A CAUSAL BOOTSTRAP

GUIDO IMBENS[♯]   KONRAD MENZEL[†]

ABSTRACT. The bootstrap, introduced by Efron (1982), has become a very popular method for estimating variances and constructing confidence intervals. A key insight is that one can approximate the properties of estimators by using the empirical distribution function of the sample as an approximation for the true distribution function. This approach views the uncertainty in the estimator as coming exclusively from sampling uncertainty. We argue that for causal estimands the uncertainty arises entirely, or partially, from a different source, corresponding to the stochastic nature of the treatment received. We develop a bootstrap procedure that accounts for this uncertainty, and compare its properties to that of the classical bootstrap.

JEL Classification: C1, C14, C18

Keywords: Potential Outcomes, Causality, Randomization Inference, Bootstrap, Copula, Partial Identification

1. INTRODUCTION

1.1. **Problem Description.** Using the potential outcome framework, e.g., Imbens and Rubin (2015), we are interested in the average causal effect of a binary variable $W_i \in \{0, 1\}$ (the "treatment") on an outcome variable whose potential outcomes we denote with $Y_i(0), Y_i(1)$, for a population of N units $i = 1, \ldots, N$. Implicitly we assume that the potential outcomes $Y_i(w)$ for unit i do not vary with the treatment status assigned to other units, known as the Stable Unit Treatment Value Assumption (SUTVA, Rubin (1978)). For all units in the population we observe the treatment W_i and the realized outcome $Y_i := Y_i(W_i)$. One common estimand is the average effect for the N units in the population:

$$\tau_{ATE} := \frac{1}{N} \sum_{i=1}^{N} \Big(Y_i(1) - Y_i(0) \Big). \tag{1.1}$$

We assume that the data arise from a completely randomized experiment, where $n \leq N$ units are selected at random from the population as experimental subjects, of which n_1 units are then randomly assigned to receive the treatment, and the remaining $n_0 = n - n_1$ units are assigned to the control group. We let $R_i \in \{0, 1\}$ denote an indicator whether the ith unit is included in the sample.

Specifically, for $\mathbf{R} := (r_1, \dots, r_N)'$ and $\mathbf{W} := (w_1, \dots, w_N)'$ we have

$$\mathrm{pr}(\mathbf{R} = \mathbf{r}) = \begin{cases} \binom{N}{n}^{-1} & \text{if } \mathbf{r} \in \{0,1\}^N \text{ and } \sum r_i = n \\ 0 & \text{otherwise} \end{cases}$$

$$\mathrm{pr}(\mathbf{W} = \mathbf{w} | \mathbf{R} = \mathbf{r}) = \begin{cases} \binom{n}{n_1}^{-1} & \text{if } \sum r_i w_i = n_1, \ \mathbf{w} \in \{0,1\}^N \text{ and } (1 - r_i) w_i = 0 \text{ for all } i \\ 0 & \text{otherwise} \end{cases}$$

1.2. **Sampling Uncertainty and Design Uncertainty.** We wish to distinguish between two types of uncertainty in estimators, *sampling uncertainty* arising from the stochastic nature of $\mathbf{R}$, and *design uncertainty* arising from the stochastic nature of $\mathbf{W}$.

To characterize sampling uncertainty we postulate the existence of a large, possibly infinite, population. We draw a random sample from this population, and observe for each unit in this sample a set of values, say, a pair (Y_i, W_i). We may be interested in in the difference between the population averages of Y_i for the subpopulations with $W_i = 0$ and $W_i = 1$. We can estimate this object using the difference in average outcomes by W_i values in the sample. This estimator differs from the target because we do not observe all units in the population. Had we drawn a different random sample, with different units, the value of the estimator would have been different. See Table 1, where R_i is the sampling indicator, equal to 1 for sampled units and 0 otherwise. The uncertainty arising from the randomness in $\mathbf{R}$ is captured by the conventional standard error.

TABLE 1. : SAMPLING-BASED UNCERTAINTY ($\checkmark$ IS OBSERVED, ? IS MISSING)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	Y_i	W_i	R_i	Y_i	W_i	R_i	Y_i	W_i	R_i	...
1	$\checkmark$	$\checkmark$	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	$\checkmark$	$\checkmark$	1	$\checkmark$	$\checkmark$	1	...
4	?	?	0	$\checkmark$	$\checkmark$	1	?	?	0	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	...
N	$\checkmark$	$\checkmark$	1	?	?	0	?	?	0	...

In a randomized experiment the uncertainty is not necessarily of this sampling variety. Instead we can think of the uncertainty arising from the stochastic nature of the assignment, $\mathbf{W}$. For units with $W_i = 0$ we observe $Y_i(0)$, and for units with $W_i = 1$ we observe $Y_i(1)$. In our sample units have a particular set of assignments. In a repeated sampling thought experiment the units in the sample would have remained the same, but their assignments would might been different, leading to a different value for the estimator. See Table 2.

Table 2. : Design-based Uncertainty (✓ is observed, ? is missing)

Unit	Actual Sample			Alternative Sample I			Alternative Sample II			...
	$Y_i(1)$	$Y_i(0)$	W_i	$Y_i(1)$	$Y_i(0)$	W_i	$Y_i(1)$	$Y_i(0)$	W_i	...
1	✓	?	1	✓	?	1	?	✓	0	...
2	?	✓	0	?	✓	0	?	✓	0	...
3	?	✓	0	✓	?	1	✓	?	1	...
4	?	✓	0	?	✓	0	✓	?	1	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
N	✓	?	1	?	✓	0	?	✓	0	...

1.3. **The Bootstrap.** The classical bootstrap corresponds to the case where the uncertainty is purely sampling uncertainty. The bootstrap approximates the cumulative distribution function of the pairs (Y_i, W_i), $F_{YW}(\cdot, \cdot)$ in the population by the empirical distribution function $\hat{F}_{YW}(\cdot, \cdot)$, where

$$\hat{F}_{YW}(w, y) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i \leq y, W_i \leq w\}.$$

It then calculates properties of the estimator given that approximate distribution $\hat{F}_{YW}(\cdot, \cdot)$. One can interpret the standard bootstrap as imputing all the missing values of (Y_i, W_i) in the population by replicates of the observed values, and thus constructing an artificial population from which we then draw random samples. This perspective is helpful to contrast the different approach underlying the causal bootstrap.

1.4. **Notation.** In the following, we denote the distribution of potential outcomes in the population with $F_{01}^p(y_0, y_1) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}$ and the size of that population with N. The distribution in the sample of size n is denoted with $F_{01}^s(y_0, y_1) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}$. $F_{01}^s(\cdot, \cdot)$ is not quite the empirical distribution because we only observe one of the values in each pair $(Y_i(0), Y_i(1))$. We use p superscripts throughout to indicate population quantities, and s superscripts to denote their sample analogs. The number of treated units in the sample is denoted with n_1, the number of control units with $n_0 = n - n_1$, and the respective shares of treated and control units with $p := n_1/n$, so that $1 - p = n_0/n$. We also define the empirical c.d.f. for either potential outcome given the randomized treatment as $\hat{F}_0(y_0) := \frac{1}{n_0} \sum_{i=1}^{N} R_i(1 - W_i)\mathbb{1}\{Y_i(0) \leq y_0\}$ and $\hat{F}_1(y_1) := \frac{1}{n_1} \sum_{i=1}^{N} R_i W_i \mathbb{1}\{Y_i(1) \leq y_1\}$.

3

2. The Causal Bootstrap for Average Treatment Effects

In this section we consider causal bootstrap inference for the population average treatment effect τ_{ATE} defined in (1.1). The estimator we use is the difference in sample averages by treatment status:

$$\hat{\tau}_{ATE} := \overline{Y}_1 - \overline{Y}_0,$$

where

$$\overline{Y}_1 := \frac{1}{n_1} \sum_{i=1}^{N} R_i W_i Y_i, \qquad \text{and} \ \ \overline{Y}_0 := \frac{1}{n_0} \sum_{i=1}^{N} R_i (1 - W_i) Y_i.$$

The repeated sampling perspective we take is one where the potential outcomes $(Y_i(0), Y_i(1))$ are fixed for all N units in the population. The stochastic properties of the estimator arise from the stochastic nature of the assignment and sampling, which are both sources of randomness in the average of realized outcomes by treatment status, where we regard n, n_0, n_1 as fixed.

2.1. The True Variance of the Estimator for the Average Treatment Effect.

Here we present the true variance of the estimator $\hat{\tau}_{ATE}$ under random assignment of the treatment. From the n experimental subjects, n_1 are selected at random to receive the active treatment, and the remainder are assigned to the control group. Define

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

$$S_0^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(0) - \overline{Y}(0) \right)^2, \qquad S_1^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(1) - \overline{Y}(1) \right)^2,$$

and

$$S_{01}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i(1) - Y_i(0) - \tau_{ATE} \right)^2.$$

Then the exact variance of $\hat{\tau}$, over the randomization distribution, is

$$\mathbb{V}(\hat{\tau}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{N}.$$

See, for example, Neyman (1923,1990), Aronow, Green, and Lee (2014), Ding (2017), and Abadie, Athey, Imbens, and Wooldridge (2017).

2.2. An Analytical Variance Estimator.

Define

$$\hat{S}_0^2 = \frac{1}{n_0 - 1} \sum_{i=1}^{N} R_i (1 - W_i) \left(Y_i - \overline{Y}_0 \right)^2, \qquad \text{and} \ \ \hat{S}_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{N} R_i W_i \left(Y_i - \overline{Y}_1 \right)^2.$$

Then the standard variance estimator is

$$\hat{\mathbb{V}}_{\text{Neyman}} := \frac{\hat{S}_0^2}{n_0} + \frac{\hat{S}_1^2}{n_1}.$$

This estimator ignores the third term in the variance, which is negative, so $\hat{\mathbb{V}}_{\text{Neyman}}$ in general overestimates the true variance. It is possible to give sharp bounds for S_{01}^2 given the respective marginal distributions of $Y_i(0)$ and $Y_i(1)$. Aronow, Green, and Lee (2014) proposed a consistent estimator for the resulting bounds on $\mathbb{V}(\hat{\tau})$ that can be expressed as

$$\hat{\mathbb{V}}_{\text{AGL}} := \frac{\hat{S}_0^2}{n_0} + \frac{\hat{S}_1^2}{n_1} - \frac{\hat{\underline{S}}_{01}}{N}$$

where $\hat{\underline{S}}_{01}^{*}$ is an estimator of the sharp lower bound for S_{01}^2.[1]

2.3. **The Causal Bootstrap.** Here we initially take the perspective that the uncertainty is solely arising from the stochastic nature of the assignment, as in Table 2. In the spirit of the above interpretation of the standard bootstrap, we use the observed data to impute all the missing values in the population. Then we simulate the estimator using this partly imputed population.

