Inference on Consensus Ranking of Distributions

David M. Kaplan*

October 18, 2020

Abstract

Instead of testing for unanimous agreement, I propose learning how broad of a consensus favors one distribution over another (of income, productivity, asset returns, test scores, etc.). Specifically, I propose statistical inference methods to learn about the set of utility functions for which one distribution has higher expected utility than another. With high probability, an "inner" confidence set is contained within this true set, while an "outer" confidence set contains the true set. Such confidence sets can be formed by inverting a proposed multiple testing procedure that controls the familywise error rate. Theoretical justification comes from empirical process results, given that very large classes of utility functions are generally Donsker (subject to finite moments). The theory additionally justifies a uniform (over utility functions) confidence band of expected utility differences, as well as tests with a utility-based "restricted stochastic dominance" as either the null or alternative hypothesis. Simulated and empirical examples illustrate the methodology.

JEL classification: C29

Keywords: confidence set, Donsker, expected utility, familywise error rate, multiple testing, stochastic dominance

1 Introduction

Different individuals with different preferences can disagree about which of two distributions is better, so economists have considered ways to capture consensus. Within the standard expected utility framework, variations of stochastic dominance characterize when certain groups of people all agree which distribution is better, i.e., when expected utility is higher for one distribution given any utility function in a particular set. For example, first-order stochastic dominance requires higher expected utility for any (non-decreasing) utility function, whereas second-order stochastic dominance relaxes the requirement to only concave (risk-averse) utility functions.

*kaplandm@missouri.edu. Thanks to Tim Armstrong for the initial idea, and to the Cowles Foundation more generally for their hospitality. I am also very grateful for many helpful questions, comments, and references from the Econometrics Workshop at The University of Chicago (specifically Stéphane Bonhomme, Jim Heckman, Azeem Shaikh, and Alex Torgovitsky).

This setup can be interpreted broadly. As usual, "individuals" can be replaced by other agents or organizations (or "utility function" replaced by social welfare function), and the "distributions" could be earnings, asset returns, productivity, scores, or other observable measures. Further, these could be conditional distributions, given partial information; for example, a firm could assess distributions only for firms with similar characteristics. The overall goal may be description (e.g., differences across countries in observational data) or causal inference and policy decision-making (e.g., comparing randomized treatment and control groups).

Statistical inference has focused on testing the null hypothesis of stochastic dominance, exploiting alternative characterizations in terms of cumulative distribution functions (CDFs). For example, first-order stochastic dominance is equivalent to one CDF lying below another at all points, which can be tested using statistical properties of empirical CDFs. Further, Barrett and Donald (2003) provide CDF-based bootstrap tests of the null hypothesis of any desired order of stochastic dominance.

Such methods are excellent for testing economic theories that imply stochastic dominance, but they do not provide strong evidence in favor of dominance (Davidson and Duclos, 2013, p. 87). Non-rejection of dominance may be a type II error, whose rate is not controlled. In contrast, false rejection of a non-dominance null hypothesis in favor of dominance would be a type I error, whose rate is controlled. However, with unbounded continuous distributions, the first-order stochastic dominance concept is too economically strong to ever reject non-dominance (Davidson and Duclos, 2013, p. 87). Consequently, Davidson and Duclos (2013) compare CDFs only over a specified interval.

For both economic and statistical reasons, I propose methods to learn about *the set of utility functions for which one distribution is preferred* to another. Economically, although the expected utility and CDF perspectives coincide in first-order stochastic dominance, their economic interpretations differ greatly when the consensus is not unanimous. The set of utility functions has a direct economic interpretation, showing which types of individuals would (not) prefer a particular distribution over another, and it can be examined in light of the literature's estimates of utility function parameters like risk aversion. The CDF perspective has a more particular economic interpretation in terms of headcount poverty (for income or consumption), as described by Atkinson (1987). Given a particular poverty line, the distribution whose CDF is lower at the poverty line has lower headcount poverty. Thus, interpreting the set of values at which one CDF is lower than the other, we have "the set of poverty lines for which one distribution is preferred" in terms of headcount poverty. This is essentially Atkinson's (1987) "restricted stochastic dominance" (Condition I, p. 751), in which one CDF lies below another over a range (set) of values. Despite this well-defined

economic interpretation, arguably expected utility is more broadly useful, especially for distributions besides income and consumption, such as productivity or test scores.

Statistically, this set of utility functions provides more information than a single all-or-nothing hypothesis test. Even if one distribution is not unanimously preferred, it is helpful to know whether the distribution is preferred over a large set of realistic utility functions. Like Davidson and Duclos (2013), compared to testing a null of dominance, my approach provides stronger evidence for a distribution being "better." Further, the object of interest does not need to be specified and then tested, but rather is learned inductively.

To quantify uncertainty about this set of utility functions for which one distribution is preferred, I propose "inner" and "outer" confidence sets. The inner confidence set is contained within the true set with high probability, while the outer confidence set contains the true set with high probability (similar to a confidence set for an identified set). The inner confidence set is thus conservative in the sense that it only contains utility functions from the true set with high probability, but it readily omits additional utility functions (from the true set) if there is too much uncertainty. The true set is "between" the inner and outer confidence sets with high probability.

These confidence sets can be constructed by inverting a multiple testing procedure. The inner confidence set contains all utility functions for which lower expected utility is rejected in favor of higher expected utility while controlling the familywise error rate. The probability of any false rejection is the probability of incorrectly including a utility function in the confidence set, so familywise error rate control implies correct coverage probability. The outer confidence set contains all utility functions for which higher expected utility is not rejected, using the same logic as Romano and Shaikh's (2010) confidence set for the identified set (e.g., see Lemma 2.1, pp. 172–173). To choose among the many possible valid multiple testing procedures (and thus confidence sets), I use the same pointwise asymptotic size at each point.

To justify the multiple testing procedures and confidence sets, I use empirical process theory for the sample expected utility difference, indexed by utility functions. Because utility functions are non-decreasing, this process generally has a Gaussian limit, subject to certain finite moments. Further, this limit is consistently estimated by exchangeable bootstrap.

These theoretical results in turn justify other statistical methods. Uniform (over utility functions) confidence bands for the expected utility difference can be constructed; these are more informative than the confidence sets, but more difficult to communicate and comprehend. I also describe hypothesis tests for having higher expected utility for all utility functions in a specified set; the null can be either dominance or non-dominance.

There is a high-level parallel is between my multiple testing (of the expected utility differ-

ence, over utility functions) and the multiple testing (of the CDF difference, over evaluation points) of Goldman and Kaplan (2018) and Kaplan (2019). Method 5 of Goldman and Kaplan (2018) is a multiple testing procedure for whether one CDF lies below another CDF at each possible point, i.e., it tests $H_{0r}\colon F_a(r) \leq F_b(r)$ for all r while controlling the (finite-sample) familywise error rate. Instead of multiple testing CDF inequalities point-by-point, I consider multiple testing expected utility inequalities function-by-function. Another interpretation in light of Atkinson (1987) is that Goldman and Kaplan (2018) try to learn about the set of poverty lines for which one distribution is "better" than another in terms of lower headcount poverty. Instead, I consider inference on the set of utility functions for which one distribution is "better" than another in terms of higher expected utility. The CDF approach has natural advantages when data are top-coded or when the tails suffer from measurement error.[1] The utility approach has advantages in economic interpretation and the potential to extend to utility functions of multiple variables.

My approach to learn about the consensus set can be applied to other distributional comparisons, too. For example, there are many proposed metrics to capture "inequality" within a distribution, but often they depend on a user-chosen parameter. Instead of picking one value (or several), we could learn about the set of parameter values for which one distribution is "better" (lower inequality), like such a set of ϵ in the Atkinson (1970) inequality index, or α in the Cowell and Flachaire (2017, eqn. (22) and §4.3) inequality index for ordinal variables. That is, we can try to learn about the set within which there is a consensus about one distribution being better. Other potential research questions are mentioned in the conclusion.

Paper structure Section 2 contains the core theoretical results. Section 3 describes multiple testing of negative/positive expected utility difference (over utility functions). Section 4 describes confidence sets for the set of utility functions with positive expected utility difference. Section 5 describes uniform confidence bands and hypothesis tests of utility restricted stochastic dominance. Sections 6 and 7 have empirical and simulation illustrations of the methodology and theory. Appendix A contains proofs not in the main text. Appendix B contains an example algorithm to compute bootstrap critical values.

Notation Following convention, let $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ denote the empirical process indexed by the set of functions $\mathcal{F}$, where $\mathbb{G}_n \equiv \sqrt{n}(\mathbb{P}_n - P)$ is the empirical process, P is the true population measure, $\mathbb{P}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the empirical measure (where δ is the Dirac delta function), and $Qf \equiv \int f \,\mathrm{d}Q$ for any measure Q, and let $\rightsquigarrow$ denote weak convergence. The two

[1] Thanks to Saku Aura for these points.

samples/populations are distinguished by superscript or subscript a or b: $P^a f = \mathrm{E}[f(Y^a)]$ for $Y^a \sim P^a$, $P^b f = \mathrm{E}[f(Y^b)]$ for $Y^b \sim P^b$, $\mathbb{P}_n^a f = n_a^{-1} \sum_{i=1}^{n_a} f(Y_i^a)$, $\mathbb{P}_n^b f = n_b^{-1} \sum_{i=1}^{n_b} f(Y_i^b)$, and

$$\mathbb{G}_n^a \equiv \sqrt{n_a}(\mathbb{P}_n^a - P^a), \quad \mathbb{G}_n^b \equiv \sqrt{n_b}(\mathbb{P}_n^b - P^b), \quad \mathbb{G}_n^\Delta \equiv \sqrt{n_a}[(\mathbb{P}_n^a - \mathbb{P}_n^b) - (P^a - P^b)]. \tag{1}$$

The statement $n \to \infty$ abbreviates $n_a, n_b \to \infty$. Let $\ell^\infty(\mathcal{F})$ denote the space of all bounded functions $\mathcal{F} \mapsto \mathbb{R}$, equipped with norm $\|g\|_\mathcal{F} = \sup_{f \in \mathcal{F}} |g(f)|$. Set difference $\setminus$ means $\mathcal{A} \setminus \mathcal{B} \equiv \{a : a \in \mathcal{A}, a \notin \mathcal{B}\}$. Acronyms used include those for confidence set (CS), constant relative risk aversion (CRRA), coverage probability (CP), cumulative distribution function (CDF), data-generating process (DGP), empirical CDF (ECDF), familywise error rate (FWER), multiple testing procedure (MTP), stochastic dominance (SD), and Vapnik–Chervonenkis (VC).

2 Theory

Section 2.1 concerns which classes of utility functions are Donsker. Given a Donsker class of utility functions, Section 2.2 establishes the Gaussian limit of the expected utility difference process, as well as bootstrap consistency.

2.1 Donsker classes of utility functions

This section considers the key technical condition for applying empirical process theory to expected utility: whether the specified set of utility functions is Donsker. Section 2.1.1 shows this holds very generally. Section 2.1.2 establishes the Donsker property of the popular but simple class of constant relative risk aversion (CRRA) utility functions under a weaker envelope function condition.

