Comparing distributions by multiple testing across quantiles

Matt Goldman* David M. Kaplan†

November 29, 2016

Abstract

When comparing two distributions, it is often helpful to learn at which quantiles there is a statistically significant difference. This provides more information than the binary "reject" or "do not reject" decision of a global goodness-of-fit test. Framing our question as multiple testing across the continuum of quantiles, we show that the Kolmogorov–Smirnov test (with appropriately modified interpretation) achieves strong control of the familywise error rate. However, its well-known flaw of low sensitivity in the tails remains. We provide an alternative method that retains such strong control of familywise error rate while also having even sensitivity, i.e., equal pointwise type I error rates at $n \to \infty$ quantiles across the distribution. Our method computes instantly, using our new formula that also instantly computes goodness-of-fit p-values and uniform confidence bands. To improve power, we also propose stepdown and pre-test procedures that maintain asymptotic familywise error rate control. One-sample (i.e., one known distribution, one unknown) and two-sample (i.e., two unknown distributions) cases are considered. Simulations, empirical examples, and code are provided.

JEL classification: C12, C14, C21

Keywords: Dirichlet; familywise error rate; Kolmogorov–Smirnov; probability integral transform; stepdown

*Microsoft Research, mattgold@microsoft.com.

†Corresponding author. Department of Economics, University of Missouri, kaplandm@missouri.edu.

1 Introduction

Increasingly, economists compare not only means, but entire distributions. This includes comparing income distributions (over two time periods, geographic areas, or demographic subpopulations) and a variety of economic outcomes in experiments and program evaluation, either comparing unknown "treated" and "untreated" distributions (i.e., two-sample inference), or comparing an unknown "treated" distribution to a known population distribution (i.e., one-sample inference).

To compare distributions, the most common statistical tests answer one of two questions: 1) Are the distributions identical or different? 2) Do the distributions differ at the median (or another pre-specified quantile)? Often, the question with the most economic and policy relevance is instead: 3) Across the entire distribution, at which quantiles do the distributions differ? For example, one may care not just *whether* two (sub)populations have different income distributions, but *where* (at which quantiles) they differ. In an experimental setting, the question is at which quantiles the treatment effect is statistically significant; see Section 7 for an empirical example.

We contribute to answering question (3). First, we formalize the question as multiple testing of a continuum of quantile hypotheses, which we call "quantile multiple testing." This framework appears to be new to the literature on distributional inference. Second, we show that the Kolmogorov–Smirnov (KS) test can be interpreted as answering question (3) and that it appropriately controls the probability of having at least one false rejection, i.e., controls the familywise error rate (FWER).[1] Third, we propose a new approach to answer question (3). Like the KS, our approach is nonparametric and appropriately controls FWER, without being conservative (like the Bonferroni method). Unlike the KS, our approach maintains "even sensitivity" (as quantified by pointwise size) across the continuum of quantile hypotheses. This addresses the long noted problem of the KS test's poor power against deviations in the tails (e.g., Eicker, 1979, p. 117). Fourth, we provide a new formula to instantly compute our method as well as related goodness-of-fit p-values and uniform confidence bands for an unknown CDF. Fifth, we refine our basic method with stepdown and pre-test procedures to improve power without sacrificing strong control of FWER.

Question (3) cannot necessarily be answered by methods addressing questions (1) or (2). The answer to question (1) is only "yes" or "no." Using a method that answers question (2), if separate hypothesis tests are run at many different quantiles, each with size α, then the well-known multiple testing problem is that the overall probability of making any false

[1] There are other reasonable ways to quantify control of type I errors for multiple testing, such as the false discovery rate (FDR) of Benjamini and Hochberg (1995) and the k-FWER and false discovery proportion (FDP) of Lehmann and Romano (2005a).

rejection (a.k.a. "false discovery") is above α; see, e.g., Romano, Shaikh, and Wolf (2010). Even with two identical distributions (so all quantiles are identical), such a naive procedure may falsely reject equality for at least one quantile 30% of the time even if $\alpha = 5\%$. This 30% overall false rejection probability (formalized later) is the FWER.

The KS test (Kolmogorov, 1933; Smirnov, 1939, 1948) is a goodness-of-fit (GOF) test that is only intended to answer question (1) above, but we show that it readily identifies a set of values at which the population and null distribution functions differ in a way that controls FWER. We call this the "KS-based" multiple testing procedure to distinguish it from the "KS test" for GOF. Using other GOF approaches like Cramér–von Mises, Anderson–Darling, and permutation tests, identifying *where* two distributions differ is not possible.

The KS test, however, is known to suffer from low sensitivity (i.e., low power) in the tails. More specifically, the allocation of sensitivity is uneven: it is concentrated in the middle of the distribution, as has been discussed formally in the literature since (at least) Eicker (1979); Jaeschke (1979). The KS test also has an uneven distribution of sensitivity to deviations above versus below the null distribution at any point x away from the median.

The KS test's uneven sensitivity can lead to obviously incorrect inferences. For example, let sample size $n = 20$, with null hypothesis distribution Uniform$(0, 1)$. Even if five of the 20 observations exceed one million, any of which alone clearly contradicts the null hypothesis, the KS test still fails to reject at a 10% level.[2] This is not a small-sample issue: with 500 of $n = 200\,000$ observations exceeding one million, the KS still fails to reject at a 10% level.[3]

For GOF testing, i.e., testing the global null $H_0 : F(\cdot) = F_0(\cdot)$, as well as uniform confidence bands, Buja and Rolke (2006) appear to be the first to achieve "even sensitivity" by using the probability integral transform. The probability integral transform reduces the problem to order statistics from a standard uniform distribution, whose finite-sample distribution is known. Their one-sample uniform confidence band (from Section 5.1) was eventually detailed and published by Aldor-Noiman, Brown, Buja, Rolke, and Stine (2013). Buja and Rolke (2006) also construct a two-sample permutation test for equality. They do not propose any multiple testing procedure or discuss FWER. Other papers have explored implications of the same probability integral transform, such as Moscovich, Nadler, and Spiegelman (2016) in the computer science literature, but only for GOF testing or uniform confidence bands, never quantile multiple testing.

The same probability integral transform underlies our methods. It provides finite-sample distributions while being distribution-free, and it facilitates finite-sample control of both

[2]R code: `ks.test(c(1:15/21,10^6+1:5),punif)` results in `D = 0.25, p-value = 0.1376`.

[3]R code: `n=200000;k=500;ks.test(c(1:(n-k)/(n+1),10^6+1:k),punif)` results in `D = 0.0025, p-value = 0.1641`.

overall FWER and pointwise type I error rates. The tradeoff is that iid sampling is required. However, the finite-sample sampling distribution (of the true CDF evaluated jointly at all order statistics) turns out to be equivalent to the finite-sample posterior distribution from the continuity-corrected Bayesian bootstrap in Banks (1988): in the iid case, the uniform confidence band of Aldor-Noiman et al. (2013) is also a Bayesian uniform credible band.[4] We are hopeful that further work will show that our iid assumption may be "relaxed" by using the Bayesian bootstrap to allow sampling weights (as in Lo, 1993) and clustering (as in Cameron, Gelbach, and Miller, 2008) while maintaining exact finite-sample properties in the iid case.

We provide a closed-form calibration formula that allows not only our multiple testing procedure but also the uniform confidence band and GOF p-values of Buja and Rolke (2006) to be computed instantly. This formula replaces just-in-time simulations that can last (depending on sample size) minutes or even hours.

Most of our new multiple testing results rely on viewing the problem from the quantile perspective rather than the probability perspective of Buja and Rolke (2006) and related papers. Seeing the problem as testing multiple quantiles helps us establish FWER properties and is critical for our procedures to improve power. Our strategy is to test n different quantile hypotheses using the n order statistics (i.e., ordered sample values). In contrast, papers like Buja and Rolke (2006) apply the null CDF $F_0(\cdot)$ to the order statistics $X_{n:1} < \cdots < X_{n:k} < \cdots < X_{n:n}$, comparing $F_0(X_{n:k})$ to certain critical values for $k = 1, \ldots, n$. When the true $F(\cdot)$ equals $F_0(\cdot)$, it is easy to analyze such a test's properties, and such is sufficient for GOF testing. However, if $F(x) = F_0(x)$ only over a proper subset of $\mathbb{R}$, then it is difficult to compute the false rejection probability: $X_{n:k}$ is random, so $F_0(X_{n:k}) = F(X_{n:k})$ is true in some samples but not others. The quantile perspective avoids this difficulty: each pointwise null hypothesis concerns only one fixed population quantile value, $F^{-1}(\tau)$, which is tested with one order statistic.

This quantile multiple testing perspective facilitates procedures to improve power. It leads naturally to a stepdown procedure in the spirit of Holm (1979), where if any quantile hypotheses are rejected by the initial test, then the remaining ones may be tested with a smaller critical value. Further, for one-sided testing, we propose a pre-test to determine at which quantiles the null hypothesis inequality constraint may be binding, and the pointwise test levels are recalibrated with attention restricted to this subset, similar to Linton, Song, and Whang (2010), among others.

In the literature, using the probability integral transform for GOF testing dates back to Fisher (1932), Pearson (1933), and Neyman (1937). An extension especially relevant to us

[4]This is shown formally in a prior version of this paper.

is that the joint distribution of $F(X_{n:1}), \ldots, F(X_{n:n})$ for order statistics $X_{n:1} < \cdots < X_{n:n}$ is the same as that of the order statistics from a Uniform$(0,1)$ distribution; Scheffé and Tukey (1945) seem to be the first to note this (e.g., as cited in David and Nagaraja, 2003). Using a closely related sampling distribution, nonparametric (empirical) likelihood-based GOF testing and uniform confidence bands are respectively proposed by Berk and Jones (1979) and Owen (1995). However, they do not discuss multiple testing or two-sample inference, and our methods compute faster and spread sensitivity more evenly.

For multiple testing concepts like FWER and stepdown procedures, see Chapter 9 in Lehmann and Romano (2005b), Romano et al. (2010), and references therein.

Section 3 contains results for the KS-based multiple testing procedure. Sections 4 and 5 describe our new methods and their properties. Sections 6 and 7 contain simulation results and an empirical example, respectively. Appendix A contains additional methods, Appendix B contains proofs, Appendix C contains computational details, and Appendix D contains additional simulations.

Notationally, we use α for FWER and $\tilde{\alpha}$ for pointwise type I error rate. Acronyms and abbreviations used include those for confidence interval (CI), data generating process (DGP), empirical distribution function (EDF), familywise error rate (FWER), goodness-of-fit (GOF), Kolmogorov–Smirnov (KS), multiple testing procedure (MTP), and rejection probability (RP). Random and non-random vectors are respectively typeset as, e.g., $\mathbf{X}$ and $\mathbf{x}$, while random and non-random scalars are typeset as X and x, and random and non-random matrices as $\underline{\mathbf{X}}$ and $\underline{\mathbf{x}}$; $\mathbb{1}\{\cdot\}$ is the indicator function. The Dirichlet distribution with parameters $a_1, \ldots, a_K$ is written $\text{Dir}(a_1, \ldots, a_K)$, and the beta distribution $\text{Beta}(a, b)$; in some cases these stand for random variables following such distributions. The α-quantile of the $\text{Beta}(k, n+1-k)$ distribution is denoted by $B^{\alpha}_{k,n}$.

2 Setup

First, we define multiple testing terms following Lehmann and Romano (2005b, §9.1).

Definition 1. For a family of null hypotheses H_{0h} indexed by h, let $I \equiv \{h : H_{0h} \text{ is true}\}$. The "familywise error rate" is

$$\text{FWER} \equiv \text{P}(\text{reject any } H_{0h} \text{ with } h \in I).$$

Definition 2. Given the notation in Definition 1, "weak control" of FWER at level α requires FWER $\leq \alpha$ if each H_{0h} is true. "Strong control" of FWER requires FWER $\leq \alpha$ for any I.

Given Definition 2, when we establish strong control of FWER, then weak control is

directly implied. In our results, we will establish strong control of FWER over a set of distributions, similar to establishing "size control" by showing that type I error rates are controlled over a set of distributions.[5]

Second, we maintain the following assumptions throughout.

Assumption 1. One-sample: scalar observations $X_i \overset{iid}{\sim} F$, and the sample size is n. Two-sample: scalar observations $X_i \overset{iid}{\sim} F_X$, $Y_i \overset{iid}{\sim} F_Y$, with respective sample sizes n_X and n_Y, and the samples are independent of each other: $\{X_i\}_{i=1}^{n_X} \perp\!\!\!\perp \{Y_k\}_{k=1}^{n_Y}$.

Assumption 2. One-sample: $F(\cdot)$ is continuous and strictly increasing over its support, taken to be $\mathbb{R}$. Two-sample: $F_X(\cdot)$ and $F_Y(\cdot)$ are continuous and strictly increasing over their common support, taken to be $\mathbb{R}$.

Assumption 1 is applicable in many cases (such as our empirical example), but excludes settings with sampling weights or dependence. As noted in Section 1, explorations of weakening this assumption through the connection with Banks (1988) are in progress.

Assumption 2 excludes discrete distributions; in such cases, our methods are conservative (like the KS test). The support is taken to be $\mathbb{R}$ for simplicity; any subset of $\mathbb{R}$ is fine. The continuity in Assumption 2 allows the probability integral transform to be used, $F(X_i) \overset{iid}{\sim}$ Unif$(0,1)$. The strict monotonicity implies the CDF is invertible (without having to define the generalized inverse), so $F^{-1}(\cdot)$ is the quantile function, and $F^{-1}\big(F(r)\big) = r$ as well as $F\big(F^{-1}(\tau)\big) = \tau$.

Third, we address the following tasks.

Task 1 One-sample, two-sided testing of $H_{0\tau} : F^{-1}(\tau) = F_0^{-1}(\tau)$ for $\tau \in (0,1)$ and fixed $F_0^{-1}(\cdot)$, with strong control of FWER.

Task 2 Same as Task 1 but with one-sided $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ or $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$.

Task 3 Two-sample, two-sided testing of $H_{0r} : F_X(r) = F_Y(r)$ for $r \in \mathbb{R}$, with strong control of FWER.

Task 4 Same as Task 3 but with one-sided $H_{0r} : F_X(r) \leq F_Y(r)$ or $H_{0r} : F_X(r) \geq F_Y(r)$.

Unlike with a GOF test, which has a single global hypothesis and corresponding single decision (reject or not), Tasks 1–4 each involve a continuum of pointwise hypotheses that each require a decision. Given Assumption 2, testing CDFs or quantile functions are in a sense equivalent: if $F(r) > F_0(r) = \tau$, then $F^{-1}(\tau) < F_0^{-1}(\tau)$. However, in the two-sample setting, it is easier to prove FWER control for the tests of CDF hypotheses; see Section 3 and the proofs of Lemma 2 and Proposition 3.

[5] We do not allow the population distribution to drift asymptotically to consider "uniformity," but we conjecture that at least our basic methods do not suffer such issues.

3 KS-based multiple testing procedures

The one-sample and two-sample KS GOF tests are well known, including the simulation of finite-sample exact p-values. We present the corresponding "KS-based" multiple testing procedures (MTPs) and establish their strong control of FWER. Although seemingly intuitive, we are unaware of such a presentation in the literature. Last in this section, we discuss the problem of uneven sensitivity.