The difference with the standard bootstrap is in the nature of the missing data process, and how we impute them. Consider unit 1 in Table 2. In the actual sample this unit receives the active treatment, and so we observe $Y_1(1)$, but we do not know the value of the control outcome for this unit, $Y_1(0)$.

A natural approach is to impute the missing value of $Y_1(0)$ using one of the observed values for $Y_i(0)$, that is, one of the realized values of Y_i for control units. The question is which one to use. It turns out that it matters how we choose to impute the missing values from the observed values. This issue is related to the term S_{01}^2 in the true variance of the estimator $\hat{\tau}$ for the average treatment effect, the term that is not consistently estimable, and which we typically ignore in practice.

To frame this question, it is useful to start with the joint distribution function of the pairs of potential outcomes in the population,

$$F_{01}^p(y_0, y_1) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\right\}.$$

The average treatment effect, and any other causal parameters of interest, can be written as a functional of this distribution,

$$\tau := \tau(F_{01}^p).$$

---

[1]Such an estimator is

$$\hat{\underline{S}}_{01} := \hat{S}_0^2 + \hat{S}_1^2 - 2\hat{\sigma}_N^H(y_0, y_1)$$

where $\hat{\sigma}_N^H(y_0, y_1)$ is a consistent estimator for the upper bound for $\text{Cov}(Y_i(0), Y_i(1))$, see equation (8) of their paper.

Given F_{01}^p, the assignment mechanism completely determines the distribution of any estimator, for example the difference in averages by treatment status, $\hat{\tau}$. This is similar to the way in which in the sampling case knowledge of the joint population distribution allows us to deduce the properties of any estimator.

The problem, and the main difference with the sampling case is that for each unit in the population, at most one of the two potential outcomes $Y_i(0)$ and $Y_i(1)$ is observed so that there is no consistent estimator for $F_{01}^p(\cdot, \cdot)$: In general, the joint distribution of potential values can be written as

$$F_{01}^p(y_0, y_1) = C(F_0^p(y_0), F_1^p(y_1))$$

where the copula $C : [0,1]^2 \mapsto [0,1]$ is a function that is nondecreasing in either argument for each value of x. By Sklar's theorem (e.g. stated as Theorem 2.3.3 in Nelsen (2006)), such a copula exists even though it need not be unique unless the marginal distributions F_0^p, F_1^p are continuous. In the following, we let

$$\mathcal{C} := \left\{ C : [0,1]^2 \mapsto [0,1], \ C(u,v) \text{ nondecreasing in } u \text{ and } v \right\}$$

denote the set of all possible copulae.

It is important to note that although the marginal distributions F_0^p, F_1^p can be estimated consistently from a completely randomized experiment as sample size grows, the data on realized treatments and outcomes impose no empirical restrictions on the copula $C(u,v)$ for the joint distribution of $(Y_i(0), Y_i(1))$. Hence, neither the parameter $\tau(F_{01}^s) = \tau(C(F_0^s, F_1^s))$ nor the distribution of an estimator $\hat{\tau}$ need in general not be point-identified.

In the spirit of the variance estimator in Aronow, Green, and Lee (2014), we address this challenge by simulating the distribution of $\hat{\tau}$ using an estimator for the population distribution F_{01}^p that is *conservative* with respect to the copula in a sense to be made more precise below. To illustrate the broader conceptual idea, consider an estimator

$$\hat{\tau} := \tau(\hat{F}_0, \hat{F}_1)$$

for a general functional $\tau(F_{01})$ of the distribution of potential values. Under regularity conditions,[2] such an estimator admits a stochastic expansion of the form

$$\hat{\tau} - \tau(F_{01}^p) = \mu(F_{01}^p) + n^{-1/2}\sigma(F_{01}^p)Z + n^{-1}\kappa(F_{01}^p) + o_P(n^{-1})$$

where $Z \sim N(0,1)$. The first-order "bias" term

$$\mu(F_{01}^p) := \mathbb{E}_{F_{01}^p}[\hat{\tau}] - \tau(F_{01}^p)$$

and the scale parameter

$$\sigma^2(F_{01}^p) := \lim_N n \text{Var}_{F_{01}}(\hat{\tau})$$

---

[2]See e.g. Bloznelis and Götze (2001) for regularity conditions for finite-population expansions of this type.

are deterministic functions of the unknown distribution $F_{01}^p = C(F_0^p, F_1^p)$, and the limit for the asymptotic variance is taken as N and $n := n_N$ grow large. The second-order approximation error $\kappa(F_{01}^p)$ is a tight random variable whose distribution also depends on F_{01}^p.

If the functional $\tau(F_{01}^p)$ is not point-identified, then $\mu(F_{01}^p)$ may take values in a set whose bounds may be characterized in terms of the marginal distributions F_0^p, F_1^p. Specifically, given the marginal distributions F_0^p, F_1^p we have sharp bounds of the form

$$\mu_L(F_0^p, F_1^p) := \inf_{\tilde{C} \in \mathcal{C}} \mu(\tilde{C}(F_0^s, F_1^s)) \leq \mu(F_{01}^s) \leq \sup_{\tilde{C} \in \mathcal{C}} \mu(\tilde{C}(F_0^s, F_1^s)) =: \mu_U(F_0^p, F_1^p),$$

that are generally available, see e.g. Heckman, Smith, and Clements (1997) and Manski (1997). Similarly we can form bounds for the variance,

$$\sigma_L(F_0^p, F_1^p) := \inf_{\tilde{C} \in \mathcal{C}} \sigma(\tilde{C}(F_0^s, F_1^s)) \leq \sigma(F_{01}^s) \leq \sup_{\tilde{C} \in \mathcal{C}} \sigma(\tilde{C}(F_0^s, F_1^s)) =: \sigma_U(F_0^p, F_1^p)$$

For a given inference problem, the bootstrap has to estimate these quantities conservatively with respect to the unknown copula $C(\cdot)$, which can be done iteratively as follows: we first need to determine which couplings C_0^* attain the value of $\mu(C_0^*(F_0, F_1))$ which is least favorable for the inference problem at hand. Within the (not necessarily singleton) set $\mathcal{C}_0^*$ of such couplings, we then determine the least-favorable value of $\sigma(C_1^*(F_0, F_1))$ for $C_1^* \in \mathcal{C}_0^*$. We can apply this principle recursively either until the resulting set $\mathcal{C}_k^*$ contains a unique copula, or until we reach the order of approximation desired for formal results regarding the bootstrap procedure. This results in an estimate $\hat{F}_{01}^* := C_k^*(\hat{F}_0, \hat{F}_1)$ for the population distribution F_{01}^p that is conservative regarding the inference task at hand. The causal bootstrap then approximates the distribution of $\hat{\tau}$ by sampling and randomization from a population $\hat{F}_{01}^*$ using the known sampling and assignment mechanism.

2.4. **Least Favorable Coupling for the Average Treatment Effect.** In this paper, we consider the special case of two-sided confidence intervals based on a t-ratio for the sample average treatment effect. The case of the average treatment effect has been the main focus of the previous literature. It is a special case for our problem in that the copula does not matter for estimation - by inspection, the functional

$$\tau_{ATE}(F_{01}) = \mathbb{E}_{F_{01}}[Y_i(1)] - \mathbb{E}_{F_{01}}[Y_i(0)] = \mathbb{E}_{F_1}[Y_i(1)] - \mathbb{E}_{F_0}[Y_i(0)] =: \tau(F_0, F_1)$$

does not depend on the copula, and the default estimator

$$\hat{\tau}_{ATE} := \tau(\hat{F}_0, \hat{F}_1) \equiv \frac{1}{n_1} \sum_{i=1}^{N} R_i W_i Y_i - \frac{1}{n_0} \sum_{i=1}^{N} R_i (1 - W_i) Y_i$$

is known to be unbiased for $\tau(C(F_0, F_1))$ under any coupling so that $\mu(C(F_0, F_1)) \equiv 0$ for each C. In order to ensure that the estimand is well-defined and satisfies other regularity conditions for the bootstrap, we make the following assumptions:

Assumption 2.1. *The first four moments of $F_0(y_0)$ and $F_1(y_1)$ are bounded.*

Also for a two-sided confidence interval constructed from inverting a t-test based on $\hat{\tau}_{ATE}$, the least favorable coupling must attain the upper bound for the asymptotic variance,

$$\sigma_U^2(F_0, F_1) = \sup_{C \in \mathcal{C}} \sigma^2(C(F_0, F_1)) =: \sigma^2(F_0, F_1)$$

We next show that $\sigma^2(C(F_0, F_1))$ is uniquely maximized at the joint distribution corresponding to the isotone assignment which matches values of $Y_i(0)$ to values of $Y_i(1)$ while preserving their respective marginal distributions. More formally, the joint distribution of the potential outcomes under the isotone coupling is characterized by the copula

$$C^{iso}(u, v) := \min\{u, v\}$$

We find that the upper bound on the variance is in fact uniquely attained at the isotone coupling. Therefore an estimator for the distribution of $\hat{\tau}_{ATE}$ which assumes the isotone coupling is asymptotically conservative at any order of approximation.

Proposition 2.1. (Least Favorable Coupling for the ATE). *Suppose that Assumption 2.1 holds. Then, given the marginal distributions F_0, F_1, the variance bound is uniquely attained at*

$$\sigma^2(F_0, F_1) := \lim_N n \mathrm{Var}_{F_{01}^{iso}}(\hat{\tau})$$

where $F_{01}^{iso} := C^{iso}(F_0, F_1)$ is the joint distribution corresponding to the isotone coupling.

The fact that the variance bound is attained at the isotone coupling is widely known (see e.g. Becker (1973), Fan and Park (2010), Stoye (2010), and Aronow, Green, and Lee (2014)), for expositional purposes we provide a proof in the appendix. We establish the slightly stronger conclusion that the distribution under the isotone coupling is in fact maximal with respect to second-order stochastic dominance. For our approach it is also important to establish that this maximum is unique in the sense that the joint distribution resulting from any other coupling yields a variance that is strictly lower than $\sigma_U^2(F_0, F_1)$. In particular, for confidence intervals based on the Gaussian asymptotic distribution, the isotone coupling does indeed constitute the least favorable coupling.

