

Inference on Winners*

Isaiah Andrews[†] Toru Kitagawa[‡] Adam McCloskey[§]

December 22, 2021

Abstract

Many empirical questions concern target parameters selected through optimization. For example, researchers may be interested in the effectiveness of the best policy found in a randomized trial, or the best-performing investment strategy based on historical data. Such settings give rise to a winner’s curse, where conventional estimates are biased and conventional confidence intervals are unreliable. This paper develops optimal confidence intervals and median-unbiased estimators that are valid conditional on the target selected and so overcome this winner’s curse. If one requires validity only on average over targets that might have been selected, we develop hybrid procedures that combine conditional and projection confidence intervals to offer further performance gains relative to existing alternatives.

KEYWORDS: WINNER’S CURSE, SELECTIVE INFERENCE

JEL CODES: C12, C13

*We thank Tim Armstrong, Stéphane Bonhomme, Raj Chetty, Gregory Cox, Áureo de Paula, Nathaniel Hendren, Patrick Kline, Hannes Leeb, Anna Mikusheva, Magne Mogstad, José Luis Montiel Olea, Mikkel Plagborg-Møller, Jack Porter, Adam Rosen, Frank Schoerfheide, Jesse Shapiro, and participants at numerous seminars and conferences for helpful comments. We also thank Raj Chetty and Nathaniel Hendren for extremely generous assistance on the application using data from Chetty et al. (2018), and thank Jeff Rowley, Peter Ruhm, and Nicolaj Thor for outstanding research assistance. Andrews gratefully acknowledges financial support from the NSF under grant number 1654234. Kitagawa gratefully acknowledges financial support from the ESRC through the ESRC Centre for Microdata Methods and Practice (CeMMAP) (grant number RES-589-28-0001) and the European Research Council (Starting grant No. 715940). Initial version posted May 10, 2018.

[†]Department of Economics, Harvard University, iandrews@fas.harvard.edu

[‡]CeMMAP and Department of Economics, University College London, t.kitagawa@ucl.ac.uk

[§]Department of Economics, University of Colorado, adam.mccloskey@colorado.edu

1 Introduction

A wide range of empirical questions involve inference on target parameters selected through optimization over a finite collection of candidates. In a randomized trial considering multiple treatments, for instance, one might want to learn about the true average effect of the treatment that performed best in the experiment. In finance, one might want to learn about the expected return of the trading strategy that performed best in a backtest.

Estimators that do not account for data-driven selection of the target parameter can be badly biased, and conventional t -test-based confidence intervals may severely under-cover. To illustrate the problem, consider inference on the true average effect of the treatment that performed best in a randomized trial.¹ Since it ignores the data-driven selection of the treatment of interest, the conventional estimate for this average effect will be biased upwards. Similarly, the conventional confidence interval will under-cover, particularly when the number of treatments considered is large. This gives rise to a form of winner’s curse, where follow-up trials will be systematically disappointing relative to what we would expect based on conventional estimates and confidence intervals. This form of winner’s curse has previously been discussed in contexts including genome-wide association studies (e.g. Zhong and Prentice, 2009; Xu et al., 2011; Ferguson et al., 2013) and online A/B tests (Lee and Shen, 2018).

This paper develops estimators and confidence intervals that eliminate the winner’s curse. There are two distinct perspectives from which to consider bias and coverage. The first requires validity conditional on the target selected, for example on the identity of the best-performing treatment, while the second is unconditional and requires validity on average over possible target parameters. Conditional validity is more demanding but may be desirable in some settings, for example when one wants to ensure validity conditional on the recommendation made to a policy maker. Both perspectives differ from inference on the effectiveness of the “true” best treatment, as in e.g. Chernozhukov et al. (2013) and Rai (2018), in that we consider inference on the effectiveness of the (observed) best-performing treatment

¹Such a scenario seems to be empirically relevant, as a number of recently published randomized trials in economics either were designed with the intent of recommending a policy or represent a direct collaboration with a policy maker. For example, Khan et al. (2016) assess how incentives for property tax collectors affect tax revenues in Pakistan, Banerjee et al. (2018) evaluate the efficacy of providing information cards to potential recipients of Indonesia’s *Raskin* programme, and Duflo et al. (2018) collaborate with the Gujarat Pollution Control Board (an Indian regulator tasked with monitoring industrial emissions in the state) to evaluate how more frequent but randomized inspection of plants performs relative to discretionary inspection. Baird et al. (2016) find that deworming Kenyan children has substantial beneficial effects on their health and labor market outcomes into adulthood, and Björkman Nyqvist and Jayachandran (2017) find that providing parenting classes to Ugandan mothers has a greater impact on child outcomes than targeting these classes at fathers.

in the sample rather than the (unobserved) best-performing treatment in the population.²

For conditional inference, we derive optimal median-unbiased estimators and equal-tailed confidence intervals. We further show that in cases where the winner’s curse does not arise (for instance because one treatment considered is vastly better than the others) our conditional procedures coincide with conventional ones. Hence, our corrections do not sacrifice efficiency in such cases.

An alternative approach to conditional inference is sample splitting. In settings with independent observations, choosing the target parameter using one subset of the data and constructing estimates and confidence intervals using the remaining subset ensures unbiasedness of estimates and validity of conventional confidence intervals conditional on the target parameter. The split-sample target parameter is necessarily more variable than the full-data target, however. Moreover, since only a subset of the data is used for inference, split-sample procedures are inefficient within the class of procedures with the same target. In the supplement to this paper we build on our conditional inference results to develop computationally tractable confidence intervals and estimators that dominate conventional sample-splitting.

We next turn to unconditional inference. One approach to constructing valid unconditional confidence intervals is projection, applied in various settings by e.g. Romano and Wolf (2005), Berk et al. (2013), and Kitagawa and Tetenov (2018a). To obtain a projection confidence interval, we form a simultaneous confidence band for all potential targets and take the implied set of values for the target of interest. The resulting confidence intervals have correct unconditional coverage but, unlike our conditional intervals, are wider than conventional confidence intervals even when the latter are valid. On the other hand, we find in simulation that projection intervals outperform conditional intervals in cases where there is substantial randomness in the target parameter, e.g. when there is not a clear best treatment.

Since neither conditional nor projection intervals are uniformly best from an unconditional perspective, we introduce hybrid estimators and confidence intervals that combine conditioning and projection. These maintain most of the good performance of our conditional approach in cases for which the winner’s curse does not arise, while improving on conditional procedures in cases where these underperform, e.g. by limiting the maximum length of hybrid intervals relative to projection intervals. In simulations calibrated to our applications we find that hybrid intervals are typically shorter than both conditional and projection intervals, often by a large margin.

²See Dawid (1994) for an early discussion of this distinction, and an argument in favor of inference on the best-performing treatment in the sample.

We derive our main results in the context of a finite-sample normal model with an unknown mean vector and a known covariance matrix. This model can be viewed as an asymptotic approximation to many different non-normal finite-sample settings. To formalize this connection, we note, and prove in the appendix, that feasible versions of our procedures, based on non-normal data and plugging in estimated variances, are uniformly asymptotically valid over a large class of data-generating processes.

We illustrate our results with two applications. The first uses data from Karlan and List (2007) to conduct inference on the effect of the best-performing treatment in an experiment studying the impact of matching incentives on charitable giving. Simulations calibrated to these data show that conventional estimates ignoring selection are substantially upward biased, while our corrections reduce bias and increase coverage. Applied to the original Karlan and List (2007) data, our corrections suggest substantially less optimism about the effect of the best-performing treatment than conventional approaches, with point estimates below the lower bound of the conventional confidence intervals.

For our second application, we consider the problem of targeting neighborhoods based on estimated economic mobility. In cooperation with the Seattle and King County public housing authorities, Bergman et al. (2020) conduct an experiment encouraging housing voucher recipients to move to high-opportunity neighborhoods, which are selected based on census-tract level estimates of economic mobility from Chetty et al. (2018). We consider an analogous exercise in the 50 largest commuting zones in the US, selecting top tracts based on estimated economic mobility and examining conventional and corrected inference on the average mobility in selected tracts. Calibrating simulations to the Chetty et al. (2018) data, we find that conventional approaches suffer from severe bias in many commuting zones. These biases are reduced, but not eliminated, by the empirical Bayes corrections used by Chetty et al. (2018) and many others in the applied literature. Intuitively, commonly-applied empirical Bayes approaches correspond to a normal prior on unit-level causal effects conditional on covariates. Bayesian arguments (discussed in Appendix E) imply that these methods correct the winner’s curse when the normal prior matches the distribution of true effects, but not in general otherwise. Turning to the original Chetty et al. (2018) data, our corrected estimates imply lower mobility, and higher uncertainty, for selected tracts than conventional approaches, but nonetheless strongly indicate gains from moving to selected tracts. Our confidence intervals likewise suggest substantially higher uncertainty than empirical Bayes credible sets, though we do not find a clear ordering between our bias-corrected point estimates of economic mobility and the empirical Bayes point estimates of Chetty et al. (2018).

The choice between conditional and unconditional inference methods is necessarily context-specific, as it depends on the extent to which we care about validity conditional on selecting a given target. We report results for both conditional and unconditional approaches in each application, but view conditional inference as particularly natural for the first application. In this setting the “winning” treatment is easily interpretable, raising the question of what we can conclude conditional on identity of this treatment. In our second application, by contrast, our primary goal is to assess the efficacy of targeting the top third of census tracts in each commuting zone, with less focus on the precise collection of tracts selected. We therefore view unconditional inference as more natural in this context.

It is important to emphasize that our goal is to evaluate the effectiveness of a recommended policy or treatment, taking the rule for selecting a recommendation as given, rather than to improve the rule. Our procedures thus play a role similar to that of ex-post policy evaluations, with the difference that we can produce estimates without waiting for a policy to be implemented. Like ex-post evaluations, these estimates may be useful for a variety of purposes, including understanding the true effectiveness of a selected policy and forecasting the effects of future implementations. Our results are also useful in settings where ex-post evaluation is possible, since comparison of our estimates with ex-post results can shed light on whether differences between observed performance and conventional ex-ante estimates can be explained solely by the winner’s curse.

Related Literature This paper is related to the literature on tests of superior predictive performance (e.g. White (2000); Hansen (2005); Romano and Wolf (2005)). That literature studies the problem of testing whether some strategy or policy beats a benchmark, while we consider the complementary question of inference on the effectiveness of the estimated “best” policy. Our conditional inference results combine naturally with the results of this literature, allowing one to condition inference on e.g. rejecting the null hypothesis that no policy outperforms a benchmark.

Our results build upon, and contribute to, the rapidly growing literature on selective inference. Fithian et al. (2017) describe a general approach to constructing optimal conditional confidence sets in a wide range of settings, while a rapidly growing literature including e.g. Harris et al. (2016), Lee et al. (2016), Tian and Taylor (2018), and our own follow-up work in Andrews et al. (2020b), works out the details of this approach in particular settings. More specifically, this literature primarily focuses on inference after regressor selection in the linear regression model using various model selection criteria, while Andrews et al. (2020b) focuses on inference after estimating a break location in a break or threshold

regression model. Like this literature, our analysis of conditional confidence intervals examines the implications of the conditional approach in our setting. Our results are also related to the growing literature on unconditional post-selection inference, including Berk et al. (2013), Bachoc et al. (2020), and Kuchibhotla et al. (2020). This literature considers analogs of our projection confidence intervals for inference following model selection (see also Laber and Murphy, 2011). Recent work by Guo and He (2021) proposes tightening projection confidence intervals in the context of a winner’s curse for selected subgroups via a sequence of tuning parameters that drifts as the sample size grows.

Beyond the new setting considered, we make three main theoretical contributions relative to the selective and post-selection inference literatures. First, when one only requires unconditional validity, we introduce the class of hybrid inference and estimation procedures. We find that hybrid procedures offer large gains in unconditional performance relative both to conditional procedures and to existing unconditional alternatives.³ Two of our own follow-up papers, Andrews et al. (2020b) and McCloskey (2020), adapt the hybrid procedures we introduce here to different settings, relying on the theoretical results established in this paper. Second, for settings where conditional inference is desired, we observe that the same structure used in the literature to develop optimal conditional confidence intervals also allows construction of optimal quantile unbiased estimators, using results from Pfanzagl (1994) on optimal estimation in exponential families.⁴ Third, our uniform asymptotic results are the first of their kind in the conditional inference literature.⁵

Finally, there is a distinct but complementary literature that studies inference on ranks based on some measure of interest. For example, this literature allows one to form valid confidence intervals for the identification of best performing unit, rather than for the performance of the unit selected as best by the data. Conventional inference procedures for these problems fail for similar reasons that give rise to a winner’s curse. Recent work by Mogstad et al. (2020) overcomes this inference failure and studies, among other settings, inference on ranks in neighborhood targeting, as in our second application.

In the next section, we begin by introducing the problem we consider and the techniques

³A related hybridization, combining conditional and unconditional inference, is used in Andrews et al. (2019) to improve power for tests of parameters identified by moment inequalities.

⁴Elias (2004) previously used results from Pfanzagl (1994) to study quantile-unbiased estimation in a different setting, targeting coefficients on highly persistent regressors.

⁵McCloskey (2020) uses our uniformity results to establish uniform asymptotic validity for hybrid confidence intervals for inference after model selection, while Tibshirani et al. (2018) and Andrews et al. (2020b) establish uniform asymptotic validity for conditional confidence intervals in different settings from ours, but only under particular local sequences.

we propose in the context of a stylized example. Section 3 introduces the normal model, develops our conditional procedures, and briefly discusses sample splitting. Section 4 introduces projection confidence intervals and our hybrid procedures. Section 5 discusses practical implementation in and translates our normal model results to uniform asymptotic results. Finally, Sections 6 and 7 discuss applications to data from Karlan and List (2007) and Bergman et al. (2020), respectively. The supplement to this paper contains proofs of our theoretical results and additional theoretical, numerical and empirical results.