2.1.1 A large nonparametric class of utility functions

Corollary 3.1 of van der Vaart (1996) establishes that the class of essentially all utility functions is Donsker, subject to a common lower (or upper) bound and an envelope function with $2 + \delta$ moments. Below, the lower bound is stated as zero without loss of generality, because adding a constant to a utility functions does not affect the expected utility difference when comparing two distributions. That is, if utility function $f(x) = \tilde{f}(x) + c$ for constant c, and Y^a and Y^b are random variables, then

$$\mathrm{E}[f(Y^a)] - \mathrm{E}[f(Y^b)] = \mathrm{E}[\tilde{f}(Y^a) + c] - \mathrm{E}[\tilde{f}(Y^b) + c] = \mathrm{E}[\tilde{f}(Y^a)] - \mathrm{E}[\tilde{f}(Y^b)].$$

Lemma 1 (van der Vaart (1996), Cor. 3.1)**.** *The set $\mathcal{F}$ of non-decreasing utility functions $f\colon \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is P-Donsker if $PF^{2+\delta} < \infty$ for some $\delta > 0$, where F is an envelope function satisfying $0 \leq f \leq F$ for all $f \in \mathcal{F}$.*

Proof. This is Corollary 3.1 of van der Vaart (1996). □

2.1.2 A parametric utility function class

Consider the simple but popular CRRA utility

$$f_\theta(x) = \begin{cases} \dfrac{x^{1-\theta} - 1}{1-\theta} & \text{if } \theta \geq 0 \text{ and } \theta \neq 1, \\ \ln(x) & \text{if } \theta = 1. \end{cases} \tag{2}$$

The parameter θ represents risk aversion. With $\theta = 0$, $f_0(x) = x - 1$, which is risk-neutral. With $\theta > 0$, $f_\theta(x)$ is concave in x, which is risk-averse.

Lemma 2 states that the VC dimension is one. For any θ, $f_\theta(1) = 0$. For any other $x \neq 1$, $f_\theta(x)$ is decreasing in θ. Thus, the subgraphs are nested, so the VC dimension is 1. The Donsker property follows if the envelope function is square integrable, or if the somewhat weaker condition in Theorem 2.6.8 of van der Vaart and Wellner (1996) holds. Depending on the distribution (P), the envelope function condition may require bounding θ. Whether such a bound is economically restrictive depends on the value of the bound and the economic setting.

Lemma 2. *Given* (2)*, let $\mathcal{F} = \{f_\theta : 0 \leq \theta < \bar{\theta}\}$, allowing $\bar{\theta} = \infty$. The VC dimension of $\mathcal{F}$ is one, and it is P-Donsker if $PF^2 < \infty$, where F is the envelope function.*

2.2 Asymptotic distributions

This section establishes the core theoretical results that justify the methods proposed later. Section 2.2.1 provides the Gaussian limit of the expected utility difference process (indexed by utility functions). Section 2.2.2 extends this result to suprema. Section 2.3 establishes exchangeable bootstrap consistency.

2.2.1 Limiting Gaussian process for expected utility difference

Theorem 3 states the Gaussian limit of the expected utility empirical process (indexed by utility functions), given the utility function class is Donsker and given Assumption A1. Assumption A1 states there are two iid samples, independent of each other, whose sample

sizes are of the same magnitude; it is similar to Assumption 2 (page 75) in Barrett and Donald (2003), among others.

Assumption A1. Random variables $Y_1^a, Y_2^a, \ldots$ are iid realizations of random variable Y^a with distribution P^a. Independent of the Y_i^a are $Y_1^b, Y_2^b, \ldots$, iid realizations of Y^b with distribution P^b. The respective data samples are Y_i^a for $i = 1, \ldots, n_a$ and Y_j^b for $j = 1, \ldots, n_b$, where $n_a/n_b = \lambda + o(1)$ as $n_a, n_b \to \infty$.

Theorem 3. *Let $\mathcal{F}$ be a P^a-Donsker and P^b-Donsker class of utility functions. Let Assumption A1 hold. Then, $\mathbb{G}_n^a \rightsquigarrow \mathbb{G}^a$ and $\mathbb{G}_n^b \rightsquigarrow \mathbb{G}^b$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}^a$ and $\mathbb{G}^b$ are respectively P^a- and P^b-Brownian bridges, i.e., mean-zero Gaussian processes with covariance functions $\mathrm{E}[\mathbb{G}^a f \mathbb{G}^a g] = P^a fg - P^a f P^a g$ and $\mathrm{E}[\mathbb{G}^b f \mathbb{G}^b g] = P^b fg - P^b f P^b g$, respectively, and they are independent ($\mathbb{G}^a \perp\!\!\!\perp \mathbb{G}^b$). Further, the convergence is joint: $(\mathbb{G}_n^a, \mathbb{G}_n^b) \rightsquigarrow (\mathbb{G}^a, \mathbb{G}^b)$.*

Proof. Because $\mathcal{F}$ is P^a-Donsker and P^b-Donsker, the marginal results are immediate; e.g., see (2.1.1) and (2.1.2) in van der Vaart and Wellner (1996). There are only two processes, so the joint convergence follows from Theorem 1.4.8 of van der Vaart and Wellner (1996). □

Corollary 4 extends Theorem 3 to expected utility differences.

Corollary 4. *Under the conditions of Theorem 3, $\mathbb{G}_n^\Delta \rightsquigarrow \mathbb{G}^\Delta$ in $\ell^\infty(\mathcal{F})$, where $\mathbb{G}^\Delta$ is a mean-zero Gaussian process with covariance function*

$$\mathrm{E}[\mathbb{G}^\Delta f \mathbb{G}^\Delta g] = P^a fg - P^a f P^a g + \lambda^2 [P^b fg - P^b f P^b g].$$

Proof. Rearranging with algebra and applying Theorem 3,

$$\overbrace{\sqrt{n_a}[(\mathbb{P}_n^a - \mathbb{P}_n^b) - (P^a - P^b)]}^{\mathbb{G}_n^\Delta} = \overbrace{\sqrt{n_a}(\mathbb{P}_n^a - P^a)}^{=\mathbb{G}_n^a \rightsquigarrow \mathbb{G}^a} - \overbrace{\sqrt{n_a/n_b}}^{\to\lambda} \overbrace{\sqrt{n_b}(\mathbb{P}_n^b - P^b)}^{=\mathbb{G}_n^b \rightsquigarrow \mathbb{G}^b} \rightsquigarrow \mathbb{G}^\Delta \equiv \mathbb{G}^a - \lambda \mathbb{G}^b,$$

where the Δ superscript stands for "difference." Both $\mathbb{G}^a$ and $\mathbb{G}^b$ are mean-zero, so $\mathbb{G}^\Delta$ is, too. Since additionally $\mathbb{G}^a \perp\!\!\!\perp \mathbb{G}^b$, the covariance function is

$$\begin{aligned}
\mathrm{E}[\mathbb{G}^\Delta f \mathbb{G}^\Delta g] &= \mathrm{E}[(\mathbb{G}^a - \lambda \mathbb{G}^b) f (\mathbb{G}^a - \lambda \mathbb{G}^b) g] \\
&= \mathrm{E}[\mathbb{G}^a f \mathbb{G}^a g] + \lambda^2 \mathrm{E}[\mathbb{G}^b f \mathbb{G}^b g] - \lambda[\mathrm{E}(\mathbb{G}^a f \mathbb{G}^b g) + \mathrm{E}(\mathbb{G}^b f \mathbb{G}^a g)] \\
&= P^a fg - P^a f P^a g + \lambda^2 [P^b fg - P^b f P^b g] \\
&\quad - \lambda[\overbrace{\mathrm{E}(\mathbb{G}^a f)}^{=0} \overbrace{\mathrm{E}(\mathbb{G}^b g)}^{=0} + \overbrace{\mathrm{E}(\mathbb{G}^b f)}^{=0} \overbrace{\mathrm{E}(\mathbb{G}^a g)}^{=0}].
\end{aligned}$$

□

2.2.2 Distributions of suprema

The distributions of maximal t-statistics are useful for inference. Denote the pointwise (scalar) variance of $\mathbb{G}^\Delta f$ as

$$\sigma_f^2 \equiv \mathrm{Var}(\mathbb{G}^\Delta f) = P^a f^2 - (P^a f)^2 + \lambda^2[P^b f^2 - (P^b f)^2], \tag{3}$$

with $\hat{\sigma}_f$ an estimator like (8). For $\mathcal{S} \subseteq \mathcal{F}$, using notation from (1), define

$$T_f \equiv \mathbb{G}^\Delta f/\sigma_f, \quad T^{\mathcal{S}\vee} \equiv \sup_{f\in\mathcal{S}} T_f, \quad T^{\mathcal{S}\wedge} \equiv \inf_{f\in\mathcal{S}} T_f, \quad |T|^{\mathcal{S}\vee} \equiv \sup_{f\in\mathcal{S}}|T_f|, \tag{4}$$

$$\hat{T}_f \equiv \mathbb{G}_n^\Delta f/\hat{\sigma}_f, \quad \hat{T}^{\mathcal{S}\vee} \equiv \sup_{f\in\mathcal{S}} \hat{T}_f, \quad \hat{T}^{\mathcal{S}\wedge} \equiv \inf_{f\in\mathcal{S}} \hat{T}_f, \quad |\hat{T}|^{\mathcal{S}\vee} \equiv \sup_{f\in\mathcal{S}}|\hat{T}_f|. \tag{5}$$

The corresponding limit distributions are given in Corollary 5, under Assumptions A2 and A3. Assumption A3 is satisfied by the bootstrap estimator in (8), among other possibilities.

Assumption A2. The pointwise variances σ_f^2 from (3) are uniformly (over $f \in \mathcal{F}$) bounded away from zero.

Assumption A3. The estimators $\hat{\sigma}_f$ are uniformly consistent: $\sup_{f\in\mathcal{F}}|\hat{\sigma}_f - \sigma_f| \xrightarrow{p} 0$.

Corollary 5. *Let the conditions for Corollary 4 hold along with Assumptions A2 and A3. Then,*

$$\hat{T}_f \xrightarrow{d} T_f \sim \mathrm{N}(0,1) \text{ for each } f \in \mathcal{F}, \quad \hat{T}^{\mathcal{S}\vee} \xrightarrow{d} T^{\mathcal{S}\vee}, \quad \hat{T}^{\mathcal{S}\wedge} \xrightarrow{d} T^{\mathcal{S}\wedge}, \quad |\hat{T}|^{\mathcal{S}\vee} \xrightarrow{d} |T|^{\mathcal{S}\vee}.$$

Proof. This follows from $\hat{T}_f \equiv \mathbb{G}_n^\Delta f/\hat{\sigma}_f = \mathbb{G}_n^\Delta f/\sigma_f + o_p(1)$ uniformly over $f \in \mathcal{F}$ (by A2 and A3) and the continuous mapping theorem applied to Corollary 4. □

2.3 Exchangeable bootstrap consistency

The exchangeable bootstrap consistently estimates the limiting Gaussian process $\mathbb{G}^\Delta$ from Corollary 4. This provides an alternative to simulating from a version of $\mathbb{G}^\Delta$ with explicitly estimated covariance function. Special cases of exchangeable bootstrap include the empirical bootstrap (multinomial weights), Bayesian bootstrap, m-out-of-n bootstrap, and subsampling. See Theorem 3.6.13 and equation (3.6.8) of van der Vaart and Wellner (1996, p. 354–355) for conditions on measurability and general weights. The conditions on the weights are stated in A4.