We also want to stress that there are other causal estimands of interest (including the distribution of treatment effects and its quantile) for which the isotone assignment is not the least favorable coupling (see e.g. Heckman, Smith, and Clements (1997), Fan and Wu (2010), Fan and Park (2010), Lu, Ding, and Dasgupta (2018)). Cambanis, Simmons, and

Stout (1976) give conditions under which the isotone assignment does in fact constitute the least favorable bound for a functional of the joint distribution.

2.5. **Related Literature.** Worst case bounds on the distributions of potential outcomes and treatment effects and their quantiles have been analyzed by Heckman, Smith, and Clements (1997), Manski (1997), Firpo and Ridder (2008), Fan and Park (2010), Fan and Park (2012), Fan and Wu (2010), and Lu, Ding, and Dasgupta (2018). This literature uses theoretical results on dependency bounds for functions of several random variables which were developed among others by Cambanis, Simmons, and Stout (1976), Makarov (1982), Frank, Nelsen, and Schweizer (1987), and Williamson and Downs (1990). Stoye (2010) establishes that a class of spread parameters is monotone with respect to conventional stochastic orders of distribution, and shows how to derive parameter bounds for causal inference. Several of these studies also propose inference procedures that account for sampling uncertainty rather than randomization error. In contrast, for our problem we need to explicitly construct the respective couplings that achieve the lower and upper bounds to the parameter, and in addition the largest randomization variance for an estimator of either bound.

Robins (1988) proposes a confidence interval for a causal parameter based on the least-favorable coupling for a binary outcome variable. Aronow, Green, and Lee (2014) propose an estimator of the sharp upper bound for the randomization variance of the average treatment effect in completely randomized experiments. Our approach of embedding the finite-population randomization distribution into an asymptotic sequence of sampling experiments closely follows Abadie, Athey, Imbens, and Wooldridge (2017). Our results make use of a finite-population CLT for the empirical process developed by Bickel (1969) for the two-sample problem. Finite-sample central limit theorems for randomization inference were also provided by Li and Ding (2017). Bootstrap methods for sampling from finite populations (without replacement) have been proposed by Bickel and Freedman (1984) and Booth, Butler, and Hall (1994). For this problem the main challenge in generating the finite bootstrap population is that the size of the super-population N may be a non-integer multiple of n. We propose a new alternative for estimating the potential outcome distribution for a super-population of exact size N and for which the marginal distributions coincide with their empirical analogs up to rounding error.

2.6. **Comparison to Fisher's Exact Test.** Bootstrap inference on the average treatment effect as proposed in this paper bears some conceptual similarities with Fisher's exact test of the sharp null of no unit-level treatment effect (see e.g. Rosenbaum (2002), Imbens and Rubin (2015), Ding (2017)), $Y_i(0) = Y_i(1)$ with probability 1. One important distinction is that the justification for our procedure is only asymptotic, whereas the Fisher exact test is valid in finite samples.

Furthermore, the Fisher exact test evaluates the randomization distribution of the estimated ATE under the sharp null of no or a constant unit-level treatment effect. The sharp null not only implies that the joint distribution of $Y_i(0)$ and $Y_i(1)$ corresponds to the isotone assignment, but also equality of the marginal distributions $F_0(y) = F_1(y)$, which may in fact be rejected by the data under the null of a zero *average* treatment effect. In that case even a conservative estimator of the randomization variance may in fact be smaller than that implied by zero, or constant, unit-level effects. More generally, when Fisher's sharp null fails and $F_0(y) \neq F_1(y)$, the bootstrap estimate of the randomization variance can in several important scenarios be smaller than that implicit in Fisher's exact test, in which case our procedure is asymptotically more powerful.

Specifically, standard variance calculations (see e.g. Ding (2017)) imply that the implicit variance estimate for Fisher's exact test under the null of no average effect is

$$\mathbb{V}_{Fisher}(\hat{\tau}_{ATE}) = \text{Var}(Y_i)\left(\frac{1}{n_1} + \frac{1}{n_0}\right) = \frac{n_0 S_0^2 + n_1 S_1^2}{n}\left(\frac{1}{n_1} + \frac{1}{n_0}\right).$$

We can compare this to the actual variance stated in Section 2.1,

$$\mathbb{V}(\hat{\tau}) = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{N}.$$

Our bootstrap procedure implies a conservative estimate, i.e. a sharp lower bound for S_{01}^2 from the isotone coupling of the potential outcomes, which is strictly positive whenever the marginal distributions of $Y_i(0)$ and $Y_i(1)$ are not the same. The comparison between the terms $\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$ and $\text{Var}(Y_i)\left(\frac{1}{n_1} + \frac{1}{n_0}\right)$ is generally ambiguous - Ding (2017) describes several cases in which the randomization variance implied by Fisher's test is strictly larger, and his conclusions carry over to the bootstrap procedure in this paper. On the other hand it is important to note that when $\frac{S_1^2}{n_1} + \frac{S_0^2}{n_0} > \text{Var}(Y_i)\left(\frac{1}{n_1} + \frac{1}{n_0}\right)$, Fisher's exact test over-rejects under the null hypothesis of no average treatment effect, so that this potential power advantage for the Fisher test only arises in situations in which the exact test does not provide a valid test of that null. We illustrate this possibility using Monte Carlo simulations in Section 4.

The relationship between Fisher's sharp null and Neyman's null hypothesis of no average effect is clarified in Ding (2017), who also shows that Neyman's test of the null of no average effect is weakly more powerful against alternatives than Fisher's exact test. Fisher's exact null also implies that the distribution of Δ_i is degenerate at a constant, however the power comparison for the ATE does not carry over to set-identified objects like quantiles or the c.d.f. of Δ_i since the bounds for the identified set are typically not attained at the isotone coupling that is implied by the sharp null.

3. Bootstrap Procedure

This section describes the bootstrap procedure for confidence intervals for the average treatment effect, in which case the least favorable coupling is the isotone (rank-preserving) assignment by Proposition 2.1. The method allows for sampling and randomization uncertainty, where we consider a sampling experiment under which the researcher observes n units that are selected at random out of a population of N units. For the purposes of asymptotic approximations, we assume that the population of interest in turn consists of N i.i.d. draws from an encompassing distribution F_{01}.

Assumption 3.1. *(Sampling Experiment)* *The population consists of N units with potential values $(Y_i(0), Y_i(1))_{i=1}^{N}$ which are i.i.d. draws from the distribution $F_{01}(y_0, y_1)$. The n observed units are sampled at random and without replacement from the population,*

$$Y_i(0), Y_i(1) \perp\!\!\!\perp R_i$$

where we denote $q := \frac{n}{N} \in (0, 1]$.

We assume throughout that the treatment $W_i \in \{0, 1\}$ is binary, and that the outcome $Y_i(W_i)$ for unit i does not vary with the treatment status assigned to other units. The latter requirement is also known as individualistic treatment response, or Stable Unit Treatment Value Assumption (SUTVA). We assume furthermore that the experiment is completely randomized:

Assumption 3.2. *(Complete Randomization)* *Treatment assignment is completely randomized, that is for each unit with $R_i = 1$ we have*

$$(Y_i(0), Y_i(1)) \perp\!\!\!\perp W_i$$

where $W_i = 1$ for n_1 units selected at random and without replacement from the n observations with $R_i = 1$, and the propensity score $p := \frac{n_1}{n}$ satisfies $0 < p < 1$.

For greater clarity of exposition we also assume that the researcher observes no further covariate information. The approach of this paper can be generalized to observational studies under unconfoundedness, and experiments with imperfect compliance for which unconfoundedness fails, but intention to treat is (conditionally) independent of potential outcomes and can serve as an instrumental variable to identify causal effects on a population of compliers.

Given a sample generated according to Assumptions 3.1 and 3.2, we denote the point estimate for the average treatment effect

$$\hat{\tau} := \tau(\hat{F}_0, \hat{F}_1)$$

and the upper variance bound

$$\hat{\sigma} := \sigma(\hat{F}_0, \hat{F}_1)$$

For the purposes of this paper, the main target of interest for the causal bootstrap is the distribution of the t-ratio

$$T := \sqrt{n}\frac{\hat{\tau} - \tau}{\hat{\sigma}}$$

3.1. **Bootstrap Algorithm.** The proposed bootstrap algorithm proceeds in four main steps:

(1) We obtain nonparametric estimates of the potential outcome distributions $F_0(y_0)$ and $F_1(y_1)$ from the units for which $W_i = 0$ ($W_i = 1$, respectively) in the actual experiment.

(2) We create an empirical population of size N, $\left(\tilde{Y}_i, \tilde{W}_i\right)_{i=1}^N$ by generating an appropriate number of replicas of the sample of n draws for W_i, Y_i. If the sample is the population, $n = N$, we can skip this step.

(3) We then impute potential values $\tilde{Y}_i(0), \tilde{Y}_i(1)$ for each unit $i = 1, \ldots, N$, where $\tilde{Y}_i(\tilde{W}_i) = \tilde{Y}_i$ and $\tilde{Y}_i(1 - \tilde{W}_i)$ is obtained from the estimated potential outcome distributions and the least-favorable copula for the parameter of interest.

(4) Finally, we simulate the randomization distribution by repeatedly drawing n units $Y_i^*(0), Y_i^*(1)$ out of that empirical population without replacement and generating randomization draws $W_1^*, \ldots, W_n^*$. We then evaluate the sample average treatment effect for the bootstrap sample $(Y_i^*(W_i^*), W_i^*)_{i=1}^n$ obtained using the imputed potential outcomes.

Given the simulated randomization distribution for the estimated bounds, we can estimate the percentiles of the t-ratios that are needed to construct confidence intervals for the functional. We next describe each of these steps in greater detail.

3.2. **Generating the Empirical Population.** To obtain the empirical population of size N, we generate replicates of the n observed units, however not necessarily of the same number for each observation when N is not an integer multiple of n. We propose the following procedure for doing so:

- We create the samples $\left(Y_j^0, \hat{U}_j^0\right)_{j=1}^{n_0}$ of values for Y_i for the n_0 units with $W_i = 0$, and $\left(Y_j^1, \hat{U}_j^1\right)_{j=1}^{n_1}$ with values Y_i for the n_1 units with $W_i = 1$. We assume that each sample is ordered, $Y_k^w \leq Y_{k+1}^w$ for all k, and the rank variable $\hat{U}_j^w = \frac{j}{n_w}$ for $w = c, t$.

