likelihood for Fisher's null) is the assignment that yields the covariate imbalance of 1. However, randomisation implies that the probability of this assignment is low, 1/70. By contrast, when Perfect discrimination is true, all assignments yield perfect covariate balance and, hence, a high likelihood (1) for the true

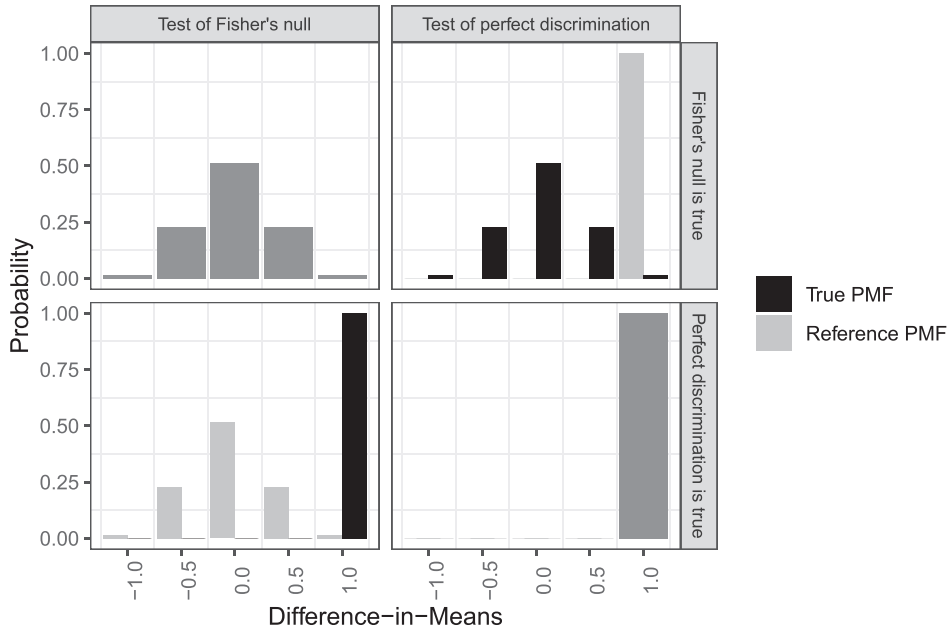

FIGURE 2. *PMFs of true and reference test statistics for tests of Fisher's null and Perfect discrimination.*

Table 3. *Distributions of likelihoods for test of Fisher's null against alternative of Perfect discrimination.*

Fisher's null is true and Perfect discrimination is false					
		Fisher's null		Perfect discrimination	
Likelihood	1/70	16/70	36/70	0	1
Probability	2/70	32/70	36/70	69/70	1/70
Fisher's null is false and Perfect discrimination is true					
		Fisher's null		Perfect discrimination	
Likelihood		1/70		1	
Probability		1		1	

causal hypothesis (Perfect discrimination) and a low likelihood $(1/70)$ for the false causal hypothesis (Fisher's null).

Consequently, for an agent who is neutral about the plausibility of these two causal hypotheses, Table 3 implies that when Fisher's null is true, its posterior probability is almost always greater than the posterior probability of Perfect discrimination. As Table 3 also implies, when Perfect discrimination is true, its posterior probability is always greater than or equal to the posterior probability of Fisher's null. In more general terms and with slight technical adjustments, Leavitt (2023) shows that, for an ex ante neutral agent, the expected value of the 1-dimensional causal effect that maximises the posterior distribution is equal to the true 1-dimensional causal effect or, if effects are heterogeneous, to the true ATE.

However, this general property does not necessarily hold for an agent who is not initially neutral about the plausibility of different causal hypotheses, nor is this property especially useful if there is high variance (over possible assignments) in the causal hypothesis that maximises the posterior. Above and beyond this property in a finite experiment, randomisation ensures another

limiting property under a mild condition on an agent's prior distribution. With minor technical adjustments, Leavitt (2023) also shows that, as the size of an experiment increases indefinitely, an agent's posterior distribution will concentrate on the true hypothesis with probability tending to 1 so long as the true effect is in the support of the prior.