2 A Stylized Example

We begin by illustrating the problem we consider, along with the solutions we propose, in a stylized example. Suppose we have data from a randomized trial of a binary treatment (e.g. participation in a job training program), where individuals $i \in \{1, \dots, n\}$ were randomly assigned to treatment ($D_i=1$) or control ($D_i=0$), with $\frac{n}{2}$ individuals in each group. We are interested in an outcome Y_i (e.g. a dummy for employment in the next year), and compute the treatment and control means,

$$(X_n^*(1), X_n^*(0)) = \left(\frac{2}{n} \sum_{i=1}^n D_i Y_i, \frac{2}{n} \sum_{i=1}^n (1-D_i) Y_i \right).$$

If trial participants are a random sample from some population, then for $Y_{i,1}$ and $Y_{i,0}$ equal to the potential outcomes for i under treatment and control, respectively, $(X_n^*(1), X_n^*(0))$ unbiasedly estimate the average potential outcomes $(\mu^*(1), \mu^*(0)) = (E[Y_{i,1}], E[Y_{i,0}])$ in the population.

For policymakers and researchers interested in maximizing the average outcome, it is natural to focus on the treatment that performed best in the experiment. Formally, let $\Theta = \{0, 1\}$ denote the set of policies (just control and treatment in this example) and define $\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} X_n^*(\theta)$ as the policy yielding the highest average outcome in the experiment. While $X_n^*(\theta)$ unbiasedly estimates $\mu^*(\theta)$ for fixed policies θ , $X_n^*(\hat{\theta}_n)$ systematically over-estimates $\mu^*(\hat{\theta}_n)$ since we are more likely to select a given policy when the experiment over-estimates its effectiveness. Likewise, confidence intervals for $\mu^*(\hat{\theta}_n)$ that ignore estimation of θ may cover $\mu^*(\hat{\theta}_n)$ less often than we intend. Hence, if a policymaker deploys the treatment $\hat{\theta}_n$, or a researcher examines it in a follow-up experiment, the results will be systematically disappointing relative to the original trial. This is a form of winner's curse: estimation error leads us to over-predict the benefits of our chosen policy and to misstate our uncertainty about its effectiveness.

To simplify the analysis and develop corrected inference procedures, we turn to asymptotic approximations. For $X_n = \sqrt{n}X_n^* = \sqrt{n}(X_n^*(0), X_n^*(1))$ and $(\mu_n(1), \mu_n(0)) = \sqrt{n}(\mu^*(1), \mu^*(0))$, provided the potential outcomes $(Y_{i,0}, Y_{i,1})$ have finite variance,

$$\begin{pmatrix} X_n(0) - \mu_n(0) \\ X_n(1) - \mu_n(1) \end{pmatrix} \Rightarrow N\left(0, \begin{pmatrix} \Sigma(0) & 0 \\ 0 & \Sigma(1) \end{pmatrix}\right), \quad (1)$$

where \Rightarrow denotes convergence in distribution and the asymptotic variance Σ can be consistently estimated while the scaled average outcomes μ_n cannot be. Motivated by (1), let us abstract from approximation error and assume that we observe

$$\begin{pmatrix} X(0) \\ X(1) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu(0) \\ \mu(1) \end{pmatrix}, \begin{pmatrix} \Sigma(0) & 0 \\ 0 & \Sigma(1) \end{pmatrix}\right)$$

for $\Sigma(0)$ and $\Sigma(1)$ known, and that $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} X(\theta)$ with $\Theta = \{0, 1\}$.⁶

As discussed above, $X(\hat{\theta})$ is biased upwards as an estimator of $\mu(\hat{\theta})$. This bias arises both conditional on $\hat{\theta}$ and unconditionally. To see this note that $\hat{\theta} = 1$ if $X(1) > X(0)$, where ties occur with probability zero. Conditional on $\hat{\theta} = 1$ and $X(0) = x(0)$, however, $X(1)$ follows a normal distribution truncated below at $x(0)$. Since this holds for all $x(0)$, $X(1)$ has positive median bias conditional on $\hat{\theta} = 1$.⁷

$$Pr_{\mu}\left\{X(\hat{\theta}) \geq \mu(\hat{\theta}) | \hat{\theta} = 1\right\} > \frac{1}{2} \text{ for all } \mu.$$

Since the same argument holds for $\hat{\theta} = 0$, $\hat{\theta}$ is also biased upwards unconditionally:

$$Pr_{\mu}\left\{X(\hat{\theta}) \geq \mu(\hat{\theta})\right\} > \frac{1}{2} \text{ for all } \mu.$$

Similarly, conventional t -statistic-based confidence intervals need not have correct coverage.

To illustrate these issues, Figure 1 plots the coverage of conventional confidence intervals, as well as the median bias of conventional estimates, in an example with $\Sigma(1) = \Sigma(0) = 1$. For comparison we also consider cases with ten and fifty policies (e.g. additional treatments) $|\Theta| = 10$ and $|\Theta| = 50$, where we again set Σ to be diagonal with $\Sigma(\theta) = 1$ for all θ and, for

⁶Finite-sample results in this normal model correspond to asymptotic results for cases where the difference in outcomes $E[Y_{i,1}] - E[Y_{i,0}]$ is of order $\frac{1}{\sqrt{n}}$, so the optimal policy $\theta^* = \operatorname{argmax}_{\theta \in \Theta} \mu^*(\theta)$ is weakly identified. We defer an in-depth discussion of asymptotics to Section 5 and Appendix D.

⁷It also has positive mean bias, but we focus on median bias for consistency with our later results.

ease of reporting, assume that all the policies other than the first (policy θ_1) are equally effective, with average outcome $\mu(\theta_{-1})$. The first panel of Figure 1 shows that while the conventional confidence interval has reasonable coverage when there are only two policies, its coverage can fall substantially when $|\Theta|=10$ or $|\Theta|=50$. The second panel shows that the median bias of the conventional estimator $\hat{\mu} = X(\hat{\theta})$, measured as the deviation of the exceedance probability $Pr_\mu\{X(\hat{\theta}) \geq \mu(\hat{\theta})\}$ from $\frac{1}{2}$, can be quite large. The third panel shows that the same is true when we measure bias as the median of $X(\hat{\theta}) - \mu(\hat{\theta})$. In all cases we find that performance is worse when we consider a larger number of policies, as is natural since a larger number of policies allows more scope for selection.

Our results correct these biases. Returning to the case with $\Theta = \{0,1\}$ for simplicity, let $F_{TN}(x(1); \mu(1), x(0))$ denote the truncated normal distribution function for $X(1)$, truncated below at $x(0)$, when the true mean is $\mu(1)$. This function is strictly decreasing in $\mu(1)$, and for $\hat{\mu}_\alpha$ the solution to $F_{TN}(X(1); \hat{\mu}_\alpha, X(0)) = 1 - \alpha$, Proposition 2 below shows that

$$Pr_\mu\left\{\hat{\mu}_\alpha \geq \mu(\hat{\theta}) \mid \hat{\theta} = 1\right\} = \alpha \quad \text{for all } \mu.$$

Hence, $\hat{\mu}_\alpha$ is α -quantile unbiased for $\mu(\hat{\theta})$ conditional on $\hat{\theta} = 1$, and the analogous statement holds conditional on $\hat{\theta} = 0$. Indeed, Proposition 2 shows that $\hat{\mu}_\alpha$ is the optimal α -quantile unbiased estimator conditional on $\hat{\theta}$.

Using this result, we can eliminate the biases discussed above. The estimator $\hat{\mu}_{1/2}$ is median unbiased and the equal-tailed confidence interval $CS_{ET} = [\hat{\mu}_{\alpha/2}, \hat{\mu}_{1-\alpha/2}]$ has conditional coverage $1 - \alpha$, where we say that a confidence interval CS has conditional coverage $1 - \alpha$ if

$$Pr\left\{\mu(\hat{\theta}) \in CS \mid \hat{\theta} = \tilde{\theta}\right\} \geq 1 - \alpha \quad \text{for } \tilde{\theta} \in \Theta \text{ and all } \mu. \quad (2)$$

By the law of iterated expectations, CS_{ET} also has unconditional coverage $1 - \alpha$:

$$Pr_\mu\left\{\mu(\hat{\theta}) \in CS\right\} \geq 1 - \alpha \quad \text{for all } \mu. \quad (3)$$

Unconditional coverage is easier to attain, so relaxing the coverage requirement from (2) to (3) may allow shorter confidence intervals in some cases. Conditional and unconditional coverage requirements address different questions, however, and which is more appropriate depends on the problem at hand. For instance, if a researcher recommends the policy $\hat{\theta}$ to a policymaker, it may also be natural to report a confidence interval that is valid conditional on the recommendation, which is precisely the conditional coverage requirement (2).

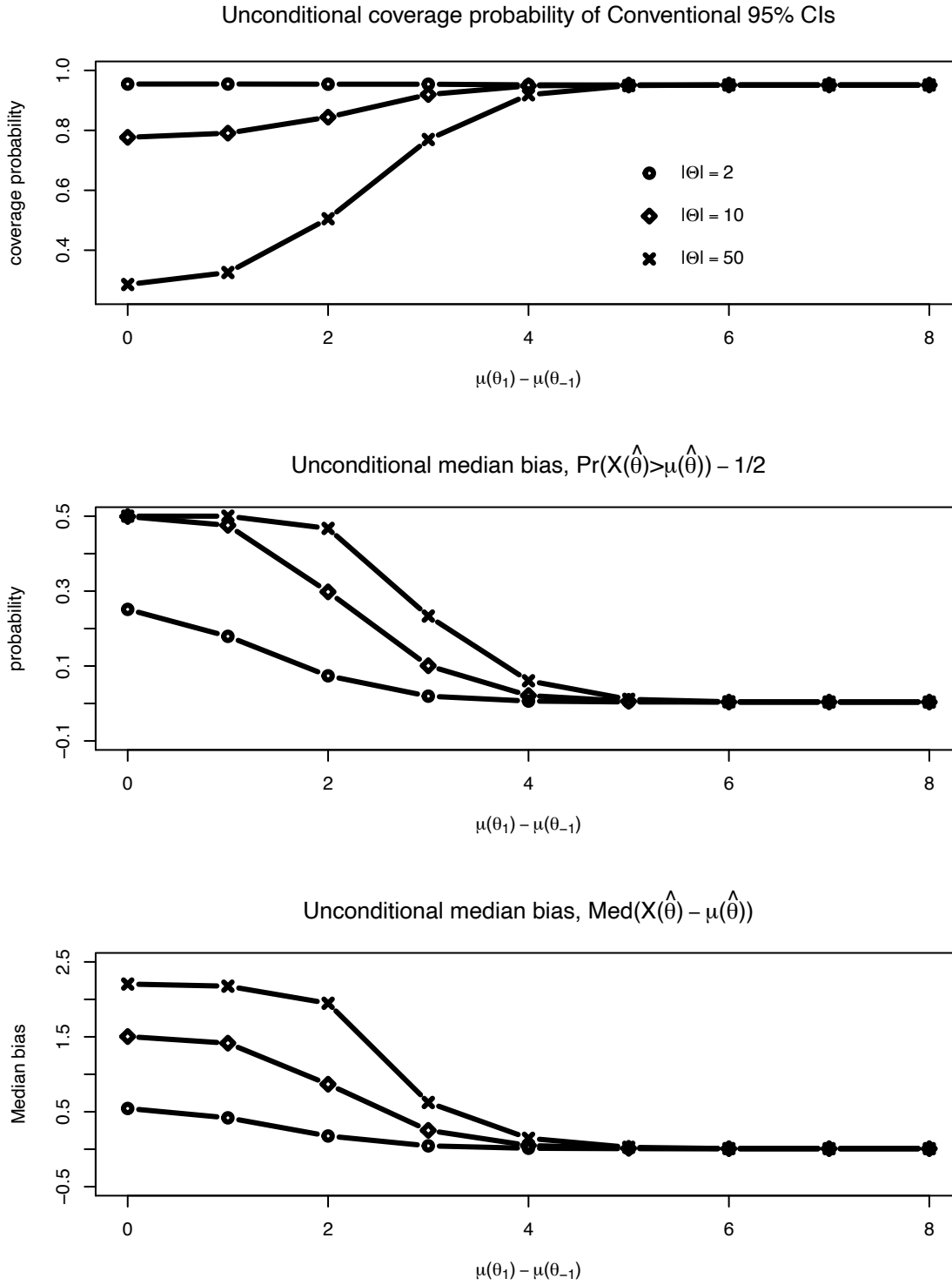


Figure 1: Performance of conventional procedures in examples with 2, 10, and 50 policies.

Conditional coverage ensures that if one considers repeated instances in which researchers recommend a particular course of action (e.g. departure from the status quo), reported confidence intervals will in fact cover the true effects a fraction $1-\alpha$ of the time. On the other hand, if we only want to ensure that our confidence intervals cover the true value with probability at least $1-\alpha$ on average across the distribution of recommendations, it suffices to impose the unconditional coverage requirement (3).

We are unaware of alternatives in the literature that ensure conditional coverage (2). For unconditional coverage (3), however, one can form an unconditional confidence interval by projecting a simultaneous confidence set for μ . In particular, let c_α denote the $1-\alpha$ quantile of $\max_j |\xi_j|$ for $\xi = (\xi_1, \xi_2)' \sim N(0, I_2)$ a two-dimensional standard normal random vector. If we define CS_P as

$$CS_P = \left[X(\hat{\theta}) - c_\alpha \sqrt{\Sigma(\hat{\theta})}, X(\hat{\theta}) + c_\alpha \sqrt{\Sigma(\hat{\theta})} \right],$$

this set has correct unconditional coverage (3).

Figure 2 plots the (unconditional) median length of 95% confidence intervals CS_{ET} and CS_P , along with the conventional confidence interval, again in cases with $|\Theta| \in \{2, 10, 50\}$. We focus on median length, rather than mean length, because the results for Kivaranovic and Leeb (2020) imply that CS_{ET} has infinite expected length. As Figure 2 illustrates, the median length of CS_{ET} is shorter than the (nonrandom) length of CS_P in all cases when $|\mu(\theta_1) - \mu(\theta_{-1})|$ exceeds four, and converges to the length of the conventional interval as $|\mu(\theta_1) - \mu(\theta_{-1})|$ grows larger. When $|\mu(\theta_1) - \mu(\theta_{-1})|$ is small, on the other hand, CS_{ET} can be substantially wider than CS_P . This reflects that in these cases, $X(\hat{\theta})$ is frequently close to the next-best treatment. For a truncated normal distribution, an observation close to the lower endpoint provides evidence of a small mean, but with little precision about the exact value, leading to long confidence intervals.