- Let $N_0 = \lceil \frac{n_0}{n}N \rceil$ and $N_1 = N - N_0$. We generate the empirical population $\left(\tilde{Y}_i, \tilde{W}_i\right)_{i=1}^N$ by including $M_j^0 := \lceil \hat{U}_{j+1}^0 N_0 \rceil - \lceil \hat{U}_j^0 N_0 \rceil$ copies of Y_j^0 with $\tilde{W}_j = 0$ and $M_j^1 := \lceil \hat{U}_{j+1}^1 N_1 \rceil - \lceil \hat{U}_j^1 N_1 \rceil$ copies of Y_j^1 with $\tilde{W}_j = 1$.

Since the respective maxima of $\hat{U}_j^0, \hat{U}_j^1$ are equal to 1 for either of the two strata (corresponding to $W_i = 0$ and $W_i = 1$, respectively), $\sum_{j=1}^n ((1 - \tilde{W}_{(j)})M_{(j)}^0 + \tilde{W}_{(j)}M_{(j)}^1) = \lceil N_0 \rceil + \lceil N_1 \rceil = N$ so that this procedure ensures that the empirical population has size

equal to N. Also, for n and N sufficiently large, the respective empirical distributions of $\tilde{Y}_i$ among units with $\tilde{W}_i = 0$ and $\tilde{Y}_i$ among units with $\tilde{W}_i = 1$ are, up to an approximation error of the order n^{-1}, equal to $\hat{F}_0$ and $\hat{F}_1$, respectively.

3.3. Imputing Missing Counterfactuals.

For the specific case of two-sided inference for the average treatment effect, Proposition 2.1 shows that the least favorable coupling corresponds to the isotone assignment $C^{iso}(u, v) := \min\{u, v\}$. For other inference problems, the missing counterfactuals would have to be imputed by drawing from the appropriate least-favorable coupling, following the strategy outlined in Section 2.3.

In order to generate an empirical population with joint distribution $\hat{F}_{01}^{iso} := C^{iso}(\hat{F}_0, \hat{F}_1)$, we can simply impute the missing counterfactuals according to:

$$
\begin{aligned}
\tilde{Y}_i(0) &:= \begin{cases} \tilde{Y}_i & \text{if } \tilde{W}_i = 0 \\ \hat{F}_0^{-1}\left(\hat{F}_1(\tilde{Y}_i)\right) & \text{otherwise} \end{cases} \\[2mm]
\tilde{Y}_i(1) &:= \begin{cases} \tilde{Y}_i & \text{if } \tilde{W}_i = 1 \\ \hat{F}_1^{-1}\left(\hat{F}_0(\tilde{Y}_i)\right) & \text{otherwise} \end{cases}
\end{aligned}
\tag{3.1}
$$

For functionals $\tau(F_{01})$ of the potential outcome distribution other that the average treatment effect, or inference problems other than two-sided confidence intervals, the least favorable coupling will be of a different form, so this step would have to be replaced by a procedure imputing the missing counterfactuals from a different coupling.

3.4. Resampling Algorithm.

For the bth bootstrap replication, we initially draw n units $(Y_{ib}^*(0), Y_{ib}^*(1))$ from the empirical population at random and without replacement.

Given a known propensity score $p := P(W_i | R_i = 1)$, for the bth bootstrap replication we can generate $W_{1b}^*, \ldots, W_{nb}^*$ as independent Bernoulli draws with success probability $P(W_{ib}^* = 1) = p$ and obtain the bootstrap sample $Y_{1b}^*, \ldots, Y_{nb}^*$, where $Y_{ib}^* := Y_{ib}^*(W_{ib}^*)$.

We can then compute the bootstrap analogs of the estimated c.d.f.s $\hat{F}_{0b}^*(y_0) := \frac{1}{n_0} \sum_{i=1}^N R_{ib}^*(1 - W_{ib}^*)\mathbb{1}\{Y_{ib}^* \le y_0\}$ and $\hat{F}_{1b}^* := \frac{1}{n_1} \sum_{i=1}^N R_{ib}^* W_{ib}^* \mathbb{1}\{Y_{ib}^* \le y_1\}$, the corresponding estimates of the average treatment effect, and the variance bound,

$$
\hat{\tau}_b^* := \tau(\hat{F}_{0b}^*, \hat{F}_{1b}^*)
$$

$$
\hat{\sigma}_b^* := \sigma(\hat{F}_{0b}^*, \hat{F}_{1b}^*)
$$

We then record the studentized values of the bootstrap estimates,

$$
T_b^* := \sqrt{n}\frac{\hat{\tau}_b^* - \hat{\tau}}{\hat{\sigma}_b^*}
$$

Repeating the resampling step B times, we obtain a sample $(T_1^*, \ldots, T_B^*)$ that constitutes independent draws from the bootstrap estimator of the randomization distribution and can be used to construct critical values for tests or confidence intervals.

13

3.5. **Confidence Intervals.** We consider confidence intervals constructed by inverting a t-test based on the point estimate $\hat{\tau} := \tau(\hat{F}_0, \hat{F}_1)$ and given the variance bound $\hat{\sigma} := \sigma(\hat{F}_0, \hat{F}_1)$ introduced before. The proposed confidence intervals for τ are then of the form

$$CI(1 - \alpha) := \left[\hat{\tau} - n^{-1/2}\hat{\sigma}\hat{c}(1 - \alpha), \hat{\tau} - n^{-1/2}\hat{\sigma}\hat{c}(\alpha)\right] \qquad (3.2)$$

We use bootstrap approximations to the randomization distribution of the t-ratio $n^{1/2}(\hat{\tau} - \tau)/\hat{\sigma}$ under the least favorable coupling in order to determine the critical values. Specifically, let $\hat{G}(z) := \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{T_b^* \leq z\}$ denote the empirical distribution for the bootstrap samples obtained from the previous step. We then estimate the critical values using $\hat{c}(\alpha) := \hat{G}^{-1}(\alpha)$ and $\hat{c}(1 - \alpha) := \hat{G}^{-1}(1 - \alpha)$.

4. Monte Carlo Simulations

We next compare the performance of this causal bootstrap with the standard bootstrap and other alternative methods based on sampling or randomization designs. Specifically, we consider confidence intervals using Gaussian critical values the respective analytic estimators of the sampling variance $\hat{\mathbb{V}}_{\text{Neyman}}$ and the causal variance $\hat{\mathbb{V}}_{\text{AGL}}$ given in Section 2.2. We also consider Gaussian inference using the variance estimators $\hat{\mathbb{V}}_{\text{s-boot}}$ and $\hat{\mathbb{V}}_{\text{c-boot}}$ obtained from the classical (sampling) bootstrap, and the causal bootstrap proposed in this paper. We compare these to confidence intervals from inverting Fisher's exact test, and confidence intervals from the standard and the causal bootstrap for the t-statistic based on either sampling or causal variance estimate. Throughout we will restrict our attention to the case $n = N$, i.e. when the full population of interest is observed.

We first consider three different simulation designs to illustrate the main points of comparison between the causal bootstrap and its main alternatives.

- Design I sets $n_0 = n_1 = 100$ and draws potential outcomes according to $Y_i(0) \sim \mathcal{N}(0, 1), Y_i(1) = Y_i(0)$. In this setting, treatment effects are constant at $Y_i(1) - Y_i(0) \equiv 0$ and the marginal distributions $F_0^p(y) \equiv F_1^p(y)$, so that all procedures should be expected to do well.
- For Design II we again have $n_0 = n_1 = 100$, but generate potential outcomes as $Y_i(0) \sim \mathcal{N}(0, 1)$, and $Y_i(1) = 0$. In that case, the marginal distributions $F_0^p(y)$ and $F_1^p(y)$ are different, so that causal standard errors and the causal bootstrap should do better than their sampling analogs.
- For Design III, $n_0 = n_1 = 10$, and we generate non-Gaussian potential outcomes where $Y_i(0)$ is mixture, with probability 0.9 it is $\mathcal{N}(0, 1)$ and with probability 0.1 it is $\mathcal{N}(0, 16)$, $Y_i(1) = 0$. This design highlights the difference between the bootstrap and the Gaussian distribution, which is no longer exact for this design.

14

Simulation results are shown in Table 4, where we compare coverage rates of nominal 95% confidence intervals, and the corresponding standard errors for each of the three designs. If a particular method does not directly calculate standard errors, we calculate the standard errors by taking the ratio of the difference between the upper and lower limit of the confidence interval and dividing by 2 times 1.96. As expected, all methods do well for the first design. In the second design, only inference based on $\hat{\mathbb{V}}_{\mathrm{AGL}}$ and the causal bootstrap (pivotal or not) do well. In the third design the pivotal causal bootstrap does better than Gaussian inference using either $\hat{\mathbb{V}}_{\mathrm{AGL}}$ or $\hat{\mathbb{V}}_{\mathrm{c-boot}}$ as an estimator for the asymptotic variance.