Assumption A4. For $n = n_a$ or $n = n_b$, let $(\tilde{W}_{n1}, \ldots, \tilde{W}_{nn})$ denote the exchangeable random vector of nonnegative weights, independent of the data. Denote the average $\bar{W}_n \equiv n^{-1} \sum_{i=1}^n \tilde{W}_{ni}$. Further, as in (3.6.8) of van der Vaart and Wellner (1996, p. 354),

$$\sup_n \|\tilde{W}_{n1} - \bar{W}_n\|_{2,1} \equiv \sup_n \int_0^\infty \sqrt{\mathrm{P}(|\tilde{W}_{n1} - \bar{W}_n| > x)}\, \mathrm{d}x < \infty,$$
$$n^{-1/2}\, \mathrm{E} \max_{1 \le i \le n} |\tilde{W}_{ni} - \bar{W}_n| \xrightarrow{p} 0, \quad n^{-1} \sum_{i=1}^n (\tilde{W}_{ni} - \bar{W}_n)^2 \xrightarrow{p} c^2 > 0.$$

Notationally, denote the bootstrap empirical process (difference) adjusted for c as

$$\begin{aligned} \tilde{\mathbb{G}}_n^\Delta &\equiv \sqrt{n_a}[(\tilde{\mathbb{P}}_n^a - \tilde{\mathbb{P}}_n^b) - (\bar{W}^a \mathbb{P}_n^a - \bar{W}^b \mathbb{P}_n^b)]/c, \\ \tilde{\mathbb{P}}_n^a &\equiv \frac{1}{n_a} \sum_{i=1}^{n_a} \tilde{W}_i^a \delta_{Y_i^a}, \quad \tilde{\mathbb{P}}_n^b \equiv \frac{1}{n_b} \sum_{i=1}^{n_b} \tilde{W}_i^b \delta_{Y_i^b}, \end{aligned} \tag{6}$$

where $(\tilde{W}_1^a, \ldots, \tilde{W}_{n_a}^a)$ and $(\tilde{W}_1^b, \ldots, \tilde{W}_{n_b}^b)$ are the exchangeable random vectors of bootstrap weights with respective averages $\bar{W}^a$ and $\bar{W}^b$. Analogous to (4) and (5), let

$$\tilde{T}_f \equiv \tilde{\mathbb{G}}_n^\Delta f / \hat{\sigma}_f, \quad \tilde{T}^{\mathcal{S}\vee} \equiv \sup_{f \in \mathcal{S}} \tilde{T}_f, \quad \tilde{T}^{\mathcal{S}\wedge} \equiv \inf_{f \in \mathcal{S}} \tilde{T}_f, \quad |\tilde{T}|^{\mathcal{S}\vee} \equiv \sup_{f \in \mathcal{S}} |\tilde{T}_f|. \tag{7}$$

A bootstrap estimator of σ_f suggested by Chernozhukov, Fernández-Val, and Melly (2013, p. 2222–2223) is the scaled interquartile range. Letting z_α denote the α-quantile of the standard normal distribution and letting $(\tilde{\mathbb{G}}_n^\Delta f)_\alpha$ denote the α-quantile of $\tilde{\mathbb{G}}_n^\Delta f$,

$$\hat{\sigma}_f = [(\tilde{\mathbb{G}}_n^\Delta f)_{0.75} - (\tilde{\mathbb{G}}_n^\Delta f)_{0.25}]/(z_{0.75} - z_{0.25}). \tag{8}$$

Theorem 6 relies on Theorem 3.6.13 of van der Vaart and Wellner (1996, p. 355).

Theorem 6. *Let the assumptions of Corollary 5 hold, as well as Assumption A4 with the same c for both samples. Also, as in Theorem 3.6.13 of van der Vaart and Wellner (1996, p. 355), assume measurability of the functions in $\mathcal{F}$ such that $\mathcal{F}_\delta$ is measurable for each $\delta > 0$, where $\mathcal{F}_\delta \equiv \{f - g : f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$ as in van der Vaart and Wellner (1996, p. 350) for both $P = P^a$ and $P = P^b$, with $\rho_P(f) \equiv \sqrt{P(f - Pf)^2}$ the variance seminorm as on page 89 of van der Vaart and Wellner (1996). Then, conditional on almost all sequences of data, $\tilde{\mathbb{G}}_n^\Delta \rightsquigarrow \mathbb{G}^\Delta$ in $\ell^\infty(\mathcal{F})$. Further,*

$$\tilde{T}^{\mathcal{S}\vee} \xrightarrow{d} T^{\mathcal{S}\vee}, \quad \tilde{T}^{\mathcal{S}\wedge} \xrightarrow{d} T^{\mathcal{S}\wedge}, \quad |\tilde{T}|^{\mathcal{S}\vee} \xrightarrow{d} |T|^{\mathcal{S}\vee}.$$

Consequently, the bootstrap critical values (quantiles) are consistent: letting subscript $1 - \alpha$

denote the $(1-\alpha)$*-quantile of a random variable,*

$$\tilde{T}_{1-\alpha}^{\mathcal{S}\vee} \xrightarrow{p} T_{1-\alpha}^{\mathcal{S}\vee}, \quad \tilde{T}_{1-\alpha}^{\mathcal{S}\wedge} \xrightarrow{p} T_{1-\alpha}^{\mathcal{S}\wedge}, \quad |\tilde{T}|_{1-\alpha}^{\mathcal{S}\vee} \xrightarrow{p} |T|_{1-\alpha}^{\mathcal{S}\vee}.$$

Proof. For $\tilde{\mathbb{G}}_n^{\Delta}$, the result is from Theorem 3.6.13 of van der Vaart and Wellner (1996, p. 355), with the adjustment for c in (6) above instead of in the limiting process.

For the suprema, the results hold by the continuous mapping theorem.

For the critical values (quantiles), consistency follows because the limit distributions are all continuous. □

Appendix B contains an example algorithm for computing bootstrap critical values.

3 Multiple testing

The results in Section 2 help justify a multiple testing procedure (MTP). The same notation continues: $\mathcal{F}$ is a large set of utility functions, and for element $f \in \mathcal{F}$, $P^a f \equiv \mathrm{E}[f(Y^a)]$ and $P^b f \equiv \mathrm{E}[f(Y^b)]$, so an individual with utility f prefers Y^a if $P^a f - P^b f > 0$. Define

$$H_{0f}\colon P^a f - P^b f \leq 0, \quad H_{1f}\colon P^a f - P^b f > 0, \quad \text{for each } f \in \mathcal{F}. \tag{9}$$

That is, given utility function f, the null hypothesis H_{0f} is true if expected utility is higher for Y^b than Y^a, and the null is false (and the alternative H_{1f} true) if expected utility is higher for Y^a. Define the set of utility functions for which Y^b is preferred (so the null is true) as

$$\mathcal{F}_T \equiv \{f : P^a f - P^b f \leq 0\} = \{f : H_{0f} \text{ is true}\}. \tag{10}$$

As in Lehmann and Romano (2005, §9.1), the familywise error rate (FWER) is

$$\text{FWER} \equiv \mathrm{P}(\text{reject any } H_{0f} \text{ with } f \in \mathcal{F}_T), \tag{11}$$

and strong control of FWER at level α means FWER $\leq \alpha$ given any $\mathcal{F}_T$ (as opposed to "weak control" for only $\mathcal{F}_T = \mathcal{F}$).

Section 3.1 describes a basic procedure for testing each H_{0f} in (9) with strong control of FWER. Section 3.2 discusses how to apply refinements from the multiple testing literature.

3.1 Basic multiple testing procedure

Proposition 7 shows the strong FWER control of the MTP in Method 1.

Method 1. *Reject H_{0f} when $\hat{T}_f > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$, with notation from* (5) *and Theorem 6 and critical value computed by Method 7.*

Proposition 7. *Given Proposition 9, Method 1 has strong control of FWER at level α.*

3.2 Power refinements

3.2.1 Stepdown procedure

A stepdown procedure in the spirit of Holm (1979) (see also Lehmann and Romano, 2005, Ch. 9) can improve the basic MTP's power while maintaining strong control of FWER. Although stepdown procedures are not uniformly better because in some cases FWER increases (but never above α), they are often preferred.

Method 2 describes a stepdown procedure whose asymptotic strong control of FWER is given in Proposition 8. The general argument is the same as usual: if any initially-rejected H_{0f} is true, then a familywise error is already committed, so further false rejections do not affect FWER; and if all initially-rejected H_{0f} are false, then the critical value can be appropriately adjusted to reflect $\mathcal{F}_T$ being smaller than initially thought (because type I errors are only possible for H_{0f} with $f \in \mathcal{F}_T$, and only the number of true H_{0f} determines the necessary multiple testing adjustment).

Method 2 (stepdown procedure)**.** *The stepdown MTP proceeds as follows.*

1. *Run Method 1: compute $\hat{T}_f$ as in* (5) *for all f, and reject all H_{0f} for which $\hat{T}_f > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$. Let $\hat{K}^{(0)} = \mathcal{F}$. Let iteration counter $i = 1$.*
2. *Let $\hat{K}^{(i)} = \{f : H_{0f}$ not yet rejected$\}$. If $\hat{K}^{(i)} = \emptyset$ or $\hat{K}^{(i)} = \hat{K}^{(i-1)}$, then stop.*
3. *Using the same bootstrap draws from Step 1, use Method 7 with $\mathcal{S} = \hat{K}^{(i)}$ to compute bootstrap critical value $\tilde{T}_{1-\alpha}^{\hat{K}^{(i)}\vee}$.*
4. *Reject any additional H_{0f} for which $\hat{T}_f > \tilde{T}_{1-\alpha}^{\hat{K}^{(i)}\vee}$.*
5. *Increment i by one and return to Step 2.*

Proposition 8. *Method 2 tests* (9) *with strong control of asymptotic FWER at level α.*

3.2.2 Other multiple testing refinements and approaches

Besides the stepdown procedure, a pre-test could be used. As usual, the objective is to identify H_{0f} that are not close to binding and can be removed from consideration, which allows a smaller critical value and thus better power. Here, $H_{0f}\colon P^a f - P^b f \leq 0$, so removing f where $P^a f - P^b f \ll 0$ can improve power; e.g., one could remove all H_{0f} for which $f \notin \hat{\mathcal{D}}_2$,

where $\hat{\mathcal{D}}_2$ is a confidence set from Proposition 9 but with confidence level $1-\tilde{\alpha}$ for small $\tilde{\alpha}$. A pre-test could be further combined with the stepdown procedure in Method 2, starting with $\hat{K} = \mathcal{F} \setminus \hat{\mathcal{D}}_2$ in the first iteration.

Besides FWER, other error rates could be used. For example, k-FWER is the probability of at least k familywise errors; the FWER used above implicitly has $k = 1$. The false discovery rate or other measures could also be used, as well as Bayesian approaches. However, only FWER-controlling MTPs can be inverted into confidence sets as in Section 4.

4 Confidence sets

An MTP from Section 3 can be inverted into a confidence set (CS).[2] Define the subset $\mathcal{D}$ of utility functions preferring Y^a over Y^b as

$$\mathcal{D} \equiv \{f \in \mathcal{F} : P^a f > P^b f\}. \tag{12}$$

This $\mathcal{D}$ captures how big (or small) a consensus prefers Y^a.