These features become still more pronounced as we increase the number of policies considered, and are still more pronounced for higher quantiles of the length distribution. To illustrate, Figure 3 plots the 95th percentile of the distribution of length in the case with $|\Theta| = 50$ policies, while results for other quantiles and specifications are reported in Appendix F.

Figure 4 plots the median absolute error $Med_\mu(|\hat{\mu} - \mu(\hat{\theta})|)$ for different estimators $\hat{\mu}$, and shows that the median-unbiased estimator likewise exhibits larger median absolute error than the conventional estimator $X(\hat{\theta})$ when $|\mu(\theta_1) - \mu(\theta_{-1})|$ is small. This feature is again more pronounced as we increase the number of policies considered, or if we consider

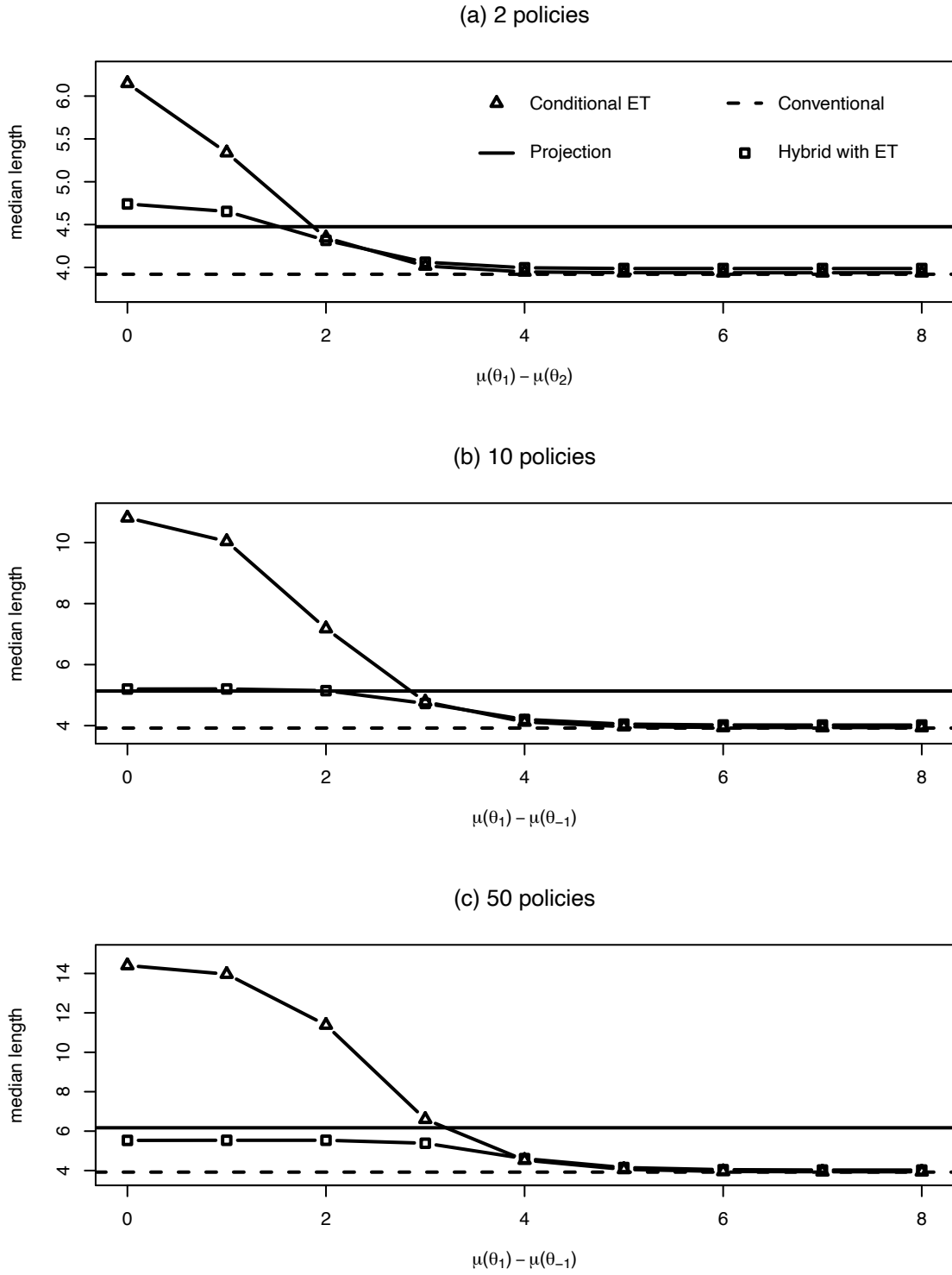


Figure 2: Median length of confidence intervals for $\mu(\hat{\theta})$ in cases with 2, 10, and 50 policies.

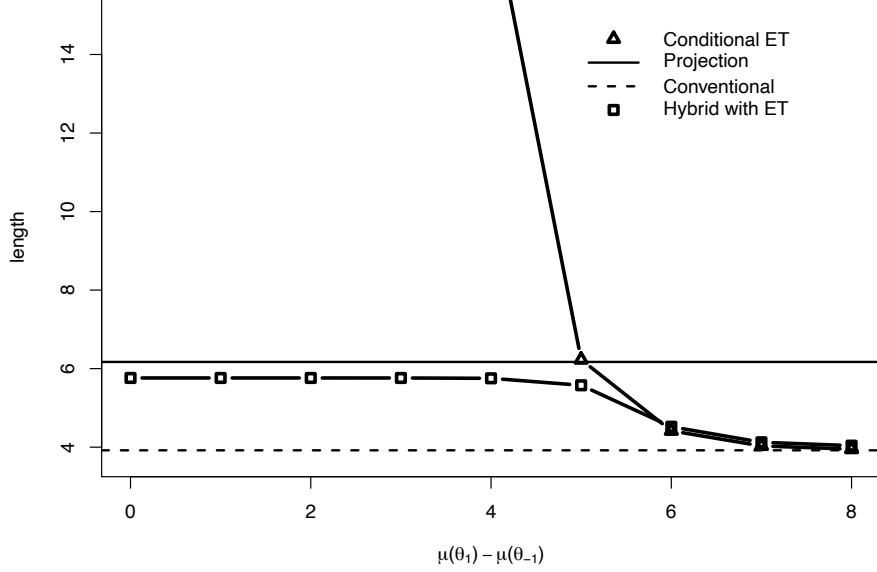


Figure 3: 95th percentile of length of confidence intervals for $\mu(\hat{\theta})$ in case with 50 policies.

higher quantiles as in Appendix F.

Recall that $\hat{\mu}_{\frac{1}{2}}$ and the endpoints of CS_{ET} are optimal quantile unbiased estimators. So long as we impose median unbiasedness and correct conditional coverage, there is hence little scope to improve conditional performance. If we instead focus on unconditional bias and coverage, by contrast, improved performance is possible.

To improve performance, we consider hybrid inference, which combines the conditional and unconditional approaches. Hybrid inference first computes a level $\beta < \alpha$ projection interval CS_P^β , and then considers conditional inference given $\hat{\theta}$ and $\mu(\hat{\theta}) \in CS_P^\beta$. In the case with $\Theta = \{0, 1\}$, for instance, if $\hat{\theta} = 1$ and the true mean is $\mu(1)$ then the conditional distribution of $X(1)$ given $\hat{\theta} = 1$, $X(0) = x(0)$, and $\mu(1) \in CS_P^\beta$ is a $N(\mu(1), \Sigma(1))$ distribution truncated to the interval

$$\left[\max\left\{x(0), \mu(1) - c_\beta \sqrt{\Sigma(1)}\right\}, \mu(1) + c_\beta \sqrt{\Sigma(1)} \right].$$

For the corresponding distribution function $F_{TN}^H(x(1); \mu(1), x(0))$, the hybrid estimator $\hat{\mu}_\alpha^H$ solves $F_{TN}^H(X(1); \hat{\mu}_\alpha^H, X(0)) = 1 - \alpha$. Arguments analogous to those in the conditional case imply that $\hat{\mu}_\alpha^H$ is α -quantile unbiased conditional on the (potentially false) event

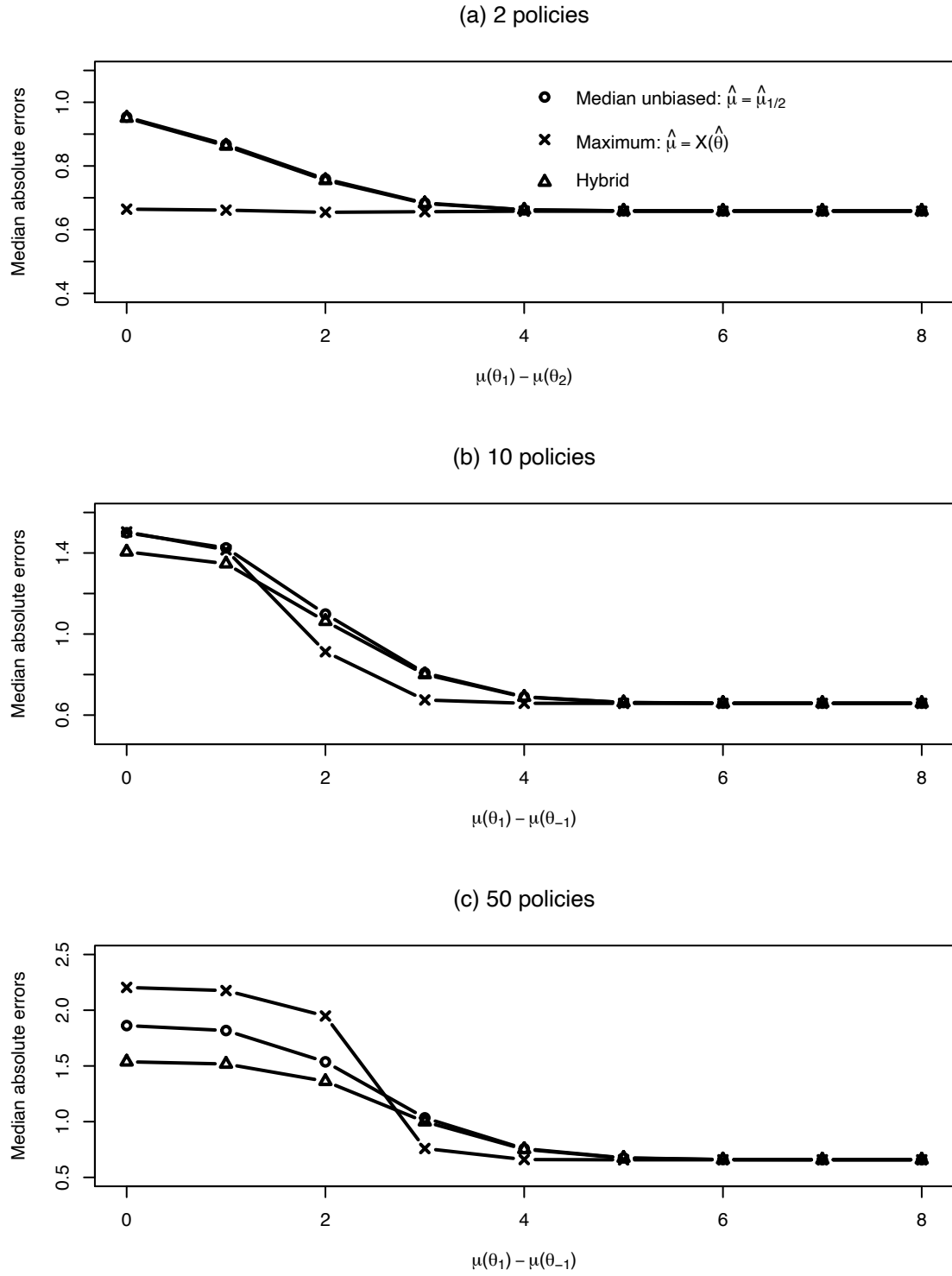


Figure 4: Median absolute error of estimators of $\mu(\hat{\theta})$ in cases with 2, 10, and 50 policies.

$\{\hat{\theta}=1, \mu(\hat{\theta}) \in CS_P^\beta\}$. Since $Pr_\mu\{\mu(\hat{\theta}) \in CS_P^\beta\} \geq 1 - \beta$ one can further show that the unconditional quantile bias of $\hat{\mu}_\alpha^H$ is bounded, in the sense that

$$\left|Pr_\mu\{\hat{\mu}_\alpha^H \geq \mu(\hat{\theta})\} - \alpha\right| \leq \beta \cdot \max\{\alpha, 1 - \alpha\}.$$

We again form level $1 - \alpha$ equal-tailed confidence intervals based on these estimates, where to account for the dependence on the projection interval we adjust the quantile considered and take $CS_{ET}^H = \left[\hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)}}^H, \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)}}^H\right]$. See Section 4.2 for details on this adjustment. By construction, hybrid intervals are never longer than the level $1 - \beta$ projection interval CS_P^β .

Due to their dependence on the projection interval, hybrid intervals do not in general have correct conditional coverage (2). By relaxing the conditional coverage requirement, however, we obtain major improvements in unconditional performance, as illustrated in Figure 2. In particular, we see that in the case with 50 policies, the hybrid confidence intervals have shorter median length than the unconditional interval CS_P for all parameter values considered. The gains relative to conditional confidence intervals are large for many parameter values, and are still more pronounced for higher quantiles of the length distribution, as in Figure 3 and Appendix F. In Figure 4 we report results for the hybrid estimator $\hat{\mu}_{\frac{1}{2}}^H$, and again find substantial performance improvements.