For the one-sample, two-sided KS-based MTP, given notation in Assumptions 1 and 2, let

$$\hat{F}(x) \equiv \sum_{i=1}^{n} (1/n)\, \mathbb{1}\{X_i \leq x\}, \quad D_n^{x,0} \equiv \sqrt{n}\big|\hat{F}(x) - F_0(x)\big|, \tag{1}$$

for all $x \in \mathbb{R}$. Let $c_n(\alpha)$ denote the exact critical value with sample size n, so

$$\mathrm{P}\big(D_n > c_n(\alpha)\big) = \alpha, \quad D_n \equiv \sup_{x\in\mathbb{R}} D_n^x, \quad D_n^x \equiv \sqrt{n}\big|\hat{F}(x) - F(x)\big|. \tag{2}$$

It is well known that $c_n(\alpha)$ is distribution-free, depending only on n and α. Alternatively, the asymptotic $c_\infty(\alpha)$ can be used, such that

$$\mathrm{P}\left(\sup_{t\in[0,1]} |B(t)| > c_\infty(\alpha)\right) = \alpha$$

for standard Brownian bridge $B(\cdot)$.

The KS test proper is a GOF test that rejects $H_0 : F(\cdot) = F_0(\cdot)$ when $\sup_{x\in\mathbb{R}} D_n^{x,0} > c_n(\alpha)$. Under H_0, this occurs with probability α.

The corresponding MTP addressing Task 1 is intuitive: reject $H_{0x} : F(x) = F_0(x)$ for any $x \in \mathbb{R}$ such that $D_n^{x,0} > c_n(\alpha)$. (To directly address Task 1: if H_{0x} is rejected, then $H_{0\tau}$ is rejected for $\tau = F_0(x)$.) Weak control of FWER is immediate from the GOF test's size control: when $F(\cdot) = F_0(\cdot)$, the probability of at least one pointwise rejection is equivalent to the probability of $\sup_{x\in\mathbb{R}} D_n^{x,0} > c_n(\alpha)$, which is exactly α. Strong control of FWER is also straightforward to establish.

Proposition 1. *Let Assumptions 1 and 2 hold, as well as the definitions in* (1) *and* (2)*. The two-sided exact (or asymptotic) KS-based MTP that rejects $H_{0x} : F(x) = F_0(x)$ for any $x \in \mathbb{R}$ where $D_n^{x,0}$ exceeds the critical value has strong control of exact (or asymptotic) FWER. The corresponding one-sided KS-based MTPs of $H_{0x} : F(x) \leq F_0(x)$ or $H_{0x} : F(x) \geq F_0(x)$ also have strong control of FWER.*

Proof. As in Definition 1, let $I \equiv \{x : H_{0x} \text{ is true}\} \subseteq \mathbb{R}$. For the two-sided case, using (1)

and (2),

$$\text{FWER} \equiv \mathrm{P}\left(\sup_{x\in I} D_n^{x,0} > c_n(\alpha)\right) = \mathrm{P}\left(\sup_{x\in I} D_n^{x} > c_n(\alpha)\right) \leq \overbrace{\mathrm{P}\left(\sup_{x\in \mathbb{R}} D_n^{x} > c_n(\alpha)\right) = \alpha}^{\text{by (2)}}.$$

The one-sided case is similar and shown in the appendix. □

In the two-sample case, let

$$\hat{F}_X(r) \equiv \sum_{i=1}^{n_X} (1/n_X)\, \mathbb{1}\{X_i \leq r\}, \quad \hat{F}_Y(r) \equiv \sum_{i=1}^{n_Y} (1/n_Y)\, \mathbb{1}\{Y_i \leq r\}, \tag{3}$$

$$D^r_{n_X,n_Y} \equiv \sqrt{\frac{n_X n_Y}{n_X + n_Y}} \big|\hat{F}_X(r) - \hat{F}_Y(r)\big|, \tag{4}$$

for all $r \in \mathbb{R}$. Under $H_0 : F_X(\cdot) = F_Y(\cdot)$, the critical value $c_{n_X,n_Y}(\alpha)$ is again distribution-free, and it converges to the same $c_\infty(\alpha)$ as for the one-sample test as $n_X, n_Y \to \infty$ at the same rate. However, in finite samples, we only have an inequality: under H_0,

$$\mathrm{P}\big(D_{n_X,n_Y} > c_n(\alpha)\big) \leq \alpha, \quad D_{n_X,n_Y} \equiv \sup_{r\in\mathbb{R}} D^r_{n_X,n_Y}. \tag{5}$$

Equality is impossible for most α because the distribution of D_{n_X,n_Y} is discrete: it depends only on the ordering (i.e., permutation) of the X_i and Y_i, and with finite n_X and n_Y the number of such orderings is finite.

The corresponding two-sample MTP addressing Task 3 is intuitive: reject $H_{0r} : F_X(r) = F_Y(r)$ for any $r \in \mathbb{R}$ such that $D^r_{n_X,n_Y} > c_{n_X,n_Y}(\alpha)$. Weak control of FWER is again immediate from the GOF test's size control. Strong control of FWER can also be established, as in Proposition 3.[6] The key is that, given α, n_X, and n_Y, rejection of H_{0r} depends only on $\hat{F}_X(r)$ and $\hat{F}_Y(r)$, whose distributions are independent (by Assumption 1 and (3)) and depend only on $F_X(r)$ and $F_Y(r)$. Such a property extends over multiple r values jointly, too. This allows us to link the FWER with a probability under $F_X(\cdot) = F_Y(\cdot)$, which is bounded by the size of the global GOF test. (Implicitly, this was actually the one-sample argument for Proposition 1, too.) Since this general proof structure is used later, part of the argument is given in the following lemma.

Lemma 2. *Let Assumptions 1 and 2 hold. Consider any MTP for Task 3 or Task 4. Assume it has weak control of FWER at level α. Assume that, given α, n_X, and n_Y, rejection of*

[6] Asymptotically, and usually not framed in terms of FWER, stronger results in more complex models exist, such as the nonparametric, uniform (over τ) confidence band for the difference of two conditional quantile processes in Qu and Yoon (2015, §6.2), or the "uniform inference" on the quantile treatment effect process in Firpo and Galvao (2015, §4).

H_{0r} depends only on $\hat{F}_X(r)$ and $\hat{F}_Y(r)$, for any $r \in \mathbb{R}$. Then, the MTP has strong control of FWER at level α.

Proposition 3. *Let Assumptions 1 and 2 hold, as well as the definitions in (3)–(5). The two-sided exact (or asymptotic) KS-based MTP that rejects $H_{0r} : F_X(r) = F_Y(r)$ for any $r \in \mathbb{R}$ where $D^r_{n_X,n_Y}$ exceeds the critical value has strong control of exact (or asymptotic) FWER. The corresponding one-sided KS-based MTPs of $H_{0r} : F_X(r) \leq F_Y(r)$ or $H_{0r} : F_X(r) \geq F_Y(r)$ also have strong control of FWER.*

Although strong control of FWER is helpful, the KS-based MTPs suffer from uneven sensitivity to deviations from the null. One symptom of this was seen in the example in Section 1, where the one-sample KS test could not reject that the population was Uniform$(0, 1)$ even with five out of $n = 20$ observations exceeding one million. More generally, the one-sample, two-sided KS test does not reject at a 10% level even if $F_0(X_{20:16}) = 1$ or if $F_0(X_{20:5}) = 0$, which any reasonable test should and which our test does. This uneven sensitivity results in "low power in the tails" and is well documented in the literature. For example, Eicker (1979) says that the KS is "sensitive asymptotically only in the central range given by $\{\tau : (\log\log n)^{-1} < F(\tau) < 1 - (\log\log n)^{-1}\}$" (p. 117).

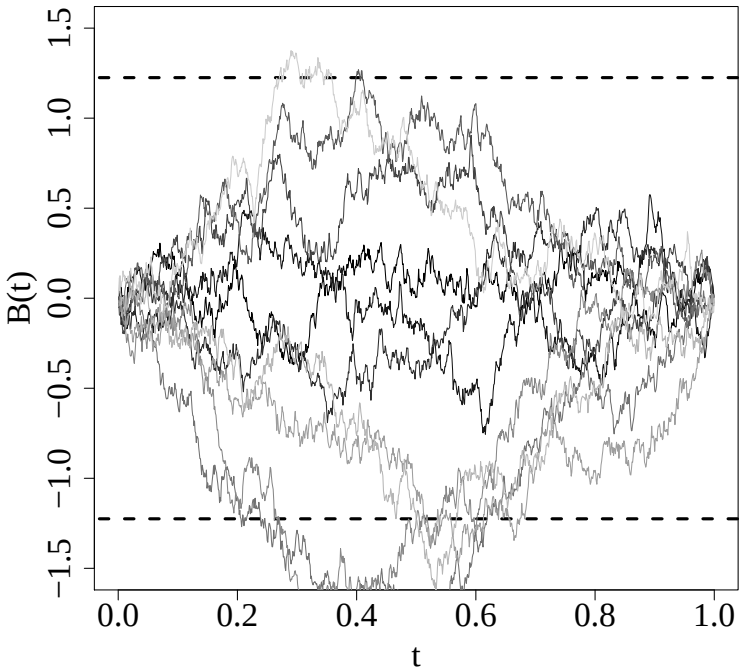

Figure 1: Sample paths of standard Brownian bridge $B(\cdot)$, with asymptotic 10% two-sided KS critical value ± 1.225. Six paths are (randomly) chosen among those exceeding the threshold; four are chosen that do not.

Figure 1 visualizes the intuition for the KS test's low sensitivity in the tails. The figure

shows sample paths (realizations) of a standard Brownian bridge along with the two-sided asymptotic 10% KS critical value, 1.225. That is, $\mathrm{P}\big(\sup_{t\in[0,1]}|B(t)| > 1.225\big) = 0.1$, so only 10% of sample paths wander above 1.225 or below -1.225, leading to (asymptotic) weak control of FWER. (The figure oversamples paths exceeding the critical value to avoid a visual mess.) However, as can be seen, such excesses are much more likely to occur near the median ($t = 0.5$) than in the tails. In the figure, the six paths deviating beyond the critical value do so only in $t \in [0.25, 0.75]$, and the clumping of sample paths near $B(t) = 0$ for t near zero or one shows the difficulty of having a large deviation in the tails. This is due to the pointwise variance being highest at $t = 0.5$, with $\mathrm{Var}\big(B(0.5)\big) = t(1-t) = 0.25$, and lowest as $t \to 0$ or $t \to 1$, where the variance approaches zero.

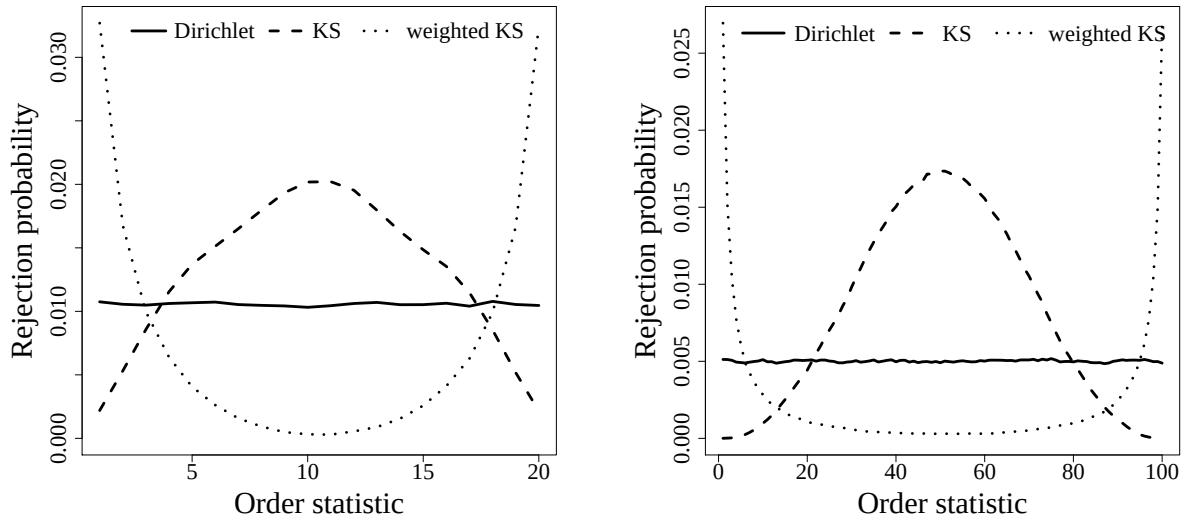

Figure 2: Simulated "pointwise" RP ($\tilde{\alpha}$) at each order statistic, $H_0 : F(\cdot) = F_0(\cdot)$, $n = 20$ (left) and $n = 100$ (right), 10^6 replications. Overall FWER for all MTPs is exactly $\alpha = 0.1$.

Figure 2 visualizes (un)even sensitivity from a different perspective. The KS-based MTP is labeled "KS" in the legend; the "weighted KS" is a weighted version described below. "Dirichlet" is our new MTP, detailed in Section 4. FWER is exactly 10% for all methods shown, but sensitivity is allocated differently across the distribution. The probability of $X_{n:k}$ causing some $H_{0r} : F(r) = F_0(r)$ to be rejected is simulated[7] for $k = 1, \ldots, n$. The resulting pattern is a more systematic account of the intuition in Figure 1: the pointwise rejection probability (RP) due to central order statistics is much higher than the RP due to extreme order statistics, and RP goes to almost zero at the sample minimum and maximum. This pattern is already clear with $n = 20$ (left) and becomes more exaggerated with $n = 100$ (right). Practically, this shows that although one is technically testing across the entire

[7]The simulation uses a standard uniform distribution for $F(\cdot)$, but the results are distribution-free: one could simply transform the data by $F_0(\cdot)$ and test against a standard uniform.

distribution, the KS-based MTP (implicitly) weights the middle of the distribution much more than the tails, which may not be desired. The corresponding uneven allocation of KS pointwise power is illustrated in Appendix D.1.

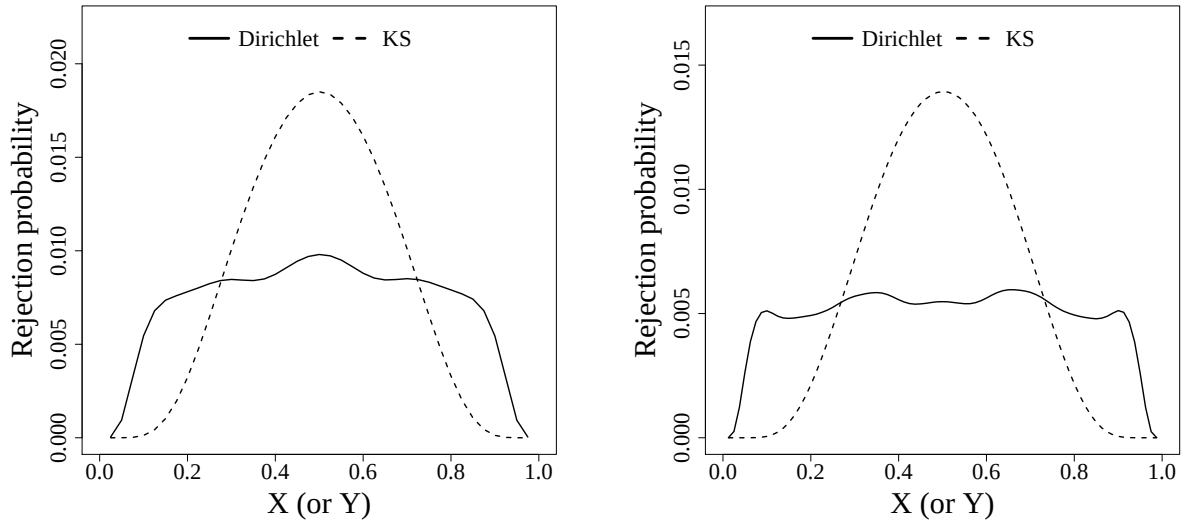

Figure 3: Simulated pointwise RP ($\tilde{\alpha}$), $n_X = n_Y = 40$ (left) and $n_X = n_Y = 80$ (right), FWER for both MTPs is exactly $\alpha = 0.1$, 10^6 replications, $F_X = F_Y = \text{Uniform}(0, 1)$.

