The Annals of Statistics
2021, Vol. 49, No. 3, 1460–1488
https://doi.org/10.1214/20-AOS2009

A CAUSAL BOOTSTRAP

BY GUIDO IMBENS[1] AND KONRAD MENZEL[2]

[1]*Department of Economics and Graduate School of Business, Stanford University, imbens@stanford.edu*
[2]*Department of Economics, New York University, km125@nyu.edu*

The bootstrap, introduced by *The Jackknife, the Bootstrap and Other Resampling Plans* ((1982), SIAM), 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 for inference regarding the average treatment effect that accounts for this uncertainty, and compare its properties to that of the classical bootstrap. We consider completely randomized and observational designs as well as designs with imperfect compliance.

1. Introduction. The bootstrap, introduced by [12], has become a very popular method for constructing hypothesis tests or confidence intervals. This popularity stems in part from the fact that it provides approximations to the distribution of an estimator or statistic that are in certain cases superior to those obtained from using a Gaussian asymptotic approximation together with estimated standard errors (asymptotic refinement). While the classical bootstrap is designed to approximate distributions that result from repeated sampling from a large population, this paper shows how to adapt the bootstrap principle when the estimand of interest is a causal parameter, and the data is generated by a randomized experiment. We also consider observational studies and designs with imperfect compliance when the population of interest may be finite.

Permutation tests, such as Fisher's exact test (see, e.g., [28]), can yield exact p-values under the auxiliary hypothesis that treatment effects are constant across units; however, we argue below that those methods are not suitable for forming confidence intervals for parameters describing the distribution of causal effects in a given population. For the average treatment effect, causal standard errors have been proposed by [3], as well as [1]. These methods impose no restrictions on treatment effect heterogeneity but their use generally relies on a Gaussian limiting approximation. We propose a bootstrap approach to causal inference which also does not restrict treatment heterogeneity, but improves on the Gaussian asymptotic approximation in samples of small or moderate size.

Using the potential outcome framework, for example, [28], 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 $i = 1, \dots, N$. 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, [38]).

Received November 2019; revised July 2020.
MSC2020 subject classifications. 62D05, 62D20.
Key words and phrases. Causal inference, bootstrap, partial identification.