One can begin to see this limiting property as the size of an experiment increases indefinitely from the Lady tasting tea with only $N = 8$. As Figure 2 and Table 3 show, when Fisher's null is true, the probability that Perfect discrimination's likelihood is equal to 0 is 69/70. Since the likelihood of Fisher's null when it is true is bounded away from 0, the probability that the posterior of Fisher's null is equal to 1 is 69/70. This probability will only become higher as the size of the experiment increases indefinitely, holding the experiment's other features constant. Ultimately, given any prior whose support includes the true causal hypothesis, so long as the experiment is sufficiently large, the true causal hypothesis (whether Perfect discrimination or Fisher's null) will receive a high posterior probability (however defined) with probability arbitrarily close to 1.

6 Conclusion

Although significance testing and Bayesian inference differ, randomisation plays an essential role in ensuring that both modes of inference satisfy epistemic properties on their own terms. For significance testing, properties of *p*-values are derived from the CDFs of the true and reference distributions. By contrast, for Bayesian inference, properties of likelihoods are derived from the PMFs of the true and reference distributions. But both functions are implied by the same randomisation procedure and, in both modes of inference, randomisation is valuable for the same reasons.

In general, randomisation implies that, in expectation and with probability limiting to 1 as the size of an experiment increases indefinitely, the Difference-in-Means will be equal to the mean of the reference distribution implied by a true causal effect and will lie in one of the tails of the reference distribution implied by a false causal effect. With slight technical adjustments (Leavitt, 2023), regardless of whether one calculates *p*-values or likelihoods, evidence that lies in the tail of a reference distribution provides less evidential support for a causal hypothesis than evidence that lies in the center of a reference distribution. Consequently, randomisation ensures that false causal hypotheses receive less evidential support than true causal hypotheses in the frameworks of both significance testing and Bayesian inference.

Fisherian balance is crucial to these epistemic properties implied by randomisation. Fisherian balance controls the probability of chance imbalances that would lead to erroneous conclusions. As this paper has aimed to establish, the value of Fisherian balance need not be understood against the background of one specific mode of drawing such conclusions.

Notes

[1]The 'expected distance' and the 'expected data' refer to two different sources of randomness. The former refers to a subjective probability measure (i.e. a probability measure derived from a rational agent's preference structure) over possible states of the world that imply particular values of the unknown causal target. The latter refers to the act itself, which, under deterministic rather than random assignment, would be the expectation of a degenerate distribution.

[2]Under Fisher's null, $\boldsymbol{y}(\boldsymbol{1}) = \boldsymbol{y}(\boldsymbol{0})$, so there is effectively only one vector of potential outcomes; hence, covariate imbalance refers to the mean difference between treated and control groups in this one potential outcomes vector. By contrast, under Perfect discrimination, there are two vectors of potential outcomes, $\boldsymbol{y}(\boldsymbol{1})$ and $\boldsymbol{y}(\boldsymbol{0})$. However, all allowable assignments yield the same mean difference between treated and control groups on both $\boldsymbol{y}(\boldsymbol{1})$ and $\boldsymbol{y}(\boldsymbol{0})$. Hence, covariate

imbalance refers to this single mean difference between treated and control groups for both of these vectors of potential outcomes.

[3]Rosenbaum (2002, 64) refers to this high rejection probability of Fisher's null when Perfect discrimination is true as the 'surprising power of the Lady tasting tea'.

REFERENCES

Bai, Y. (2023). Why randomize? Minimax optimality under permutation invariance. *J. Econ.*, **232**(2), 565–575.

Banerjee, A., Chassang, S., Montero, S. & Snowberg, E. (2020). A theory of experimenters: Robustness, randomization, and balance. *Am. Econ. Rev.*, **110**(4), 1206–1230.

Banerjee, A., Chassang, S. & Snowberg, E. 2017. Decision theoretic approaches to experiment design and external validity. In *Handbook of Field Experiments*, Eds. Duflo, E. & Banerjee, A., Vol. **1**, North-Holland: Amsterdam, NL, pp. 141–174.

Basu, D. (1980). Rejoinder. *J. Am. Stat. Assoc.*, **75**(371), 593–595.

Bertsimas, D., Johnson, M. & Kallus, N. (2015). The power of optimization over randomization in designing experiments involving small samples. *Oper. Res.*, **63**(4), 868–876.

Box, J.F. (1978). *R. A. Fisher, the Life of a Scientist*. Wiley: New York, NY.

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

Ding, P. & Guo, T. (2023). Posterior predictive propensity scores and *p*-values. *Observ. Stud.*, **9**(1), 3–18.

Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58**(3), 403–417.

Fedorov, V.V. (1972). *Theory of Optimal Experiments*. Academic Press: New York, NY.