Figure 3 shows the same qualitative pattern for the two-sample KS-based MTP. The horizontal axis now just shows the value $r \in [0, 1]$ for which $H_{0r} : F_X(r) = F_Y(r)$ is rejected. As in Figure 2, the pointwise RP peaks near the median and goes to zero in the tails.

If this uneven sensitivity stems from Figure 1 having uneven pointwise variance $t(1-t)$, then the simple solution is to divide by the standard deviation $\sqrt{t(1-t)}$ to achieve equal, unit variance at each t. In the one-sample KS context, this means dividing by $\sqrt{F_0(t)[1-F_0(t)]}$. Anderson and Darling (1952) gave exactly this solution in their Example 2 (p. 202), where they note it was first suggested by L. J. Savage (Footnote 2); they say, "In a certain sense, this function assigns to each point of the distribution $F(x)$ equal weights" (pp. 202–203).[8] However, they note that their results require the tails to have zero weight (pp. 210–211), which undermines the goal of even sensitivity. If nonetheless the weight is applied even in the tails, then the tails become overly sensitive. This is characterized by Eicker (1979, p. 117) as the weighted KS being "sensitive only in the moderate tails given by, e.g., $\left\{\tau : n^{-1} \log n < F(\tau) < \left((\log \log n) \log \log \log n\right)^{-1}\right\}$." Other discussions of the unintended (bad) consequences of this weighting scheme are found in Jaeschke (1979) and Lockhart (1991), among others. For the corresponding MTP, the "weighted KS" line

[8]The same paper includes a weighted Cramér–von Mises test that is most commonly called the Anderson–Darling test.

in Figure 2 shows that, indeed, the pointwise RP is much higher in the tails than near the median.

Figures 2 and 3 each show a line labeled "Dirichlet" that achieves a great degree of even sensitivity. These are the basic MTPs we propose in Sections 4 and 5.

4 One-sample Dirichlet approach

We propose methods for multiple testing of quantiles based on the probability integral transform and Dirichlet distribution, including stepdown and pre-test procedures to improve power. The Dirichlet distribution is used for GOF testing and uniform confidence bands in Buja and Rolke (2006), but our methods, their properties, and the quantile multiple testing framework itself are novel.

The Dirichlet approach uses the same pointwise type I error rate, $\tilde{\alpha}$, for multiple quantile tests across the distribution, while choosing the value of $\tilde{\alpha}$ to ensure strong control of finite-sample FWER at level α.

4.1 Basic method, FWER, and computation

All of our methods use the probability integral transform. The following results are from Wilks (1962, pp. 236–238), with some notational changes.

Theorem 4 (Wilks **8.7.1**, **8.7.2**, **8.7.4**). *The following are true under Assumptions 1 and 2. Denote the order statistics by $X_{n:1} < \cdots < X_{n:n}$. Then $F(X_{n:1})$, $F(X_{n:2}) - F(X_{n:1})$, ..., $F(X_{n:n}) - F(X_{n:n-1})$, $1 - F(X_{n:n})$ are random variables jointly following the $(n+1)$-variate Dirichlet distribution $Dir(1, \ldots, 1)$. That is, the random variables $F(X_{n:1})$, ..., $F(X_{n:n})$ have the ordered n-variate Dirichlet distribution $Dir^*(1, \ldots, 1; 1)$, with marginals $F(X_{n:k}) \sim Beta(k, n+1-k)$.*

Theorem 4 determines the finite-sample size of a single quantile test based on an order statistic. Specifically, consider the test of $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ that rejects when $X_{n:k} < F_0^{-1}(\tau)$ for some k. Under H_0, the type I error rate is bounded (tightly) by

$$\mathrm{P}\big(X_{n:k} < F_0^{-1}(\tau)\big) \leq \mathrm{P}\big(X_{n:k} < F^{-1}(\tau)\big) = \mathrm{P}\big(F(X_{n:k}) < \tau\big) = \mathrm{P}\big(\mathrm{Beta}(k, n+1-k) < \tau\big), \tag{6}$$

i.e., the Beta$(k, n+1-k)$ CDF evaluated at τ. This CDF can be computed immediately by any modern statistical software. The only difficulty is if a specific α is desired for a specific τ, in which case one cannot find an exact, non-randomized test (but for solutions using interpolation, see Beran and Hall, 1993; Goldman and Kaplan, 2016a).

Here, instead of testing a specific τ, we presume that the econometrician desires to test a wide range of quantiles. By choosing τ values that allow exact testing using order statistics $X_{n:k}$, we can get finite-sample results for a growing number (n) of quantiles. For comparison, Goldman and Kaplan (2016b) provide a confidence set for a fixed number of exactly pre-specified τ (and α), with $O(n^{-1})$ coverage probability error.

Our strategy is to use each of the n order statistics to test a different τ-quantile null hypothesis. Pointwise, let $B_{k,n}^{\tilde{\alpha}}$ denote the $\tilde{\alpha}$-quantile of the Beta$(k, n+1-k)$ distribution. For any $\tilde{\alpha} \in (0, 0.5)$, let the tested quantiles be

$$\ell_k \equiv B_{k,n}^{\tilde{\alpha}}, \quad u_k \equiv B_{k,n}^{1-\tilde{\alpha}}, \tag{7}$$

so $\mathrm{P}\big(X_{n:k} < F^{-1}(\ell_k)\big) = \tilde{\alpha}$ and $\mathrm{P}\big(X_{n:k} > F^{-1}(u_k)\big) = \tilde{\alpha}$, using (6). A one-sided test of $H_{0\ell_k} : F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k)$ that rejects when $X_{n:k} < F_0^{-1}(\ell_k)$ thus has exact size $\tilde{\alpha}$, and similarly for the test of $H_{0u_k} : F^{-1}(u_k) \leq F_0^{-1}(u_k)$ that rejects when $X_{n:k} > F_0^{-1}(u_k)$. By using the same $\tilde{\alpha}$ across the entire distribution, we achieve even sensitivity. More precisely, we achieve the same finite-sample size $\tilde{\alpha}$ for the pointwise tests at the n quantile indices ℓ_k or u_k, $k = 1, \ldots, n$.

Moreover, using the Dirichlet distribution in Theorem 4, one can solve for the $\tilde{\alpha}$ value such that

$$\alpha = 1 - \mathrm{P}\Big(\bigcap_{k=1}^{n} X_{n:k} \geq F^{-1}(\ell_k)\Big) = 1 - \mathrm{P}\Big(\bigcap_{k=1}^{n} F(X_{n:k}) \geq \ell_k\Big). \tag{8}$$

This choice of $\tilde{\alpha}$ achieves strong control of finite-sample FWER at level α. To see this, let $K \equiv \{k : F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k)\}$, the set of true hypotheses. Then,

$$\begin{aligned}
\mathrm{FWER} &= \overbrace{1 - \mathrm{P}(\text{no rejections among } k \in K)}^{\text{by definition of FWER}} = \overbrace{1 - \mathrm{P}\Big(\bigcap_{k \in K} X_{n:k} \geq F_0^{-1}(\ell_k)\Big)}^{\text{by definition of } H_{0\ell_k}} \\
&\leq \overbrace{1 - \mathrm{P}\Big(\bigcap_{k \in K} X_{n:k} \geq F^{-1}(\ell_k)\Big)}^{\text{because } F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k) \text{ for all } k \in K, \text{ by definition of } K} \leq \overbrace{1 - \mathrm{P}\Big(\bigcap_{k=1}^{n} X_{n:k} \geq F^{-1}(\ell_k)\Big)}^{\text{because } K \subseteq \{1,2,\ldots,n\}} \\
&= \overbrace{\alpha}^{\text{from (8)}}
\end{aligned}$$

A parallel argument applies to the other one-sided case with u_k.

To extend testing of the n quantiles to the continuum of $H_{0\tau}$ for all $\tau \in (0, 1)$, without affecting FWER, monotonicity of the quantile function is sufficient. If $X_{n:k} < F_0^{-1}(\ell_k)$, then $H_{0\ell_k} : F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k)$ is rejected. If $X_{n:k} < F_0^{-1}(\tau)$ for another $\tau < \ell_k$, then $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ is also rejected. If we add this event, or rather its complement, into

(8), however, it disappears because

$$\{F(X_{n:k}) \geq \ell_k\} \cap \{F(X_{n:k}) \geq \tau\} = \{F(X_{n:k}) \geq \ell_k\}$$

since the event $\{F(X_{n:k}) \geq \tau\} \supset \{F(X_{n:k}) \geq \ell_k\}$ for any $\tau < \ell_k$.

The full one-sided and two-sided MTPs are now described, followed by their strong control of exact FWER, and finally a computational improvement.

Method 1. For Task 2, consider $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$. Let $\ell_k \equiv B_{k,n}^{\tilde{\alpha}}$. Solve for $\tilde{\alpha}$ from (8), using Theorem 4.[9] For every $\tau \in (0,1)$, reject $H_{0\tau}$ if and only if $F_0^{-1}(\tau) > \min\{X_{n:k} : \ell_k \geq \tau\}$. For $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$, replace ℓ_k with $u_k \equiv B_{k,n}^{1-\tilde{\alpha}}$ and reverse all inequalities. □

Method 2. For Task 1, let $\ell_k \equiv B_{k,n}^{\tilde{\alpha}}$ and $u_k \equiv B_{k,n}^{1-\tilde{\alpha}}$. Using Theorem 4, solve for $\tilde{\alpha}$ from

$$\alpha = 1 - \mathrm{P}\Big(\bigcap_{k=1}^{n}\{F^{-1}(\ell_k) \leq X_{n:k} \leq F^{-1}(u_k)\}\Big) = 1 - \mathrm{P}\Big(\bigcap_{k=1}^{n}\{\ell_k \leq F(X_{n:k}) \leq u_k\}\Big), \qquad (9)$$

or use the $\tilde{\alpha}$ approximation in Proposition 6. For every $\tau \in (0,1)$, reject $H_{0\tau}$ if and only if $F_0^{-1}(\tau) > \min\{X_{n:k} : \ell_k \geq \tau\}$ or $F_0^{-1}(\tau) < \max\{X_{n:k} : u_k \leq \tau\}$. □

Theorem 5. *Under Assumptions 1 and 2, Methods 1 and 2 have strong control of finite-sample FWER.*

Additionally, we contribute a new, fast approximation that can be used not only for Methods 1 and 2 but also to compute GOF p-values as well as uniform confidence bands for $F(\cdot)$. Solving (9) requires approximating the n-variate Dirichlet distribution by either numerical integration or simulation (i.e., drawing standard uniform order statistics), which can be slow for large n. Extensive simulations have revealed a closed-form formula to approximate the necessary $\tilde{\alpha}$ as a function of α and n with a high degree of accuracy. Computation now takes only a couple seconds even for $n = 100\,000$.

Proposition 6. *Under Assumptions 1 and 2, for* $\alpha \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 0.9\}$ *and* $n \in [4, 10^6]$*, for two-sided testing,*

$$\tilde{\alpha} = \exp\Big\{-c_1(\alpha) - c_2(\alpha)\sqrt{\ln[\ln(n)]} - c_3(\alpha)[\ln(n)]^{c_4(\alpha)}\Big\},$$

with $c_1(\alpha) = -2.75 - 1.04\ln(\alpha)$*,* $c_2(\alpha) = 4.76 - 1.20\alpha$*,* $c_3(\alpha) = 1.15 - 2.39\alpha$*, and* $c_4(\alpha) = -3.96 + 1.72\alpha^{0.171}$*, provides an approximate solution to* (9)*. Define the relative approximation error to be* $(\alpha^* - \alpha)/\alpha$*, where* α *is the nominal FWER and* α^* *is the true FWER (i.e.,*

[9] Alternatively, use the $\tilde{\alpha}$ approximation in Proposition 6 after adjusting the one-sided α to two-sided $\alpha_2 = 2\alpha - \alpha^2$ per Theorem 5.1 of Moscovich et al. (2016); details below.

simulated with 10^6 *replications). Then, across all* α *and* n *listed above, the relative approximation error never exceeds* 20% *in absolute value. Excluding* $\alpha = 0.001$*, absolute relative approximation error never exceeds* 11% *(e.g., worst-case FWER is* 0.111 *when* $\alpha = 0.1$*).*

For one-sided multiple testing, to apply Proposition 6, the initial α can be adjusted using Theorem 5.1 of Moscovich et al. (2016), which is asymptotically exact and slightly conservative in finite samples. Specifically, the two-sided FWER α_2 and one-sided FWER α_1 (either lower or upper) are related by $\alpha_2 = 2\alpha_1 - \alpha_1^2$ asymptotically; in finite samples, $2\alpha_1 \geq \alpha_2 \geq 2\alpha_1 - \alpha_1^2$.

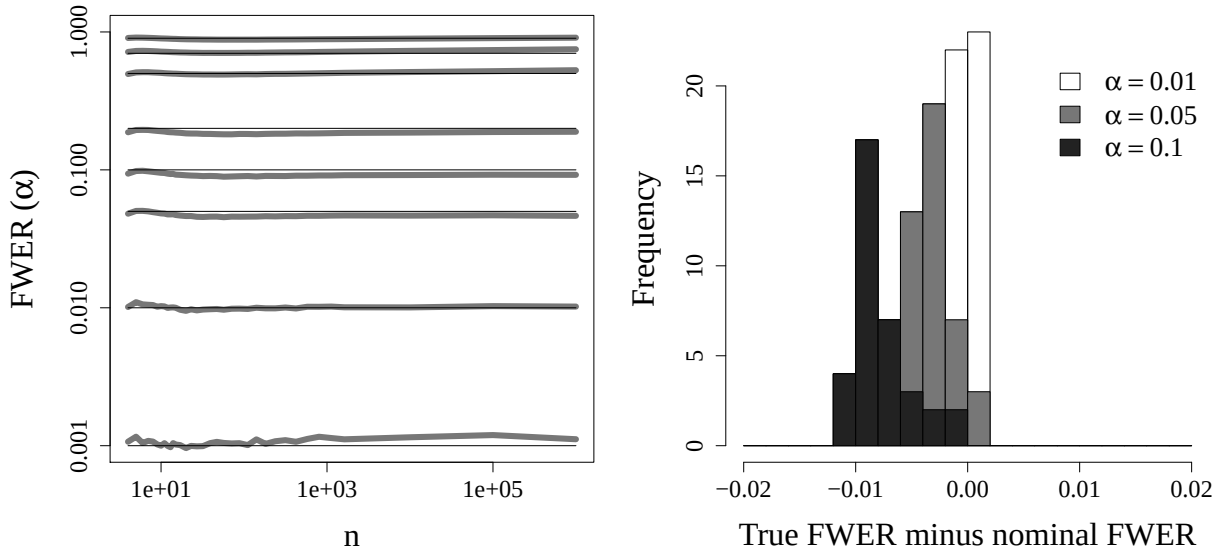

Figure 4: Histograms showing the accuracy of Proposition 6 for sample sizes $n \in \{4, 5, \ldots, 14\}$, $n = \lfloor \exp\{\exp(\kappa)\} \rfloor$ for $\kappa = 1.00, 1.05, \ldots, 2.00$, and $n \in \{10^4, 10^5, 10^6\}$. "True" FWER is from 10^6 simulation replications. Left: thick gray lines show true FWER; thin black lines show nominal FWER. Right: FWER error (true minus nominal) for $\alpha \in \{0.01, 0.05, 0.1\}$; note that the total height of each bar is the total across all three α values.