TABLE 3. 95% Confidence Intervals And Standard Errors

Variance Estimator	Bootstrap Version	Pivotal Statistic	Design I Cov Rate	Design I Med s.e.	Design II Cov Rate	Design II Med s.e.	Design III Cov Rate	Design III Med s.e.
$\hat{\mathbb{V}}_{\mathrm{Neyman}}$	N/A	No	0.9928	0.0997	0.9524	0.1410	0.9588	0.4487
$\hat{\mathbb{V}}_{\mathrm{AGL}}$	N/A	No	0.9480	0.0705	0.9476	0.1394	0.8858	0.3173
$\hat{\mathbb{V}}_{\mathrm{s-boot}}$	Standard	No	0.9922	0.0993	0.9502	0.1403	0.9506	0.4255
$\hat{\mathbb{V}}_{\mathrm{c-boot}}$	Causal	No	0.9472	0.0703	0.9480	0.1390	0.8792	0.3090
Fisher's Exact Test			0.9584	0.0997	0.9772	0.1410	0.9644	0.4494
$\hat{\mathbb{V}}_{\mathrm{Neyman}}$	Standard	Yes	0.9938	0.1011	0.9600	0.1421	0.9758	0.6212
$\hat{\mathbb{V}}_{\mathrm{AGL}}$	Standard	Yes	0.9938	0.1011	0.9576	0.1444	0.9758	0.6212
$\hat{\mathbb{V}}_{\mathrm{Neyman}}$	Causal	Yes	0.9492	0.0714	0.9564	0.1407	0.9224	0.4375
$\hat{\mathbb{V}}_{\mathrm{AGL}}$	Causal	Yes	0.9492	0.0714	0.9560	0.1423	0.9224	0.4375

We next illustrate the role of the coupling of the potential values where we draw $(Y_i(0), Y_i(1))$ from a bivariate Gaussian distribution with variances $\mathrm{Var}(Y_i(0)) = 0.5$ and $\mathrm{Var}(Y_i(1)) = 2$ and correlation coefficient of the two potential values, $\varrho_{01} \in \{-1, 0, 1\}$. From our theoretical results, we should expect Gaussian inference using causal standard errors and the causal bootstrap to have asymptotically exact coverage under the isotonic coupling $\varrho_{01} = 1$ and be conservative when $\varrho_{01} < 1$. Furthermore, for any coupling this design implies heterogeneous treatment effects, so that Fisher's exact test does not in general control nominal confidence size for the average treatment effect. Given the calculations in Section 2.6 we designed the experiment deliberately to illustrate the potential of Fisher's exact procedure to underestimate the spread of the randomization distribution, where $n_0 > n_1$ and $\mathrm{Var}(Y_i(1)) > \mathrm{Var}(Y_i(0))$. Since the potential outcomes follow a Gaussian distribution, we should not expect refinements for the bootstrap relative to Gaussian inference.

In Table 4 we report simulated coverage rates for nominal 95% confidence intervals for the average treatment effect, where for either sample size we report the lowest coverage rate across the three different couplings in a separate column. The simulation results broadly confirm the theoretical predictions. Overcoverage from using sampling-based, rather than causal estimators for the variance or the bootstrap is not evident from the design with smaller

Table 4. Coverage of nominal 95% Confidence Intervals, Gaussian Potential Outcomes with Different Couplings

Variance Estimator	Bootstrap Version	Pivotal Statistic	$(n_0, n_1) = (50, 20)$				$(n_0, n_1) = (200, 80)$			
			$\varrho_{01} = 1$	$\varrho_{01} = 0$	$\varrho_{01} = -1$	minimum	$\varrho_{01} = 1$	$\varrho_{01} = 0$	$\varrho_{01} = -1$	minimum
$\hat{\mathbb{V}}_{\text{Neyman}}$	N/A	No	0.9560	0.9656	0.9832	0.9560	0.9650	0.9796	0.9880	0.9650
$\hat{\mathbb{V}}_{\text{AGL}}$	N/A	No	0.9352	0.9510	0.9730	0.9352	0.9462	0.9664	0.9818	0.9462
$\hat{\mathbb{V}}_{\text{s-boot}}$	Standard	No	0.9508	0.9616	0.9804	0.9508	0.9636	0.9778	0.9878	0.9636
$\hat{\mathbb{V}}_{\text{c-boot}}$	Causal	No	0.9308	0.9490	0.9706	0.9308	0.9452	0.9654	0.9838	0.9452
Fisher's Exact Test			0.9332	0.9112	0.8948	0.8948	0.8624	0.8638	0.8616	0.8616
$\hat{\mathbb{V}}_{\text{Neyman}}$	Standard	Yes	0.9652	0.9754	0.9878	0.9652	0.9660	0.9792	0.9886	0.9660
$\hat{\mathbb{V}}_{\text{AGL}}$	Standard	Yes	0.9632	0.9744	0.9878	0.9632	0.9656	0.9786	0.9886	0.9656
$\hat{\mathbb{V}}_{\text{Neyman}}$	Causal	Yes	0.9444	0.9610	0.9776	0.9444	0.9492	0.9684	0.9836	0.9492
$\hat{\mathbb{V}}_{\text{AGL}}$	Causal	Yes	0.9432	0.9608	0.9774	0.9432	0.9490	0.9684	0.9832	0.9490

sample sizes ($n_0 = 50, n_1 = 20$), but becomes clearly visible once we move to the design with a larger number of units ($n_0 = 200, n_1 = 80$). The confidence interval based on Fisher's exact test has coverage that is consistently below the nominal 95% level.

Next we compare coverage rates of these confidence intervals as the size of the sample increases, where we choose a design with non-Gaussian distributions for the potential outcomes. Specifically, we let $Y_i(0) \equiv 0$ and $Y_i(1)|S_i(1) \sim N(0, S_i^2)$, where $S_i = 1$ with probability 0.9, and $S_i = 4$ with probability 0.1. Since the marginal distributions for $Y_i(0)$ and $Y_i(1)$ are different, the difference between sampling variance and the upper bound for the causal variance is nontrivial. Furthermore, while we do not give formal results, under certain regularity conditions the pivotal causal bootstrap should be expected to provide refinements over the Gaussian limiting approximation to the randomization distribution.

Table 5. Coverage of nominal 95% Confidence Intervals, non-Gaussian Potential Values with Isotone Coupling

Variance Estimator	Bootstrap Version	Pivotal Statistic	(n_0, n_1) (20, 20)	(n_0, n_1) (50, 50)	(n_0, n_1) (100, 100)	(n_0, n_1) (200, 200)	(n_0, n_1) (500, 500)
$\hat{\mathbb{V}}_{\text{Neyman}}$	N/A	No	0.9768	0.9866	0.9914	0.9932	0.9924
$\hat{\mathbb{V}}_{\text{AGL}}$	N/A	No	0.9186	0.9358	0.9396	0.9450	0.9436
$\hat{\mathbb{V}}_{\text{s-boot}}$	Standard	No	0.9752	0.9864	0.9912	0.9928	0.9924
$\hat{\mathbb{V}}_{\text{c-boot}}$	Causal	No	0.9144	0.9336	0.9378	0.9436	0.9436
Fisher's Exact Test			0.9752	0.9652	0.9672	0.9560	0.9592
$\hat{\mathbb{V}}_{\text{Neyman}}$	Standard	Yes	0.9870	0.9912	0.9940	0.9942	0.9934
$\hat{\mathbb{V}}_{\text{AGL}}$	Standard	Yes	0.9870	0.9912	0.9940	0.9942	0.9934
$\hat{\mathbb{V}}_{\text{Neyman}}$	Causal	Yes	0.9470	0.9532	0.9582	0.9548	0.9482
$\hat{\mathbb{V}}_{\text{AGL}}$	Causal	Yes	0.9470	0.9532	0.9582	0.9548	0.9482

Table 4 shows simulated coverage rates for the different confidence intervals at the nominal 95% significance level under this design. The results show that coverage rates for both the sampling-based variance estimators and bootstrap are higher throughout than for their causal analogs. The comparison between Gaussian confidence intervals using the causal variance estimators, $\hat{\mathbb{V}}_{\mathrm{AGL}}$ and $\hat{\mathbb{V}}_{\mathrm{c-boot}}$, respectively, to the pivotal causal bootstrap is also indicative of refinements, where the confidence interval based on the pivotal causal bootstrap has coverage rates much closer to the nominal level for small sample sizes, but that advantage vanishes as n_0, n_1 grow large.

5. Large Sample Theory

To characterize the asymptotic properties of the bootstrap procedure, we can cast the statistical experiment of sampling from a finite population with subsequent randomization of treatment among the sampled units as a two-stage scheme of sampling without replacement from nested finite populations. Specifically, in a first step we draw n units without replacement from the population of N units. In a second step, we draw n_1 units at random and without replacement from that sample to receive the treatment $W_i = 1$, whereas the remaining $n_0 = n - n_1$ units are assigned $W_i = 0$. This second step is conditionally independent of the first.

To characterize the contribution of sampling uncertainty to the distribution of the functional we define

$$F_{01}^p(y_0, y_1) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}$$

$$F_{01}^s(y_0, y_1) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}$$

with corresponding marginals $F_0^p, F_1^p, F_0^s, F_1^s$. In particular,

$$F_{01}^s(y_0, y_1) - F_{01}^p(y_0, y_1) = \frac{1}{n} \sum_{i=1}^{N} (R_i - q) \, \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\} \tag{5.1}$$

Turning to the contribution of design uncertainty, we define

$$\hat{F}_0(y_0) := \frac{1}{n(1-p)} \sum_{i=1}^{n} R_i(1 - W_i)\mathbb{1}\{Y_i(0) \leq y_0\}$$

$$\hat{F}_1(y_1) := \frac{1}{np} \sum_{i=1}^{n} R_i W_i \mathbb{1}\{Y_i(1) \leq y_1\}$$

17

where we can rewrite

$$\hat{F}_0(y_0) = \frac{p}{n(1-p)} \sum_{i=1}^{N} R_i \left(1 - \frac{W_i}{p}\right) \mathbb{1}\{Y_i(0) \leq y_0\}$$

Hence, we have

$$\begin{pmatrix} \hat{F}_0(y_0) - F_0^s(y_0) \\ \hat{F}_1(y_1) - F_1^s(y_1) \end{pmatrix} = \frac{1}{np} \sum_{i=1}^{N} R_i(W_i - p) \begin{pmatrix} -\frac{p}{1-p}\mathbb{1}\{Y_i(0) \leq y_0\} \\ \mathbb{1}\{Y_i(1) \leq y_1\} \end{pmatrix} \qquad (5.2)$$

Taken together, (5.1) and (5.2) characterize the uncertainty from sampling and randomization in estimating the respective marginal distributions of $Y_i(0)$ and $Y_i(1)$ as a two-stage process of drawing without replacement from nested finite populations. An asymptotic Donsker Theorem for empirical processes based on sampling without replacement from a finite population is available from Bickel (1969).

We now state the limiting properties of the bootstrap as N and n grow large. Specifically, we derive the limits of the randomization and bootstrap distributions. We then show that the latter is an asymptotically conservative estimator of the former for the purposes of forming confidence intervals.

5.1. **Consistency and Randomization CLT.** Consistency of the estimated bounds follows from consistency of $\hat{F}_0(y_0)$ and $\hat{F}_1(y_1)$ for $F_0(y_0)$ and $F_1(y_1)$, respectively, and the continuous mapping theorem, noting that the conditions in Assumption 2.1 are sufficient for the parameter bounds to be continuous functions of $F_0(y_0)$ and $F_1(y_1)$.