The improved unconditional performance of the hybrid confidence intervals is achieved by requiring only unconditional, rather than conditional, coverage. To illustrate, Figure 5 plots the conditional coverage given $\hat{\theta} = \theta_1$ in the case with two policies. As expected, the conditional interval has correct conditional coverage, while coverage distortions appear for the hybrid and projection intervals when $\mu(\theta_1) \ll \mu(\theta_2)$. In this case $\hat{\theta} = \theta_2$ with high probability but the data will nonetheless sometimes realize $\hat{\theta} = \theta_1$. Conditional on this event, $X(\theta_1)$ will be far away from $\mu(\theta_1)$ with high probability, so projection and hybrid confidence intervals under-cover. As $\mu(\theta_1) - \mu(\theta_2)$ diverges to $-\infty$, their conditional coverage probabilities given $\hat{\theta} = \theta_1$ approach 0.

3 Conditional Inference

This section introduces our general setting, which extends the stylized example of the previous section in several directions, and develops conditional inference procedures. We then discuss sample splitting as an inefficient conditional inference method and briefly discuss the construction of dominating procedures. Finally, we show that our conditional procedures converge to conventional ones when the latter are valid.

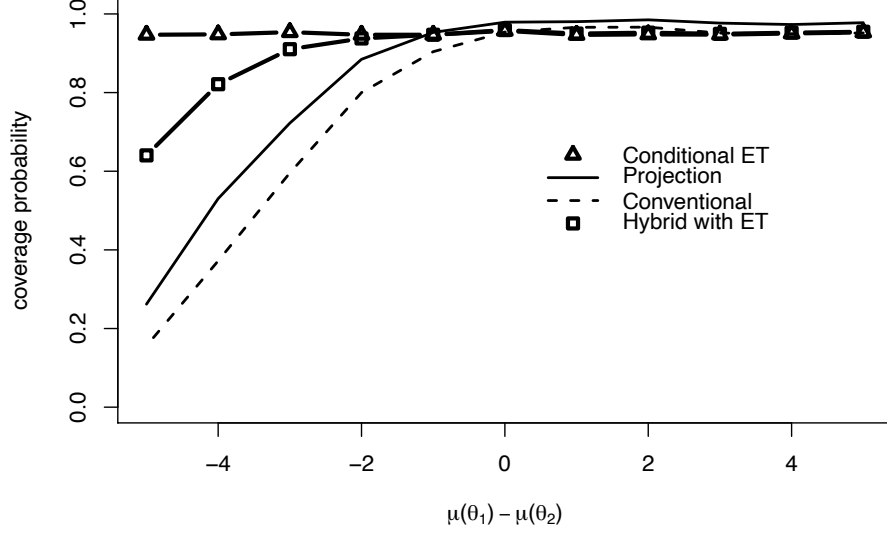


Figure 5: Coverage conditional on $\hat{\theta} = \theta_1$ in case with two policies.

3.1 Setting

Suppose we observe a collection of normal random vectors $(X(\theta), Y(\theta))' \in \mathbb{R}^2$ for $\theta \in \Theta$ where Θ is a finite set. For $\Theta = \{\theta_1, \dots, \theta_{|\Theta|}\}$, let $X = (X(\theta_1), \dots, X(\theta_{|\Theta|}))'$ and $Y = (Y(\theta_1), \dots, Y(\theta_{|\Theta|}))'$. Then

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(\mu, \Sigma) \quad (4)$$

for

$$E \left[\begin{pmatrix} X(\theta) \\ Y(\theta) \end{pmatrix} \right] = \mu(\theta) = \begin{pmatrix} \mu_X(\theta) \\ \mu_Y(\theta) \end{pmatrix},$$

$$\Sigma(\theta, \tilde{\theta}) = \begin{pmatrix} \Sigma_X(\theta, \tilde{\theta}) & \Sigma_{XY}(\theta, \tilde{\theta}) \\ \Sigma_{YX}(\theta, \tilde{\theta}) & \Sigma_Y(\theta, \tilde{\theta}) \end{pmatrix} = Cov \left(\begin{pmatrix} X(\theta) \\ Y(\theta) \end{pmatrix}, \begin{pmatrix} X(\tilde{\theta}) \\ Y(\tilde{\theta}) \end{pmatrix} \right).$$

We assume that Σ is known, while μ is unknown and unrestricted unless noted otherwise. For brevity of notation, we abbreviate $\Sigma(\theta, \theta)$ to $\Sigma(\theta)$. We assume throughout that $\Sigma_Y(\theta) > 0$ for all $\theta \in \Theta$, since the inference problem we study is trivial when $\Sigma_Y(\theta) = 0$. As discussed in Section 5 below, this model arises naturally as an asymptotic approximation.

We are interested in inference on $\mu_Y(\hat{\theta})$, where $\hat{\theta}$ is determined based on X . We define

$\hat{\theta}$ through the level maximization,⁸

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} X(\theta). \quad (5)$$

In a follow-up paper, Andrews et al. (2020b), we develop results on inference when $\hat{\theta}$ instead maximizes $\|X(\theta)\|$ and $X(\theta)$ may be vector-valued.

We are interested in constructing estimates and confidence intervals for $\mu_Y(\hat{\theta})$ that are valid either conditional on the value of $\hat{\theta}$ or unconditionally. In many cases, as in Section 2 above, we are interested in the mean of the same variable that drives selection, so $X=Y$ and $\mu_X=\mu_Y$. In other settings, however, we may select on one variable but want to do inference on the mean of another. Continuing with the example discussed in Section 2, for instance, we might select $\hat{\theta}$ based on outcomes for all individuals, but want to conduct inference on average outcomes for some subgroup defined using covariates. In this case, $Y(\theta)$ corresponds to the estimated average outcome for the group of interest under treatment θ .

3.2 Conditional Inference

We first consider conditional inference, seeking estimates of $\mu_Y(\hat{\theta})$ which are quantile unbiased conditional on $\hat{\theta}$:

$$Pr_{\mu} \left\{ \hat{\mu}_{\alpha} \geq \mu_Y(\hat{\theta}) \mid \hat{\theta} = \tilde{\theta} \right\} = \alpha \text{ for all } \tilde{\theta} \in \Theta \text{ and all } \mu. \quad (6)$$

Since $\hat{\theta}$ is a function of X , we can re-write the conditioning event in terms of the sample space of X as $\{X : \hat{\theta} = \tilde{\theta}\} = \mathcal{X}(\tilde{\theta})$.⁹ Thus, for conditional inference we are interested in the distribution of (X, Y) conditional on $X \in \mathcal{X}(\tilde{\theta})$. Our results below imply that the elements of Y other than $Y(\tilde{\theta})$ do not help in constructing a quantile-unbiased estimate or confidence interval for $\mu_Y(\hat{\theta})$ conditional on $X \in \mathcal{X}(\tilde{\theta})$. Hence, we limit attention to the conditional distribution of $(X, Y(\tilde{\theta}))$ given $X \in \mathcal{X}(\tilde{\theta})$.

Since $(X, Y(\tilde{\theta}))$ is jointly normal unconditionally, it has a multivariate truncated normal distribution conditional on $X \in \mathcal{X}(\tilde{\theta})$. Correlation between X and $Y(\tilde{\theta})$ implies that the conditional distribution of $Y(\tilde{\theta})$ depends on both the parameter of interest $\mu_Y(\hat{\theta})$ and μ_X . To eliminate dependence on the nuisance parameter μ_X , we condition on a sufficient

⁸For simplicity of notation we assume $\hat{\theta}$ is unique almost surely unless noted otherwise.

⁹If $\hat{\theta}$ is not unique we change the conditioning event from $\hat{\theta} = \tilde{\theta}$ to $\tilde{\theta} \in \operatorname{argmax} X(\theta)$.

statistic. Without truncation and for any fixed $\mu_Y(\tilde{\theta})$, a minimal sufficient statistic for μ_X is

$$Z_{\tilde{\theta}} = X - \left(\Sigma_{XY}(\cdot, \tilde{\theta}) / \Sigma_Y(\tilde{\theta}) \right) Y(\tilde{\theta}), \quad (7)$$

where we use $\Sigma_{XY}(\cdot, \tilde{\theta})$ to denote $Cov(X, Y(\tilde{\theta}))$. $Z_{\tilde{\theta}}$ corresponds to the part of X that is (unconditionally) orthogonal to $Y(\tilde{\theta})$ which, since $(X, Y(\tilde{\theta}))$ are jointly normal, means that $Z_{\tilde{\theta}}$ and $Y(\tilde{\theta})$ are independent. Truncation breaks this independence, but $Z_{\tilde{\theta}}$ remains minimal sufficient for μ_X . The conditional distribution of $Y(\hat{\theta})$ given $\{\hat{\theta} = \tilde{\theta}, Z_{\tilde{\theta}} = z\}$ is truncated normal:

$$Y(\hat{\theta}) | \hat{\theta} = \tilde{\theta}, Z_{\tilde{\theta}} = z \sim \xi | \xi \in \mathcal{Y}(\tilde{\theta}, z), \quad (8)$$

where $\xi \sim N(\mu_Y(\tilde{\theta}), \Sigma_Y(\tilde{\theta}))$ is normally distributed and

$$\mathcal{Y}(\tilde{\theta}, z) = \left\{ y : z + \left(\Sigma_{XY}(\cdot, \tilde{\theta}) / \Sigma_Y(\tilde{\theta}) \right) y \in \mathcal{X}(\tilde{\theta}) \right\} \quad (9)$$

is the set of values for $Y(\tilde{\theta})$ such that the implied X falls in $\mathcal{X}(\tilde{\theta})$ given $Z_{\tilde{\theta}} = z$. Thus, conditional on $\hat{\theta} = \tilde{\theta}$, and $Z_{\tilde{\theta}} = z$, $Y(\hat{\theta})$ follows a one-dimensional truncated normal distribution with truncation set $\mathcal{Y}(\tilde{\theta}, z)$.

The following result, based on Lemma 5.1 of Lee et al. (2016), characterizes $\mathcal{Y}(\tilde{\theta}, z)$.

Proposition 1

Let $\Sigma_{XY}(\tilde{\theta}) = Cov(X(\tilde{\theta}), Y(\tilde{\theta}))$. Define

$$\mathcal{L}(\tilde{\theta}, Z_{\tilde{\theta}}) = \max_{\theta \in \Theta : \Sigma_{XY}(\tilde{\theta}) > \Sigma_{XY}(\tilde{\theta}, \theta)} \frac{\Sigma_Y(\tilde{\theta}) (Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta}))}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta}, \theta)},$$

$$\mathcal{U}(\tilde{\theta}, Z_{\tilde{\theta}}) = \min_{\theta \in \Theta : \Sigma_{XY}(\tilde{\theta}) < \Sigma_{XY}(\tilde{\theta}, \theta)} \frac{\Sigma_Y(\tilde{\theta}) (Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta}))}{\Sigma_{XY}(\tilde{\theta}) - \Sigma_{XY}(\tilde{\theta}, \theta)},$$

and

$$\mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) = \min_{\theta \in \Theta : \Sigma_{XY}(\tilde{\theta}) = \Sigma_{XY}(\tilde{\theta}, \theta)} - (Z_{\tilde{\theta}}(\theta) - Z_{\tilde{\theta}}(\tilde{\theta})).$$

If $\mathcal{V}(\tilde{\theta}, z) \geq 0$, then $\mathcal{Y}(\tilde{\theta}, z) = [\mathcal{L}(\tilde{\theta}, z), \mathcal{U}(\tilde{\theta}, z)]$. If $\mathcal{V}(\tilde{\theta}, z) < 0$, then $\mathcal{Y}(\tilde{\theta}, z) = \emptyset$.

Thus, $\mathcal{Y}(\tilde{\theta}, z)$ is an interval bounded above and below by functions of z . While we must have $\mathcal{V}(\tilde{\theta}, z) \geq 0$ for this interval to be non-empty, $Pr_{\mu} \left\{ \mathcal{V}(\tilde{\theta}, Z_{\tilde{\theta}}) < 0 \right\} = 0$ for all μ so this

constraint holds almost surely when we consider the value $\hat{\theta}$ observed in the data. Hence, in applications we can safely ignore this constraint and calculate only $\mathcal{L}(\hat{\theta}, Z_{\hat{\theta}})$ and $\mathcal{U}(\hat{\theta}, Z_{\hat{\theta}})$.

Using this result, it is straightforward to construct quantile-unbiased estimators for $\mu_Y(\hat{\theta})$. Let $F_{TN}(y; \mu_Y(\tilde{\theta}), \tilde{\theta}, z)$ denote the distribution function for the truncated normal distribution (8). This function is strictly decreasing in $\mu_Y(\tilde{\theta})$. Define $\hat{\mu}_\alpha$ as the unique solution to $F_{TN}(Y(\hat{\theta}); \hat{\mu}_\alpha, \tilde{\theta}, Z_{\hat{\theta}}) = 1 - \alpha$. Proposition 2 below shows that $\hat{\mu}_\alpha$ is conditionally α -quantile-unbiased in the sense of (6), so $\hat{\mu}_{\frac{1}{2}}$ is median-unbiased while the equal-tailed interval $CS_{ET} = [\hat{\mu}_{\alpha/2}, \hat{\mu}_{1-\alpha/2}]$ has conditional coverage $1 - \alpha$

$$Pr\left\{\mu_Y(\hat{\theta}) \in CS_{ET} | \hat{\theta} = \tilde{\theta}\right\} \geq 1 - \alpha \text{ for } \tilde{\theta} \in \Theta \text{ and all } \mu. \quad (10)$$

Moreover results in Pfanzagl (1979) and Pfanzagl (1994) on optimal estimation for exponential families imply that $\hat{\mu}_\alpha$ is optimal in the class of quantile-unbiased estimators.

To establish optimality, we add the following assumption:

Assumption 1

If $\Sigma = \text{Cov}((X', Y')')$ has full rank, then the parameter space for μ is $\mathbb{R}^{2|\Theta|}$. Otherwise, there exists some μ^* such that the parameter space for μ is $\left\{\mu^* + \Sigma^{\frac{1}{2}}v : v \in \mathbb{R}^{2|\Theta|}\right\}$, where $\Sigma^{\frac{1}{2}}$ is the symmetric square root of Σ .