The accuracy of the Proposition 6 formulas is shown in Figure 4. As stated in Proposition 6, across all α and n, the relative FWER error never exceeds 20% (and is usually close to zero), or 11% when excluding $\alpha = 0.001$. For example, for $\alpha = 0.1$, true FWER is always between 0.089 and 0.111. To see more specific values, the right panel of Figure 4 shows a histogram of true minus nominal FWER differences. For the most commonly used $\alpha \in \{0.01, 0.05, 0.1\}$, these are seen to err slightly on the conservative side (i.e., true FWER is below nominal) and are often close to zero. In our code, we use a version of Proposition 6 with coefficients specific to α, which further increases the accuracy.

By monotonicity of the mapping $\tilde{\alpha}(\alpha, n)$ in α and n, additional approximation error from interpolation between the given values of n and α is small. We conjecture that the

formulas are accurate even outside the ranges given, especially in n; moreover, few economic applications require $n > 10^6$, $\alpha < 0.001$, or $\alpha > 0.9$.

Proposition 6 may also be used to quickly compute GOF p-values and uniform confidence bands; see Buja and Rolke (2006), Aldor-Noiman et al. (2013), our code, and an earlier version of this paper for details.

4.2 Procedures to improve power

4.2.1 Stepdown procedure

Within our quantile multiple testing framework, a stepdown procedure to improve power is possible. The general stepdown strategy dates back to Holm (1979); see also Lehmann and Romano (2005b, Ch. 9). Although the stepdown procedure strictly improves power (against false hypotheses), it does not "dominate" in the general decision-theoretic sense since FWER also increases in some cases, though never exceeding α.

In our context, consider one-sided multiple testing with ℓ_k. If at least one $H_{0\ell_k}$ is rejected by the basic MTP, then we proceed to test the remaining hypotheses as if the rejected ones are indeed false. In that case, we can remove such ℓ_k from the calibration equation (8). Intuitively, with fewer true quantile hypotheses, we can test the remainder with greater pointwise size while maintaining the same overall FWER control. Mechanically, this is accomplished by changing which order statistic is used to test a hypothesis, to make it more likely to reject.[10] As in other stepdown procedures, the "trick" is that if one of the initially rejected hypotheses was in fact true, then a "familywise error" has already been made, so rejecting additional hypotheses has no effect on the FWER.

Method 3. For Task 2, let $\hat{K}_0 \equiv \{1, \ldots, n\}$, and let $r_{k,0} = k$ for $k \in \hat{K}_0$. Consider $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$. Let $\ell_k = B_{k,n}^{\tilde{\alpha}}$, where (as in Method 1) $\tilde{\alpha}$ satisfies

$$\alpha = 1 - \mathrm{P}\left(\bigcap_{k \in \hat{K}_0} X_{n:r_{k,0}} \geq F^{-1}(\ell_k) \right) = 1 - \mathrm{P}\left(\bigcap_{k \in \hat{K}_0} F(X_{n:r_{k,0}}) \geq \ell_k \right). \tag{10}$$

Theorem 4 determines the joint distribution of the $F(X_{n:k})$ in (11) and thus the probability. Reject $H_{0\tau}$ if $F_0^{-1}(\tau) > \min\{X_{n:k} : \ell_k \geq \tau\}$ (as in Method 1). Then, increment i to $i = 1$ and iterate the following.

Step 1. Let $\hat{K}_i = \{k : H_{0\ell_k} \text{ not yet rejected}\}$. If $\hat{K}_i = \emptyset$ or $\hat{K}_i = \hat{K}_{i-1}$, then stop.

[10] If instead $\tilde{\alpha}$ were increased, then the ℓ_k would change. This may be reasonable practically, but it would complicate matters and require a different strategy to prove strong control of FWER.

Step 2. Choose integers $r_{k,i} \leq r_{k,i-1}$ (based only on $\hat{K}_i$) satisfying:[11]

$$\alpha \geq 1 - \mathrm{P}\left(\bigcap_{k \in \hat{K}_i} F(X_{n:r_{k,i}}) \geq \ell_k \right). \tag{11}$$

Step 3. Reject any additional $H_{0\tau}$ for which $F_0^{-1}(\tau) > \min\{X_{n:r_{k,i}} : \ell_k \geq \tau, k \in \hat{K}_i\}$.
Step 4. Increment i by one and return to Step 1.

For $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$, replace ℓ_k with $u_k = B_{k,n}^{1-\tilde{\alpha}}$ and reverse the inequalities. □

Method 6 in Appendix A describes a two-sided stepdown procedure.

Theorem 7. *Under Assumptions 1 and 2, Methods 3 and 6 have strong control of finite-sample FWER.*

4.2.2 Pre-test procedure

For one-sided multiple testing, a pre-test (actually "pre-MTP") can improve pointwise power. It can also improve power of the corresponding global test, which in the one-sided case has first-order stochastic dominance as the null hypothesis.[12] Intuitively, the basic MTP for $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ over $\tau \in (0,1)$ must control FWER even for the most difficult $F(\cdot)$, where $F^{-1}(\tau) = F_0^{-1}(\tau)$ for all $\tau \in (0,1)$. In the GOF context, this $F^{-1}(\cdot) = F_0^{-1}(\cdot)$ is commonly called the "least favorable configuration": it is the $F(\cdot)$ that maximizes the type I error rate. If (by pre-testing) $F(\cdot)$ can be restricted to a subset that excludes the least favorable configuration, then we can increase RPs while still controlling FWER.

Specifically, the pre-test determines at which τ the constraint $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ appears slack, i.e., where we can reject $F^{-1}(\tau) \leq F_0^{-1}(\tau)$ in favor of $F^{-1}(\tau) > F_0^{-1}(\tau)$. Then, we recalibrate $\tilde{\alpha}$ using only the unrejected $H_{0\tau}$ to improve power.

Falsely inferring that the constraint is slack leads to over-rejection of the resulting MTP, so the probability of doing so should be small. This probability is the FWER of the pre-test. If α_p is the FWER level of the pre-test, then $\alpha_p \to 0$ ensures zero asymptotic size distortion.[13] Of course, in any finite sample, $\alpha_p > 0$, so α_p should be tolerably small. We suggest $\alpha_p = \alpha / \ln[\ln(\max\{n, 15\})]$.

[11] This leaves many possibilities for $r_{k,i}$. In our code, we use a "greedy" algorithm (e.g., Sedgewick and Wayne, 2011, §4.3), iteratively decreasing (by one) whichever $r_{k,i}$ achieves the biggest pointwise RP increase, until none can be decreased without violating (11).

[12] Alternatively, one may test a null of non-dominance as suggested by Davidson and Duclos (2013); there are also big differences between frequentist and Bayesian inference for first-order stochastic dominance, even asymptotically and with nonparametric methods, as pointed out by Kaplan and Zhuo (2016).

[13] This idea is found in Linton et al. (2010), whose (13) has $c_N \to 0$, and in Donald and Hsu (2016), whose (3.4) has $a_N \to -\infty$, among others.

The pre-test implemented in our code is described in Method 7 (and its strong control of FWER in Proposition 10) in Appendix A. The overall method (of which the pre-test is the first step) is described in Method 4.

Method 4. For Task 2, consider $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ over $\tau \in (0,1)$. First pre-test $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$ for $\tau \in (0,1)$ using Method 7 with strong control of FWER at level $\alpha_p = \alpha / \ln[\ln(\max\{n, 15\})]$. Let $\hat{K}$ denote the set of k such that $H_{0\ell_k}$ was not rejected by the pre-test, defining ℓ_k as in (7). Then choose integers $r_k \geq k$ such that

$$\alpha \geq 1 - \mathrm{P}\left(\bigcap_{k\in\hat{K}} X_{n:r_k} \geq F^{-1}(\ell_k)\right) = 1 - \mathrm{P}\left(\bigcap_{k\in\hat{K}} F(X_{n:r_k}) \geq \ell_k\right),$$

computing the probability using Theorem 4.[14] Reject $H_{0\tau}$ when $\min\{X_{n:k} : \ell_k \geq \tau, k \in \hat{K}\} < F_0^{-1}(\tau)$.

For $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$, reverse inequalities and replace ℓ_k with u_k (from (7)). □

Theorem 8. *Under Assumptions 1 and 2, Method 4 has strong control of finite-sample FWER at level $\alpha + \alpha_p$, approaching α as $n \to \infty$.*

The FWER upper bound $\alpha+\alpha_p$ in Theorem 8 is usually far from binding. It assumes that a false pre-test rejection always leads to a false rejection, whereas in reality the probability is only somewhat increased. Simulations show the FWER level to be much closer to α than $\alpha + \alpha_p$.

5 Two-sample Dirichlet approach

5.1 Basic method and FWER

Similar to the two-sample KS-based MTP and the two-sample GOF test in Buja and Rolke (2006, §5.2), our two-sample MTP depends only on the ordering of X and Y observations (see below). The difference is which orderings trigger rejections of which hypotheses. Compared to the KS-based MTP, as seen in Figure 3, our MTP allocates pointwise size (and thus power) more evenly across the distribution.

Our two-sample MTP differs from the two-sample GOF test in Buja and Rolke (2006). Like our one-sample MTP, our two-sample MTP determines a particular $\tilde{\alpha}$ that depends only on the sample sizes and α, after which order statistics are compared to different beta distribution quantiles to determine rejection. The Buja and Rolke (2006, §5.2) GOF test

[14] Similar remarks to Footnote 11 apply.

uses permutations of the observed data. Our approach has two advantages. First, computationally, our MTP's pointwise $\tilde{\alpha}$ may be pre-computed (given α, n_X, and n_Y, as we have done in a large reference table), whereas permutations of observed data require just-in-time computation for each new dataset. Second, regarding FWER control, it is not clear that the MTP based on the Buja and Rolke (2006) GOF test satisfies the assumption of Lemma 2; it may still have strong control of FWER, but it would be more difficult to prove.

Method 5. Using (3) and (7), given $\tilde{\alpha}$, let

$$\begin{aligned} \hat{\ell}_X(r) &\equiv B^{\tilde{\alpha}}_{n_X \hat{F}_X(r), n_X}, \quad \hat{u}_X(r) \equiv B^{1-\tilde{\alpha}}_{n_X \hat{F}_X(r)+1, n_X}, \\ \hat{\ell}_Y(r) &\equiv B^{\tilde{\alpha}}_{n_Y \hat{F}_Y(r), n_Y}, \quad \hat{u}_Y(r) \equiv B^{1-\tilde{\alpha}}_{n_Y \hat{F}_Y(r)+1, n_Y}, \end{aligned} \tag{12}$$

defining $B^{\tilde{\alpha}}_{0,n} \equiv 0$ and $B^{\tilde{\alpha}}_{n+1,n} \equiv 1$ for any $\tilde{\alpha}$. For Task 3, reject $H_{0r} : F_X(r) = F_Y(r)$ when either $\hat{\ell}_X(r) > \hat{u}_Y(r)$ or $\hat{\ell}_Y(r) > \hat{u}_X(r)$. Using many simulated samples with $X_i \overset{iid}{\sim} \text{Unif}(0,1)$ for $i = 1, \ldots, n_X$ and independent $Y_j \overset{iid}{\sim} \text{Unif}(0,1)$ for $j = 1, \ldots, n_Y$, choose the largest $\tilde{\alpha}$ such that the (simulated) probability of rejecting any H_{0r} (i.e., FWER) is less than or equal to α, and then subtract 0.0001 to get the final $\tilde{\alpha}$.

For Task 4, reject $H_{0r} : F_X(r) \leq F_Y(r)$ when $\hat{\ell}_X(r) > \hat{u}_Y(r)$, or reject $H_{0r} : F_X(r) \geq F_Y(r)$ when $\hat{\ell}_Y(r) > \hat{u}_X(r)$. As above, simulate independent standard uniform datasets and choose the largest $\tilde{\alpha}$ such that the (simulated) probability of rejecting any H_{0r} (i.e., FWER) is less than or equal to α, and then subtract 0.0001 from $\tilde{\alpha}$. □

As seen in Method 5, whether or not there is at least one H_{0r} rejected (vs. zero rejected) depends only on the ordering of X and Y values in the sample. For example, if $n_X = n_Y = 2$, any sample with $X_{2:1} < X_{2:2} < Y_{2:1} < Y_{2:2}$ has the same ordering, XXYY; either all samples with that ordering reject at least one H_{0r} (possibly with different r), or all samples accept all H_{0r}. Consequently, there is only a finite number of α (FWER level) that may be achieved exactly. Equivalently, the GOF p-value distribution is discrete. The same issue of a finite number of attainable α applies to the two-sample KS approach since it also depends on (only) the ordering.

Similar to Methods 1 and 2, the key to FWER control for Method 5 is the choice of $\tilde{\alpha}$. Method 5 chooses $\tilde{\alpha}$ such that the FWER is no greater than α under the global null $H_0 : F_X(\cdot) = F_Y(\cdot)$; i.e., weak control of FWER is ensured by construction. Computationally, we now describe two alternative ways to simulate the mapping from $\tilde{\alpha}$ to α (given n_X and n_Y) when $F_X(\cdot) = F_Y(\cdot)$.

The first strategy for simulating α given $\tilde{\alpha}$ (and n_X and n_Y) employs a convenient, order-preserving transformation. From the probability integral transform (and Assumption 1), $F_Y(Y_i) \overset{iid}{\sim} \text{Unif}(0,1)$, and under $H_0 : F_X(\cdot) = F_Y(\cdot)$, then $F_Y(X_i) = F_X(X_i) \overset{iid}{\sim} \text{Unif}(0,1)$,

too. From Assumption 2, $F_Y(\cdot)$ is order-preserving. Thus, we may simulate independent standard uniform samples of sizes n_X and n_Y to compute the FWER of Method 5 given any $\tilde{\alpha}$, as suggested in Method 5. Since FWER is monotonic in $\tilde{\alpha}$, which is a scalar, a simple numerical search finds the $\tilde{\alpha}$ that leads to a desired α. These simulations may be done ahead of time to generate a reference table of $\tilde{\alpha}$ values, as we provide along with our code.

The second strategy for simulating α given $\tilde{\alpha}$ uses permutations. The distribution of $(X_1, \ldots, X_{n_X}, Y_1, \ldots, Y_{n_Y})$ under $H_0 : F_X(\cdot) = F_Y(\cdot)$ is the same as that of any permutation of that vector, satisfying the "randomization hypothesis" in Definition 15.2.1 of Lehmann and Romano (2005b), for example. Given this, Buja and Rolke (2006) propose a GOF test based on permutations of the observed data, implicitly following Theorem 15.2.1 of Lehmann and Romano (2005b). Alternatively, we follow Theorem 15.2.2 of Lehmann and Romano (2005b) and use the fact that each of the $\binom{n_X+n_Y}{n_X}$ orderings is equally likely under H_0. This argument is again distribution-free, so our $\tilde{\alpha}$ is only a function of α, n_X, and n_Y and can be computed ahead of time.

The strong control of FWER in Theorem 9 comes from the weak control of FWER (by construction) combined with Lemma 2. For the one-sided case, as shown formally in the proofs, $F_X(\cdot) = F_Y(\cdot)$ is the least favorable configuration, so FWER is only lower for other distributions satisfying $H_0 : F_X(\cdot) \leq F_Y(\cdot)$.