Fisher, R.A. (1925). *Statistical Methods for Research Workers*, 1 *Edited by* Crew, F.A.E. & Cutler, D.W., Biological Monographs and Manuals. Oliver and Boyd: Edinburgh, SCT.

Fisher, R.A. (1926). The arrangement of field experiments. *J. Ministry Agricul. Great Britain*, **33**, 503–513.

Fisher, R.A. (1935). *The Design of Experiments*. Oliver and Boyd: Edinburgh, SCT.

Hall, N.S. (2007). R. A. Fisher and his advocacy of randomization. *J. Hist. Biol.*, **40**(2), 295–325.

Harshaw, C., Sävje, F., Spielman, D.A. & Zhang, P. (2024). Balancing covariates in randomized experiments with the Gram–Schmidt walk design. *J. Am. Stat. Assoc.*, **2024**, 1–13.

Harville, D.A. (1975). Experimental randomization: Who needs it? *The Am. Stat.*, **29**(1), 27–31.

Howson, C. & Urbach, P. (2006). *Scientific Reasoning: The Bayesian Approach*, 3. Open Court Publishing: Chicago, IL.

Jeffrey, R. 1983. Bayesianism with a human face. In *Testing Scientific Theories*, Ed. Earman, J., Minnesota Studies in the Philosophy of Science, Vol. **X**, University of Minnesota Press, pp. 133–156.

Johansson, P., Rubin, D.B. & Schultzberg, M. (2021). On optimal rerandomization designs. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, **83**(2), 395–403.

Kadane, J.B. & Seidenfeld, T. (1990). Randomization in a Bayesian perspective. *J. Stat. Plan. Infer.*, **25**(3), 329–345.

Kallus, N. (2018). Optimal a priori balance in the design of controlled experiments. *J. R. Stat. Soc.: Ser. B (Stat. Methodol.)*, **80**(1), 85–112.

Kallus, N. 2020. On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. arXiv Preprint, https://arxiv.org/pdf/2005.03151

Kallus, N. (2021). On the optimality of randomization in experimental design: How to randomize for minimax variance and design-based inference. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **83**(2), 404–409.

Kapelner, A., Krieger, A.M., Sklar, M. & Azriel, D. (2022). Optimal rerandomization designs via a criterion that provides insurance against failed experiments. *J. Stat. Plan. Infer.*, **19**, 63–84.

Kapelner, A., Krieger, A.M., Sklar, M., Shalit, U. & Azriel, D. (2021). Harmonizing optimized designs with classic randomization in experiments. *The Am. Stat.*, **75**(2), 195–206.

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Polit. Anal.*, **24**(3), 324–338.

Kiefer, J. (1959). Optimum experimental designs. *J. R. Stat. Soc.. Ser. B (Methodol.)*, **21**(2), 272–319.

Leavitt, T. (2023). Randomization-based, Bayesian inference of causal effects. *J. Causal Infer.*, **11**(1), 20220025.

Lindley, D.V. (1982). The role of randomization in inference. In *Psa: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Vol. **2: Symposia and Invited Papers**, pp. 431–446. University of Chicago Press, Philosophy of Science Association: Chicago, IL.

Martinez, M. & Teira, D. (2024). Why experimental balance is still a reason to randomize. *The British J. Phil. Sci.*, **75**(2), 519–535.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course With Examples in R and Stan*, 2. Chapman & Hall/CRC: Boca Raton, FL.