This assumption requires that the parameter space for μ be sufficiently rich. When Σ is degenerate (for example when $X = Y$, as in Section 2), this assumption further implies that (X, Y) have the same support for all values of μ . This rules out cases in which a pair of parameter values μ_1, μ_2 can be perfectly distinguished based on the data. Under this assumption, $\hat{\mu}_\alpha$ is an optimal quantile-unbiased estimator.

Proposition 2

For $\alpha \in (0, 1)$, $\hat{\mu}_\alpha$ is conditionally α -quantile-unbiased in the sense of (6). If Assumption 1 holds, then $\hat{\mu}_\alpha$ is the uniformly most concentrated α -quantile-unbiased estimator, in that for any other conditionally α -quantile-unbiased estimator $\hat{\mu}_\alpha^*$ and any loss function $L(d, \mu_Y(\tilde{\theta}))$ that attains its minimum at $d = \mu_Y(\tilde{\theta})$ and is quasiconvex in d for all $\mu_Y(\tilde{\theta})$,

$$E_\mu \left[L\left(\hat{\mu}_\alpha, \mu_Y(\tilde{\theta})\right) | \hat{\theta} = \tilde{\theta} \right] \leq E_\mu \left[L\left(\hat{\mu}_\alpha^*, \mu_Y(\tilde{\theta})\right) | \hat{\theta} = \tilde{\theta} \right]$$

for all μ and all $\tilde{\theta} \in \Theta$.

Proposition 2 shows that $\hat{\mu}_\alpha$ is optimal in the strong sense that it has lower expected loss than any other quantile-unbiased estimator for a large class of loss functions.

Other Selection Events We have discussed inference conditional on $\hat{\theta} = \tilde{\theta}$, but the same approach applies, and is optimal, for more general conditioning events. For instance, in the context of Section 2 a researcher might deliver a recommendation to a policymaker only when a statistical test indicates that the best-performing treatment outperforms some benchmark (see Tetenov, 2012). In this case, it is natural to also condition inference on the result of this test. Analogously, one may wish to conduct inference on the performance of an estimated best trading strategy or forecasting rule after finding a rejection when testing for superior predictive ability according to methods of e.g. White (2000), Hansen (2005) or Romano and Wolf (2005). Appendix A discusses the conditional approach in this more general case and derives the additional conditioning event in the context of the example just described.

Uniformly Most Accurate Unbiased Confidence Intervals In addition to equal-tailed confidence intervals, classical results on testing in exponential families discussed in Fithian et al. (2017) also permit the construction of uniformly most accurate unbiased confidence intervals. A level $1 - \alpha$ confidence set is unbiased if its probability of covering a false parameter value is bounded above by $1 - \alpha$, and uniformly most accurate unbiased confidence intervals minimize the probability of covering all incorrect parameter values over the class of unbiased confidence sets. Details of how to construct these confidence intervals are deferred to Appendix A for brevity.

3.3 Comparison to Sample Splitting

An alternative remedy for winner’s curse bias is to split the sample. If we have iid observations and select $\hat{\theta}^1$ based on the first half of the data, conventional estimates and confidence intervals for $\mu_Y(\hat{\theta}^1)$ that use only the second half of the data will be conditionally valid given $\hat{\theta}^1$. Hence, it is natural to ask how the analog of our conditioning approach applied to inference on $\mu_Y(\hat{\theta}^1)$, conditional on the realization of $\hat{\theta}^1$, compares to this conventional sample splitting approach.

Asymptotically, even sample splits yield a pair of independent and identically distributed normal draws (X^1, Y^1) and (X^2, Y^2) , both of which follow (4), albeit with a different scaling for (μ, Σ) than in the full-sample case.¹⁰ Sample splitting procedures calculate $\hat{\theta}^1$ as

¹⁰Appendix C considers cases with general sample splits and describes the scaling for (μ, Σ) . Intuitively, the scope for improvement over conventional split-sample inference is increasing in the fraction of the data used to construct X_1 .

in (5), replacing X by X^1 . Inference on $\mu_Y(\hat{\theta}^1)$ is then conducted using Y^2 . In particular, the conventional 95% sample-splitting confidence interval for $\mu_Y(\hat{\theta}^1)$,

$$\left[Y^2(\hat{\theta}^1) - 1.96\sqrt{\Sigma_Y(\hat{\theta}^1)}, Y^2(\hat{\theta}^1) + 1.96\sqrt{\Sigma_Y(\hat{\theta}^1)} \right],$$

has correct (conditional) coverage, and $Y^2(\hat{\theta}^1)$ is median-unbiased for $\mu_Y(\hat{\theta}^1)$.

Conventional sample splitting resolves the winner’s curse, but comes at a cost. First, $\hat{\theta}^1$ is based on less data than in the full-sample case, which is unappealing since a policy recommendation estimated with a smaller sample size leads to a lower expected welfare (see, e.g., Theorems 2.1 and 2.2 in Kitagawa and Tetenov (2018b)). Moreover, even after conditioning on $\hat{\theta}^1$, the full-sample average $\frac{1}{2}(X^1, Y^1) + \frac{1}{2}(X^2, Y^2)$ remains minimal sufficient for μ . Hence, using only Y^2 for inference sacrifices information.

Fithian et al. (2017) formalize this point and show that conventional sample splitting tests (and thus confidence intervals) are inefficient.¹¹ Motivated by this result, in Appendix C we derive optimal estimators and confidence intervals for $\mu_Y(\hat{\theta}^1)$ that are valid conditional on $\hat{\theta}^1$. These optimal split-sample procedures involve distributions that are difficult to compute, however, so we also propose computationally straightforward alternatives. These alternatives dominate conventional split-sample methods for inference on $\mu_Y(\hat{\theta}^1)$, but are in turn dominated by the (intractable) optimal split-sample procedures. The split sample methods we introduce in Appendix C are related to conditionally valid methods for inference in adaptive clinical trial designs proposed in the biostatistics literature (e.g., Cohen and Sackrowitz, 1989; Sampson and Sill, 2005). See Appendix C for details.

3.4 Behavior When $Pr_\mu\{\hat{\theta} = \tilde{\theta}\}$ is Large

As discussed in Section 2, if we ignore selection and compute the conventional (or “naive”) estimator $\hat{\mu}_N = Y(\hat{\theta})$ and the conventional confidence interval

$$CS_N = \left[Y(\hat{\theta}) - c_{\alpha/2, N}\sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_{\alpha/2, N}\sqrt{\Sigma_Y(\hat{\theta})} \right],$$

where $c_{\alpha, N}$ is the $1 - \alpha$ -quantile of the standard normal distribution, $\hat{\mu}_N$ is biased and CS_N has incorrect coverage conditional on $\hat{\theta} = \tilde{\theta}$. These biases are mild when $Pr_\mu\{\hat{\theta} = \tilde{\theta}\}$ is close to one, however, since in this case the conditional distribution is close to the unconditional

¹¹Corollary 1 of Fithian et al. (2017) applied in our setting shows that for any sample splitting test based on Y^2 , there exists a test that uses the full data and has weakly higher power against all alternatives and strictly higher power against some alternatives.

one. Intuitively, $Pr_\mu\{\hat{\theta}=\tilde{\theta}\}$ is close to one for some $\tilde{\theta}$ when $\mu_X(\theta)$ has a well-separated maximum. Our procedures converge to conventional ones in this case.

Proposition 3

Consider any sequence of values μ_m such that $Pr_{\mu_m}\{\hat{\theta}=\tilde{\theta}\}\rightarrow 1$. Then under μ_m we have $CS_{ET}\rightarrow_p CS_N$ and $\hat{\mu}_{\frac{1}{2}}\rightarrow_p Y(\tilde{\theta})$ both conditional on $\hat{\theta}=\tilde{\theta}$ and unconditionally, where for confidence intervals \rightarrow_p denotes convergence in probability of the endpoints.

This result provides an additional argument for using our procedures: they remain valid when conventional procedures fail, but coincide with conventional procedures when the latter are valid. On the other hand, as we saw in Section 2, there are cases where our conditional procedures have poor unconditional performance.

4 Unconditional Inference

Rather than requiring validity conditional on $\hat{\theta}$, one might instead require coverage only on average, yielding the unconditional coverage requirement

$$Pr\{\mu_Y(\hat{\theta})\in CS\}\geq 1-\alpha \text{ for all } \mu. \quad (11)$$

All confidence intervals with correct conditional coverage in the sense of (10) also have correct unconditional coverage provided $\hat{\theta}$ is unique with probability one.

Proposition 4

Suppose that $\hat{\theta}$ is unique with probability one for all μ . Then any confidence interval CS with correct conditional coverage (10) also has correct unconditional coverage (11).

Uniqueness of $\hat{\theta}$ implies that the conditioning events $\mathcal{X}(\tilde{\theta})$ partition the support of X with measure zero overlap. The result then follows from the law of iterated expectations.

A sufficient condition for almost sure uniqueness of $\hat{\theta}$ is that Σ_X has full rank. A weaker sufficient condition is given in the next lemma. Cox (2018) gives sufficient conditions for uniqueness of a global optimum in a much wider class of problems.

Lemma 1

Suppose that for all $\theta, \tilde{\theta}\in\Theta$ such that $\theta\neq\tilde{\theta}$, $X(\theta)$ and $X(\tilde{\theta})$ are not perfectly (positively) correlated. Then $\hat{\theta}$ is unique with probability one for all μ .

While the conditional confidence intervals derived in the last section are unconditionally valid, unconditional coverage is less demanding than conditional coverage. Hence, if we

are only concerned with unconditional coverage, relaxing the coverage requirement may allow us to obtain shorter confidence intervals in some settings.

This section explores the benefits of such a relaxation. We begin by introducing unconditional confidence intervals based on projections of simultaneous confidence bands for μ . We then introduce hybrid estimators and confidence intervals that combine projection intervals with conditioning arguments.

4.1 Projection Confidence Intervals

One approach to obtain an unconditional confidence interval for $\mu_Y(\hat{\theta})$ is to start with a joint confidence interval for μ and project on the dimension corresponding to $\hat{\theta}$. This approach was used by Romano and Wolf (2005) in the context of multiple testing, and by Kitagawa and Tetenov (2018a) for inference on an estimated optimal policy. This approach has also been used in a large and growing statistics literature on post-selection inference including e.g. Berk et al. (2013), Bachoc et al. (2020) and Kuchibhotla et al. (2020).

To formally describe the projection approach, let c_α denote the $1 - \alpha$ quantile of $\max_{\theta} |\xi(\theta)| / \sqrt{\Sigma_Y(\theta)}$ for $\xi \sim N(0, \Sigma_Y)$. If we define

$$CS_\mu = \left\{ \mu_Y : |Y(\theta) - \mu_Y(\theta)| \leq c_\alpha \sqrt{\Sigma_Y(\theta)} \text{ for all } \theta \in \Theta \right\},$$

then CS_μ is a level $1 - \alpha$ confidence set for μ_Y . If we then define

$$CS_P = \left\{ \tilde{\mu}_Y(\hat{\theta}) : \exists \mu \in CS_\mu \text{ such that } \mu_Y(\hat{\theta}) = \tilde{\mu}_Y(\hat{\theta}) \right\} = \left[Y(\hat{\theta}) - c_\alpha \sqrt{\Sigma_Y(\hat{\theta})}, Y(\hat{\theta}) + c_\alpha \sqrt{\Sigma_Y(\hat{\theta})} \right]$$

as the projection of CS_μ on the parameter space for $\mu_Y(\hat{\theta})$, then since $\mu_Y \in CS_\mu$ implies $\mu_Y(\hat{\theta}) \in CS_P$, CS_P satisfies the unconditional coverage requirement (11). As noted in Section 2, however, CS_P does not generally have correct conditional coverage.

The width of the confidence interval CS_P depends on the variance $\Sigma_Y(\hat{\theta})$ but does not otherwise depend on the data.¹² To account for the randomness of $\hat{\theta}$, the critical value c_α is typically larger than the conventional two-sided normal critical value. Hence, CS_P will be conservative in cases where $\hat{\theta}$ takes a given value $\tilde{\theta}$ with high probability. To improve performance in such cases, we propose a hybrid inference approach.

¹²One could consider alternative projection intervals, for instance optimized to have shorter length at some $\hat{\theta}$ values in exchange for greater length at others. See Freyberger and Rai (2018) and Frandsen (2020).

4.2 Hybrid Inference

As shown in Section 2, the conditional and projection approaches each have good unconditional performance in some cases, but neither is fully satisfactory. Hybrid inference combines the approaches to obtain good performance over a wide range of parameter values.

To construct hybrid estimators, we condition both on $\hat{\theta} = \tilde{\theta}$ and on the event that $\mu_Y(\hat{\theta})$ lies in the level $1 - \beta$ projection confidence interval CS_P^β for $0 \leq \beta < \alpha$. Hence, the conditioning event becomes

$$\mathcal{Y}^H(\tilde{\theta}, \mu_Y(\tilde{\theta}), z) = \mathcal{Y}(\tilde{\theta}, z) \cap \left[\mu_Y(\tilde{\theta}) - c_\beta \sqrt{\Sigma_Y(\tilde{\theta})}, \mu_Y(\tilde{\theta}) + c_\beta \sqrt{\Sigma_Y(\tilde{\theta})} \right].$$

Let $F_{TN}^H(y; \mu_Y(\tilde{\theta}), \tilde{\theta}, z)$ denote the conditional distribution function of $Y(\tilde{\theta})$, and define $\hat{\mu}_\alpha^H$ to solve $F_{TN}^H(Y(\hat{\theta}); \hat{\mu}_\alpha^H, \hat{\theta}, Z_{\hat{\theta}}) = 1 - \alpha$. The hybrid estimator $\hat{\mu}_\alpha^H$ is α -quantile unbiased conditional on $\mu_Y(\hat{\theta}) \in CS_P^\beta$.