Method 3 describes "inner" and "outer" CSs for $\mathcal{D}$. Asymptotically, the inner CS $\hat{\mathcal{D}}_1$ is contained inside $\mathcal{D}$ with at least $1-\alpha$ probability, whereas the outer CS $\hat{\mathcal{D}}_2$ contains $\mathcal{D}$ with at least $1-\alpha$ probability. Mathematically,

$$\lim_{n\to\infty} \mathrm{P}(\hat{\mathcal{D}}_1 \subseteq \mathcal{D}) \geq 1-\alpha, \quad \lim_{n\to\infty} \mathrm{P}(\hat{\mathcal{D}}_2 \supseteq \mathcal{D}) \geq 1-\alpha. \tag{13}$$

"Two-sided" CS pairs are described in terms of a uniform band (Section 5.1), satisfying

$$\lim_{n\to\infty} \mathrm{P}(\hat{\mathcal{D}}_1 \subseteq \mathcal{D} \subseteq \hat{\mathcal{D}}_2) \geq 1-\alpha. \tag{14}$$

Method 3. *Inner CS: run an MTP of $H_{0f}\colon P^a f - P^b f \leq 0$ over $f \in \mathcal{F}$ with strong control of FWER at level α (like Methods 1 and 2); then, $\hat{\mathcal{D}}_1 = \{f \in \mathcal{F} : \text{MTP rejects } H_{0f}\}$. Outer CS: run an MTP of $H_{0f}\colon P^a f - P^b f \geq 0$ and let $\hat{\mathcal{D}}_2 = \{f \in \mathcal{F} : \text{MTP does not reject } H_{0f}\}$. Joint/two-sided CS pair: compute two-sided uniform confidence band $\hat{b}_1(\cdot)$ and $\hat{b}_2(\cdot)$ using Method 4, then let $\hat{\mathcal{D}}_1 = \{f \in \mathcal{F} : \hat{b}_1(f) > 0\}$ and $\hat{\mathcal{D}}_2 = \{f \in \mathcal{F} : \hat{b}_2(f) > 0\}$.*

Proposition 9. *Given* (12) *and Propositions 7 and 8, the inner and outer CSs in Method 3 satisfy* (13)*. Given* (12) *and Proposition 10, the joint CS pair in Method 3 satisfies* (14)*.*

[2] Thanks to Tim Armstrong for pointing this out.

5 Other methods

Section 5.1 describes uniform (over utility functions) confidence bands for the expected utility difference. Section 5.2 describes hypothesis tests for utility restricted stochastic dominance, with either dominance or non-dominance as the null.

5.1 Uniform confidence bands

This section considers uniform (over $f \in \mathcal{F}$) confidence bands for the expected utility difference $P^a f - P^b f$. Let $\hat{b}_1 \colon \mathcal{F} \mapsto \mathbb{R}$ and $\hat{b}_2 \colon \mathcal{F} \mapsto \mathbb{R}$ denote the lower and upper confidence band functions. The goal is uniform $1-\alpha$ asymptotic coverage:

$$1 - \alpha = \lim_{n\to\infty} \mathrm{P}\{\hat{b}_1(f) \leq P^a f - P^b f \leq \hat{b}_2(f) \text{ for all } f \in \mathcal{F}\}. \tag{15}$$

Such bands contain the most information among all inference methods proposed in this paper, but they are also the most difficult to comprehend (and thus communicate). Methods 1, 3, and 5 can all be characterized as summaries of the uniform confidence band. Such summaries improve comprehension at the cost of information loss. At one extreme of the information–comprehension spectrum, the uniform band consists of a range of values for each $f \in \mathcal{F}$; at the other extreme, the tests in Section 5.2 consist of only a binary decision (reject or not). The multiple testing procedure and confidence sets provide a balance, improving comprehension while retaining information across $f \in \mathcal{F}$.

Method 4 constructs uniform bands whose asymptotic coverage is in Proposition 10.

Method 4 (uniform confidence bands)**.** *First, run Method 7, with the value of α needed below and $\mathcal{S} = \mathcal{F}$. (As in Method 7, in practice usually $\mathcal{S}$ is replaced by a fine grid.) For a symmetric two-sided uniform $1-\alpha$ confidence band,*

$$\hat{b}_1(f) = (\mathbb{P}_n^a - \mathbb{P}_n^b)f - |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}, \quad \hat{b}_2(f) = (\mathbb{P}_n^a - \mathbb{P}_n^b)f + |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}, . \tag{16}$$

For a one-sided uniform $1-\alpha$ confidence band, either

$$\begin{aligned} &\hat{b}_1(f) = (\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}, \quad \hat{b}_2(f) = \infty, \quad \textit{or} \\ &\hat{b}_2(f) = (\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{\alpha}^{\mathcal{F}\wedge}\hat{\sigma}_f/\sqrt{n_a}, \quad \hat{b}_1(f) = -\infty. \end{aligned} \tag{17}$$

For an "equal-tailed" two-sided band (that may be conservative),

$$\hat{b}_1(f) = (\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{1-\alpha/2}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}, \quad \hat{b}_2(f) = (\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{\alpha/2}^{\mathcal{F}\wedge}\hat{\sigma}_f/\sqrt{n_a}, . \tag{18}$$

Proposition 10. *Given Theorem 6, the bands in* (16) *and* (17) *in Method 4 have exact*

asymptotic uniform coverage probability (i.e., satisfy (15)*), and the band in* (18) *has asymptotic uniform coverage probability of at least* $1-\alpha$.

5.2 Testing utility restricted dominance and non-dominance

Define pointwise t-statistics centered at zero (instead of at $P^a f - P^b f$ as with $\hat{T}_f$ in (5)),

$$\hat{T}_f^0 \equiv \sqrt{n_a}(\mathbb{P}_n^a - \mathbb{P}_n^b)f/\hat{\sigma}_f. \tag{19}$$

These are the t-statistics for testing $H_{0f}\colon P^a f - P^b f = 0$ (or ≤ 0 or ≥ 0).

Section 5.2.1 considers testing the null hypothesis of utility restricted stochastic dominance, whereas Section 5.2.2 takes dominance as the alternative hypothesis (and thus non-dominance as the null). As noted by Davidson and Duclos (2013, pp. 88–89), strong and weak dominance cannot be distinguished statistically, so in this paper the choice of strict or weak inequality is made to optimize intuition or notational convenience.

5.2.1 Testing the null of utility restricted stochastic dominance

Define the null hypothesis of stochastic dominance restricted to $\mathcal{F}$ (and the alternative of non-dominance) as

$$\begin{aligned} &H_0\colon Y^b \,\mathrm{SD}_{\mathcal{F}}\, Y^a \iff P^a f - P^b f \leq 0 \text{ for all } f \in \mathcal{F}, \\ &H_1\colon Y^b \,\textit{non}\mathrm{SD}_{\mathcal{F}}\, Y^a \iff P^a f - P^b f > 0 \text{ for some } f \in \mathcal{F}. \end{aligned} \tag{20}$$

Method 5 describes a test whose asymptotic size control is given in Proposition 11.

Method 5 (test of $\mathrm{SD}_{\mathcal{F}}$)**.** *Compute* $\hat{T}_f^0$ *from* (19) *for each* $f \in \mathcal{F}$*. In the notation of* (7) *and* (20)*, reject* $H_0\colon Y^b \,\mathrm{SD}_{\mathcal{F}}\, Y^a$ *in favor of* $H_1\colon Y^b \,\textit{non}\mathrm{SD}_{\mathcal{F}}\, Y^a$ *when* $\sup_{f\in\mathcal{F}} \hat{T}_f^0 > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$*. Equivalently: construct a one-sided* $1-\alpha$ *uniform confidence function* $\hat{b}_1(\cdot)$ *as in Method 4 and reject when* $\hat{b}_1(f) > 0$ *for some* $f \in \mathcal{F}$*.*

Proposition 11. *Given Corollary 5 and Theorem 6, Method 5 has asymptotic size* α*.*

5.2.2 Testing the null of non-dominance

Now define the null and alternative hypotheses as

$$\begin{aligned} &H_0\colon Y^a \,\textit{non}\mathrm{SD}_{\mathcal{F}}\, Y^b \iff P^a f - P^b f \leq 0 \text{ for some } f \in \mathcal{F}, \\ &H_1\colon P^a f - P^b f > 0 \text{ for all } f \in \mathcal{F}. \end{aligned} \tag{21}$$

Testing (21) using the band in Method 4 is valid but conservative. If $\hat{b}_1(f) > 0$ for all $f \in \mathcal{F}$, then $non\mathrm{SD}_{\mathcal{F}}$ can be rejected at level α in favor of dominance. The band excludes the true $P^a f - P^b f$ with probability α, so given $P^a f - P^b f < 0$ for some f, there is only α probability of rejecting, i.e., of a band with $\hat{b}_1 \geq 0$. The condition $\hat{b}_1(f) > 0$ for all $f \in \mathcal{F}$ is equivalent to $\inf_{f\in\mathcal{F}}(\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a} > 0$, which is equivalent to $\inf_{f\in\mathcal{F}}\sqrt{n_a}(\mathbb{P}_n^a - \mathbb{P}_n^b)f/\hat{\sigma}_f > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$, or $\inf_{f\in\mathcal{F}}\hat{T}_f^0 > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$.

Power can be improved by decreasing the critical value $\tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$ to the standard normal $z_{1-\alpha}$, while still controlling asymptotic size. The idea is similar to Theorem 2.2 (pages 853–854) of Kaur, Prakasa Rao, and Singh (1994) for testing second-order restricted non-dominance over a range of CDF evaluation points (instead of utility functions).

Method 6 describes the test, whose asymptotic size control is given in Proposition 12.

Method 6 (test of $non\mathrm{SD}_{\mathcal{F}}$)**.** *Compute* $\hat{T}_f^0$ *from* (19)*, for* $f \in \mathcal{F}$*. In the notation of* (21)*, reject* $H_0\colon Y^a$ $non\mathrm{SD}_{\mathcal{F}}$ Y^b *in favor of* $H_1\colon Y^a$ $\mathrm{SD}_{\mathcal{F}}$ Y^b *when* $\inf_{f\in\mathcal{F}}\hat{T}_f^0 > z_{1-\alpha}$*, where* $z_{1-\alpha}$ *is the* $(1-\alpha)$*-quantile of the* $\mathrm{N}(0,1)$ *distribution.*

Proposition 12. *Given Corollary 5, the asymptotic size of the test in Method 6 is* α*.*

6 Empirical

6.1 Setting and methods

The following empirical example illustrates the new methodology, specifically the inner confidence set (CS) in Method 3. Two earnings distributions are considered. The object of interest is the set of utility functions for which the first distribution has higher expected utility, i.e., $\mathcal{D}$ in (12). The inner CS reports a set of utility functions $\hat{\mathcal{D}}_1$ that are a subset of the true set with high probability, i.e., $\mathrm{P}(\hat{\mathcal{D}}_1 \subseteq \mathcal{D}) \geq 1 - \alpha + o(1)$ for confidence level $1-\alpha$. That is, we can feel fairly confident that the first wage distribution is preferred for all utility functions in the inner CS. There are likely additional utility functions in $\mathcal{D}$, but there is not enough data to distinguish these statistically. If $\hat{\mathcal{D}}_1$ is reasonably large, then there is evidence of a broad consensus favoring the first distribution.