Theorem 9. *Under Assumptions 1 and 2, given an arbitrarily large number of simulations, Method 5 has strong control of finite-sample FWER at level α.*

In the proof of Theorem 9, the use of Lemma 2 would not be possible with quantile (rather than CDF) hypotheses. Rejection of quantile $H_{0\tau}$ depends on order statistics $X_{n_X:k}$ and $Y_{n_Y:m}$ for some k and m, but the finite-sample sampling distributions of the order statistics depend on more than just $F_X^{-1}(\tau)$ and $F_Y^{-1}(\tau)$. It may be possible to show strong control of FWER for quantiles with an argument more complicated than ours, but the marginal benefit of such an additional result seems minor in practice.

In the appendix, we propose methods to improve power via stepdown and pre-test procedures. Although performance in simulations is reasonable, these methods are not as elegant as the one-sample methods. They test fewer than n quantiles, and we provide only heuristic justification.

6 Simulations

All simulations may be replicated with code from the latter author's website. For comparison with our Dirichlet methods, we use KS-based MTPs as described in Section 3. The

unweighted KS-based MTP uses the KS implementation `ks.test` in the `stats` package in R (R Core Team, 2013). For the weighted KS, asymptotic critical values from Jaeschke (1979) and Chicheportiche and Bouchaud (2012) were inaccurate,[15] so we simulate exact critical values. However, this simulation is time-consuming, which is a practical disadvantage.

Earlier, Figures 2 and 3 showed simulation results on the uneven sensitivity of weighted and unweighted KS-based MTPs and the (relatively) even sensitivity of the Dirichlet MTPs, in terms of pointwise type I error rates. Intuitively, those differences translate into differences in pointwise power, as shown in Appendix D.1. The Dirichlet's more even sensitivity also achieves generally better global power than the (GOF) KS test. This is illustrated in Appendix D.1, as well as in Table 1 and Figure 8 of Aldor-Noiman et al. (2013), who also show a power advantage over the Anderson–Darling (i.e., weighted Cramér–von Mises) test for a variety of distributions. That is, there is not a tradeoff between even sensitivity and global power; the Dirichlet approach has both more even sensitivity and better global power.

In this section, we focus on our methods' strong control of FWER, the power improvements of stepdown and pre-test procedures, and the computational benefit of Proposition 6.

6.1 FWER

Table 1 shows weak control of FWER for one-sample, two-sided MTPs, i.e., it shows FWER simulated under $F(\cdot) = F_0(\cdot)$. Since all MTPs considered are distribution-free under Assumptions 1 and 2, the DGP is $X_i \overset{iid}{\sim} \text{Uniform}(0, 1)$. For our Method 2 ("Dirichlet"), this is exact by construction, up to the approximation error in Proposition 6. This error is negligible in Table 1. Earlier, Figure 4 showed simulated FWER for additional n and nominal α when using Proposition 6.

Table 1: Simulated FWER, one-sample, two-sided, $F(\cdot) = F_0(\cdot)$, 10^6 replications.

α	n	Dirichlet	KS	KS (exact)	weighted KS (exact)
0.10	20	0.101	0.100	0.100	0.099
0.10	100	0.101	0.094	0.100	0.098
0.05	20	0.050	0.050	0.050	0.053
0.05	100	0.050	0.045	0.050	0.049

Table 2 shows strong control of FWER for one-sample, one-sided MTPs: the basic Dirichlet test in Method 1, as well as the stepdown and pre-test procedures in Methods 3 and 4, respectively. The null distribution F_0 is Uniform$(-1, 1)$ and $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$. Fig-

[15] Jaeschke (1979, p. 108) appropriately warns, "Since... the rate of convergence... is very slow, we would not encourage anyone to use the confidence intervals based on the asymptotic analysis."

ure 5 shows $F_0^{-1}(\cdot)$ and $F^{-1}(\cdot)$ for each row in Table 2. The Dirichlet MTP in Method 1 always controls FWER, but well below the required level $\alpha = 0.1$ when $F(\cdot) \neq F_0(\cdot)$. The FWERs for the stepdown method and combined pre-test/stepdown method are higher but still below $\alpha = 0.1$, as desired. Of course, all else equal, a higher error rate is never desired, but (as seen later) there is a corresponding gain in power, so the tradeoff may be beneficial from a minimax risk sort of perspective: a slight increase in FWER when FWER is near zero is not very costly, while improving worst-case power near zero is very beneficial.

$H_{0\tau}$ true	$F^{-1}(\tau) = F_0^{-1}(\tau)$	Dirichlet	Stepdown	Pre+Step
$\tau \in [0, 1]$	$\tau \in [0, 1]$	0.101	0.101	0.101
$\tau \in [0, 0.5]$	$\tau \in [0, 0.5]$	0.048	0.083	0.082
$\tau \in [0, 1]$	$\tau \in [0.5, 1]$	0.068	0.068	0.079
$\tau \in [0, 0.5]$	$\tau \in \{0.5\}$	0.004	0.017	0.024

Table 2: FWER, nominal level $\alpha = 0.1$, $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$, $F_0 = \text{Unif}(-1, 1)$ so $F_0^{-1}(\tau) = 2(\tau - 0.5)$, $n = 100$, 1000 replications. For τ where $F^{-1}(\tau) \neq F_0^{-1}(\tau)$, $F^{-1}(\tau) = 4(\tau - 0.5)$; see Figure 5. "Dirichlet" is Method 1, "Stepdown" is Method 3, and "Pre+Step" is Methods 3 and 4 combined.

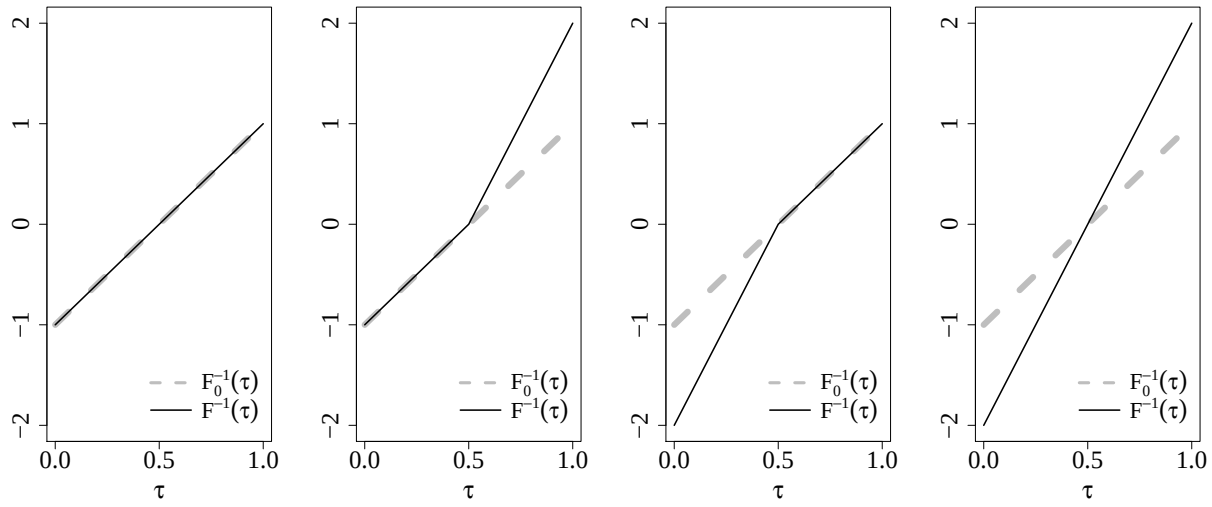

Figure 5: Null and true quantile functions, $F_0^{-1}(\cdot)$ and $F^{-1}(\cdot)$, for the four rows in Table 2.

Table 3 shows weak control of FWER for two-sample, two-sided MTPs, i.e., FWER under $H_0 : F_X(\cdot) = F_Y(\cdot)$. Since all MTPs shown are distribution-free in this case, both samples are iid Uniform$(0, 1)$.

Table 3 shows our MTP's nearly exact FWER. The asymptotic KS-based MTP is somewhat conservative in these cases, as is the "exact" KS-based MTP. The exact and asymptotic

Table 3: Simulated FWER, two-sample, two-sided, $F_X(\cdot) = F_Y(\cdot)$, 10^6 replications.

α	n_X	n_Y	Dirichlet	KS	KS (exact)
0.05	25	500	0.050	0.039	0.049
0.10	25	500	0.100	0.082	0.095
0.10	30	30	0.101	0.071	0.071
0.10	29	30	0.101	0.079	0.099
0.10	100	100	0.101	0.078	0.078
0.10	99	100	0.106	0.090	0.099

KS can be identical due to the discreteness of the GOF p-value distributions (as discussed in Section 5.1), if the exact and asymptotic p-values lie on the same side of $\alpha = 0.1$ for every possible data ordering (permutation). This discreteness makes the exact KS notably conservative when $n_X = n_Y = 30$ and even $n_X = n_Y = 100$, but the effect vanishes when reducing n_X by one so that $n_X \neq n_Y$. The effect of discreteness on the Dirichlet MTP is negligible in all cases.

Table 4 is the two-sample analog of Table 2, showing strong control of FWER for one-sided MTPs. Figure 5 again visualizes the quantile functions for each row of the table, but now $F_Y = F_0$ and $F_X = F$. We compare MTPs for Tasks 4 and 6 with $\alpha = 0.05$: the basic Dirichlet MTP in Method 5, the joint quantile difference MTP in Appendix A.2, the stepdown procedure in Method 8, and the combined pre-test/stepdown procedure in Method 9. For Method 5, we forgo the adjustment of α for one-sided testing in favor of using our $\tilde{\alpha}$ reference table for faster computation.

H_{0r} true	$H_{0\tau}$ true	$F_X^{-1}(\tau) = F_Y^{-1}(\tau)$	Basic	Joint	Stepdown	Pre+Step
$r \in [-1, 1]$	$\tau \in [0, 1]$	$\tau \in [0, 1]$	0.049	0.044	0.044	0.044
$r \in [-1, 0]$	$\tau \in [0, 0.5]$	$\tau \in [0, 0.5]$	0.031	0.031	0.044	0.044
$r \in [-1, 1]$	$\tau \in [0, 1]$	$\tau \in [0.5, 1]$	0.013	0.026	0.026	0.032
$r \in [-1, 0]$	$\tau \in [0, 0.5]$	$\tau \in \{0.5\}$	0.003	0.000	0.002	0.006

Table 4: FWER, $\alpha = 0.05$, $H_{0r} : F_X(r) \geq F_Y(r)$ or $H_{0\tau} : F_X^{-1}(\tau) \leq F_Y^{-1}(\tau)$, $F_Y = \text{Unif}(-1, 1)$ so $F_Y^{-1}(\tau) = 2(\tau - 0.5)$, $n_X = n_Y = 200$, 1000 replications. For τ where $F_X^{-1}(\tau) \neq F_Y^{-1}(\tau)$, $F_X^{-1}(\tau) = 4(\tau - 0.5)$; see Figure 5, where $F_Y = F_0$ and $F_X = F$. "Basic" is Method 5, "Joint" uses only iteration $i = 0$ from Method 8, "Stepdown" is Method 8, and "Pre+Step" is Method 9. The Basic test is evaluated at $r = -0.99, -0.98, \ldots, 0.99$.

Table 4 shows strong control of FWER for all four methods. The stepdown and pre-test procedures' strong control of FWER in Table 4 supports our heuristic arguments. Overall, the patterns are similar to those in Table 2.

6.2 Power comparison

We now illustrate the power improvement from the stepdown and pre-test procedures. For pointwise and global power comparisons with the KS-based MTP, see Appendix D.1.

For one-sample multiple testing, Figure 6 compares the pointwise (by τ) RPs of the same methods shown in Table 2. The DGPs are the same as the rows in Table 2 where $\{\tau : H_{0\tau} \text{ is true}\} = [0, 0.5]$, visualized in the second and fourth panels in Figure 5. Since all methods (correctly) have RP near zero for $\tau < 0.5$, only larger τ are shown. Compared with the basic Dirichlet MTP, the stepdown procedure weakly increases pointwise power, and adding the pre-test does, too. The pre-test is only helpful in the right panel where the null hypothesis constraint is slack for $\tau < 0.5$.

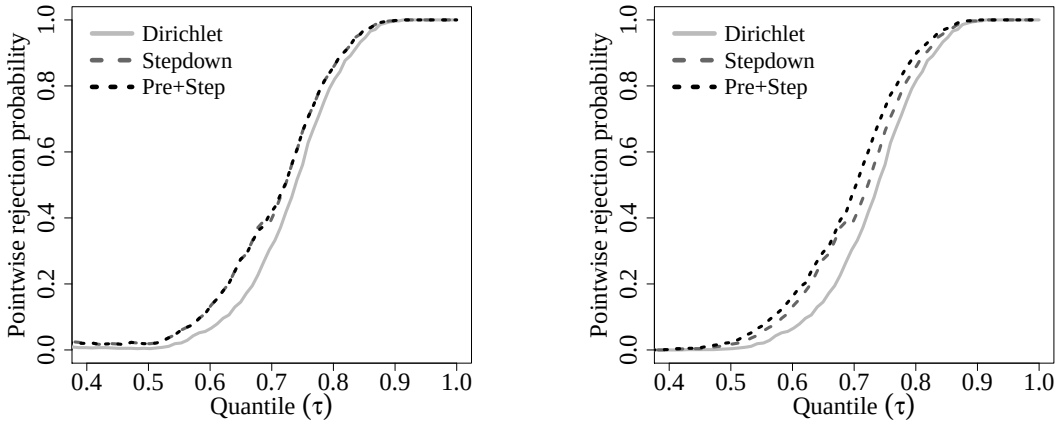

Figure 6: Simulated pointwise RP by quantile (τ), same DGP and methods as Table 2. Left: $F^{-1}(\tau) = F_0^{-1}(\tau)$ for $\tau \leq 0.5$, $F^{-1}(\tau) = 4(\tau - 0.5)$ otherwise. Right: $F^{-1}(\tau) = 4(\tau - 0.5)$.

For two-sample multiple testing, Figure 7 compares the pointwise (by τ) RPs of the methods shown in Table 4. For the basic MTP that tests H_{0r} (with $r \in \mathbb{R}$) instead of $H_{0\tau}$, we plot the RP of H_{0r} at $\tau = F_Y(r) = (r + 1)/2$. The DGPs are the same as the rows in Table 4 where $\{\tau : H_{0\tau} \text{ is true}\} = [0, 0.5]$. The power improvements due to the stepdown and pre-test procedures are similar to Figure 6: modest but noticeable over a range of $\tau > 0.5$. A bigger power difference is between the basic MTP and the joint quantile difference MTP (iteration $i = 0$ of Method 8). These MTPs' powers differ because the latter MTP explicitly focuses on fewer quantiles, in this case only 10, so more power can be focused on each quantile. This may be a reasonable way to improve power, especially in small samples, or if one assumes the quantile differences do not vary too quickly with τ. One could further increase pointwise power by examining yet fewer quantiles, but the choice of quantiles becomes arbitrary and subject to manipulation.