Proposition 5

For $\alpha \in (0, 1)$, $\hat{\mu}_\alpha^H$ is unique and $\hat{\mu}_\alpha^H \in CS_P^\beta$. If $\hat{\theta}$ is unique almost surely for all μ , $\hat{\mu}_\alpha^H$ is α -quantile unbiased conditional on $\mu_Y(\hat{\theta}) \in CS_P^\beta$:

$$Pr_\mu \left\{ \hat{\mu}_\alpha^H \geq \mu_Y(\hat{\theta}) \mid \mu_Y(\hat{\theta}) \in CS_P^\beta \right\} = \alpha \text{ for all } \mu.$$

Proposition 5 implies several notable properties for the hybrid estimator. First, since $Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CS_P^\beta \right\} \geq 1 - \beta$ by construction, one can show that

$$\left| Pr_\mu \left\{ \hat{\mu}_\alpha^H \geq \mu_Y(\hat{\theta}) \right\} - \alpha \right| \leq \beta \cdot \max\{\alpha, 1 - \alpha\} \text{ for all } \mu.$$

This implies that the absolute median bias of $\hat{\mu}_{\frac{1}{2}}^H$ (measured as the deviation of the exceedance probability from $1/2$) is bounded above by $\beta/2$. On the other hand, since $\hat{\mu}_{\frac{1}{2}}^H \in CS_P^\beta$ we have $\left| \hat{\mu}_{\frac{1}{2}}^H - Y(\hat{\theta}) \right| \leq c_\beta \sqrt{\Sigma_Y(\tilde{\theta})}$, so the difference between $\hat{\mu}_{\frac{1}{2}}^H$ and the conventional estimator $Y(\hat{\theta})$ is bounded above by half the width of CS_P^β . As β varies, the hybrid estimator interpolates between the median-unbiased estimator $\hat{\mu}_{\frac{1}{2}}$ and the conventional estimator $Y(\hat{\theta})$.

As with the quantile-unbiased estimator $\hat{\mu}_\alpha$, we can form confidence intervals based on hybrid estimators. In particular, the set $[\hat{\mu}_{\alpha/2}^H, \hat{\mu}_{1-\alpha/2}^H]$ has coverage $1 - \alpha$ conditional on $\mu_Y(\hat{\theta}) \in CS_P^\beta$. This is not fully satisfactory, however, as $Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CS_P^\beta \right\} < 1$. Hence, to ensure correct coverage, we define the level $1 - \alpha$ hybrid confidence interval

as $CS_{ET}^H = \left[\hat{\mu}_{\frac{\alpha-\beta}{2(1-\beta)}}^H, \hat{\mu}_{1-\frac{\alpha-\beta}{2(1-\beta)}}^H \right]$. With this adjustment, hybrid confidence intervals have coverage at least $1-\alpha$ both conditional on $\mu_Y(\hat{\theta}) \in CS_P^\beta$ and unconditionally.

Proposition 6

Provided $\hat{\theta}$ is unique with probability one for all μ , the hybrid confidence interval CS_{ET}^H has coverage $\frac{1-\alpha}{1-\beta}$ conditional on $\mu_Y(\hat{\theta}) \in CS_P^\beta$.

$$Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CS_{ET}^H \mid \mu_Y(\hat{\theta}) \in CS_P^\beta \right\} = \frac{1-\alpha}{1-\beta} \text{ for all } \mu.$$

Moreover, the unconditional coverage is between $1-\alpha$ and $\frac{1-\alpha}{1-\beta} \leq 1-\alpha+\beta$:

$$\inf_\mu Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CS_{ET}^H \right\} \geq 1-\alpha, \quad \sup_\mu Pr_\mu \left\{ \mu_Y(\hat{\theta}) \in CS_{ET}^H \right\} \leq \frac{1-\alpha}{1-\beta}.$$

Hybrid confidence intervals strike a balance between the conditional and projection approaches. The maximal length of hybrid confidence intervals is bounded above by the length of CS_P^β . For small β , hybrid confidence intervals will be close to conditional confidence intervals, and thus to conventional confidence intervals, when $\hat{\theta} = \tilde{\theta}$ with high probability. For $\beta > 0$, however, hybrid confidence intervals do not fully converge to conventional confidence intervals as $Pr_\mu \left\{ \hat{\theta} = \tilde{\theta} \right\} \rightarrow 1$. Nevertheless, our simulations in Section 2 find similar performance for the hybrid and conditional approaches in well-separated cases.

While hybrid confidence intervals combine the conditional and projection approaches, they can yield overall performance more appealing than either. In Section 2 we found that hybrid confidence intervals had a shorter median length for many parameter values than did either the conditional or projection approaches used in isolation. Our simulation results below provide further evidence of outperformance in realistic settings.

Choice of β To use the hybrid approach we must select the coverage β of the initial projection interval CS_P^β . Intuitively this choice trades off the length of CS_P^β , which bounds the worst-case length of CS_{ET}^H in the poorly-separated case, against the length of CS_{ET}^H in the well-separated case. For a given Σ we can precisely quantify this tradeoff, calculating the length of CS_P^β and the length of CS_{ET}^H in the well-separated case for each β and selecting a point on the resulting frontier. This frontier is Σ -specific, however, so this analysis does not deliver a general recommendation.

As an alternative, we note that the length of CS_{ET}^H in the well-separated case is bounded above by that of the level $\frac{1-\alpha}{1-\beta}$ conventional confidence interval. Specifically, for the standard

choice of $\alpha=5\%$, choosing $\beta=\frac{\alpha}{10}=0.5\%$ implies that the CS_{ET}^H has half-length no more than 2.0025 standard errors in the well-separated case. We view this as a small increase relative to the half-length of the conventional 5% interval, 1.96 standard errors, and so suggest this as a default choice, focusing on $\beta=\frac{\alpha}{10}$ in our simulations and applications.¹³

Comparison to Bonferroni Adjustment It is worth contrasting our hybrid approach with Bonferroni corrections as in e.g. Romano et al. (2014) and McCloskey (2017). A simple Bonferroni approach for our setting intersects a level $1-\beta$ projection confidence interval CS_P^β with a level $1-\alpha+\beta$ conditional interval that conditions only on $\hat{\theta}=\tilde{\theta}$. Bonferroni intervals differ from our hybrid approach in two respects. First, they use a level $1-\alpha+\beta$ conditional confidence interval, while the hybrid approach uses a level $\frac{1-\alpha}{1-\beta}$ conditional interval, where $\frac{1-\alpha}{1-\beta} \leq 1-\alpha+\beta$. Second, the conditional interval used by the Bonferroni approach does not condition on $\mu_Y(\tilde{\theta}) \in CS_P^\beta$, while that used by the hybrid approach does. Consequently, one can show that hybrid confidence intervals exclude the endpoints of CS_P^β almost surely, while the same is not true of Bonferroni intervals.

5 Feasible Inference and Large-Sample Results

Our results have so far assumed that (X,Y) are jointly normal with known variance Σ . While exact normality is rare in practice, researchers commonly use estimators that are asymptotically normal with consistently estimable asymptotic variance. Our results for the finite-sample normal model translate to asymptotic results in this case.

Specifically, suppose that for sample size n we construct a vector of statistics X_n , that $\hat{\theta}_n = \operatorname{argmax}_{\theta \in \Theta} X_n(\theta)$, and that we are interested in the mean of $Y_n(\hat{\theta}_n)$. In the treatment choice example discussed in Section 2, for instance, θ indexes treatments, $X_n(\theta)$ is \sqrt{n} times the sample average outcome under treatment θ , and $Y_n(\theta) = X_n(\theta)$. We suppose that (X_n, Y_n) are jointly asymptotically normal once recentered by vectors $(\mu_{X,n}, \mu_{Y,n})$,

$$\begin{pmatrix} X_n - \mu_{X,n} \\ Y_n - \mu_{Y,n} \end{pmatrix} \Rightarrow N(0, \Sigma).$$

In the treatment choice example $\mu_{X,n}(\theta) = \mu_{Y,n}(\theta) = \sqrt{n}E[Y_{i,\theta}]$ is the average potential outcome under treatment θ , scaled by \sqrt{n} . We further assume that we have a consistent estimator $\hat{\Sigma}$ for the asymptotic variance Σ . In the treatment choice example, for instance, we can take $\hat{\Sigma}$ to be the matrix with the sample variance of the outcome for each the

¹³Romano et al. (2014) and McCloskey (2017) likewise find this choice to perform well in two different settings when using a Bonferroni correction

treatment group along the diagonal and zeros elsewhere.

More broadly, (X_n, Y_n) can be any vectors of asymptotically normal estimators, and we can calculate $\hat{\Sigma}$ as we would for inference on $(\mu_{X,n}, \mu_{Y,n})$, including corrections for clustering, serial correlation, and the like in the usual way.¹⁴ Feasible inference based on our approach simply substitutes (X_n, Y_n) and $\hat{\Sigma}$ in place of (X, Y) and Σ in all expressions. Appendix D shows that this plug-in approach yields asymptotically valid inference on $\mu_{Y,n}(\hat{\theta}_n)$. This result is trivial when the sequence of vectors $\mu_{X,n}$ has a well-separated maximizer $\theta^* = \operatorname{argmax}_{\theta \in \Theta} \mu_{X,n}(\theta)$ with $\mu_{X,n}(\theta^*) \gg \max_{\theta \in \Theta \setminus \theta^*} \mu_{X,n}(\theta)$ for large n , since in this case $\hat{\theta}_n = \theta^*$ with high probability, and the selection problem vanishes. In Section 2, for instance, if we fix a data-generating process with $E[Y_{i,1}] > E[Y_{i,0}]$ and take $n \rightarrow \infty$, then $Pr\{\hat{\theta}_n = 1\} \rightarrow 1$.

Based on this argument, it could be tempting to conclude that inference ignoring the winner’s curse will be approximately valid so long as there is not an exact tie for the treatment yielding the highest average outcome. In finite samples, however, near-ties yield very similar behavior to exact ties. Moreover, no matter how large the sample size, we can have near-ties sufficiently close that inference ignoring selection remains unreliable. Hence, what matters for inference is neither whether there are exact ties, nor the sample size as such (beyond the minimum needed to justify the normal approximation), but instead how close the best-performing treatments are to each other *relative* to the degree of sampling uncertainty. Depending on the data generating process, selection issues can thus remain important no matter how large the sample. To obtain reliable large-sample approximations, we thus seek *uniform* asymptotic results, which for sufficiently large samples guarantee performance over a wide class of data generating processes. Appendix D establishes that plug-in versions of our proposed procedures are uniformly asymptotically valid in this sense.

6 Application: Charitable Giving

Karlan and List (2007) partner with a political charity to conduct a field experiment examining the effectiveness of matching incentives at increasing charitable giving. In matched donations, a lead donor pledges to ‘match’ any donations made by other donors up to some threshold, effectively lowering the price of political activism for other donors.

Karlan and List (2007) use a factorial design. Potential donors, who were previous donors to the charity, were mailed a four page letter asking for a donation. The contents of the letter were randomized, with one third of the sample assigned to a control group

¹⁴We have scaled (X_n, Y_n) by \sqrt{n} for expositional purposes, but dropping this scaling yields inference on the correspondingly scaled version of $\mu_{Y,n}(\hat{\theta}_n)$. Hence, one can use estimators and estimated variances with the natural scale in a given setting.

Index	Treatment	Description
0	0	Control group with no matched donations
Match ratio		
1	1:1	An additional dollar up to the match limit
2	2:1	Two additional dollars up to the match limit
3	3:1	Three additional dollars up to the match limit
Match size		
1	\$25,000	Up to \$25,000 is pledged
2	\$50,000	Up to \$50,000 is pledged
3	\$100,000	Up to \$100,000 is pledged
4	Unstated	The pledged amount is not stated
Ask amount		
1	Same	The individual is asked to give as much as their largest past donation
2	25% more	The individual is asked to give 25% more than their largest past donation
3	50% more	The individual is asked to give 50% more than their largest past donation

Table 1: Treatment arms for Karlan and List (2007). Individuals were assigned to the control group or to the treatment group, in the ratio 1:2. Treated individuals were randomly assigned a match ratio, a match size and an ask amount with equal probability. There are 36 possible combinations, plus the control group. The leftmost column specifies a reference index used throughout this section for convenience.

that received a standard letter with no match. The remaining two thirds received a letter with the line “now is the time to give!” and details for a match. Treated individuals were randomly assigned with equal probability to one of 36 separate treatment arms. Treatment arms are characterized by a match ratio, a match size, and an ask amount, for which further details are given in Table 1. The outcome of interest is the average dollar amount that individuals donated to the charity in the month following the solicitation.

In total, 50,083 individuals were contacted, of which 16,687 were randomly assigned to the control group, while 33,396 were randomly assigned to one of the 36 treatment arms. The (unconditional) average donation was \$0.81 in the control group and \$0.92 in the treatment group. Conditional on giving, these figures were \$45.54 and \$44.35, respectively. The discrepancy reflects the low response rate; only 1,034 of 50,083 individuals donated.

Treatment	Average donation	Standard error	95% CI
(1,3,2)	1.52	0.35	[0.83,2.20]
(2,1,3)	1.51	0.46	[0.61,2.41]
(2,1,1)	1.42	0.39	[0.66,2.19]
(3,1,3)	1.40	0.36	[0.70,2.11]

Table 2: The average donations for the four best treatment arms according to the data, $n=50,083$. Treatments are indexed by the indicators for (Match ratio, Match size, Ask amount) defined in Table 1. The reported 95% confidence intervals are the conventional ones that do not take selection into account.

Table 2 reports average revenue from the four best-performing treatment arms, along with standard errors and conventional confidence intervals. Taken at face value, the results for the best-performing arm suggest that a similarly-situated nonprofit considering a campaign that promises a dollar-for-dollar match up to \$100,000 in donations and asks individuals to donate 25% more than their largest past donation could expect to raise \$1.52 per potential donor, on average, with a confidence interval of \$0.83 to \$2.20. This estimate and confidence interval are clearly subject to winner’s curse bias, however: we are picking the best-performing arm out of 37 in the experiment, which will bias our estimates and confidence intervals upward.

Simulation Results To investigate the extent of winner’s curse bias and the finite-sample performance of our corrections, we calibrate simulations to this application. We simulate datasets by resampling observations with replacement from the Karlan and List (2007) data (i.e. by drawing nonparametric bootstrap samples). In each simulated sample we re-estimate the effectiveness of each treatment arm, pick the best-performing arm, and study the performance of estimates and confidence intervals, treating the estimates for the original Karlan and List (2007) data as the true values. The underlying data here are non-normal and we re-estimate the variance in each simulation draw. Hence, these results also speak to the finite-sample performance of the normal approximation. We report results based on 10,000 simulation draws.