The universe $\mathcal{F}$ contains shifted CRRA utility functions. Specifically,

$$\mathcal{F} = \{f_{\theta,s} : 0 \leq \theta \leq 3, 0 \leq s \leq 110\}, \quad f_{\theta,s}(y) = \begin{cases} \ln(y-s) & \text{if } \theta = 1, \\ \frac{(y-s)^{1-\theta}-1}{1-\theta} & \text{if } \theta \neq 1. \end{cases} \tag{22}$$

The shift $y - s$ can be thought of as having subsistence level s. The θ is the usual risk aversion parameter.

For comparison, I also report an inner CS for the range of values on which the first distribution's CDF lies below the second distribution's CDF. This can also be interpreted as the range of poverty lines for which the first distribution has lower headcount poverty (Atkinson, 1987). To define such an inner CS formally, let $F_a(\cdot)$ and $F_b(\cdot)$ be the two CDFs, and let

$$\mathcal{V} \equiv \{y : F_a(y) < F_b(y), y \in \mathbb{R}\}.$$

That is, $\mathcal{V}$ is the set of *values* on which there is CDF-based restricted first-order stochastic dominance in the sense of Atkinson (1987, Condition I, p. 751), parallel to how $\mathcal{D}$ is the set of *utility functions* on which there is restricted stochastic dominance in the sense of this paper. Parallel to $\hat{\mathcal{D}}_1$, the inner CS $\hat{\mathcal{V}}_1$ should satisfy $\mathrm{P}(\hat{\mathcal{V}}_1 \subseteq \mathcal{V}) \geq 1-\alpha+o(1)$ for confidence level $1-\alpha$.

An inner CS with finite-sample coverage $\mathrm{P}(\hat{\mathcal{V}}_1 \subseteq \mathcal{V}) \geq 1-\alpha$ is produced by Method 5 of Goldman and Kaplan (2018, p. 153). Although they frame the method in terms of multiple testing and strong control of FWER, it is equivalent to an inner CS for the same reasons given in Sections 3 and 4, most formally in the proof of Proposition 7. Explicitly, Theorem 9 of Goldman and Kaplan (2018, p. 155) establishes strong control of FWER for their multiple testing procedure (Method 5) of $H_{0y}\colon F_a(y) \geq F_b(y)$ over $y \in \mathbb{R}$. The corresponding inner CS is

$$\hat{\mathcal{V}}_1 = \{y : H_{0y} \text{ rejected}\}.$$

The finite-sample coverage probability follows:

$$\begin{aligned}\mathrm{P}(\hat{\mathcal{V}}_1 \subseteq \mathcal{V}) &= \mathrm{P}(\text{only false } H_{0y} \text{ rejected}) = 1 - \mathrm{P}(\text{falsely reject at least one true } H_{0y}) \\ &= 1 - \overbrace{\mathrm{FWER}}^{\leq\alpha} \geq 1-\alpha.\end{aligned}$$

Code and data to replicate the following results are publicly available. The code is on my website[3] and uses the `wage2` dataset in the `wooldridge` package in R (Shea, 2018), which comes from Wooldridge (2020), who in turn got it from Blackburn and Neumark (1992).

6.2 Results and interpretation

Figure 1 shows the empirical CDF (ECDF) for urban monthly earnings and non-urban monthly earnings (in 1980 US dollars). The urban ECDF lies below the non-urban ECDF at most points. However, it does not lie below at all points: although difficult to see visually, the two lowest earnings observations are for urban individuals, and the urban ECDF briefly

[3] `https://faculty.missouri.edu/kaplandm`

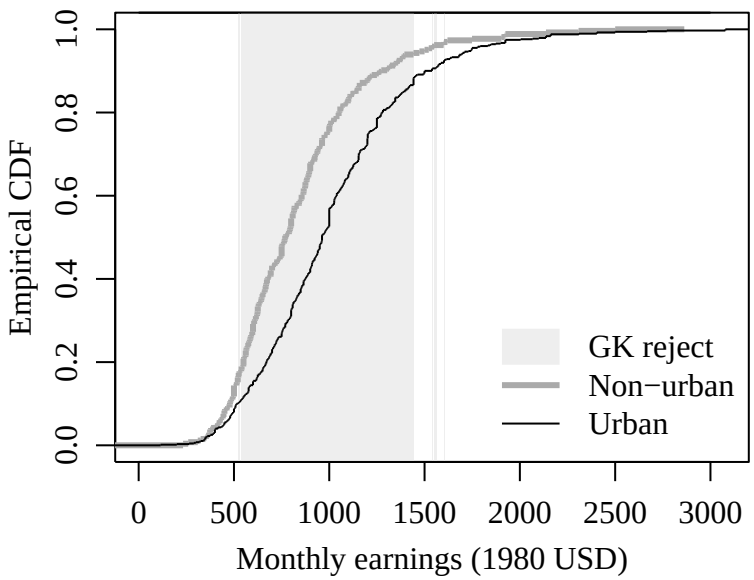

Figure 1: Empirical CDFs and Goldman and Kaplan (2018) rejected ranges of y.

jumps above the non-urban ECDF after \$350. Overall, the intuitive impression is that the urban earnings distribution is better, but with some uncertainty in the tails, especially the lower tail that matters more when risk aversion is high.

Figure 1 also shows the results using the CDF-based method of Goldman and Kaplan (2018). The shaded regions indicate $\hat{\mathcal{V}}_1$: the values y for which $H_{0y}\colon F_{\text{urban}}(y) \geq F_{\text{non-urban}}(y)$ is rejected, controlling the FWER at 5%. This provides statistical evidence in favor of CDF-based restricted stochastic dominance (of urban over non-urban earnings) over a wide range of values, but excluding the tails where there is too much uncertainty. This evidence is stronger than failing to reject a null hypothesis of stochastic dominance, which could happen simply from low power.

Figure 2 shows the inner CS using this paper's new methodology. The darker shading shows the values of (θ, s) corresponding to utility functions $f_{\theta,s} \in \hat{\mathcal{D}}_1$, the inner CS. As in Section 6.1, s can be interpreted as a subsistence level of earnings (in 1980 USD/mo) and θ as the usual risk aversion parameter.

Figures 1 and 2 agree at a high level but have different economic interpretations. At a high level, they agree that the urban distribution generally seems better, but there is still some uncertainty. However, Figure 1 assesses the degree of statistical uncertainty across values of earnings (potential poverty lines), whereas Figure 2 assesses uncertainty across utility functions. At a 95% confidence level, most utility functions in the grid correspond to larger expected utility for the urban distribution. Still, there are some utility functions for which the uncertainty is too great. Specifically, given the data, it is more difficult to support

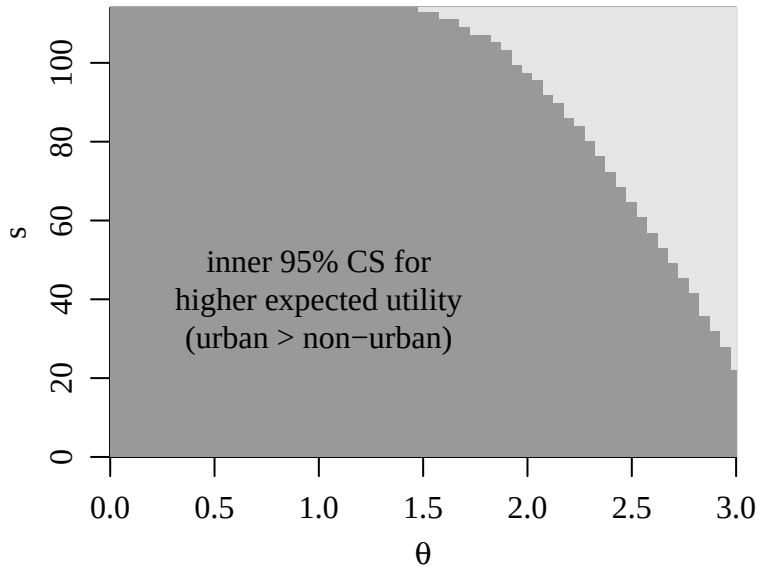

Figure 2: Inner CS $\hat{\mathcal{D}}_1$ in terms of (θ, s).

the superiority of the urban distribution when the subsistence level s is higher and the risk aversion parameter θ is higher, as quantified in Figure 2.

The utility perspective of Figure 2 complements the CDF perspective of Figure 1. As in Atkinson (1987), the CDF perspective can be interpreted as which distribution is preferred in terms of headcount poverty, given different possible poverty lines. The utility perspective shows which distribution is preferred in terms of expected utility, given different possible utility functions. Although headcount poverty is important, expected utility is arguably a more common way to quantify preferences over distributions. Further, results like in Figure 2 can be combined with results in the literature estimating the risk aversion parameter or other utility function properties, to see if the utility functions in the inner CS seem to cover a large segment of the population or not.

7 Simulation

The uniform confidence band and confidence sets are examined in the following simulation. Replication code in R (R Core Team, 2020) is on my website.[4] Along with other details, the numbers of simulation and bootstrap replications are in the table notes.

In the data-generating process, both Y^a and Y^b are log-normal with respective parameters (μ_a, σ_a) and (μ_b, σ_b). The value of (μ_b, σ_b) varies as shown in the table, whereas $(\mu_a, \sigma_a) = (0, 1)$ in every case. Sampling is iid, with sample sizes n_a (for Y^a) and n_b (for Y^b) shown in

[4] `https://faculty.missouri.edu/kaplandm`

the table.

I consider utility functions

$$\mathcal{F} = \{f_\theta : 0 \leq \theta \leq 3\}, \quad f_\theta(y) = \begin{cases} \ln(y + 0.1) & \text{if } \theta = 1, \\ \frac{(y+0.1)^{1-\theta}-1}{1-\theta} & \text{if } \theta \neq 1. \end{cases} \tag{23}$$

This is equivalent to shifting Y^a and Y^b up by 0.1 and using the usual constant relative risk aversion (CRRA) utility function. Without the small shift, $\lim_{y \downarrow 0} f_\theta(y) = -\infty$ for $\theta \geq 1$, which naturally causes problems in finite samples. The small shift makes the envelope function of $\mathcal{F}$ bounded. For computation, the grid $\theta = 0.0, 0.1, \ldots, 3.0$ is used.

The following coverage probability (CP) results are reported. The nominal confidence level is $1 - \alpha = 0.9$. For each simulated dataset, the exchangeable bootstrap uniform confidence band for $\Delta(f) \equiv P^a f - P^b f$ (over $f \in \mathcal{F}$) is computed, using Method 4. The reported CP is the proportion of simulation replications (datasets) in which the band contained $P^a f - P^b f$ for all $f \in \mathcal{F}$, i.e., in which $\hat{b}_1(f) \leq \Delta(f) \leq \hat{b}_2(f)$ for all $f \in \mathcal{F}$. From this band, the (joint) inner confidence set $\hat{\mathcal{D}}_1$ and outer confidence set $\hat{\mathcal{D}}_2$ are computed, using Method 3. The column "both sets" shows the proportion of replications in which $\hat{\mathcal{D}}_1 \subseteq \mathcal{D} \subseteq \hat{\mathcal{D}}_2$, which should be at least as big as the nominal $1 - \alpha$ (recall the $\geq$ in Proposition 9). For additional detail, this is further split into simulated CP of $\hat{\mathcal{D}}_1 \subseteq \mathcal{D}$ and of $\mathcal{D} \subseteq \hat{\mathcal{D}}_2$; note that the sum of these two simulated probabilities always equals one plus the "both sets" simulated probability.