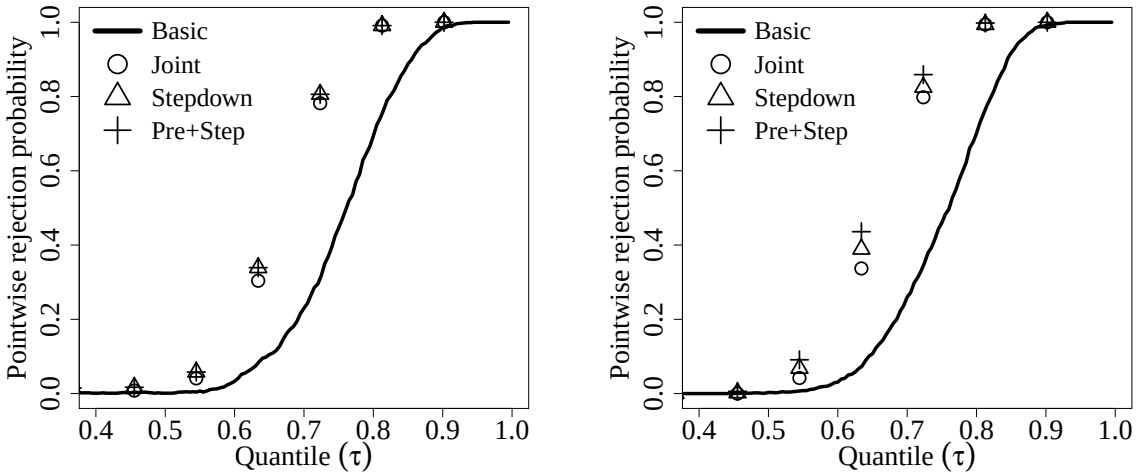

Figure 7: Simulated pointwise RP by quantile, same DGP and methods as Table 4. For the Basic method, the RP is plotted for $\tau = F_Y(r) = (r+1)/2$. Left: $F_X^{-1}(\tau) = F_Y^{-1}(\tau)$ for $\tau \leq 0.5$, $F_X^{-1}(\tau) = 4(\tau - 0.5)$ otherwise. Right: $F_X^{-1}(\tau) = 4(\tau - 0.5)$.

6.3 Computation time

Table 5 shows computation times for one-sample, two-sided methods: the Dirichlet MTP, the asymptotic KS test, and the exact KS test. Each value in the table has been averaged over at least four repetitions, using a standard desktop computer (8GB RAM, 3.2GHz processor). The time to simulate $\tilde{\alpha}$ (to the same degree of precision as Proposition 6) is also shown; this is the time saved by Proposition 6 compared with just-in-time simulation as in Buja and Rolke (2006). The simulation time depends on the starting value of $\tilde{\alpha}$ in the numerical search; we use five search iterations to be comparable to Aldor-Noiman et al. (2013, p. 254), who report a runtime of 10 seconds for $n = 100$ (compared to 9.47 seconds in our table).

Table 5: Computation times (in seconds) for one-sample, two-sided inference, $\alpha = 0.1$.

$\log_{10}(n)$	Proposition 6	Buja and Rolke (2006)	KS	KS (exact)
2	0.00	9.47	0.00	0.00
3	0.02	14.84	0.00	0.00
4	0.23	82.48	0.00	0.08
5	2.20	851.14	0.01	25.25

In Table 5, the asymptotic KS test runs instantly even for $n = 100\,000$. The exact KS slows significantly around $n = 100\,000$, requiring over 20 seconds per test. With Proposition 6, the Dirichlet MTP only takes a few seconds even with $n = 100\,000$, faster than the exact KS and orders of magnitude faster than just-in-time simulation.

7 Empirical example

We revisit data from Gneezy and List (2006, Tables I and V). Results may be replicated using code from the latter author's website. The global p-values are from the method proposed by Buja and Rolke (2006), implemented in our code with our faster computation.

The experiment of Gneezy and List (2006) pays control group individuals an advertised hourly wage and treatment group individuals an unexpectedly larger "gift" wage upon arrival. The "gift exchange" question is whether the higher wage induces higher effort in return. The experiment is run separately for library data entry and door-to-door fundraising tasks. The sample sizes are small: 10 and 9 for control and treatment (respectively) for the library task, and 10 and 13 for fundraising. With small samples, our methods' finite-sample FWER control is especially desirable.

The main finding of Gneezy and List (2006) is that the "gift wage" treatment raises productivity significantly in the first time period but not thereafter. Complementing the original results, we examine heterogeneity in the period 1 treatment effect, testing across the productivity distribution.

Figure 8 shows the two-sided bands used by Method 5, $[\hat{\ell}_X(\cdot), \hat{u}_X(\cdot)]$ and $\hat{\ell}_Y(\cdot), \hat{u}_Y(\cdot)]$. Wherever the bands do not overlap, the pointwise null hypothesis H_{0r} is rejected. With such small sample sizes, discreteness precludes an exact 10% FWER level, so exact 8.5% (library) and 9.3% (fundraising) FWER levels are used instead.

For the library task, with 8.5% FWER level, our MTP does not reject equality of the treatment and control productivity CDFs at any point. However, there is almost a rejection near 56–58 books, around the upper quartile; increasing the FWER level to 14% triggers rejection here. With one-sided global testing, i.e., testing first-order stochastic dominance, the Dirichlet test cannot reject that the treatment distribution dominates the control distribution ($p = 0.996$), whereas it does reject at a 10% level that the control distribution dominates the treatment distribution ($p = 0.076$) because of the pointwise rejection near the 0.8-quantile. In contrast, the KS test fails to reject at a 10% level: its one-sided p-value is 0.13.

For the fundraising task, with 9.3% FWER level, Figure 8 shows two ranges near the lower quartile of the distributions where a zero treatment effect is rejected: 8–14 and 21–26 dollars raised. This suggests the gift wage most strongly affects low-productivity individuals in fundraising, opposite the library task. Testing first-order stochastic dominance, the Dirichlet test cannot reject that the treatment dominates the control distribution ($p = 1$), whereas it can reject at a 5% level that the control distribution dominates the treatment distribution ($p = 0.021$). The one-sided KS $p = 0.034$: higher, but still below 0.05. For two-sided testing

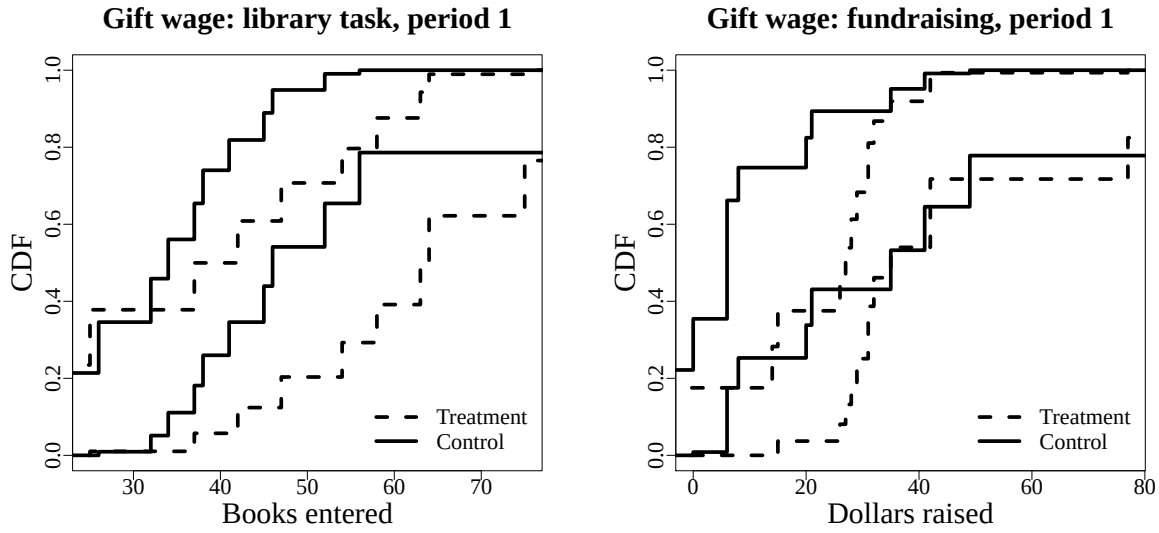

Figure 8: Comparison of treatment and control group productivity in the first period of the library (left) and fundraising (right) tasks in Gneezy and List (2006): bands for two-sided MTP (rejecting wherever there is no overlap) with exact FWER levels 8.5% (left) and 9.3% (right).

of zero treatment effect, the Dirichlet test rejects at a 5% level while the KS cannot: the Dirichlet $p = 0.041$, while the KS $p = 0.069$.

Table 6: Summary of intervals of r where H_{0r} is rejected in the empirical example. Units are books entered (library) or dollars raised (fundraising). With $F_T(\cdot)$ the treated population CDF and $F_C(\cdot)$ the control CDF, "2-sided" means $H_{0r} : F_T(r) = F_C(r)$, and "1-sided" means $H_{0r} : F_T(r) \geq F_C(r)$.

	Library		Fundraising	
Method	2-sided	1-sided	2-sided	1-sided
FWER level: $\alpha = 0.05$				
Dirichlet	none	none	$(8, 14) \cup (21, 26)$	$(8, 14) \cup (21, 26)$
KS-based	none	none	none	$(21, 26)$
FWER level: $\alpha = 0.1$				
Dirichlet	none	$(56, 58)$	$(8, 14) \cup (21, 26)$	$(6, 15) \cup (21, 27)$
KS-based	none	none	$(21, 26)$	$(8, 13) \cup (21, 26)$

Table 6 summarizes the results from running the Dirichlet and KS-based MTPs. In addition to the Dirichlet rejecting some H_{0r} in many cases where the KS-based MTP cannot, the Dirichlet MTP rejects more H_{0r} in cases where both MTPs reject at some r.

8 Conclusion

We have considered the question, "At which quantiles do two distributions differ?" Framed as multiple testing across the continuum of quantiles $\tau \in (0, 1)$, we have shown KS-based multiple testing procedures to have strong control of FWER, for both one-sided and two-sided, one-sample and two-sample inference. Our newly proposed Dirichlet-based procedures also have strong control of finite-sample FWER, along with other advantages: more even sensitivity than KS, improved global power, stepdown and pre-test power improvements, and fast computation.

Future work may include conditional versions of these tests, as well as exploration of the connection with the continuity-corrected Bayesian bootstrap of Banks (1988).

References

Aldor-Noiman, S., L. D. Brown, A. Buja, W. Rolke, and R. A. Stine (2013). The power to see: A new graphical test of normality. *The American Statistician 67*(4), 249–260.

Anderson, T. W. and D. A. Darling (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *Annals of Mathematical Statistics 23*(2), 193–212.

Banks, D. L. (1988). Histospline smoothing the Bayesian bootstrap. *Biometrika 75*(4), 673–684.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*(1), 289–300.

Beran, R. and P. Hall (1993). Interpolated nonparametric prediction intervals and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 55*(3), 643–652.

Berk, R. H. and D. H. Jones (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 47*(1), 47–59.

Buja, A. and W. Rolke (2006). Calibration for simultaneity: (re)sampling methods for simultaneous inference with applications to function estimation and functional data. Working paper, available at `http://stat.wharton.upenn.edu/~buja/PAPERS/paper-sim.pdf`.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008). Bootstrap-based improvements for inference with clustered errors. *Review of Economics and Statistics 90*(3), 414–427.

Chicheportiche, R. and J.-P. Bouchaud (2012). Weighted Kolmogorov–Smirnov test: accounting for the tails. *Physical Review E 86*(4), 041115.

David, H. A. and H. N. Nagaraja (2003). *Order Statistics* (3rd ed.). New York: Wiley.

Davidson, R. and J.-Y. Duclos (2013). Testing for restricted stochastic dominance. *Econometric Reviews 32*(1), 84–125.

Donald, S. G. and Y.-C. Hsu (2016). Improving the power of tests of stochastic dominance. *Econometric Reviews 35*(4), 553–585.

Eicker, F. (1979). The asymptotic distribution of the suprema of the standardized empirical processes. *Annals of Statistics 7*(1), 116–138.

Firpo, S. and A. F. Galvao (2015). Uniform inference on functionals of quantiles of potential outcomes. Working paper.

Fisher, R. A. (1932). *Statistical Methods for Research Workers* (4th ed.). Edinburg: Oliver and Boyd.

Gneezy, U. and J. A. List (2006). Putting behavioral economics to work: testing for gift exchange in labor markets using field experiments. *Econometrica 74*(5), 1365–1384.

Goldman, M. and D. M. Kaplan (2016a). Fractional order statistic approximation for nonparametric conditional quantile inference. *Journal of Econometrics XXX*(X), XXX–XXX. Forthcoming.

Goldman, M. and D. M. Kaplan (2016b). Nonparametric inference on conditional quantile differences, linear combinations, and vectors, using L-statistics. Working paper, available at `http://faculty.missouri.edu/~kaplandm`.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Jaeschke, D. (1979). The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. *Annals of Statistics 7*(1), 108–115.

Kaplan, D. M. and L. Zhuo (2016). Bayesian and frequentist inequality tests. Working paper, available at `http://faculty.missouri.edu/~kaplandm`.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari 4*(1), 83–91.

Lehmann, E. L. and J. P. Romano (2005a). Generalizations of the familywise error rate. *Annals of Statistics 33*(3), 1138–1154.

Lehmann, E. L. and J. P. Romano (2005b). *Testing Statistical Hypotheses* (3rd ed.). Springer Texts in Statistics. Springer.

Linton, O., K. Song, and Y.-J. Whang (2010). An improved bootstrap test of stochastic dominance. *Journal of Econometrics 154*(2), 186–202.

Lo, A. Y. (1993). A Bayesian method for weighted sampling. *Annals of Statistics 21*(4), 2138–2148.

Lockhart, R. (1991). Overweight tails are inefficient. *Annals of Statistics 19*(4), 2254–2258.

Moscovich, A., B. Nadler, and C. Spiegelman (2016). On the exact Berk–Jones statistics and their p-value calculation. *Electronic Journal of Statistics 10*(2), 2329–2354.

Neyman, J. (1937). »Smooth test» for goodness of fit. *Skandinavisk Aktuarietidskrift 20*(3–4), 149–199.

Owen, A. B. (1995). Nonparametric likelihood confidence bands for a distribution function. *Journal of the American Statistical Association 90*(430), 516–521.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika 25*, 379–410.

Qu, Z. and J. Yoon (2015). Nonparametric estimation and inference on conditional quantile processes. *Journal of Econometrics 185*(1), 1–19.

R Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Romano, J. P., A. M. Shaikh, and M. Wolf (2010). Multiple testing. In S. N. Durlauf and

L. E. Blume (Eds.), *The New Palgrave Dictionary of Economics* (Online ed.). Palgrave Macmillan.

Scheffé, H. and J. W. Tukey (1945). Non-parametric estimation. I. validation of order statistics. *Annals of Mathematical Statistics 16*(2), 187–192.

Sedgewick, R. and K. Wayne (2011). *Algorithms* (4th ed.). Addison-Wesley Professional.

Smirnov, N. V. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bulletin Mathématique de l'Université de Moscou 2*(2), 3–16.

Smirnov, N. V. (1948). Table for estimating the goodness of fit of empirical distributions. *Annals of Mathematical Statistics 19*(2), 279–281.

Stigler, S. M. (1977). Fractional order statistics, with applications. *Journal of the American Statistical Association 72*(359), 544–550.

Wilks, S. S. (1962). *Mathematical Statistics*. New York: Wiley.