Since revenue does not account for the cost of the fund-raising campaign, it is impossible for the solicitation to raise a negative amount. We therefore set the parameter space for $\mu(\hat{\theta})$ to \mathbb{R}_+ , and trim the point estimators and the confidence intervals at zero, $\hat{\mu}^{trim} \equiv \max\{0, \hat{\mu}\}$ and $CS^{trim} = [0, \infty) \cap CS$. This trimming does not affect the coverage of the confidence intervals, and also preserves the α -quantile unbiasedness of the estimators

Winner											
(1,3,2)	(1,4,2)	(1,4,3)	(2,1,1)	(2,1,3)	(2,2,2)	(2,3,3)	(2,4,1)	(2,4,2)	(3,1,1)	(3,1,3)	(3,3,1)
16.0%	11.4%	1.3%	13.0%	18.9%	10.8%	1.3%	1.5%	2.8%	5.1%	10.0%	3.6%

Table 3: Frequency of simulation replications where each treatment is estimated to perform best in simulations calibrated to Karlan and List (2007). Treatments are indexed by the indicators for (Match ratio, Match size, Ask amount) defined in Table 1. 31 of the 37 treatments are best in at least one replication; those that won in at least 1% of simulated samples are reported.

	Estimate		
	Conventional	Median unbiased	Hybrid
Median bias	0.61	-0.18	-0.18
Probability bias	0.50	-0.07	-0.07
Median absolute error	0.61	0.65	0.64

Table 4: Performance measures for alternative estimators in simulations calibrated to Karlan and List (2007). Probability bias is $Pr\{\hat{\mu}^{trim} > \mu(\hat{\theta})\} - \frac{1}{2}$.

so long as the true value $\mu(\hat{\theta})$ is greater than zero.

There is substantial variability in the “winning” arm: 31 of the 37 treatments won in at least one simulation draw and 12 treatment arms won in at least 1% of simulated samples. Table 3 lists these 12 treatments. The variability of the winning arm suggests that there is scope for a winner’s curse in this setting.

Table 4 examines the performance of naive, median unbiased, and hybrid estimates, reporting (unconditional) median bias, probability bias ($Pr\{\hat{\mu}^{trim} > \mu(\hat{\theta})\} - \frac{1}{2}$), and median absolute error. Trimming the estimators at zero does not affect the reported performance measures. Naive estimates suffer from substantial bias in this setting: they have a median bias of \$0.61, and over-estimate the revenue generated by the selected arm 100% of the time, up to rounding. The median unbiased and hybrid estimators substantially improve both measures of bias, though given the finite-sample setting they do not eliminate it completely and are both somewhat downward biased, though to a lesser degree.¹⁵ All three estimators perform similarly in terms of median absolute error.

Tables 5 and 6 report results for confidence intervals. Specifically, we consider the naive,

¹⁵This is a particularly challenging setting for the normal approximation, as the outcomes distribution is highly skewed due to the large number of zeros. In particular, there are on average only 20 nonzero outcomes per non-control treatment (out of approximately 930 observations in each treatment group).

projection, conditional, and hybrid confidence intervals with nominal coverage 95%. Table 5 reports unconditional coverage and median length, while Table 6 reports conditional coverage probabilities given $\hat{\theta}$ values among the 12 treatments listed in Table 3. Naive confidence intervals slightly undercover unconditionally, with coverage 92%. Their conditional coverage varies depending on which treatment is the winner. If the winning treatment is one of the six best-performing treatments, the conditional coverage is at least 95%, while otherwise the naive confidence intervals under-cover with coverage probability as low as 65%.

	Unconditional coverage	Median length	
		Trimmed	Untrimmed
Naive CS	0.92	1.88	1.88
CS_P	1.00	3.08	3.08
CS_{ET}	0.97	2.69	5.91
CS_{ET}^H	0.97	2.52	2.56

Table 5: Unconditional coverage probabilities of the confidence intervals in simulations calibrated to Karlan and List (2007). Unconditional median lengths are reported for the trimmed and untrimmed confidence intervals.

Treatment θ	Average donation $\mu(\theta)$	Conditional coverage			
		Naive CS	CS_P	CS_{ET}	CS_{ET}^H
(1,3,2)	1.52	0.95	1	0.98	0.98
(2,1,3)	1.51	0.97	1	0.97	0.97
(2,1,1)	1.42	0.94	1	0.97	0.97
(3,1,3)	1.40	0.95	1	0.97	0.97
(2,2,2)	1.34	0.96	1	0.97	0.98
(1,4,2)	1.27	0.99	1	0.97	0.97
(3,3,1)	1.26	0.84	1	0.96	0.97
(3,1,1)	1.24	0.89	1	0.97	0.97
(2,4,2)	1.22	0.79	1	0.99	0.99
(2,3,3)	1.12	0.65	1	0.98	0.98
(2,4,1)	1.10	0.81	1	0.97	0.97
(1,4,3)	1.03	0.78	1	0.96	0.97

Table 6: Conditional coverage probabilities, $Pr\{\mu(\hat{\theta}) \in CS^{trim} | \hat{\theta} = \theta\}$, of the confidence intervals for each of the 12 treatments in Table 3. The treatments are sorted according to the average donation.

Treatment (1,3,2)		Estimates	Equal-tailed CI
Naive		1.52	[0.83,2.20]
Projection		–	[0.40,2.63]
Conditional	– trimmed	0	[0, 1.42]
	– untrimmed	-7.49	[-47.66,1.42]
Hybrid		0.20	[0.19,1.47]

Table 7: Naive and bias-corrected estimates and confidence intervals for best-performing treatment in Karlan and List (2007) data.

Projection confidence intervals over-cover unconditionally and conditionally for these treatments, with coverage 100%. Conditional and hybrid confidence intervals slightly over-cover, with unconditional and conditional coverage about 97%, and have unconditional median (trimmed) length around 35% larger than naive intervals and around 20% shorter than projection intervals. It is important to emphasize, however, that the conditional coverage for projection and hybrid intervals is particular to the data generating process considered here: as illustrated in Figure 5, these intervals do not ensure conditional coverage in general.

The median length of conditional intervals more than doubles if we leave their lower bound untrimmed. In contrast, the median length of the hybrid confidence intervals is basically unaffected by trimming. This is because despite the similarity of their upper bounds, the lower bound of the conditional confidence intervals tends to be negative and substantially lower than the lower bound of the hybrid confidence intervals. In other words, if the parameter space is unconstrained, the hybrid confidence intervals are substantially shorter than conditional confidence intervals. The good performance of the hybrid approach in applications with unconstrained parameter space is encouraging, and in line with the results in Section 2.

Empirical results Returning to the Karlan and List (2007) data, Table 7 reports corrected estimates and confidence intervals for the best-performing treatment in the experiment. We repeat the naive estimate and confidence interval for comparison. The median unbiased estimate makes an aggressive downwards correction to the naive estimate, suggesting negative revenue (-\$7.49) from the winning arm if not trimmed. The conditional confidence interval is tight, ranging from 0 to \$1.42, if trimmed at zero, and otherwise extremely wide, ranging from -\$47.66 to \$1.42. The hybrid estimate also shifts the conventional estimate downwards, but much less so. Moreover, the hybrid confidence interval is no wider than the naive interval, and excludes both zero and the naive estimate.

These results suggest that future fundraising campaigns deploying the winning strategy in the experiment are likely to raise some revenue, but substantially less than would be expected based on the naive estimates.

Conditional inference seems potentially natural in this application. The data highlight an interpretable combination of treatment parameters (1:1 match, \$100,000 pledged, with an ask 25% above an individual’s highest past donation) as best-performing, raising the question of what we can conclude about this particular treatment, given that it was the best in the experiment. This is precisely the question answered by the conditional approach. By contrast, while the hybrid approach ensures correct coverage on average across different “winning” treatments which could arise, it offers no guarantees given the particular winner observed in the Karlan and List (2007) data.

7 Application: Neighborhood Effects

⚡ We next discuss simulation and empirical results based on Chetty et al. (2018) and Bergman et al. (2020). Earlier work, including Chetty and Hendren (2016a) and Chetty and Hendren (2016b) argues that the neighborhood in which a child grows up has a long-term causal impact on income in adulthood, and, moreover, that these impacts are closely related to the adult income of children who spend their entire childhood in a given neighborhood.

Motivated by these findings, Bergman et al. (2020) partnered with the public housing authorities in Seattle and King County in Washington State in an experiment helping housing voucher recipients move to a set of higher-opportunity target neighborhoods. Bergman et al. (2020) choose target neighborhoods based on the Chetty et al. (2018) “Opportunity Atlas.” This atlas compiles census-tract level estimates of economic mobility for communities across the United States. Bergman et al. (2020) define target neighborhoods by selecting approximately the top third of tracts in Seattle and King County based on estimated economic mobility.¹⁶ They then make “relatively minor” adjustments to the set of target tracts based on other criteria (Bergman et al., 2020, Appendix A).

A central question in this setting is whether families moving to the target tracts will in fact experience the positive outcomes predicted based on the Opportunity Atlas estimates and the hypothesis of neighborhood effects. Once long-term outcomes for the experimental sample are available, one can begin address this question by comparing outcomes for children in treated families to the Opportunity Atlas estimates used to select the target tracts in

¹⁶They measure economic mobility in terms of the average household income rank in adulthood for children growing up at the 25th percentile of the income distribution. See Chetty et al. (2018) for details.

the first place. Such a comparison is complicated by the winner’s curse, however: the Atlas estimates were already used to select the target tracts, so the naive estimate for the causal effect of the selected tracts will be systematically biased upwards. It is therefore useful to examine the extent of the winner’s curse, and the impact of our corrections, in this setting.

Motivated by related issues, Chetty et al. (2018) and Bergman et al. (2020) do not focus on naive estimates, but instead adopt a shrinkage or empirical Bayes approach. Their estimates correspond to Bayesian posterior means under a prior that takes tract-level economic mobility to be normally distributed conditional on a vector of observable tract characteristics, and then estimates mean and variance hyperparameters from the data. If one takes this prior seriously and abstracts from estimation of the hyperparameters (for instance because the number of tracts is large and we plug in consistent estimates), the posterior median for average economic mobility over selected tracts will be median-unbiased under the prior, and Bayesian credible sets will have correct coverage, again under the prior. See Section E in the supplement for further discussion. The efficacy of Bayesian approaches for correcting selection issues hinges crucially on correct specification of the prior, however, whereas our results ensure correct coverage and controlled median bias for all possible distributions of economic mobility across tracts. Throughout this section, we therefore include empirical Bayes procedures in our analysis as a point of comparison.¹⁷

Simulation Results To examine the extent of winner’s curse bias and the performance of different corrections, we calibrate simulations to the Opportunity Atlas data. For each of the 50 largest commuting zones (CZs) in the United States we treat the (un-shrunk) tract-level Opportunity Atlas estimates as the true values. We then simulate estimates by adding normal noise with standard deviation equal to the Opportunity Atlas standard error.¹⁸

We select the top third of tracts in each commuting zone based on these simulated estimates.¹⁹ To cast this into our setting, let \mathcal{T} be the set of tracts in a given CZ and

¹⁷Armstrong et al. (2020) propose an approach to robustify empirical Bayes confidence intervals to the choice of priors. Applied in the present setting, this approach would ensure correct coverage for tract-level economic mobility on average across all tracts in a given commuting zone. This approach does not focus on high-ranking tracts, however, and so does not, and is not intended to, address the winner’s curse. Hence, we do not report results for this approach.

¹⁸We base our estimates in this setting on the public Opportunity Atlas estimates and standard errors since we do not have access to the underlying microdata. We also do not have access to the correlation structure of the estimate across tracts. Such correlations arise from individuals who move across tracts, and there are few movers between most pairs of tracts, so we expect that these omitted correlations are small.

¹⁹We select the target tracts based on the un-shrunk estimates, rather than shrunk estimates as in Bergman et al. (2020). We do this because we find that selecting based on un-shrunk estimates yields slightly higher average quality for selected tracts than selecting on shrunk estimates, and because selection

Θ the set of selections from \mathcal{T} containing one third of tracts, $\Theta = \{\theta \subset \mathcal{T} : |\theta| = \lfloor |\mathcal{T}|/3 \rfloor\}$. For $\hat{\mu}_t$ the estimated effect of tract t , define $X(\theta)$ as the average estimate over tracts in θ , $X(\theta) = \frac{1}{|\theta|} \sum_{t \in \theta} \hat{\mu}_t$. Target tracts are selected as $\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} X(\theta)$, and the naive estimate for the average effect over selected tracts is $X(\hat{\theta})$. Correspondingly, for μ_t the neighborhood effect for tract t , let $\mu_X(\theta)$ be the average neighborhood effect over tracts in θ , $\mu_X(\theta) = \frac{1}{|\theta|} \sum_{t \in \theta} \mu_t$. We are interested in the difference between the average quality of selected tracts and the average over all tracts in the same commuting zone, $\mu_Y(\hat{\theta}) = \frac{1}{|\hat{\theta}|} \sum_{t \in \hat{\theta}} \mu_t - \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mu_t$, and so define $Y(\theta) = \frac{1}{|\theta|} \sum_{t \in \theta} \mu_t - \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mu_t$.²⁰ We study the performance of naive estimates and confidence intervals, empirical Bayes estimates and credible sets, and our corrected estimates and confidence intervals.

Figure 6 reports results based on ten thousand simulation draws. Panel (a) plots the average true upward mobility for selected tracts $E\left[\frac{1}{|\hat{\theta}|} \sum_{t \in \hat{\theta}} \mu_t\right]$ less the average over all tracts in the same CZ $\frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \mu_t$, $E\left[\mu_Y(\hat{\theta})\right]$, across the 50 CZs considered. Selected tracts are better than average across all 50 CZs, though the precise degree of improvement varies. Panel (b) shows median bias for the estimators we consider, where the quantity of interest is again the difference between average upward mobility for selected tracts, less the average over all tracts in the same CZ. As expected the naive estimator is biased upwards, while the sign of the bias for empirical Bayes differs across CZs. The conditional estimator is median unbiased up to simulation error, while the hybrid estimator is very close to median unbiased. Panel (c) plots the median absolute estimation error across the four estimators. The naive estimator has the largest median absolute estimation error in most CZs, while the empirical Bayes typically has the smallest. The conditional and hybrid estimators are in the middle, with quite similar median absolute estimation errors for this application. Finally, panels (d) and (e) plot the coverage and median length of confidence intervals. We see that the naive confidence interval severely under-covers, with coverage close to zero in all 50 CZs. The coverage of empirical Bayes intervals differs widely across CZs,

based on shrunk estimates introduces nonlinearity (due to estimation of the hyperparameters) which complicates conditional and hybrid inference.