Table 1 shows the confidence band's CP performing reasonably across DGPs. For the smallest sample size, CP ranges from 0.855 to 0.938, depending on the parameters. For the DGP with the lowest CP, increasing the sample size leads to CP near the nominal level: increasing from $n_a = n_b = 40$ to 100 to 250, CP changes from 0.855 to 0.908 to 0.893. The same pattern is true for the DGP with the next-lowest CP, which increases from 0.861 to 0.887 to 0.892.

Table 1 also shows that the CS coverage probabilities are conservative, as expected (per the $\geq$ in Proposition 9). They are always at least 0.95 (even though $1 - \alpha = 0.9$), sometimes 1.000 (up to simulation error). The CSs are all valid, but there is room for improvement in future work, like constructing $\hat{\mathcal{D}}_1$ by inverting the stepdown MTP in Method 2.

8 Conclusion

I have considered learning about the consensus set of utility functions for which one distribution is preferred to another (higher expected utility), which opens other areas for future research. For example, although the modified bootstrap computation is straightforward,

Table 1: Simulated coverage probability.

					Probability of			
n_a	n_b	σ_b	μ_b	$\{\theta : f_\theta \in \mathcal{D}\}$	$\hat{b}_1 \leq \Delta \leq \hat{b}_2$	$\hat{\mathcal{D}}_1 \subseteq \mathcal{D} \subseteq \hat{\mathcal{D}}_2$	$\hat{\mathcal{D}}_1 \subseteq \mathcal{D}$	$\mathcal{D} \subseteq \hat{\mathcal{D}}_2$
40	40	0.7	−0.3	[0.0, 2.8]	0.873	0.960	0.968	0.992
40	40	0.7	0.0	[0.0, 1.1]	0.865	0.972	0.990	0.982
40	40	0.7	0.3	[ ]	0.855	0.998	0.998	1.000
40	40	1.0	−0.3	[0.0, 3.0]	0.920	0.999	1.000	0.999
40	40	1.0	0.0	[ ]	0.938	0.972	0.972	1.000
40	40	1.0	0.3	[ ]	0.922	0.995	0.995	1.000
40	40	1.3	−0.3	[0.2, 3.0]	0.896	0.965	0.967	0.998
40	40	1.3	0.0	[1.2, 3.0]	0.883	0.976	0.988	0.988
40	40	1.3	0.3	[2.5, 3.0]	0.861	0.962	0.994	0.968
100	100	0.7	−0.3	[0.0, 2.8]	0.907	0.968	0.975	0.993
100	100	0.7	0.0	[0.0, 1.1]	0.897	0.977	0.993	0.984
100	100	0.7	0.3	[ ]	0.908	0.999	0.999	1.000
100	100	1.0	−0.3	[0.0, 3.0]	0.934	1.000	1.000	1.000
100	100	1.0	0.0	[ ]	0.929	0.965	0.965	1.000
100	100	1.0	0.3	[ ]	0.922	1.000	1.000	1.000
100	100	1.3	−0.3	[0.2, 3.0]	0.901	0.974	0.979	0.995
100	100	1.3	0.0	[1.2, 3.0]	0.900	0.983	0.987	0.996
100	100	1.3	0.3	[2.5, 3.0]	0.887	0.964	0.992	0.972
250	250	0.7	−0.3	[0.0, 2.8]	0.920	0.978	0.983	0.995
250	250	0.7	0.0	[0.0, 1.1]	0.912	0.981	0.995	0.986
250	250	0.7	0.3	[ ]	0.893	0.998	0.998	1.000
250	250	1.0	−0.3	[0.0, 3.0]	0.920	1.000	1.000	1.000
250	250	1.0	0.0	[ ]	0.937	0.968	0.968	1.000
250	250	1.0	0.3	[ ]	0.942	1.000	1.000	1.000
250	250	1.3	−0.3	[0.2, 3.0]	0.927	0.976	0.978	0.998
250	250	1.3	0.0	[1.2, 3.0]	0.902	0.979	0.988	0.991
250	250	1.3	0.3	[2.5, 3.0]	0.892	0.974	0.994	0.980

Nominal level $1 - \alpha = 0.9$ for band and "two-sided" confidence set; $\mu_a = 0$, $\sigma_a = 1$, 1000 simulation replications, 199 bootstrap draws, $\theta = 0, 0.1, \ldots, 3$, $f(y) = [(y + 0.1)^{1-\theta} - 1]/(1 - \theta)$. In the header, $\Delta = P^a - P^b$, and $\hat{b}_1 \leq \Delta \leq \hat{b}_2$ means $\hat{b}_1(f) \leq P^a f - P^b f \leq \hat{b}_2(f)$ for all $f \in \mathcal{F}$.

extending the theory to non-iid sampling would be valuable, especially for the complex sampling designs often used for income, wealth, and consumption surveys. Also valuable would be extensions to increase the flexibility of a class of utility functions while retaining computational feasibility and economic interpretation, whether inspired by economic theory or convenient basis functions or both. Other extensions include simultaneous ranking of more than two distributions or more than one pair of distributions. For the former, it may be helpful to combine my approach with that of Mogstad, Romano, Shaikh, and Wilhelm (2020), who consider ranking many distributions based on a scalar summary statistic. There may be ways to increase the power of the proposed hypothesis tests, for example by the bootstrap approach of Davidson and Duclos (2013) for testing non-dominance. My general approach can also be considered for ranking distributions using inequality measures indexed by utility functions or other parameters, or for stochastic monotonicity. It can also extend to utility functions of multiple variables (like income and health). Finally, this approach could be considered within choice models or in light of observed choices. For example, if it is known that an individual prefers the first distribution, then the set of utility functions for which the first distribution has higher expected utility can be interpreted as the identified set for that individual's utility function, and the outer confidence set is a confidence set for the identified set.

References

Atkinson, A. B. 1987. "On the Measurement of Poverty." *Econometrica* 55 (4):749–764. URL https://www.jstor.org/stable/1911028.

Atkinson, Anthony B. 1970. "On the Measurement of Inequality." *Journal of Economic Theory* 2 (3):244–263. URL https://doi.org/10.1016/0022-0531(70)90039-6.

Barrett, Garry F. and Stephen G. Donald. 2003. "Consistent tests for stochastic dominance." *Econometrica* 71 (1):71–104. URL https://www.jstor.org/stable/3082041.

Blackburn, McKinley and David Neumark. 1992. "Unobserved Ability, Efficiency Wages, and Interindustry Wage Differentials." *Quarterly Journal of Economics* 107 (4):1421–1436. URL https://www.jstor.org/stable/2118394.

Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. 2013. "Inference on Counterfactual Distributions." *Econometrica* 81 (6):2205–2268. URL https://www.jstor.org/stable/23524318.

Cowell, Frank A. and Emmanuel Flachaire. 2017. "Inequality with Ordinal Data." *Economica* 84 (334):290–321. URL https://doi.org/10.1111/ecca.12232.

Davidson, Russell and Jean-Yves Duclos. 2013. "Testing for Restricted Stochastic Dominance." *Econometric Reviews* 32 (1):84–125. URL https://doi.org/10.1080/07474938.2012.690332.

Goldman, Matt and David M. Kaplan. 2018. "Comparing distributions by multiple testing

across quantiles or CDF values." *Journal of Econometrics* 206 (1):143–166. URL https://doi.org/10.1016/j.jeconom.2018.04.003.

Holm, Sture. 1979. "A Simple Sequentially Rejective Multiple Test Procedure." *Scandinavian Journal of Statistics* 6 (2):65–70. URL https://www.jstor.org/stable/4615733.

Kaplan, David M. 2019. "distcomp: Comparing distributions." *Stata Journal* 19 (4):832–848. URL https://doi.org/10.1177/1536867x19893626.

Kaur, Amarjot, B. L. S. Prakasa Rao, and Harshinder Singh. 1994. "Testing for Second-Order Stochastic Dominance of Two Distributions." *Econometric Theory* 10 (5):849–866. URL https://doi.org/10.1017/S0266466600008884.

Lehmann, E. L. and Joseph P. Romano. 2005. *Testing Statistical Hypotheses*. Springer Texts in Statistics. Springer, 3rd ed. URL https://books.google.com/books?id=Y7vSVW3ebSwC.

Mogstad, Magne, Joseph P. Romano, Azeem M. Shaikh, and Daniel Wilhelm. 2020. "Inference for Ranks with Applications to Mobility across Neighborhoods and Academic Achievement across Countries." Working paper, available at https://home.uchicago.edu/amshaikh/webfiles/rankingsconf.pdf.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Romano, Joseph P. and Azeem M. Shaikh. 2010. "Inference for the Identified Set in Partially Identified Econometric Models." *Econometrica* 78 (1):169–211. URL https://www.jstor.org/stable/25621400.

Shea, Justin M. 2018. *wooldridge: 111 Data Sets from "Introductory Econometrics: A Modern Approach, 6e" by Jeffrey M. Wooldridge*. URL https://CRAN.R-project.org/package=wooldridge. R package version 1.3.1.

van der Vaart, Aad. 1996. "New Donsker Classes." *Annals of Probability* 24 (4):2128–2140. URL https://doi.org/10.1214/aop/1041903221.

van der Vaart, Aad W. and Jon A. Wellner. 1996. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. New York: Springer. URL https://doi.org/10.1007/978-1-4757-2545-2.

Wooldridge, Jeffrey M. 2020. *Introductory Econometrics: A Modern Approach*. Cengage, 7th ed.

A Proofs

A.1 Proof of Lemma 2

Proof. First, the VC dimension is shown to be one by showing the subgraphs are ordered by inclusion. This is most readily apparent by graphing[5] but formally shown by the fact that

[5]For example, https://www.wolframalpha.com/input/?i=plot+%28x%5E%281-0%29-1%29%2F%281-0%29%2C+%28x%5E%281-0.5%29-1%29%2F%281-0.5%29%2C+%28x%5E%281-2%29-1%29%2F%281-2%29%2C+%28x%5E%281-4%29-1%29%2F%281-4%29+from+x%3D0..6

$f_\theta(x)$ is (weakly) decreasing in θ for any x. Taking the derivative with respect to θ,

$$\begin{aligned}\frac{\partial}{\partial\theta}\frac{x^{1-\theta}-1}{1-\theta} &= \frac{(1-\theta)\frac{\partial}{\partial\theta}(x^{1-\theta}-1)-(x^{1-\theta}-1)(-1)}{(1-\theta)^2} \\ &= \frac{(\theta-1)x^{1-\theta}\ln(x)-1+x^{1-\theta}}{(1-\theta)^2}. \end{aligned} \tag{24}$$

The numerator is now shown to be weakly negative (because the denominator is positive). If $x=1$, then the numerator in (24) equals $(\theta-1)1^{1-\theta}\ln(1)-1+1^{1-\theta}=0$ for any θ. Further, the derivative with respect to x of the numerator in (24) is negative for $x>1$ and positive for $x<1$:

$$\begin{aligned}\frac{\partial}{\partial x}[(\theta-1)x^{1-\theta}\ln(x)-1+x^{1-\theta}] &= (\theta-1)[(1-\theta)x^{-\theta}\ln(x)+x^{1-\theta}(1/x)]+(1-\theta)x^{-\theta} \\ &= -(\theta-1)^2x^{-\theta}\ln(x)+(\theta-1)x^{-\theta}+(1-\theta)x^{-\theta} \\ &= [-(\theta-1)^2\ln(x)+(\theta-1)+(1-\theta)]x^{-\theta} \\ &= \overbrace{(\theta-1)^2}^{>0}[-\ln(x)]\overbrace{x^{-\theta}}^{>0}\end{aligned}$$

has the same sign as $-\ln(x)$, which is strictly positive for $0<x<1$ and strictly negative for $x>1$. Thus, the zero value of (24) at $x=1$ is the global maximum for any θ, so (24) is non-positive. That is, $f_\theta(x)$ is decreasing in θ for any x.