Theorem 5.1. (Consistency) *Suppose Assumptions 2.1, 3.1, and 3.2 hold. Then $\hat{\tau}$ and $\hat{\sigma}$ are consistent for $\tau(F_0^p, F_1^p)$ and $\sigma(F_0^p, F_1^p)$, respectively.*

For a randomization CLT for the estimated bounds we first establish a functional CLT for the randomization processes

$$\begin{aligned} \hat{G}_0 &:= \sqrt{n}(\hat{F}_0 - F_0^p) \\ \hat{G}_1 &:= \sqrt{n}(\hat{F}_1 - F_1^p) \end{aligned}$$

for conditional distributions of potential outcomes. We argue that Assumption 2.1 is sufficient to establish Hadamard differentiability of the functionals $\tau(\hat{F}_0, \hat{F}_1)$, $\sigma(\hat{F}_0, \hat{F}_1)$ so that asymptotic normality of $\sqrt{n}\frac{\hat{\tau}-\tau}{\hat{\sigma}}$ follows from the functional Delta rule and Slutsky's theorem.

Theorem 5.2. (Randomization CLT) *Suppose Assumptions 2.1, 3.1, and 3.2 hold. Then the asymptotic distribution of the t-ratio for $\hat{\tau}_{ATE}$ is given by*

$$\sqrt{n}\frac{\hat{\tau}-\tau}{\hat{\sigma}} \xrightarrow{d} N\left(0, \frac{\sigma^2(F_{01}^p)}{\sigma^2(F_0^p, F_1^p)}\right)$$

where $\sigma(F_{01})^2 := \lim_n n\mathrm{Var}_{F_{01}}(\hat{\tau})$.

The proof of this result is given in the appendix. The formal argument adapts a finite-population CLT for the empirical process developed by Bickel (1969) for the two-sample problem to the case of sampling and randomization in a finite population.

5.2. **Bootstrap CLT.** For a bootstrap replication, denote the empirical distributions of $Y_i^* | W_i^* = 0$ and $Y_i^* | W_i^* = 1$ with $\hat{F}_0^*$ and $\hat{F}_1^*$, respectively. Also, let $\hat{\tau}^* = \tau(\hat{F}_0^*, \hat{F}_1^*)$ and $\hat{\sigma}^* = \sigma(\hat{F}_0^*, \hat{F}_1^*)$

We then establish a CLT for the bootstrap analogs

$$\hat{G}_0^* := \sqrt{n}(\hat{F}_0^* - \hat{F}_0)$$
$$\hat{G}_1^* := \sqrt{n}(\hat{F}_1^* - \hat{F}_1).$$

A CLT for the bootstrapped bounds $\sqrt{n}\frac{\hat{\tau}^* - \hat{\tau}}{\hat{\sigma}^*}$ then relies again on Hadamard differentiability of the variance bounds and the Delta rule for the bootstrap.

A bootstrap CLT can be shown using analogous steps as in a proof for Theorem 5.2, where the randomization distribution is generated based on an estimator for F_{01}^p based on the estimated distributions $\hat{F}_0^p(y_0)$, $\hat{F}_1^p(y_1)$ and the respective least-favorable coupling $C^{iso}(\cdot)$.

Theorem 5.3. **(Bootstrap CLT)** *Suppose Assumptions 2.1, 3.1, and 3.2 hold. Then the asymptotic distribution of the bootstrapped t-ratio for $\hat{\tau}_{ATE}$ is given by*

$$\sqrt{n}\frac{\hat{\tau}^* - \hat{\tau}}{\hat{\sigma}^*} \xrightarrow{d} N(0,1)$$

Most importantly, by Theorem 5.1 the bootstrap estimator for the randomization distribution for $\frac{\hat{\tau} - \tau}{\hat{\sigma}}$ has asymptotic variance equal to 1, whereas the asymptotic variance of the randomization distribution is $\frac{\sigma^2(F_{01}^p)}{\sigma^2(F_0^p, F_1^p)}$ which is less than 1 by construction. That is, the bootstrap algorithm in section 3.1 converges to a "least-favorable" limiting experiment in an appropriate sense. Note also that the formal argument in the proofs of Theorems 5.1-5.3 immediately apply to any other functional $\tau(F_0, F_1)$ that is Hadamard-differentiable in F_0, F_1, and for which the variance bound $\sigma^2(F_0, F_1)$ is continuous in F_0, F_1.

5.3. **Asymptotic Validity of Confidence Intervals.** It remains to show that confidence intervals of the form (3.2) that are constructed under the "least-favorable" limiting experiment are indeed conservative given the CLT under the true randomization distribution in Theorem 5.2. This can be proven by combining the randomization and bootstrap CLTs, replacing the unidentified randomization variance with an estimate of the bound $\sigma(F_0, F_1) \geq \sigma(F_{01})$.

Corollary 5.1. *(Asymptotic Validity of Confidence Intervals)* *Under Assumptions 2.1, 3.1, and 3.2, the $1 - \alpha$ confidence interval (3.2) using bootstrap critical values is asymptotically valid,*

$$\liminf_{n} \inf_{F_{01}^p} \mathbb{P}_{F_{01}^p}\left(\tau(F_{01}^p) \in CI(1-\alpha)\right) \geq 1 - \alpha \quad \text{a.s.}$$

Given Theorems 5.2 and 5.3 this result follows immediately from the definition of the variance bound $\sigma(F_0, F_1)$.

APPENDIX A. RANDOMIZATION DISTRIBUTION FOR $\hat{F}_0(y_0), \hat{F}_1(y_1)$

We first compute the randomization covariance $\text{Cov}_{W,R}^p(\hat{F}_0(y_0), \hat{F}_1(y_1))$ given the population distribution $F_{01}^p(y_0, y_1)$, where

$$\hat{F}_0(y_0) = \frac{1}{n(1-p)} \sum_{i=1}^{N} R_i(1-W_i)\mathbb{1}\{Y_i(0) \leq y_0\}$$

$$\hat{F}_1(y_1) = \frac{1}{np} \sum_{i=1}^{N} R_i W_i \mathbb{1}\{Y_i(1) \leq y_1\}$$

In the following we write $A_{0i} := \mathbb{1}\{Y_i(0) \leq y_0\}$ and $A_{1i} := \mathbb{1}\{Y_i(1) \leq y_1\}$, and take any moments to be with respect to the distribution of R_i and W_i and conditional on the values of $(Y_i(0), Y_i(1))_{i=1}^{N}$ in the population. We then have

$$\text{Cov}(\hat{F}_0(y_0), \hat{F}_1(y_1)) = \frac{1}{n^2 p(1-p)}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} R_i R_j(1-W_i)W_j A_{0i} A_{1j}\right]$$

$$= \frac{1}{n^2 p(1-p)}\sum_{i=1}^{N}\sum_{j=1}^{N} \mathbb{E}\left[R_i R_j(1-W_i)W_j\right] A_{0i} A_{1j}$$

$$= \frac{1}{n^2 p(1-p)}\sum_{i=1}^{N}\sum_{j \neq i} \mathbb{E}\left[R_i R_j\right] \mathbb{E}\left[(1-W_i)W_j\right] A_{0i} A_{1j}$$

$$= \frac{1}{n^2 p(1-p)}\sum_{i=1}^{N}\sum_{j \neq i} \frac{n(n-1)}{N^2}\mathbb{E}\left[(1-W_i)W_j\right] \frac{n^2 p(1-p)}{n(n-1)} A_{0i} A_{1j}$$

$$= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j \neq i} A_{0i} A_{1j} = \frac{1}{N^2}\left(\left[\sum_{i=1}^{N} A_{0i}\right]\left[\sum_{j=1}^{N} A_{1j}\right] - \sum_{i=1}^{N} A_{0i} A_{1i}\right)$$

$$= -\frac{1}{N}\left(F_{01}^p(y_0, y_1) - F_0^p(y_0)F_1^p(y_1)\right)$$

20

To evaluate $\text{Cov}\left(\hat{F}_0(y_0), \hat{F}_0(y_1)\right)$, let $B_{0i} := \mathbb{1}\{Y_i(0) \leq y_0\} - F_0^p(y_0)$ and $B_{0i} := \mathbb{1}\{Y_i(0) \leq y_1\} - F_0^p(y_1)$. We can then write

$$
\begin{aligned}
\text{Cov}\left(\hat{F}_0(y_0), \hat{F}_0(y_1)\right) &= \frac{1}{n^2(1-p)^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \mathbb{E}\left[R_i R_j (1-W_i)(1-W_j)\right] B_{0i} B_{1j} \\
&= \frac{1}{n^2(1-p)^2} \left[\sum_{i=1}^{n} \frac{n}{N} \frac{n(1-p)}{n} B_{0i} B_{1i} + \sum_{i=1}^{N} \sum_{j \neq i} \frac{n(n-1)}{N^2} \frac{n(1-p)(n(1-p)-1)}{n^2} B_{0i} B_{1j} \right] \\
&= \frac{1}{n(1-p)N} \sum_{i=1}^{N} B_{0i} B_{1i} + \frac{(n-1)(n(1-p)-1)}{n^2(1-p)} \left[\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j \neq i} B_{0i} B_{1j} \right] \\
&= \left[\frac{1}{N} \sum_{i=1}^{N} B_{0i} B_{1i} \right] \left(\frac{1}{n(1-p)} - \frac{(n-1)(n(1-p)-1)}{Nn^2(1-p)} \right) \\
&= \left(\min\{F_0^p(y_0), F_0^p(y_1)\} - F_0^p(y_0)F_0^p(y_1) \right) \left(\frac{1}{n(1-p)} - \frac{1}{N} + O\left(\frac{1}{nN}\right) \right)
\end{aligned}
$$

Similarly,

$$
\text{Cov}\left(\hat{F}_1(y_0), \hat{F}_1(y_1)\right) = \left(\min\{F_1^p(y_0), F_1^p(y_1)\} - F_1^p(y_0)F_1^p(y_1) \right) \left(\frac{1}{np} - \frac{1}{N} + O\left(\frac{1}{nN}\right) \right)
$$

Furthermore,

$$
\begin{aligned}
\text{Cov}\left(\hat{F}_0(y_0), \hat{F}_0(y_1)\right) &= \frac{1}{n} \min\{F_0^p(y_0), F_0^p(y_1)\} - F_0^p(y_0)F_0^p(y_1) \\
\text{Cov}\left(\hat{F}_1(y_0), \hat{F}_1(y_1)\right) &= \frac{1}{n} \min\{F_1^p(y_0), F_1^p(y_1)\} - F_1^p(y_0)F_1^p(y_1)
\end{aligned}
$$

We let $\mathbf{H}$ denote the covariance kernel of the randomization process with elements

$$
\begin{aligned}
H_{00}(y_0, y_0') &= \lim_n n\text{Cov}(\hat{F}_0(y_0), \hat{F}_0(y_0')) = \left(\frac{1}{1-p} - \frac{n}{N} \right) \left(\min\{F_0^p(y_0), F_0^p(y_0')\} - F_0^p(y_0)F_0^p(y_0') \right) \\
H_{01}(y_0, y_1) &= \lim_n n\text{Cov}(\hat{F}_0(y_0), \hat{F}_1(y_1')) = \lim_n \frac{n}{N} \left(F_{01}^p(y_0, y_1) - F_0^p(y_0)F_1^p(y_1) \right) \qquad\qquad (\text{A.1}) \\
H_{11}(y_1, y_1') &= \lim_n n\text{Cov}(\hat{F}_1(y_1), \hat{F}_1(y_1')) = \left(\frac{1}{p} - \frac{n}{N} \right) \left(\min\{F_1^p(y_1), F_1^p(y_1')\} - F_1^p(y_1)F_0^p(y_1') \right)
\end{aligned}
$$

Note also that $\frac{1}{1-p} \geq 1 \geq \frac{n}{N} \geq 0$ and $\frac{1}{p} \geq 1 \geq \frac{n}{N} \geq 0$, so that $H_{00}(\cdot, \cdot)$ and $H_{11}(\cdot, \cdot)$ are nonnegative.