Nordin, M. & Schultzberg, M. (2022). Properties of restricted randomization with implications for experimental design. *J. Causal Infer.*, **10**(1), 227–245.
Ramsey, F.P. 1931. Truth and probability. In *The Foundations of Mathematics and Other Logical Essays*, Ed. Braithwaite, R.B., Kegan, Paul, Trench, Trubner & Co.: London, UK, pp. 156–198.
Ramsey, F.P. 1990. Knowledge. In *F.P. Ramsey: Philosophical Papers*, Ed. Mellor, D.H., Cambridge University Press: New York, NY, pp. 110–111. 1929.
Rosenbaum, P.R. (2002). *Observational Studies*, 2. Springer: New York, NY.
Rosenbaum, P.R. (2010). *Design of Observational Studies*. Springer: New York, NY.
Rubin, D.B. 1976. Bayesian inference for causality: The importance of randomization. In *American Statistical Association: 1975 Proceedings of the Social Statistics Section*, Ed. Goldfield, E.D., American Statistical Association: Washington, D. C., pp. 233–239.
Rubin, D.B. (1978). Bayesian inference for causal effects: The role of randomization. *The Ann. Stat.*, **6**(1), 34–58.
Rubin, D.B. (1980). Comment on 'Randomization analysis of experimental data in the Fisher randomization test' by Basu, D. *J. Am. Stat. Assoc.*, **75**(371), 591–593.
Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *The Ann. Stat.*, **12**(4), 1151–1172.
Rubin, D.B. (1986). Which ifs have causal answers? (Comment on 'Statistics and causal inference' by Paul W. Holland). *J. Am. Stat. Assoc.*, **81**(396), 961–962.
Savage, L.J. (1954). *The Foundations of Statistics*. John Wiley & Sons: Hoboken, NJ.
Savage, L.J. (1962a). On the foundations of statistical inference: Discussion. *J. Am. Stat. Assoc.*, **57**(298), 307–308.
Savage, L.J. 1962b. Subjective probability and statistical practice. In *The Foundations of Statistical Inference: A Discussion*, Ed. Bartlett, M.S., John Wiley & Sons: New York, NY, pp. 9–35.
Senn, S. (1994). Fisher's game with the devil. *Stat. Med.*, **13**(3), 217–230.
Senn, S. (2013). Seven myths of randomisation in clinical trials. *Stat. Med.*, **32**(9), 1439–1450.
Stone, M. (1969). The role of experimental randomization in Bayesian statistics: Finite sampling and two Bayesians. *Biometrika*, **56**(3), 681–683.
Strevens, M. 2012. Notes on Bayesian confirmation theory.
Suppes, P. (1982). Arguments for randomizing. *PSA: Proc. Biennial Meeting Phil. Sci. Assoc.*, **1982**(2), 464–475.
Urbach, P. (1985). Randomization and the design of experiments. *Phil. Sci.*, **52**(2), 256–273.
Worrall, J. (2002). What evidence in evidence-based medicine? *Phil. Sci.*, **69**(S3), S316–S330.
Worrall, J. (2007a). Evidence in medicine and evidence-based medicine. *Phil. Compass*, **2**(6), 981–1022.
Worrall, J. (2007b). Why there's no cause to randomize. *The British J. Phil. Sci.*, **58**(3), 451–488.
Worrall, J. (2008). Evidence and ethics in medicine. *Perspect. Biol. Med.*, **51**(3), 418–431.
Wu, C.-F. (1981). On the robustness and efficiency of some randomized designs. *The Ann. Stat.*, **9**(6), 1168–1177.

APPENDIX A: Proof of Proposition 1

Proof *Suppose SUTVA and that an assignment, $\mathbf{z}$, yields balance in the treated and control groups' means of treated potential outcomes, that is, following Equation 2:*

$$\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} z_i y_i(1) - \left(\frac{1}{n_0}\right)\sum_{i=1}^{N}(1-z_i)y_i(1) \quad = 0. \tag{A1}$$

Then write the mean of treated potential outcomes under assignment $\mathbf{z}$ as

$$\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} z_i y_i(1) = \left[\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} z_i y_i(1)\right]\left[\left(\frac{n_1}{N}\right) + \left(\frac{N-n_1}{N}\right)\right] \tag{A2}$$

$$= \left(\frac{n_1}{N}\right)\left[\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} z_i y_i(1)\right] + \left(\frac{N-n_1}{N}\right)\left[\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} z_i y_i(1)\right]. \tag{A3}$$

Equation A1 then implies that the mean of treated potential outcomes under assignment $\mathbf{z}$ *in (A3) is*

$$\left(\frac{n_1}{N}\right)\left[\left(\frac{1}{n_1}\right)\sum_{i=1}^{N} z_i y_i(1)\right] + \left(\frac{N-n_1}{N}\right)\left[\left(\frac{1}{N-n_1}\right)\sum_{i=1}^{N}(1-z_i)y_i(1)\right], \tag{A4}$$

which is equal to the true mean of treated potential outcomes, denoted by $\bar{y}(1)$*, since (A4) is the average of treated potential outcomes in the treated and control groups, weighted by the respective proportions of units in treatment and control. That is, if means of treated potential outcomes are balanced between treatment and control groups, then the mean of (observed) treated potential outcomes in the treated group must be equal to the mean of all treated potential outcomes (both observed and unobserved). Analogous logic applies to control potential outcomes, thereby implying that, if balance in means holds for two special covariates (treated and control potential outcomes), then the difference-in-means is exactly equal to the ATE.* □

[*Received October 2023; accepted September 2024*]