A Additional methods

A.1 One-sample methods

Method 6. For Task 1, modify Method 3 as follows. Define ℓ_k and u_k as in (7). Instead of only $r_{k,i}$ corresponding to either ℓ_k or u_k, include both $r_{k,\ell,i}$ corresponding to ℓ_k and $r_{k,u,i}$ corresponding to u_k. Instead of $\hat{K}_0 = \{1,\dots,n\}$, let $\hat{K}_0^\ell = \hat{K}_0^u = \{1,\dots,n\}$, where $\hat{K}_0^\ell$ corresponds to the ℓ_k and $\hat{K}_0^u$ to the u_k. Replace (11) with

$$\begin{aligned}\alpha &\geq 1 - \mathrm{P}\Big(\bigcap_{k\in\hat{K}_i^\ell}\big\{X_{n:r_{k,\ell,i}} \geq F^{-1}(\ell_k)\big\} \cap \bigcap_{k\in\hat{K}_i^u}\big\{X_{n:r_{k,u,i}} \leq F^{-1}(u_k)\big\}\Big) \\ &\geq 1 - \mathrm{P}\Big(\bigcap_{k\in\hat{K}_i^\ell}\big\{F(X_{n:r_{k,\ell,i}}) \geq \ell_k\big\} \cap \bigcap_{k\in\hat{K}_i^u}\big\{F(X_{n:r_{k,u,i}}) \leq u_k\big\}\Big). \qquad (13)\end{aligned}$$

Check for rejections of both $F^{-1}(\tau) \geq F_0^{-1}(\tau)$ and $F^{-1}(\tau) \leq F_0^{-1}(\tau)$ as described in Method 3; either implies rejection of $H_{0\tau} : F^{-1}(\tau) = F_0^{-1}(\tau)$. □

Method 7 (Pre-test only)**.** Consider the pre-test null hypotheses $H_{0\ell_k} : F^{-1}(\ell_k) \leq F_0^{-1}(\ell_k)$, defining ℓ_k as in (7). Given $\tilde{\alpha}$, let $\underline{k} = \min\{k : \ell_k \geq B_{1,n}^{1-\tilde{\alpha}}\}$ and $r_k = \max\{k' : B_{k',n}^{1-\tilde{\alpha}} \leq \ell_k\}$ (for $k \geq \underline{k}$), where both k and k' are restricted to integers $\{1,\dots,n\}$. Using Theorem 4, calculate

$$\alpha_p(\tilde{\alpha}, n) = 1 - \mathrm{P}\Big(\bigcap_{k=\underline{k}}^{n} X_{n:r_k} \leq F^{-1}(\ell_k)\Big) = 1 - \mathrm{P}\Big(\bigcap_{k=\underline{k}}^{n} F(X_{n:r_k}) \leq \ell_k\Big).$$

Adjust $\tilde{\alpha}$ until $\alpha_p(\tilde{\alpha}, n)$ equals (approximately) the desired FWER. Reject $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$ when $\max\{X_{n:r_k} : \ell_k \leq \tau\} > F_0^{-1}(\tau)$.

To instead pre-test $H_{0u_k} : F^{-1}(u_k) \geq F_0^{-1}(u_k)$, reverse all inequalities and min/max, and replace ℓ_k with u_k (also from (7)), $B_{k,n}^{1-\tilde{\alpha}}$ with $B_{k,n}^{\tilde{\alpha}}$, $\underline{k} = \min\{k : \ell_k \geq B_{1,n}^{1-\tilde{\alpha}}\}$ with $\bar{k} = \max\{k : u_k \leq B_{n,n}^{\tilde{\alpha}}\}$, and $\bigcap_{k=\underline{k}}^{n}$ with $\bigcap_{k=1}^{\bar{k}}$. □

A.2 Procedures to improve two-sample power

The two-sample stepdown and pre-test procedures are no longer based on finite-sample distributions of order statistics. Instead, we (slightly) extend results from Goldman and Kaplan (2016b). This requires that the quantiles not be too close together. To be more explicit about how the methods work, we present modified tasks that they address.

Task 5 Testing a family of $M_n = \lfloor n^{2/5} \rfloor$ two-sample quantile equality hypotheses with strong control of FWER; specifically, for $j = 1, \ldots, M_n$, $H_{0j} : F_X^{-1}(t) = F_Y^{-1}(t)$ for all $t \in [(j-0.5)/(M_n+1), (j+0.5)/(M_n+1)]$.

Task 6 Same as Task 5 but with $F_X^{-1}(t) \leq F_Y^{-1}(t)$ or $F_X^{-1}(t) \geq F_Y^{-1}(t)$.

Consider a fixed set of M quantiles, $\tau_1, \ldots, \tau_M$, and let $\Delta_j \equiv F_Y^{-1}(\tau_j) - F_X^{-1}(\tau_j)$. Goldman and Kaplan (2016b) use "fractional order statistics" to construct a CI for each Δ_j with $1 - \alpha + O\big(n^{-2/3}\log(n)\big)$ coverage probability, and CIs for all $F_X^{-1}(\tau_j)$ or $F_Y^{-1}(\tau_j)$ that have joint (over $j = 1, \ldots, M$) coverage probability of $1 - \alpha + O(n^{-1})$. It is a small step to infer that CIs for all Δ_j can be constructed with joint $1 - \alpha + O\big(n^{-2/3}\log(n)\big)$, using the modified calibration (of $\tilde{\alpha}$) seen in our code. For a lower one-sided CI, the upper endpoints are $\hat{Q}_Y^L\big(u_{y,j}^h(\tilde{\alpha})\big) - \hat{Q}_X^L\big(u_{x,j}^l(\tilde{\alpha})\big)$, where $u_{y,j}^h(\tilde{\alpha}) \approx \tau_j + n_Y^{-1/2} z_{1-\tilde{\alpha}} \sqrt{\tau_j(1-\tau_j)}$, $u_{x,j}^l(\tilde{\alpha}) \approx \tau_j - n_X^{-1/2} z_{1-\tilde{\alpha}} \sqrt{\tau_j(1-\tau_j)}$, $z_{1-\tilde{\alpha}}$ is the standard normal distribution's $(1-\tilde{\alpha})$-quantile, $\tilde{\alpha}$ solves

$$1 - \alpha = \mathrm{P}\left(\bigcap_{j=1}^{M} \Big\{ \tilde{Q}_{U_y}^I\big(u_{y,j}^h(\tilde{\alpha})\big) - \tilde{Q}_{U_x}^I\big(u_{x,j}^l(\tilde{\alpha})\big) > 0 \Big\} \right),$$

$\tilde{Q}_{U_x}^I$ is a Dirichlet process with index measure $\nu(\cdot)$ where $\nu([0,t]) = (n_X + 1)t$ for $t \in [0,1]$ (Stigler, 1977), and $\hat{Q}_X^L(u) \equiv X_{n_X:k} + [u(n_X+1) - k]X_{n_X:k+1}$, $k = \lfloor u(n_X+1) \rfloor$, and similarly for $\hat{Q}_Y^L(u)$. The upper one-sided CI is defined similarly, and the two-sided CI is the intersection of upper and lower one-sided CIs.

Let $\widehat{\mathrm{CI}}_j$ denote the CI for Δ_j. Letting $I = \{j : H_{0j} \text{ is true}\}$,

$$\mathrm{FWER} = 1 - \mathrm{P}\left(\bigcap_{j \in I} \{\Delta_j \in \widehat{\mathrm{CI}}_j\} \right) \leq 1 - \mathrm{P}\left(\bigcap_{j=1}^{M} \{\Delta_j \in \widehat{\mathrm{CI}}_j\} \right) \to 1 - (1 - \alpha) = \alpha.$$

If $M_n \to \infty$ too quickly, then the arguments from Goldman and Kaplan (2016b) break down, but we conjecture they still hold with $M_n = O(n^{2/5})$.

Method 8. For Task 5, let $\tau_j = j/(M_n+1)$ for $j = 1, \ldots, M_n$. Let $\hat{T}_0 \equiv \{1, \ldots, M_n\}$. Given a pointwise $\tilde{\alpha}$, let $k_{X,j}^u$ and $k_{X,j}^\ell$ be such that

$$\mathrm{P}\big(\mathrm{Beta}(k_{X,j}^u, n_X + 1 - k_{X,j}^u) < \tau_j\big) = \tilde{\alpha}/2 = \mathrm{P}\big(\mathrm{Beta}(k_{X,j}^\ell, n_X + 1 - k_{X,j}^\ell) > \tau_j\big),$$

and similarly for $k_{Y,j}^u$ and $k_{Y,j}^\ell$ (with n_Y instead of n_X). These k may have fractional (non-integer) values. For iteration i, CIs with joint $1 - \alpha$ coverage probability are constructed

with $\tilde{\alpha}$ chosen such that

$$1-\alpha = \mathrm{P}\left(\bigcap_{j\in\hat{T}_i}\left\{D^{\ell}_{X,j} < D^{u}_{Y,j}, D^{\ell}_{Y,j} < D^{u}_{X,j}\right\}\right), \tag{14}$$

defining $F_X(X_{n_X:0}) \equiv 0$, $F_X(X_{n_X:n_X+1}) \equiv 1$, $X_{n_X:k} \equiv (1-k+\lfloor k\rfloor)X_{n_X:\lfloor k\rfloor}+(k-\lfloor k\rfloor)X_{n_X:\lfloor k\rfloor+1}$ for fractional k, and using the distribution

$$\begin{aligned}(D_{X,1}, D_{X,2}-D_{X,1}, \ldots, D_{X,2M_n}-D_{X,2M_n-1}, 1-D_{X,2M_n})\\ \sim \mathrm{Dir}(k_1, k_2-k_1, \ldots, k_{2M_n}-k_{2M_n-1}, n_X+1-k_{2M_n})\end{aligned}$$

with vector $k=(k_1,\ldots,k_{2M_n})$ containing all the $k^{\ell}_{X,j}$ and $k^{u}_{X,j}$ in ascending order so that $k_1 \le \cdots \le k_{2M_n}$; and defining all these objects similarly for Y, with $\mathbf{D}_X \perp\!\!\!\perp \mathbf{D}_Y$. For iteration $i=0$, reject any H_{0j} for which the CI $[Y_{n_Y:k^{\ell}_{Y,j}}-X_{n_X:k^{u}_{X,j}}, Y_{n_Y:k^{u}_{Y,j}}-X_{n_X:k^{\ell}_{X,j}}]$ does not contain zero. Then, iteratively perform the following steps, starting with $i=1$.

Step 1. Let $\hat{T}_i = \{j : H_{0j} \text{ not yet rejected}\}$. If $\hat{T}_i = \emptyset$ or $\hat{T}_i = \hat{T}_{i-1}$, then stop.
Step 2. Use $\hat{T}_i$ and (14) to construct new joint CIs.
Step 3. Reject any additional H_{0j} for which the corresponding CI does not contain zero.
Step 4. Increment i by one and return to Step 1.

For Task 6, use the above with only upper (or lower) endpoints. □

Method 9. For Task 6, using notation from Method 8, consider $H_{0j} : F_X^{-1}(\tau_j) \ge F_Y^{-1}(\tau_j)$. First run a pre-test of $H'_{0j} : F_X^{-1}(\tau_j) \le F_Y^{-1}(\tau_j)$ using iteration $i=0$ of Method 8 (i.e., the basic method without stepdown) with FWER level $\alpha_p = \alpha/\ln[\ln(\max\{n,15\})]$. Then, use Method 8 starting with $\hat{T}_0$ containing all j such that H_{0j} was not rejected by the pre-test. □

We conjecture that under Assumptions 1 and 2, Methods 8 and 9 have strong control of asymptotic FWER.

B Mathematical proofs

B.1 Proof of Proposition 1

Proof. The two-sided proof is in the main text.

The one-sided case follows the same argument (after modifying D^x_n and $D^{x,0}_n$), with the additional inequality that if $H_{0x} : F(x) \le F_0(x)$ is true, then $\hat{F}(x)-F_0(x) \le \hat{F}(x)-F(x)$. Let $D^x_n \equiv \sqrt{n}\big(\hat{F}(x)-F(x)\big)$, $D_n \equiv \sup_{x\in\mathbb{R}} D^x_n$, and $c_n(\alpha)$ now satisfies $\mathrm{P}\big(D_n > c_n(\alpha)\big) = \alpha$ in finite samples. Let

$$D^{x,0}_n \equiv \sqrt{n}\big(\hat{F}(x)-F_0(x)\big), \quad I \equiv \{x : H_{0x} \text{ is true}\}, \quad D^I_n \equiv \sup_{x\in I} D^{x,0}_n \le \sup_{x\in I} D^x_n,$$

where the last inequality follows because $F(x) \leq F_0(x)$ for $x \in I$, so $D_n^{x,0} \leq D_n^x$ (whereas before this was an equality). Then, since $I \subseteq \mathbb{R}$,

$$\text{FWER} \equiv \mathrm{P}\big(D_n^I > c_n(\alpha)\big) \leq \mathrm{P}\big(D_n > c_n(\alpha)\big) = \alpha.$$

The one-sided argument with $H_{0x} : F(x) \geq F_0(x)$ is identical when using $-D_n^x$ instead.

Alternatively, the results can be derived using the fact that the KS test can be inverted to give a uniform confidence band, and any MTP based on a uniform confidence band has strong control of FWER. □

B.2 Proof of Lemma 2

Proof. For the EDFs defined in (3), pointwise, $n_X \hat{F}_X(r) \sim \text{Binomial}\big(n_X, F_X(r)\big)$, $n_Y \hat{F}_Y(r) \sim \text{Binomial}\big(n_Y, F_Y(r)\big)$, and by Assumption 1 $\hat{F}_X(\cdot) \perp\!\!\!\perp \hat{F}_Y(\cdot)$. Since (by assumption) rejection of H_{0r} depends only on $\hat{F}_X(r)$ and $\hat{F}_Y(r)$, the RP depends only on $F_X(r)$ and $F_Y(r)$. More generally, the distribution of

$$\big(n_X \hat{F}_X(r_1), n_X[\hat{F}_X(r_2) - \hat{F}_X(r_1)], \ldots, n_X[\hat{F}_X(r_m) - \hat{F}_X(r_{m-1})]\big)$$

is multinomial with parameters n_X and $\big(F_X(r_1), F_X(r_2) - F_X(r_1), \ldots, F_X(r_m) - F_X(r_{m-1})\big)$, and similarly for Y. Even if set S is a continuum, the distribution of $\{\hat{F}_X(r), \hat{F}_Y(r)\}_{r \in S}$ depends only on n_X, n_Y, and $\{F_X(r), F_Y(r)\}_{r \in S}$. Consequently, RPs of H_{0r} over $r \in S$ depend only on n_X, n_Y, and $\{F_X(r), F_Y(r)\}_{r \in S}$, too.

As in Definition 1, let $I \equiv \{r : H_{0r} \text{ is true}\}$, so $I \subseteq \mathbb{R}$. Define $G_X(\cdot)$ such that $G_X(r) = F_X(r)$ if H_{0r} is true and $G_X(r) = F_Y(r)$ if H_{0r} is false. Thus, if we had $G_X(\cdot)$ instead of $F_X(\cdot)$, H_{0r} would be true for all $r \in \mathbb{R}$. Then,

$$\begin{aligned}
\text{FWER} &\equiv \overbrace{\mathrm{P}(\text{reject } H_{0r} \text{ for any } r \in I \mid F_X, F_Y)}^{\text{by Definition 1}} \\
&= \overbrace{\mathrm{P}(\text{reject } H_{0r} \text{ for any } r \in I \mid G_X, F_Y)}^{\text{by above properties and } F_X(r) = G_X(r) \text{ for } r \in I} \\
&\leq \overbrace{\mathrm{P}(\text{reject } H_{0r} \text{ for any } r \in \mathbb{R} \mid G_X, F_Y)}^{\text{by } I \subseteq \mathbb{R}} \\
&\leq \alpha
\end{aligned}$$

by assumption of weak control of FWER at level α, since all H_{0r} are true given $G_X(\cdot)$ and $F_Y(\cdot)$. □

B.3 Proof of Proposition 3

Proof. The method rejects H_{0r} depending only on $\hat{F}_X(r)$ and $\hat{F}(r)$, through their difference $\hat{F}_X(r) - \hat{F}_Y(r)$. It is well known that the two-sample KS GOF test controls size, which is equivalent to weak control of FWER. Thus, the assumptions of Lemma 2 are satisfied, so the method has strong control of FWER. □

B.4 Proof of Theorem 5

Proof. The one-sided proof is entirely in the main text.