Since the subgraphs are ordered by inclusion (because $f_a \le f_b$ for any $a \le b$), the VC dimension is one (VC index is two); e.g., see the proof of Lemma 2.6.16 in van der Vaart and Wellner (1996).

If additionally the envelope function is square integrable, then this implies $\mathcal{F}$ is P-Donsker by Theorems 2.5.2 and 2.6.7 in van der Vaart and Wellner (1996). □

A.2 Proof of Proposition 7

Proof. Using Corollary 5 and Theorem 6,

$$\begin{aligned}\text{FWER} &= \text{P}(\text{reject any } H_{0f} \text{ with } f \in \mathcal{F}_T) \\ &= \text{P}(\sup_{f\in\mathcal{F}_T}\hat{T}_f > \tilde{T}^{\mathcal{F}_\vee}_{1-\alpha}) \\ &\le \text{P}(\sup_{f\in\mathcal{F}}\hat{T}_f > \tilde{T}^{\mathcal{F}_\vee}_{1-\alpha}) \to \alpha.\end{aligned}$$

□

A.3 Proof of Proposition 8

Proof. First, consider the infeasible oracle critical value from Step 3 when the true set of true hypotheses $\hat{K}^{(i)} = \mathcal{F}_T$ is used, i.e., the critical value $\tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}$. Hypothetically, if this oracle critical value were used, then

$$\lim_{n\to\infty} \mathrm{P}(\overbrace{\sup_{f\in\mathcal{F}_T} \hat{T}_f}^{\hat{T}^{\mathcal{F}_T\vee}} > \tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}) = \alpha, \tag{25}$$

because $\hat{T}^{\mathcal{F}_T\vee} \xrightarrow{d} T^{\mathcal{F}_T\vee}$ by Corollary 5 and $\tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee} \xrightarrow{p} T_{1-\alpha}^{\mathcal{F}_T\vee}$, and the distribution of $T^{\mathcal{F}_T\vee}$ is continuous.

Second, the argument follows from a monotonicity property similar to (15.37) in Lehmann and Romano (2005). Specifically, $\mathcal{F} = \hat{K}^{(0)} \supset \hat{K}^{(1)} \supset \cdots$ because additional H_{0f} can be rejected in every iteration, but once rejected they can never be un-rejected. Combined with using the same bootstrap draws in every iteration, this implies the critical values are also monotonic over iterations: $\tilde{T}_{1-\alpha}^{\hat{K}^{(0)}\vee} > \tilde{T}_{1-\alpha}^{\hat{K}^{(1)}\vee} > \cdots$.

Consider a dataset in which

$$\sup_{f\in\mathcal{F}_T} \hat{T}_f \leq \tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}. \tag{26}$$

That is, given the oracle critical value $\tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}$ based on the true $\mathcal{F}_T$, none of the true H_{0f} is rejected (i.e., no familywise error is committed). By (25), the probability of such a dataset converges to $1-\alpha$.

By induction, the stepdown procedure never commits a familywise error in such datasets. Equivalently, it is shown that $\hat{K}^{(i)} \supseteq \mathcal{F}_T$ for all iterations i in such datasets, which implies $\tilde{T}_{1-\alpha}^{\hat{K}^{(i)}\vee} \geq \tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}$. In the first iteration of the stepdown procedure, $\hat{K}^{(0)} = \mathcal{F} \supseteq \mathcal{F}_T$, so the critical value satisfies $\tilde{T}_{1-\alpha}^{\hat{K}^{(0)}\vee} \geq \tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}$. Next, it is shown that $\hat{K}^{(i)} \supseteq \mathcal{F}_T \implies \hat{K}^{(i+1)} \supseteq \mathcal{F}_T$. Specifically, in iteration i, if $\hat{K}^{(i)} \supseteq \mathcal{F}_T$, then $\tilde{T}_{1-\alpha}^{\hat{K}^{(i)}\vee} \geq \tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}$, so none of the true H_{0f} are rejected:

$$\overbrace{\sup_{f\in\mathcal{F}_T} \hat{T}_f \leq \underbrace{\tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee}}_{\text{since } \hat{K}^{(i)}\supseteq\mathcal{F}_T} \leq \tilde{T}_{1-\alpha}^{\hat{K}^{(i)}\vee}}^{\text{by (26)}}.$$

Since no true H_{0f} are rejected, $\hat{K}^{(i+1)} \supseteq \mathcal{F}_T$, too. Thus,

$$\mathrm{FWER} = \mathrm{P}(\text{reject any true } H_{0f}) \leq \overbrace{\mathrm{P}(\sup_{f\in\mathcal{F}_T} \hat{T}_f > \tilde{T}_{1-\alpha}^{\mathcal{F}_T\vee})}^{\to\alpha \text{ by (25)}} \to \alpha. \qquad \square$$

A.4 Proof of Proposition 9

Proof. For the inner and outer CS, asymptotic coverage follows from FWER control. For $\hat{\mathcal{D}}_1$, with $H_{0f}\colon P^a f - P^b f \leq 0$,

$$\begin{aligned}
\mathrm{P}(\hat{\mathcal{D}}_1 \subseteq \mathcal{D}) &= \mathrm{P}(f \in \hat{\mathcal{D}}_1 \text{ only if } f \in \mathcal{D}) \\
&= \mathrm{P}(\text{MTP rejects } H_{0f} \text{ only if } P^a f - P^b f > 0) \\
&= 1 - \overbrace{\mathrm{P}(\text{MTP rejects any } H_{0f} \text{ with } P^a f - P^b f \leq 0)}^{=\text{FWER} \leq \alpha + o(1)} \\
&\geq 1 - \alpha + o(1).
\end{aligned}$$

For $\hat{\mathcal{D}}_2$, with $H_{0f}\colon P^a f - P^b f \geq 0$,

$$\begin{aligned}
\mathrm{P}(\hat{\mathcal{D}}_2 \supseteq \mathcal{D}) &= \mathrm{P}(f \in \hat{\mathcal{D}}_2 \text{ when } f \in \mathcal{D}) \\
&= \mathrm{P}(\text{MTP does not reject any } H_{0f} \text{ when } P^a f - P^b f > 0) \\
&\geq \mathrm{P}(\text{MTP does not reject any } H_{0f} \text{ when } P^a f - P^b f \geq 0) \\
&= 1 - \overbrace{\mathrm{P}(\text{MTP rejects any } H_{0f} \text{ with } P^a f - P^b f \geq 0)}^{=\text{FWER} \leq \alpha + o(1)} \\
&\geq 1 - \alpha + o(1).
\end{aligned}$$

For the joint CS, coverage follows from the uniform confidence band's coverage:

$$\begin{aligned}
\mathrm{P}(\hat{\mathcal{D}}_1 \subseteq \mathcal{D} \subseteq \hat{\mathcal{D}}_2) &= \mathrm{P}\Big(\hat{b}_1(f) \leq 0 \text{ whenever } P^a f - P^b f \leq 0, \text{ and} \\
&\qquad \hat{b}_2(f) > 0 \text{ whenever } P^a f - P^b f > 0\Big) \\
&\geq \mathrm{P}\Big(\hat{b}_1(f) \leq P^a f - P^b f \text{ whenever } P^a f - P^b f \leq 0, \text{ and} \\
&\qquad \hat{b}_2(f) > P^a f - P^b f \text{ whenever } P^a f - P^b f > 0\Big) \\
&\geq \overbrace{\mathrm{P}(\hat{b}_1(f) \leq P^a f - P^b f \leq \hat{b}_2(f) \text{ for all } f \in \mathcal{F})}^{\text{apply Proposition 10}} \\
&\geq 1 - \alpha + o(1).
\end{aligned}$$

□

A.5 Proof of Proposition 10

Proof. For the symmetric band,

$$\mathrm{P}\{\hat{b}_1(f) \leq P^a f - P^b f \leq \hat{b}_2(f) \text{ for all } f \in \mathcal{F}\}$$

$$
\begin{aligned}
&= \mathrm{P}\{\overbrace{(\mathbb{P}_n^a - \mathbb{P}_n^b)f - |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}}^{\hat{b}_1(f) \text{ from (17)}} \le P^a f - P^b f \le \overbrace{(\mathbb{P}_n^a - \mathbb{P}_n^b)f + |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}}^{\hat{b}_2(f) \text{ from (17)}},\ \forall f \in \mathcal{F}\} \\
&= \mathrm{P}\{-|\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a} \le [(P^a - P^b) - (\mathbb{P}_n^a - \mathbb{P}_n^b)]f \le |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a} \text{ for all } f \in \mathcal{F}\} \\
&= \mathrm{P}\{-|\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee} \le \overbrace{\sqrt{n_a}[(P^a - P^b) - (\mathbb{P}_n^a - \mathbb{P}_n^b)]f/\hat{\sigma}_f}^{-\hat{T}_f \text{ from (5)}} \le |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee} \text{ for all } f \in \mathcal{F}\} \\
&= \mathrm{P}\{|\hat{T}_f| \le |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee} \text{ for all } f \in \mathcal{F}\} \\
&= \mathrm{P}\{\overbrace{\sup_{f\in\mathcal{F}}|\hat{T}_f|}^{|\hat{T}|^{\mathcal{F}\vee} \text{ from (5)}} \le |\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee}\} \\
&\to 1-\alpha
\end{aligned}
$$

because $|\hat{T}|^{\mathcal{F}\vee} \xrightarrow{d} |T|^{\mathcal{F}\vee}$ by Corollary 5 and $|\tilde{T}|_{1-\alpha}^{\mathcal{F}\vee} \xrightarrow{p} |T|_{1-\alpha}^{\mathcal{F}\vee}$ by Theorem 6, and because $|T|^{\mathcal{F}\vee}$ has a continuous distribution.

For the one-sided band with $\hat{b}_1$,

$$
\begin{aligned}
&\mathrm{P}(\hat{b}_1(f) \le P^a f - P^b f \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}(\overbrace{(\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a}}^{\hat{b}_1(f) \text{ from (17)}} - (P^a f - P^b f) \le 0 \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}((\mathbb{P}_n^a - \mathbb{P}_n^b)f - (P^a f - P^b f) \le \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a} \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}(\overbrace{\sqrt{n_a}[(\mathbb{P}_n^a - \mathbb{P}_n^b) - (P^a - P^b)]f/\hat{\sigma}_f}^{\hat{T}_f \text{ from (5)}} \le \tilde{T}_{1-\alpha}^{\mathcal{F}\vee} \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}(\overbrace{\sup_{f\in\mathcal{F}}\hat{T}_f}^{\hat{T}^{\mathcal{F}\vee} \text{ from (5)}} \le \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}) \\
&\to 1-\alpha
\end{aligned}
$$

because $\hat{T}^{\mathcal{F}\vee} \xrightarrow{d} T^{\mathcal{F}\vee}$ by Corollary 5 and $\tilde{T}_{1-\alpha}^{\mathcal{F}\vee} \xrightarrow{p} T_{1-\alpha}^{\mathcal{F}\vee}$ by Theorem 6, and because $T^{\mathcal{F}\vee}$ has a continuous distribution.

