Title: Practical Methods for Uncertainty in Bounds

Authors:

Peng Ding, UC Berkeley, pengdingpku@berkeley.edu

Avi Feller, UC Berkeley, afeller@berkeley.edu
[Presenting author]

Luke W. Miratrix, Harvard University, lmiratrix@g.harvard.edu

Problem: Researchers are often interested in causal effects for subgroups defined by post-treatment outcomes. Prominent examples in education include noncompliance, attrition, and implementation fidelity. Unfortunately, estimating principal causal effects (PCEs) in these *principal strata* or *endogenous subgroups* can be difficult, as subgroup membership is not fully observed. A large and growing literature addresses the associated methodological challenges (see Page, Feller, Grindal, Miratrix, & Somers (2015) for a recent review).

One simple but powerful strategy for estimating PCEs is to bound the unknown quantities. Miratrix, Furey, Feller, Grindal, & Page (2017) give an overview of this approach, illustrating the method with a randomized experiment in education. An important practical consideration is how to account for uncertainty with bounds. Incorporating uncertainty from *unidentifiable* parameters is straightforward: the investigator simply plugs in extreme values for the unknown quantity. Counterintuitively, however, incorporating uncertainty from *identifiable* quantities is relatively difficult. For example, in standard settings, calculating a 95% confidence interval for a population mean is immediate. And yet, calculating a 95% confidence interval for a bound is both computationally and conceptually challenging. The key complication is that otherwise standard tools for incorporating uncertainty, such as the Delta method, break down in this setting. Ignoring this uncertainty could yield misleading conclusions about program impacts.

Prior methodological research: There is a small but growing literature on incorporating uncertainty into estimates of bounds. Canay & Shaikh (2016) give a recent review of existing methods, particularly those from econometrics. In general, existing methods fall into:

- *Asymptotic approximations*. These methods leverage asymptotic results about the endpoints of a bound, typically assuming that endpoints are approximately Normal (or truncated Normal). While straightforward to compute, these approximations only perform well in special cases. See Imbens & Manski (2004).

- *Resampling methods*. These methods typically employ a two-step procedure. First, the investigator uses subsampling or the bootstrap to generate distributions for carefully chosen test statistics. Second, the investigator generates confidence intervals for bounds by inverting a sequence of tests based on these resampled test statistics. See Bugni (2014). While several of these methods have strong theoretical guarantees, they can be quite cumbersome to implement and can have poor performance in realistic settings.

- *Bayesian methods*. These methods incorporate uncertainty via the Bayesian framework, typically via Markov chain Monte Carlo (MCMC). See Kline & Tamer (2016). While promising, these methods are generally under-developed and are not particularly accessible for applied researchers.

Method: This paper has two main goals. First, we review existing methods for uncertainty in bounds, with an eye toward making these results accessible to applied education researchers. In addition to clear descriptions of these (often complex) approaches, we also discuss important questions in how to even conceptualize uncertainty with bounds, as it can differ from more classical definitions.

Second, we describe two practical methods, case-resampling bootstrap and modular Bayes (Jacob, Murray, Holmes, & Robert, 2017). In particular:

- **Case-resampling bootstrap**. Constructing confidence intervals for bounds is straightforward with this approach. The investigator simply generates *B* samples (with replacement), calculates bounds for each, and sets the 95% confidence interval for the bound via the appropriate quantiles of the bootstrap distributions of the endpoints (see Miratrix et al., 2017). The challenge is that this approach can yield invalid inference when the true parameter is at the edge of the bound (Andrews, 1999). While this is a legitimate concern in theory, we investigate this via extensive simulation and argue that, while theoretically possible, it is not a meaningful concern in practice for most researchers.

- **Modular Bayes.** This approach instead uses Bayesian models to account for uncertainty in the endpoints, building off recent proposals for *modular Bayesian inference* (Jacob et al., 2017). Analogous to the bootstrap approach, the investigator uses the appropriate percentiles of the posterior to construct uncertainty intervals. We show that implementation is straightforward via the Bayesian programming language Stan. We compare this simplified approach to fully Bayesian methods, such as Chen, Christensen, & Tamer (2016).

We assess the performance of these simple methods via "calibrated simulations" (Kern, Stuart, Hill, & Green, 2016), simulation studies built from datasets constructed by imputing missing class membership and potential outcomes from real-world studies. This approach preserves important structure among the covariates and outcomes. Using these studies, we compare the performance of the different techniques under a variety of plausible circumstances, in particular under extreme scenarios where the true parameters are at or near the bound endpoints.

Setting: We finally apply these methods to two common data sets that represent the type of data increasingly available to researchers, the JOBS II study (Jo, 2002) and the Head Start Impact Study (Puma, Bell, Cook, Heid, & Shapiro, 2010). For each, we compute bounds on the relevant causal effects of interest and, importantly, assess the uncertainty on these bounds with our different approaches. For JOBS II, the two principal causal effects of interest are the effect of randomization on depression score for Compliers (those who would enroll if offered) and for Never Takers (those who would never enroll). As Jo (2002) notes, there is a concern that randomization might have a negative impact on Never Takers, even though randomization does not change enrollment behavior. We therefore do not want to assume *ex ante* that the PCE for this group is zero, the assumption necessary for using a classic IV approach. For HSIS, we also have two subgroups of interest defined by their counterfactual care setting. See Feller, Grindal, Miratrix, & Page (2016).

References

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*.

Bugni, F. A. (2014). Comparison of inferential methods in partially identified models in terms of error in coverage probability. *Econometric Theory*, *32*(01), 187–242.

Canay, I. A., & Shaikh, A. M. (2016). Practical and Theoretical Advances in Inference for Partially Identified Models.

Chen, X., Christensen, T., & Tamer, E. (2016). MCMC Confidence Sets for Identified Sets. *arXiv*.

Feller, A., Grindal, T., Miratrix, L. W., & Page, L. C. (2016). Compared to What? Variation in the Impacts of Early Childhood Education by Alternative Care Type. *Annals of Applied Statistics*.

Imbens, G. W., & Manski, C. F. (2004). Confidence Intervals for Partially Identified Parameters. *Econometrica*, *72*(6), 1845–1857.

Jacob, P. E., Murray, L. M., Holmes, C. C., & Robert, C. P. (2017, August 29). Better together? Statistical learning in models made of modules. *arXiv*.

Jo, B. (2002). Estimation of Intervention Effects with Noncompliance: Alternative Model Specifications. *Journal of Educational and Behavioral Statistics*, *27*(4), 385–409.

Kern, H. L., Stuart, E. A., Hill, J., & Green, D. P. (2016). Assessing Methods for Generalizing Experimental Impact Estimates to Target Populations. *Journal of Research on Educational Effectiveness*, *9*(1), 103–127.

Kline, B., & Tamer, E. (2016). Bayesian inference in a class of partially identified models. *Quantitative Economics*, *7*(2), 329–366.

Miratrix, L., Furey, J., Feller, A., Grindal, T., & Page, L. C. (2017). Bounding, an accessible method for estimating principal causal effects, examined and explained. *Journal of Research on Educational Effectiveness*, *78*(2).

Page, L. C., Feller, A., Grindal, T., Miratrix, L., & Somers, M. A. (2015). Principal Stratification: A Tool for Understanding Variation in Program Effects Across Endogenous Subgroups. *American Journal of Evaluation*, *36*(4), 514–531.

Puma, M., Bell, S. H., Cook, R., Heid, C., & Shapiro, G. (2010). Head Start Impact Study. Final Report. *HHS, Administration for Children & Families*.