²⁰Under the neighborhood effects model, $\mu_Y(\hat{\theta})$ corresponds to the average effect of moving a household from a randomly selected tract in the CZ to a randomly selected target tract. This need not correspond to the average treatment effect from the experiment in Bergman et al. (2020), since treatment and control households are not in general uniformly distributed across these sets of tracts. Indeed, some treated households settle in non-target tracts, and some control households settle in target tracts. Given realized location choices for treatment and control households, one could re-define $\mu_Y(\hat{\theta})$ accordingly. We do not pursue this extension, however, as data on location choice under treatment only exists for the Seattle CZ, where Bergman et al. (2020) conducted their experiment. A previous version of the paper applied an incorrect scaling when calculating $\mu_Y(\hat{\theta})$. We thank Magne Mogstad and Larry Katz flagging scaling errors.

ranging from less than 1% to over 90%.²¹ Conditional confidence intervals have coverage equal to 95% up to simulation error in all CZs, while the hybrid intervals have coverage very close to 95%, and below 96%, in all cases. Finally, projection intervals have coverage equal to 100%, up to simulation error, in all CZs. Turning to median length, we see that hybrid intervals are longer than empirical Bayes and naive confidence intervals, but are considerably shorter than conditional and projection intervals in many cases.

Empirical Results Figure 7 plots results for the Opportunity Atlas data. As in the simulations we select the top third of census tracts in each CZ based on the naive estimates and then report naive, empirical Bayes, and hybrid estimates and intervals, as well as projection intervals, for the average upward mobility across selected tracts, less the average over the commuting zone. For visibility, we defer results for conditional intervals to Figure 8 of Appendix E. From these results, we see that both the empirical Bayes and hybrid adjustments shift the naive estimates and intervals downward. There is not a clear pattern to the shifts in the estimates: in some cases the empirical Bayes estimate is below the hybrid, while in other cases the order is reversed. As expected given our simulation results, the (coverage-maintaining) hybrid intervals are wider than the (under-covering) empirical Bayes intervals, but considerably shorter than projection intervals. Hybrid and projection intervals exclude zero in all CZs, suggesting that, under the hypothesis of neighborhood effects, there is real scope for selecting better neighborhoods based on the Opportunity Atlas, albeit less than the naive estimates suggest.²²

The results for conditional procedures in Figure 8 of Appendix E are qualitatively similar, but the width of the conditional intervals is extremely variable across CZs. Specifically, while conditional intervals are quite similar to hybrid intervals in some CZs, they are much

²¹If one selects target tracts based on the empirical Bayes, rather than naive, estimates, this reduces the bias of EB estimates, with the average median bias across the 50 CZs falling from approximately -0.0023 to approximately -0.0001. Empirical Bayes credible sets continue to under-cover, however, with average coverage rising from approximately 50% to approximately 59%.

²²It is useful to compare our results with those of Mogstad et al. (2020), who study the problem of inference on ranks and consider the Opportunity Atlas data for Seattle as an example. They show that if one forms simultaneous confidence sets for individual tracts, one can say very little about which tracts are best. Hence, we can say little about the effect of moving an individual from an arbitrary non-target tract to arbitrary target tract, and can likewise say little about the average treatment effect of shifting households from one group of tracts to the other if we allow arbitrary location choices within each group of tracts. We consider a complementary exercise, inference on the average quality of selected sets of tracts, corresponding to an average treatment effect under uniformly distributed location choices. For this problem, we find strong evidence that selected tracts are, as a group, better than average. These exercises answer different questions, and the more positive results obtained in our case reflect that it is statistically easier to distinguish average mobility across groups of selected tracts than it is to rank individual tracts.

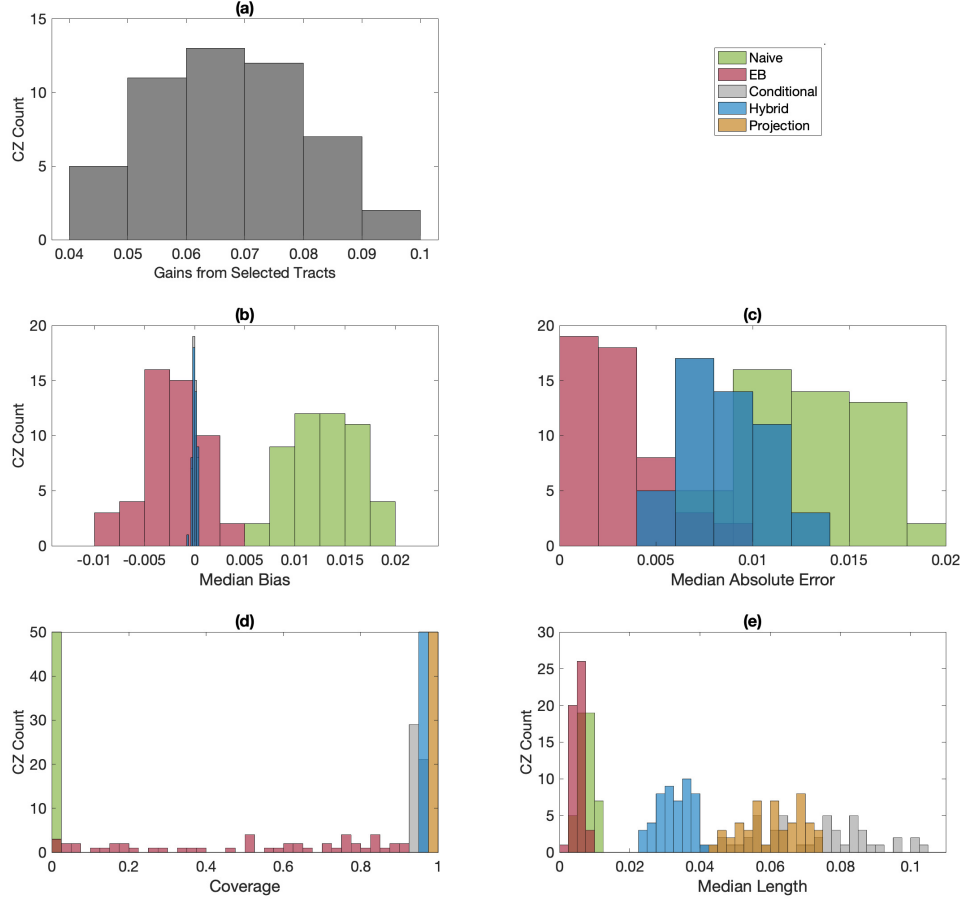


Figure 6: Simulation results from calibration to Chetty et al. (2018) Opportunity Atlas. Panel (a) shows the distribution of average improvement in economic mobility in selected tracts, relative to within-CZ average, across 50 largest CZs. An effect size of 0.01 corresponds to predicted household income rank in adulthood increasing by 1 percentile (for a child spending their entire childhood in a given tract). Panel (b) shows the median bias of different estimators across the 50 CZs. Panel (c) plots the median absolute error across the same CZs. Panel (d) shows coverage of confidence intervals across the 50 largest CZs, while panel (e) plots their median length.

longer in others.²³ Conditional intervals lie above zero in 20 of the 50 commuting zones, but include zero in the other 30. Hence, if we are satisfied with unconditional coverage we find

²³Interestingly, the hybrid interval for Seattle, the site of Bergman et al. (2020)’s experiment, is very short, and substantially shifted downwards relative to the naive and empirical Bayes intervals. This reflects that fact that the “next best” tracts in this case are extremely close to some of the included tracts. This leads to very strong downward correction by the conditional approach, and a hybrid interval concentrated near the lower bound of the projection interval CS_P^β .

strong evidence that selected tracts are better than average, while if we demand conditional coverage results are more mixed, and depend on which commuting zone we consider.

Precision aside, unconditional inference seems potentially more natural than conditional inference in this application. Within a given CZ, our primary interest is in the efficacy of targeting the top third of tracts based on estimated economic mobility, rather than in the precise combination of tracts selected. Hence, it is natural to focus on unconditional approaches, which give guarantees on average across potential selections. Moreover, in this application we consider multiple CZs, putting the focus even more squarely on average results from targeting tracts in this way, rather than the particular selection in a given CZ.

8 Conclusion

This paper considers a form of the winner’s curse that arises when we select a target parameter for inference based on optimization. We propose confidence intervals and quantile unbiased estimators for the target parameter that are optimal conditional on its selection. We hence recommend our conditional inference procedures when it is appropriate to remove uncertainty about the choice of target parameters from inferential statements. These conditionally valid procedures are also unconditionally valid, but we find that they sometimes have unappealing (unconditional) performance relative to existing alternatives. If one is satisfied with correct unconditional coverage and (in the case of estimation) a small, controlled degree of bias, we propose hybrid procedures which combine conditioning with projection confidence intervals.

Our results suggest a range of opportunities for future work. First, rather than considering inference on $\mu_Y(\hat{\theta})$, under suitable assumptions one could build on our results to forecast $Y(\hat{\theta})$. Alternatively, while conditional and projection confidence intervals have antecedents in the literature on inference after model selection, including in Berk et al. (2013) and Fithian et al. (2017), there is no analog of our hybrid approach in this literature. Our positive simulation results for the hybrid method suggest that this approach might yield appealing performance in a range of post-selection-inference settings. Even if a fully conditional approach is desired in the post-selection problem, as in Fithian et al. (2017), one could consider the analog of our optimal median-unbiased estimates that condition on the selected model. Finally, the problem of estimating the value of a dynamic treatment rule (c.f. Chakraborty and Murphy, 2014; Han, 2020) is closely related to our setting, so it seems likely that our results could prove useful there as well.

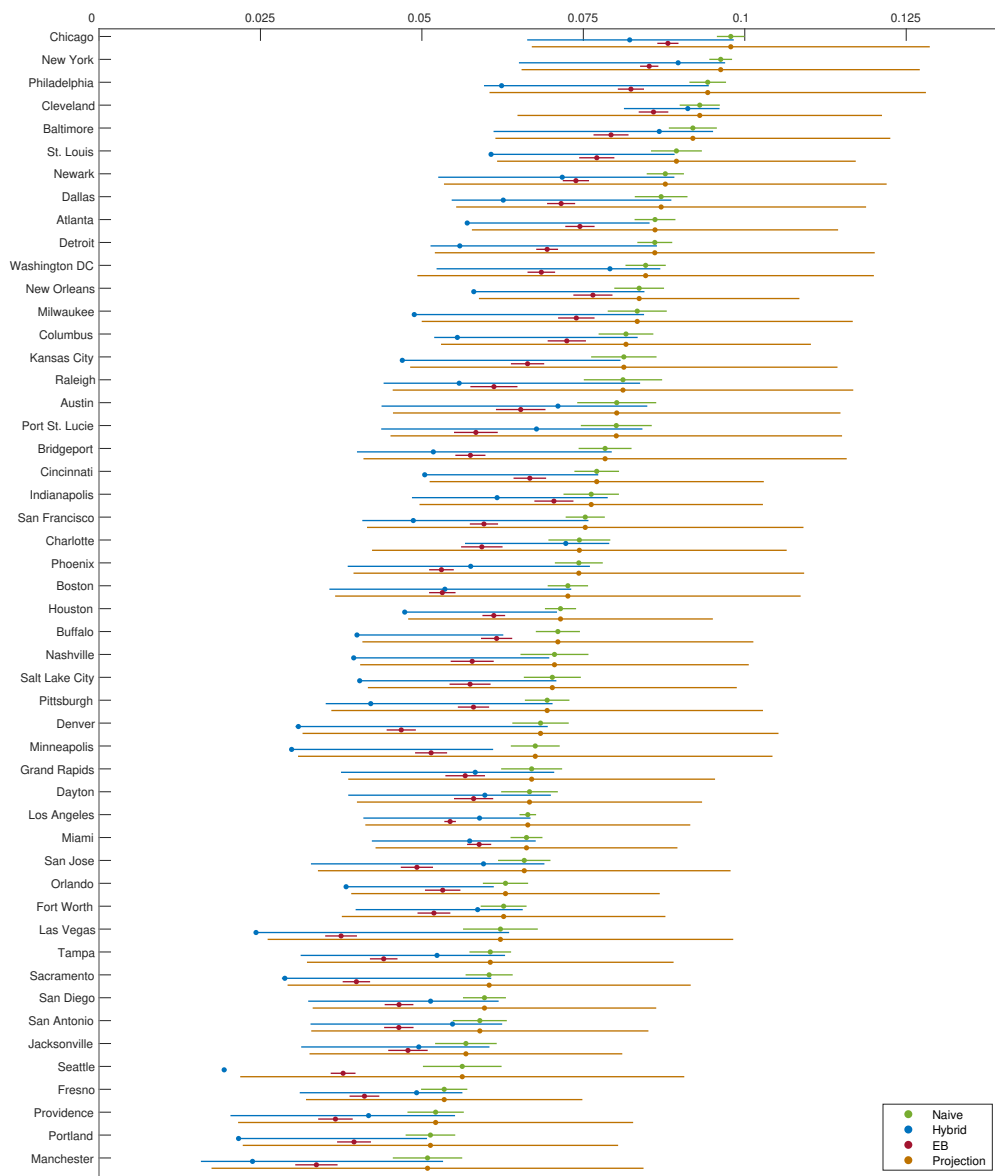


Figure 7: Estimates and confidence intervals for average economic mobility for selected census tracts based on Chetty et al. (2018) Opportunity Atlas, relative to the within-CZ average. CZs are ordered by the magnitude of the naive estimate.