APPENDIX B. PROOFS FOR SECTION 2.4

B.1. **Least Favorable Coupling for the Average Treatment Effect.** We first prove a more general result than Proposition 2.1 by showing that the isotone coupling of potential outcomes in fact results in a distribution for the ATE parameter which dominates that under any other coupling in the sense of second-order stochastic dominance (SOSD):

Lemma B.1. *(Ordering of Distributions) Let F_{01} be an arbitrary joint distribution with marginal distributions F_0 and F_1, and let $F_{01}^{iso} := C^{iso}(F_0, F_1)$ be the joint distribution under the isotone coupling. Then for any convex function, the randomization distribution for $\hat{\tau}_{ATE}$ satisfies*

$$
\mathbb{E}_{F_{01}^{iso}}[v(\hat{\tau}_{ATE})] \geq \mathbb{E}_{F_{01}}[v(\hat{\tau}_{ATE})]
$$

For any strictly convex function $v(\cdot)$ this inequality is strict whenever $F_{01} \neq F_{01}^{iso}$.

This result is a straightforward consequence of the familiar observation that the isotone (assortative) coupling of potential outcomes results in the distribution of $Y_i(0) + Y_i(1)$ which second-order stochastic dominates that resulting from any other coupling (see e.g. Becker (1973), Fan and Park (2010), and Stoye (2010)). For illustrative purposes, we give a complete proof here.

PROOF: In order to establish second-order stochastic dominance of the isotone assignment $Y_i(1) = F_1^{-1}(F_0(Y_i(0)))$, consider the expectation of $v(\hat\tau_{ATE})$ for any convex function $v(u)$. Note that for any pair of observations i, j we can write

$$\hat\tau_{ATE} = \frac{1}{n}\left(B_{-ij} + R_i W_i\left(Y_i(0)/(1-p) + Y_i(1)/p\right) + R_j W_j\left(Y_j(0)/(1-p) + Y_j(1)/p\right)\right)$$

where $B_{-ij} := \sum_{k\neq i,j} R_k\left(W_k(Y_k(0)/(1-p) + Y_k(1)/p) - Y_k(0)/(1-p)\right) - (Y_i(0) + Y_j(0))/(1-p)$.

We can now consider the change in $\mathbb{E}[v(\hat\tau_{ATE})]$ from pairwise substitutions of potential outcomes between units i and j. Specifically suppose that under the initial coupling, the potential outcomes for unit i are given by $Y_i(0), Y_i(1)$, and the potential outcomes for unit j are $Y_j(0), Y_j(1)$. We then consider the effect of switching the assignment to potential outcomes $Y_i(0), Y_j(1)$ for unit i, and potential outcomes $Y_j(0), Y_i(1)$ for unit j.

Since W_i, W_j are independent of W_k, that change leads to an increase in $\mathbb{E}[v(\hat\tau_{ATE})]$ if and only if

$$\begin{aligned}
0 \;\leq\; & P(W_i = 1, W_j = 0)\left\{v(B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p) - v(B_{-ij} + Y_i(0)/(1-p) + Y_j(1)/p)\right\} \\
& + P(W_i = 0, W_j = 1)\left\{v(B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p) - v(B_{-ij} + Y_j(0)/(1-p) + Y_i(1)/p)\right\} \\
=\; & p(1-p)\Big\{v(B_{-ij} + Y_i(0)/(1-p) + Y_i(1)/p) + v(B_{-ij} + Y_j(0)/(1-p) + Y_j(1)/p) \\
& - v(B_{-ij} + Y_i(0)/(1-p) + Y_j(1)/p) - v(B_{-ij} + Y_j(0)/(1-p) + Y_i(1)/p)\Big\}
\end{aligned}$$

for any pair of observations with $R_i = R_j = 1$. Noting that for any convex function $v(\cdot)$, $v(b + x_0 + x_1)$ is supermodular in $x = (x_0, x_1)'$, this difference is nonnegative if and only if $Y_i(0) - Y_j(0)$ and $Y_i(1) - Y_j(1)$ have the same sign. Furthermore, if in addition $v(\cdot)$ is strictly convex, the first inequality is strict.

Since any coupling of potential outcomes can be obtained from the isotone assignment by pairwise substitutions of this form, the isotone assignment maximizes the expectation

$$\mathbb{E}[v(\hat\tau_{ATE})] = \mathbb{E}\left[v\left(\frac{1}{n}\sum_{i=1}^N R_i\left\{W_i Y_i(1)/p - (1-W_i)Y_i(0)/(1-p)\right\}\right)\right]$$

for all convex functions $v(\cdot)$. Therefore the distribution of $\hat\tau_{ATE}$ under the isotone assignment dominates that under any alternative coupling, as claimed above. □

Proof of Proposition 2.1: The claim in the proposition follows immediately from Lemma B.1 and the observation that the function $v(y) = y^2$ is strictly convex □

APPENDIX C. PROOFS FOR SECTION 5

C.0.1. *Proof of Theorem 5.1.* From standard results (see e.g. Example 19.6 in van der Vaart (1998)), the class $\mathcal{F} := \{(-\infty, y] : y \in \mathbb{R}\}$ is Glivenko-Cantelli, so that $(\hat F_0 - F_0^p, \hat F_1 - F_1^p)$ converges to zero almost surely as an element of the space of bounded functions on $\mathbb{R}$. Since Assumption 2.1 is sufficient to guarantee

that the functionals $\tau(F_0, F_1)$ and $\sigma(F_0, F_1)$, are continuous in F_0, F_1, the claim of the Theorem follows immediately from the continuous mapping theorem (see e.g. Theorem 18.11 in van der Vaart (1998)) $\qquad\square$

For the proof of Theorem 5.2, we need to characterize functional convergence of the randomization process. To that end, we first introduce some standard notation from empirical process theory (see van der Vaart and Wellner (1996)). Let $\mathcal{F} := \{\mathbb{1}\{y \leq (-\infty, t]\} : t \in \mathbb{R}\}$ be the class of indicator functions for the left-open half-lines on $\mathbb{R}$ and let $\ell^\infty(\mathcal{F})$ be the space of bounded functions from $\mathcal{F}$ to $\mathbb{R}$ endowed with the norm $\|z\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |z(f)|$. Also, let BL_1 denote the set of all functions $h : \ell^\infty \mathcal{F} \mapsto [0,1]$ with $|h(z_1) - h(z_2)| \leq \|z_1 - z_2\|_{\mathcal{F}}$.

Lemma C.1. *Suppose that* $(Y_i(0), Y_i(1)) \overset{iid}{\sim} F_{01}$. *Then the randomization process*

$$\hat{G}_n := \sqrt{n} \begin{pmatrix} \hat{F}_0 - F_0^p \\ \hat{F}_1 - F_1^p \end{pmatrix}$$

converges in outer probability to $\mathbb{G}$ *under the bounded Lipschitz metric,*

$$\sup_{h \in BL_1} |\mathbb{E}_W h(\hat{G}_n) - \mathbb{E}h(\mathbb{G})| \to 0$$

in outer probability, where $\mathbb{G}$ *is a Gaussian process with covariance kernel* **H**.

PROOF: As before, denote the joint c.d.f. of potential outcomes (observed and counterfactuals) for the n units included in the sample with

$$F_{01}^s(y_0, y_1) := \frac{1}{n} \sum_{i=1}^{N} R_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}$$

and the empirical c.d.f. among the units included in the sample for which $W_i = 1$,

$$F_{01}^t(y_0, y_1) := \frac{1}{np} \sum_{i=1}^{N} R_i W_i \mathbb{1}\{Y_i(0) \leq y_0, Y_i(1) \leq y_1\}$$

Using this notation we can write

$$\sqrt{n}(F_{01}^t(y_0, y_1) - F_{01}^p(y_0, y_1)) = \sqrt{n}(F_{01}^t(y_0, y_1) - F_{01}^s(y_0, y_1)) + \sqrt{n}(F_{01}^s(y_0, y_1) - F_{01}^p(y_0, y_1))$$

Since R_i, W_i are drawn at random and without replacement, it follows from Theorem 3.1 in Bickel (1969) that

$$\sqrt{n}(F_{01}^t(y_0, y_1) - F_{01}^s(y_0, y_1)) \quad \rightsquigarrow \quad \mathbb{G}_{F_{01}^s}$$

$$\sqrt{n}(F_{01}^s(y_0, y_1) - F_{01}^p(y_0, y_1)) \quad \rightsquigarrow \quad \mathbb{G}_{F_{01}^p}$$

for Brownian bridges $\mathbb{G}_{F_{01}^s}$ and $\mathbb{G}_{F_{01}^p}$. Since for any joint distribution $F_{01}(y_0, y_1)$ the marginals satisfy $\lim_{y_1 \to \infty} F_{01}(y_0, y_1) = F_0(y_0)$ for each y_0, weak convergence of the joint process implies weak convergence of the marginal empirical processes,

$$\sqrt{n}(F_0^t - F_0^p) \quad \rightsquigarrow \quad \mathbb{G}_{F_0^s} + \mathbb{G}_{F_0^p}$$

$$\sqrt{n}(F_1^t - F_1^p) \quad \rightsquigarrow \quad \mathbb{G}_{F_1^s} + \mathbb{G}_{F_1^p}$$