For the one-sided band with $\hat{b}_2$,

$$
\begin{aligned}
&\mathrm{P}(\hat{b}_2(f) \ge P^a f - P^b f \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}(\overbrace{(\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{\alpha}^{\mathcal{F}\wedge}\hat{\sigma}_f/\sqrt{n_a}}^{\hat{b}_2(f) \text{ from (17)}} - (P^a f - P^b f) \ge 0 \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}((\mathbb{P}_n^a - \mathbb{P}_n^b)f - (P^a f - P^b f) \ge \tilde{T}_{\alpha}^{\mathcal{F}\wedge}\hat{\sigma}_f/\sqrt{n_a} \text{ for all } f \in \mathcal{F})
\end{aligned}
$$

$$
\begin{aligned}
&= \mathrm{P}(\overbrace{\sqrt{n_a}[(\mathbb{P}_n^a - \mathbb{P}_n^b) - (P^a - P^b)]f/\hat{\sigma}_f}^{\hat{T}_f \text{ from (5)}} \geq \tilde{T}_\alpha^{\mathcal{F}\wedge} \text{ for all } f \in \mathcal{F}) \\
&= \mathrm{P}(\overbrace{\inf_{f\in\mathcal{F}} \hat{T}_f}^{\hat{T}^{\mathcal{F}\wedge} \text{ from (5)}} \geq \tilde{T}_\alpha^{\mathcal{F}\wedge}) \\
&\to 1 - \alpha
\end{aligned}
$$

because $\hat{T}^{\mathcal{F}\wedge} \xrightarrow{d} T^{\mathcal{F}\wedge}$ by Corollary 5 and $\tilde{T}_{1-\alpha}^{\mathcal{F}\wedge} \xrightarrow{p} T_{1-\alpha}^{\mathcal{F}\wedge}$ by Theorem 6, and because $T^{\mathcal{F}\wedge}$ has a continuous distribution.

For the "equal-tailed" band,

$$
\begin{aligned}
&\mathrm{P}(\hat{b}_1(f) \leq P^a f - P^b f \leq \hat{b}_2(f) \text{ for all } f \in \mathcal{F}) \\
&= 1 - \mathrm{P}(\hat{b}_1(f) > P^a f - P^b f \text{ for some } f, \text{ or } \hat{b}_2(f) < P^a f - P^b f \text{ for some } f) \\
&\geq 1 - \mathrm{P}(\hat{b}_1(f) > P^a f - P^b f \text{ for some } f) - \mathrm{P}(\hat{b}_2(f) < P^a f - P^b f \text{ for some } f) \\
&\to 1 - \alpha/2 - \alpha/2 = 1 - \alpha,
\end{aligned}
$$

essentially a Bonferroni adjustment argument. □

A.6 Proof of Proposition 11

Proof. Under H_0, $(P^a - P^b)f \leq 0$ for all $f \in \mathcal{F}$, so

$$
\hat{T}_f^0 = \frac{\sqrt{n_a}(\mathbb{P}_n^a - \mathbb{P}_n^b)f}{\hat{\sigma}_f} = \frac{\overbrace{\sqrt{n_a}[(\mathbb{P}_n^a - \mathbb{P}_n^b) - (P^a - P^b)]f}^{\hat{T}_f}}{\hat{\sigma}_f} + \frac{\sqrt{n_a}\overbrace{(P^a - P^b)f}^{\leq 0}}{\hat{\sigma}_f} \leq \hat{T}_f.
$$

Thus, using notation from (4), (5), and (7) along with $\hat{T}^{\mathcal{F}\vee} \xrightarrow{d} T^{\mathcal{F}\vee}$ (Corollary 5) and $\tilde{T}_{1-\alpha}^{\mathcal{F}\vee} \xrightarrow{p} T_{1-\alpha}^{\mathcal{F}\vee}$ (Theorem 6), and the fact that $T^{\mathcal{F}\vee}$ has a continuous distribution,

$$
\begin{aligned}
\lim_{n\to\infty} \mathrm{P}(\sup_{f\in\mathcal{F}} \hat{T}_f^0 > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}) &\leq \lim_{n\to\infty} \mathrm{P}(\overbrace{\sup_{f\in\mathcal{F}} \hat{T}_f}^{\hat{T}^{\mathcal{F}\vee}} > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}) \\
&= \lim_{n\to\infty} \mathrm{P}(\overbrace{\hat{T}^{\mathcal{F}\vee}}^{\xrightarrow{d} T^{\mathcal{F}\vee}} > \overbrace{\tilde{T}_{1-\alpha}^{\mathcal{F}\vee}}^{\xrightarrow{p} T_{1-\alpha}^{\mathcal{F}\vee}}) \\
&= \mathrm{P}(T^{\mathcal{F}\vee} > T_{1-\alpha}^{\mathcal{F}\vee}) = \alpha.
\end{aligned}
$$

The version using $\hat{b}_1$ is equivalent because $\hat{b}_1(f) > 0$ for some $f \in \mathcal{F}$ if and only if $\sup_{f\in\mathcal{F}}(\mathbb{P}_n^a - \mathbb{P}_n^b)f - \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}\hat{\sigma}_f/\sqrt{n_a} > 0$, which in turn is equivalent to $\sup_{f\in\mathcal{F}} \sqrt{n_a}(\mathbb{P}_n^a -$

$\mathbb{P}_n^b)f/\hat{\sigma}_f > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$ and thus $\sup_{f\in\mathcal{F}} \hat{T}_f^0 > \tilde{T}_{1-\alpha}^{\mathcal{F}\vee}$. □

A.7 Proof of Proposition 12

Proof. Let H_0 hold, so $P^a g - P^b g \leq 0$ for at least one $g \in \mathcal{F}$. Since $\{g\} \subset \mathcal{F}$, $\inf_{f\in\mathcal{F}} \hat{T}_f^0 \leq \hat{T}_g^0$. Since $P^a g - P^b g \leq 0$,

$$\hat{T}_g^0 = \frac{\sqrt{n_a}(\mathbb{P}_n^a - \mathbb{P}_n^b)g}{\hat{\sigma}_g} = \overbrace{\frac{\sqrt{n_a}[(\mathbb{P}_n^a - \mathbb{P}_n^b) - (P^a - P^b)]g}{\hat{\sigma}_g}}^{\hat{T}_g} + \overbrace{\frac{\sqrt{n_a}(P^a - P^b)g}{\hat{\sigma}_g}}^{\leq 0} \leq \overbrace{\hat{T}_g \xrightarrow{d} \mathrm{N}(0,1)}^{\text{by Corollary 5}}.$$

Thus,

$$\lim_{n\to\infty} \mathrm{P}(\inf_{f\in\mathcal{F}} \hat{T}_f^0 > z_{1-\alpha}) \leq \lim_{n\to\infty} \mathrm{P}(\hat{T}_g^0 > z_{1-\alpha}) \leq \lim_{n\to\infty} \mathrm{P}(\hat{T}_g > z_{1-\alpha}) = \alpha.$$ □

B Computing bootstrap critical values

Method 7 is an example algorithm to compute the bootstrap critical values from Theorem 6. Parts are similar to Algorithm 3 of Chernozhukov, Fernández-Val, and Melly (2013, p. 2222). An example implementation in R (R Core Team, 2020) is in the simulation code on my website.[6]

Method 7 (bootstrap critical values)**.** *Take as given $\alpha \in (0,1)$, $\mathcal{S} \subseteq \mathcal{F}$, n_a, n_b, c (from A4), and the number of bootstrap replications R. Note $c = 1$ for empirical bootstrap, Bayesian bootstrap, and wild bootstrap with weights of unit variance. In practice, usually $\mathcal{S}$ must be replaced by a finite grid, which may be arbitrarily fine (restricted only by computation time and patience).*

1. *Draw weights $(\tilde{W}_1^a, \ldots, \tilde{W}_{n_a}^a)$. Compute $\tilde{\mathbb{P}}_n^a f = n_a^{-1} \sum_{i=1}^{n_a} f(Y_i^a)\tilde{W}_i^a$ for all $f \in \mathcal{S}$, as well as $\bar{W}^a = n_a^{-1} \sum_{i=1}^{n_a} \tilde{W}_i^a$.*
2. *Independently draw $(\tilde{W}_1^b, \ldots, \tilde{W}_{n_b}^b)$. Compute $\tilde{\mathbb{P}}_n^b f = n_b^{-1} \sum_{i=1}^{n_b} f(Y_i^b)\tilde{W}_i^b$ for all $f \in \mathcal{S}$, as well as $\bar{W}^b = n_b^{-1} \sum_{i=1}^{n_b} \tilde{W}_i^b$.*
3. *Following (6), compute $\tilde{\mathbb{G}}_n^\Delta f = \sqrt{n_a}[(\tilde{\mathbb{P}}_n^a - \tilde{\mathbb{P}}_n^b) - (\bar{W}^a\mathbb{P}_n^a - \bar{W}^b\mathbb{P}_n^b)]f/c$ for all $f \in \mathcal{S}$.*
4. *Repeat Steps 1–3 R times, generating $(\tilde{\mathbb{G}}_n^\Delta f)^{(r)}$ in bootstrap replications $r = 1, \ldots, R$.*
5. *For all $f \in \mathcal{S}$, compute $\hat{\sigma}_f$ following (8) with quantiles from $((\tilde{\mathbb{G}}_n^\Delta f)^{(1)}, \ldots, (\tilde{\mathbb{G}}_n^\Delta f)^{(R)})$.*
6. *For all $f \in \mathcal{S}$ and $r = 1, \ldots, R$, compute $\tilde{T}_f^{(r)} = (\tilde{\mathbb{G}}_n^\Delta f)^{(r)}/\hat{\sigma}_f$ as in (7).*

[6] `https://faculty.missouri.edu/kaplandm`

7. *Following* (7), *for each* $r = 1, \ldots, R$, *compute*

$$\tilde{T}^{\mathcal{S}\vee}_{(r)} = \sup_{f\in\mathcal{S}} \tilde{T}^{(r)}_f, \quad \tilde{T}^{\mathcal{S}\wedge}_{(r)} = \inf_{f\in\mathcal{S}} \tilde{T}^{(r)}_f, \quad |\tilde{T}|^{\mathcal{S}\vee}_{(r)} = \sup_{f\in\mathcal{S}} |\tilde{T}|^{(r)}_f.$$

8. *Let* $\tilde{T}^{\mathcal{S}\vee}_{1-\alpha}$ *be the sample* $(1-\alpha)$*-quantile of* $(\tilde{T}^{\mathcal{S}\vee}_{(1)}, \ldots, \tilde{T}^{\mathcal{S}\vee}_{(R)})$, $\tilde{T}^{\mathcal{S}\wedge}_{1-\alpha}$ *the sample* $(1-\alpha)$*-quantile of* $(\tilde{T}^{\mathcal{S}\wedge}_{(1)}, \ldots, \tilde{T}^{\mathcal{S}\wedge}_{(R)})$, *and* $|\tilde{T}|^{\mathcal{S}\vee}_{1-\alpha}$ *the sample* $(1-\alpha)$*-quantile of* $(|\tilde{T}|^{\mathcal{S}\vee}_{(1)}, \ldots, |\tilde{T}|^{\mathcal{S}\vee}_{(R)})$.