Finally, $\hat{F}_1(y_1) \equiv F_1^t(y_1)$ and $\hat{F}_0(y_0) \equiv \frac{1}{(p-1)}(F_0^s(y_0) - pF_0^t(y_0))$, establishing the claim, where the structure of the covariance kernel follows from the point-wise calculations in the derivation of (A.1) $\qquad\square$

C.0.2. *Proof of Theorem 5.2:* From Assumption 2.1 it is immediate that $\tau(F_0, F_1)$ is Hadamard-differentiable. Lemma C.1 and the functional delta method, see e.g. Theorem 20.8 in van der Vaart (1998), then imply asymptotic normality of $\sqrt{n}(\hat{\tau} - \tau)/\sigma(F_0, F_1)$. Theorem 5.2 then follows from Slutsky's theorem and consistency of $\hat{\sigma}$ from Theorem 5.1 $\hfill\square$

We next turn to the bootstrap distribution: Consider the bootstrap replicates

$$\hat{F}_0^*(y_0) := \frac{1}{n(1-p)} \sum_{i=1}^n R_i^*(1-W_i^*)\mathbb{1}\{Y_i^*(0) \le y_0\}, \qquad \hat{F}_1^*(y_1) := \frac{1}{np} \sum_{i=1}^n R_i^* W_i^* \mathbb{1}\{Y_i^*(1) \le y_1\}$$

by randomizing from $\hat{F}_{01}$. We also define the asymptotic covariance kernel $\mathbf{H}^{iso}$ corresponding to the coupling C^{iso} in analogy to (A.1) where F_{01} is replaced with $C^{iso}(F_0, F_1)$. We first show the two following Lemmas:

Lemma C.2. *Suppose that* $(Y_i(0), Y_i(1)) \overset{iid}{\sim} F_{01}$. *Then for any copula* $C : [0,1]^2 \to [0,1]$,

$$\sup_{y_0, y_1 \in \mathbb{R}} \left| C(\hat{F}_0, \hat{F}_1)(y_0, y_1) - C(F_0^p, F_1^p)(y_0, y_1) \right| \overset{a.s.}{\to} 0$$

PROOF: From standard results, the class $\mathcal{F} := \{(-\infty, y] : y \in \mathbb{R}\}$ is Glivenko-Cantelli, so that $(\hat{F}_0 - F_0^p, \hat{F}_1 - F_1^p)$ converges to zero almost surely as an element of the space of bounded functions on $\mathbb{R}$. Noting that any copula $C : [0,1]^2 \to [0,1]$ is a bounded nondecreasing function in each of its arguments, it follows that

$$\sup_{y_0, y_1 \in \mathbb{R}} \left| C(\hat{F}_0, \hat{F}_1)(y_0, y_1) - C(F_0^p, F_1^p)(y_0, y_1) \right| \overset{a.s.}{\to} 0$$

establishing the claim $\hfill\square$

Lemma C.3. *Suppose that* $(Y_i(0), Y_i(1)) \overset{iid}{\sim} F_{01}$. *Then the bootstrap process*

$$\hat{G}_n^* := \sqrt{n} \begin{pmatrix} \hat{F}_0^* - \hat{F}_0 \\ \hat{F}_1^* - \hat{F}_1 \end{pmatrix}$$

converges in outer probability to $\mathbb{G}$ *under the bounded Lipschitz metric, that is*

$$\sup_{h \in BL_1} \left| \mathbb{E}_W h(\hat{G}_n^*) - \mathbb{E} h(\mathbb{G}) \right| \to 0$$

in outer probability, where $\mathbb{G}$ *is a Gaussian processes with covariance kernel* $\mathbf{H}$.

PROOF: By construction of the coupling $(Y_i^*(0), Y_i^*(1))$, the marginal distributions of $Y_i^*(0)$ and $Y_i^*(1)$ are equal to $\hat{F}_0$ and $\hat{F}_1$, respectively. By construction of the bootstrap, the bootstrap replications $\hat{F}_0^*, \hat{F}_1^*$ are generated by randomization from the samples $(\tilde{Y}_i(1), \tilde{Y}_i(1))_{i=1}^n$ corresponding to the joint distribution $\hat{F}_{01} := C^{iso}(\hat{F}_0, \hat{F}_1)$.

Now let $\hat{\mathbf{H}}^{iso}$ the covariance kernel obtained from (A.1) replacing F_0 with $\hat{F}_0$, F_1 with $\hat{F}_1$, and F_{01} with $C^{iso}(\hat{F}_0, \hat{F}_1)$, respectively. By construction, the bootstrap distribution of $\hat{G}_n^*$ conditional on $\hat{F}_0, \hat{F}_1$ have covariance given by $\hat{\mathbf{H}}^{iso}$. Finally, $\hat{\mathbf{H}}^{iso}$ is a continuous function of $C^{iso}(\hat{F}_0, \hat{F}_1)$. Hence by Lemma C.2 and the continuous mapping theorem we have that

$$\|\hat{\mathbf{H}}^{iso} - \mathbf{H}^{iso}\| \overset{a.s.}{\to} 0$$

which completes the proof.

The claim of the Lemma then follows from the same arguments as in Lemma C.1 and the continuous mapping theorem $\hfill\square$

24

C.0.3. *Proof of Theorem 5.3:* It follows from Assumption 2.1 that $\tau(F_0, F_1), \sigma(F_0, F_1)$ are Hadamard differentiable, so that Theorem C.3 follows from Lemma C.3 and the functional Delta method (e.g. Theorem 20.8 in van der Vaart (1998)) $\qquad\square$

<div align="center">REFERENCES</div>

ABADIE, A., S. ATHEY, G. IMBENS, AND J. WOOLDRIDGE (2017): "Sampling-based vs. Design-based Uncertainty in Regression Analysis," working paper, MIT, Stanford, Michigan State.

ARONOW, P., D. GREEN, AND D. LEE (2014): "Sharp Bounds on the Variance in Randomized Experiments," *Annals of Statistics*, 42(3), 850–871.

BECKER, G. (1973): "A Theory of Marriage, Part I," *Journal of Political Economy*, 81(4), 813–846.

BICKEL, P. (1969): "A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case," *Annals of Mathematical Statistics*, 40(1), 1–23.

BICKEL, P., AND D. FREEDMAN (1984): "Asymptotic Normality and the Bootstrap in Stratified Sampling," *Annals of Statistics*, 12(2), 470–482.

BLOZNELIS, M., AND F. GÖTZE (2001): "Orthogonal Decomposition of Finite Population Statistics and its Applications to Distributional Asymptotics," *Annals of Statistics*, 29(3), 899–917.

BOOTH, J., R. BUTLER, AND P. HALL (1994): "Bootstrap Methods for Finite Populations," *Journal of the American Statistical Association*, 89, 1282–1289.

CAMBANIS, S., G. SIMMONS, AND W. STOUT (1976): "Inequalities for $\mathcal{E}k(X, Y)$ when the Marginals are Fixed," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 36, 285–294.

DING, P. (2017): "A Paradox From Randomization-Based Causal Inference," *Statistical Science*, forthcoming.

EFRON, B. (1982): *The Jackknife, the Bootstrap and Other Resampling Plans*, CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

FAN, Y., AND S. PARK (2010): "Sharp Bounds on the Distribution of Treatment Effects and their Statistical Inference," *Econometric Theory*, 26(3), 931–951.

——— (2012): "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments," *Journal of Econometrics*, 167, 330–344.

FAN, Y., AND J. WU (2010): "Partial Identification of the Distribution of Treatment Effects in Switching Regime Models and its Confidence Sets," *Review of Economic Studies*, 77(3), 1002–1041.

FIRPO, S., AND G. RIDDER (2008): "Bounds on Functionals of the Distribution of Treatment Effects," working paper, Insper and USC.

FRANK, M., R. NELSEN, AND B. SCHWEIZER (1987): "Best-Possible Bounds for the Distribution of a Sum - a Problem of Kolmogorov," *Probability Theory and Related Fields*, 74, 199–211.

HECKMAN, J., J. SMITH, AND N. CLEMENTS (1997): "Making the Most Out Of Programme Evaluations and Social Experiments: Accounting For Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487–535.

IMBENS, G., AND D. RUBIN (2015): *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

LI, X., AND P. DING (2017): "General forms of finite population central limit theorems with applications to causal inference," *JASA*, forthcoming.

LU, J., P. DING, AND T. DASGUPTA (2018): "Treatment Effects on Ordinal Outcomes: Causal Estimands and Sharp Bounds," *Journal of Educational and Behavioral Statistics*, 20(10), 1–28.

MAKAROV, G. (1982): "Estimates for the Distribution Function of a Sum of Two Random Variables when the Marginal Distributions are Fixed," *Theory of Probability and its Applications*, 26(4), 803–806.

MANSKI, C. (1997): "The Mixing Problem in Programme Evaluation," *Review of Economic Studies*, 64(4), 537–553.

NELSEN, R. (2006): *An Introduction to Copulas*. Springer, New York.

NEYMAN, J. (1923,1990): "On the Application of Probability Theory to Agricultural Experiments, Essays on Principles. Section 9.," translated in *Statistical Science* (with discussion), 5(4), 465-480.

ROBINS, J. (1988): "Confidence Intervals for Causal Parameters," *Statistics in Medicine*, 7, 773–785.

ROSENBAUM, P. (2002): *Observational Studies*. Springer, New York, 2 edn.

RUBIN, D. (1978): "Bayesian Inference for Causal Effects: The Role of Randomization," *Annals of Statistics*, 6, 34–58.

STOYE, J. (2010): "Partial Identification of Spread Parameters," *Quantitative Economics*, 1, 323–357.

VAN DER VAART, A. (1998): *Asymptotic Statistics*. Cambridge University Press, Cambridge.

VAN DER VAART, A., AND J. WELLNER (1996): *Weak Convergence and Empirical Processes*. Springer, New York.

WILLIAMSON, R., AND T. DOWNS (1990): "Probabilistic Arithmetic. I. Numerical Methods for Calculating Convolutions and Dependency Bounds," *International Journal of Approximate Reasoning*, 4, 89–158.