For the two-sided case, we have a parallel argument. Let

$$K^\ell \equiv \{k : F^{-1}(\ell_k) = F_0^{-1}(\ell_k)\}, \quad K^u \equiv \{k : F^{-1}(u_k) = F_0^{-1}(u_k)\},$$

the sets of true hypotheses. Then,

$$\begin{aligned}
\text{FWER} &= \overbrace{1 - \text{P}(\text{no rejections among } k \in \{K^\ell \cup K^u\})}^{\text{by definition of FWER}} \\
&= \overbrace{1 - \text{P}\Big(\bigcap_{k\in K^\ell} F_0^{-1}(\ell_k) \le X_{n:k} \cap \bigcap_{k\in K^u} X_{n:k} \le F_0^{-1}(u_k)\Big)}^{\text{by definition of } H_{0\ell_k},\ H_{0u_k}} \\
&\le \overbrace{1 - \text{P}\Big(\bigcap_{k\in K^\ell} F^{-1}(\ell_k) \le X_{n:k} \cap \bigcap_{k\in K^u} X_{n:k} \le F^{-1}(u_k)\Big)}^{\text{because } F^{-1}(\ell_k) \ge F_0^{-1}(\ell_k) \text{ for all } k \in K^\ell,\ F^{-1}(u_k) \le F_0^{-1}(u_k) \text{ for all } k \in K^u} \\
&\le \overbrace{1 - \text{P}\Big(\bigcap_{k=1}^{n} F^{-1}(\ell_k) \le X_{n:k} \le F^{-1}(u_k)\Big)}^{\text{because } K^\ell, K^u \subseteq \{1,2,\ldots,n\}} \\
&= \overbrace{\alpha}^{\text{from (9)}}.
\end{aligned}$$

□

B.5 Proof of Theorem 7

Proof. Consider the one-sided case with $H_{0\tau} : F^{-1}(\tau) \ge F_0^{-1}(\tau)$ for $\tau \in (0, 1)$. We again focus on the n hypotheses $F^{-1}(\ell_k) \ge F_0^{-1}(\ell_k)$ for $k = 1, \ldots, n$ since rejections at other τ are simply by logical implication of the monotonicity of $F_0^{-1}(\cdot)$ and thus do not affect FWER. (The same is true in the two-sided case since it essentially combines lower and upper one-sided MTPs.)

Let $K \equiv \{k : F^{-1}(\ell_k) \ge F_0^{-1}(\ell_k)\}$, the (true) set of true hypotheses. Let r_{k^*} denote the order statistic indices that would be chosen by Method 3 when attention is restricted to $k \in K$. (Many choices of r_{k^*} still control FWER, but the choice must rely only on the set K.) Thus, for $k \in K$, the r_{k^*} satisfy $r_{k^*} \le k$ and

$$\alpha \ge 1 - \text{P}\Big(\bigcap_{k\in K} \big\{X_{n:r_{k^*}} \ge F_0^{-1}(\ell_k)\big\}\Big). \tag{15}$$

The stepdown procedure specifies monotonicity in the $r_{k,i}$ and $\hat{K}_i$ over iterations $i = 0, 1, \ldots$, where $\hat{K}_0 = \{1, \ldots, n\}$ and $r_{k,0} = k$. Specifically, $\hat{K}_0 \supset \hat{K}_1 \supset \cdots$, and for each k, $r_{k,0} \ge r_{k,1} \ge \cdots$. This monotonicity is similar in spirit to (15.37) in Lehmann and Romano (2005b).

The proof is by induction. Consider any dataset where

$$1 = \prod_{k \in K} \mathbb{1}\big\{X_{n:r_{k^*}} \geq F_0^{-1}(\ell_k)\big\}.$$

In iteration i, if $\hat{K}_i \supseteq K$, then none of the true hypotheses are rejected since $r_{k,i} \geq r_{k^*}$, which implies $X_{n:r_{k,i}} \geq X_{n:r_{k^*}} \geq F_0^{-1}(\ell_k)$. Consequently, $\hat{K}_{i+1} \supseteq K$, too. Since $\hat{K}_0 \supseteq K$, the stepdown procedure does not reject any true hypothesis in such a dataset. Along with (15), this implies FWER $\leq \alpha$.

The other one-sided case with $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$ is entirely parallel, simply reversing inequalities and replacing ℓ_k with u_k.

For the two-sided case with $H_{0\tau} : F^{-1}(\tau) = F_0^{-1}(\tau)$, the key is again the monotonicity (by construction) in the $r_{k,\ell,i}$, $r_{k,u,i}$, $\hat{K}_i^\ell$, and $\hat{K}_i^u$. Specifically, $\hat{K}_0^\ell \supset \hat{K}_1^\ell \supset \cdots$, $\hat{K}_0^u \supset \hat{K}_1^u \supset \cdots$, and for each k, $r_{k,\ell,0} \geq r_{k,\ell,1} \geq \cdots$ and $r_{k,u,0} \leq r_{k,u,1} \leq \cdots$. Let $K^\ell \equiv \{k : F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k)\}$ and $K^u \equiv \{k : F^{-1}(u_k) \leq F_0^{-1}(u_k)\}$, the (true) sets of true hypotheses. For $k \in K^\ell$, let $r_{k^*,\ell}$ satisfy $r_{k^*,\ell} \leq k$; for $k \in K^u$, let $r_{k^*,u}$ satisfy $r_{k^*,u} \geq k$. Also, these satisfy

$$\alpha \geq 1 - \mathrm{P}\Big(\bigcap_{k \in K^\ell} \big\{F_0^{-1}(\ell_k) \leq X_{n:r_{k^*,\ell}}\big\} \cap \bigcap_{k \in K^u} \big\{X_{n:r_{k^*,u}} \leq F_0^{-1}(u_k)\big\}\Big). \tag{16}$$

As for the one-sided case, by induction, consider any dataset where

$$1 = \prod_{k \in K^\ell} \mathbb{1}\big\{X_{n:r_{k^*,\ell}} \geq F_0^{-1}(\ell_k)\big\} \prod_{k \in K^u} \mathbb{1}\big\{X_{n:r_{k^*,u}} \leq F_0^{-1}(u_k)\big\}.$$

In iteration i, if $\{\hat{K}_i^\ell \cup \hat{K}_i^u\} \supseteq \{K^\ell \cup K^u\}$, then none of the true hypotheses are rejected since $r_{k,\ell,i} \geq r_{k^*,\ell}$ and $r_{k,u,i} \leq r_{k^*,u}$, which implies

$$X_{n:r_{k,\ell,i}} \geq X_{n:r_{k^*,\ell}} \geq F_0^{-1}(\ell_k), \quad X_{n:r_{k,u,i}} \geq X_{n:r_{k^*,u}} \leq F_0^{-1}(u_k).$$

Consequently, $\{\hat{K}_{i+1}^\ell \cup \hat{K}_{i+1}^u\} \supseteq \{K^\ell \cup K^u\}$, too. Since $\hat{K}_0^\ell \supseteq K^\ell$ and $\hat{K}_0^u \supseteq K^u$, the stepdown procedure does not reject any true hypothesis in such a dataset. Along with (16), this implies FWER $\leq \alpha$. □

B.6 Proof of Proposition 10 (for proof of Theorem 8)

Proposition 10. *Under Assumptions 1 and 2, Method 7 has strong control of finite-sample FWER.*

Proof. The MTP is based on a one-sided uniform confidence band, so it strongly controls FWER by the same argument as in the proof of Theorem 5. That is, the uniform confidence band covers the entire $F(\cdot)$ with at least $1 - \alpha$ probability, so it covers any subset of $F(\cdot)$ with at least $1 - \alpha$ probability, too. Thus, FWER is below $1 - (1 - \alpha) = \alpha$. □

B.7 Proof of Theorem 8

Proof. The stated FWER bound is conservative, relying on the following two worst-case assumptions. First, assume that any false pre-test rejection leads to a false rejection of the overall test. Second, assume that Method 1, i.e., the test without using a pre-test, never falsely rejects when the pre-test falsely rejects. Then, the worst-case (i.e., upper bound for) FWER is $\alpha + \alpha_p$, where the α_p is guaranteed by Proposition 10. □

B.8 Proof of Theorem 9

Proof. To apply Lemma 2, the method must reject H_{0r} depending only on $\hat{F}_X(r)$ and $\hat{F}_Y(r)$, and it must have weak control of FWER. First, by construction, as seen in (12), given n_X, n_Y, and α, the method will reject H_{0r} depending only on $\hat{F}_X(r)$ (which determines the $\hat{u}_X(r)$ and $\hat{\ell}_X(r)$) and on $\hat{F}_Y(r)$ (which determines the $\hat{u}_Y(r)$ and $\hat{\ell}_Y(r)$).

Second, for the two-sided MTP, weak control of FWER is by construction (up to simulation error). Weak control of FWER is equivalent to size control of the corresponding GOF test. When $F_X(\cdot) = F_Y(\cdot)$, the distribution of the ordering of the X and Y values is distribution-free given Assumptions 1 and 2; this (finite-sample) distribution is used explicitly to control the probability of any H_{0r} being rejected at level α.

For the one-sided case, consider the GOF null $H_0 : F_X(\cdot) \leq F_Y(\cdot)$. This is rejected based on the ordering of the X_i and Y_j, which is the same as the ordering of the $F_Y(X_i)$ and $F_Y(Y_j)$. By construction, the ordering of $F_X(X_i) \overset{iid}{\sim} \text{Unif}(0,1)$ and $F_Y(Y_j)$ will lead to rejection of H_0 with less than or equal to α probability. Under H_0, $F_Y(X_i) \geq F_X(X_i)$ for any X_i, so rejection of H_0 is even less likely with $F_Y(X_i)$ and size remains below α. This corresponds to the intuition that $F_X(\cdot) = F_Y(\cdot)$ is the least favorable configuration (i.e., results in highest RP) among all distributions satisfying $H_0 : F_X(\cdot) \leq F_Y(\cdot)$.

Since the assumptions are met, Lemma 2 gives strong control of FWER. Appendix C.3 discusses the suggested 0.0001 adjustment of $\tilde{\alpha}$ in light of possible simulation error and discontinuity in the mapping from $\tilde{\alpha}$ to α. □

C Computational details

We discuss some computational details of our code's implementation of our methods, specifically the simulation of the mapping from $\tilde{\alpha}$ to α.

C.1 Calibration of $\tilde{\alpha}$

Consider a given n. The joint distribution of the uniform order statistics is

$$(U_{n:1}, U_{n:2} - U_{n:1}, U_{n:3} - U_{n:2}, \ldots, U_{n:n} - U_{n:n-1}, 1 - U_{n:n}) \sim \text{Dirichlet}\overbrace{(1, \ldots, 1)}^{n+1}.$$

We simulate this with repeated random draws $U_i^{(m)} \overset{iid}{\sim} \text{Uniform}(0,1)$ for observations $i = 1, \ldots, n$ in samples $m = 1, \ldots, M$. Given $\tilde{\alpha}$, which determines all ℓ_k and u_k, the simulated

two-sided FWER (for example) is

$$\hat{\alpha} = 1 - \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\Big\{\ell_1 < U_{n:1}^{(m)} < u_1\Big\} \times \cdots \times \mathbb{1}\big\{\ell_n < U_{n:n}^{(m)} < u_n\big\}. \tag{17}$$

While (17) alone is sufficient for global (GOF) p-value computation, we need to search for the $\tilde{\alpha}$ that leads to a specific desired α for the simulations informing Proposition 6. Given search tolerance T (see Appendix C.2), we stop the search over $\tilde{\alpha}$ if $|\hat{\alpha} - \alpha| < T$. Otherwise, if $\hat{\alpha} < \alpha$ then $\tilde{\alpha}$ is increased, and if $\hat{\alpha} > \alpha$ then $\tilde{\alpha}$ is decreased. Since $\hat{\alpha}$ is a monotonic function of $\tilde{\alpha}$, which is a scalar, this is an easy search problem. Note that the random draws do not need to be repeated each iteration, only the $2n$ beta quantile function calls; or, the simulation is easily parallelized by slicing the M samples across CPUs.

With two samples, the only difference is (17). The GOF null $H_0 : F_X(\cdot) = F_Y(\cdot)$ is rejected whenever there is at least one point where the band for one distribution lies strictly above the other band, i.e., at least one H_{0r} is rejected. This depends on $\tilde{\alpha}$ and the relative ordering of values in the two samples, but not on the sample values themselves (more below). Because of this difference, with small sample sizes, there can be jumps of bigger than T in $\hat{\alpha}$ as a function of $\tilde{\alpha}$, in which case we pick $\tilde{\alpha}$ slightly smaller than the point of discontinuity.

The fact that the test's rejection is determined only by the ordering of values from the two samples (rather than the values themselves) is apparent from the construction of the test, as discussed in the main text. Each ordering of X and Y values is equally likely under $H_0 : F_X(\cdot) = F_Y(\cdot)$ and Assumptions 1 and 2; as usual, with larger sample sizes, permutations are randomly sampled rather than fully enumerated.

C.2 Calibration accuracy

As introduced in Appendix C.1, to search for the $\tilde{\alpha}$ that maps to a desired α, the required number of Dirichlet draws (M) and the tolerance parameter (T) must be specified. They may be determined given the desired overall simulation error. Given α, we chose to determine $\tilde{\alpha}$ such that the true FWER would be within $c\alpha$ of the desired α for some small $c > 0$, like $c = 0.02$ for $\alpha = 0.05$ implying FWER of 0.05 ± 0.001. As in Appendix C.1, the search stops when $|\hat{\alpha} - \alpha| < T$. The M Dirichlet draws are iid, so the total number of draws with a familywise error follows a binomial distribution. Since M is large, the normal approximation is quite accurate. We want the simulation to have a high probability, like $1 - p = 0.95$, of estimating $\hat{\alpha} > \alpha + T$ when $\tilde{\alpha}$ yields a true FWER above $\alpha(1 + c)$. If the true FWER is $\alpha(1 + c)$, then the total number of simulated familywise errors follows a Binomial$(M, \alpha(1 + c))$ distribution, so $\hat{\alpha} \overset{a}{\sim} \mathrm{N}\big(\alpha(1 + c), \alpha(1 + c)[1 - \alpha(1 + c)]/M\big)$, and we choose T and M to equate T with the p-quantile of this distribution:

$$\begin{aligned}
\alpha + T &= \alpha(1 + c) + \Phi^{-1}(p)\sqrt{\alpha(1 + c)(1 - \alpha(1 + c))}/\sqrt{M}, \\
T &= c\alpha - \Phi^{-1}(1 - p)\sqrt{\alpha(1 + c)(1 - \alpha(1 + c))}/\sqrt{M}, \\
M &= \left(\frac{\Phi^{-1}(1 - p)\sqrt{\alpha(1 + c)(1 - \alpha(1 + c))}}{c\alpha - T}\right)^2.
\end{aligned}$$

For $\alpha \in \{0.10, 0.05\}$, we used $M = 2 \times 10^5$, $p = 0.05$, and $c = 0.02$, leading to $T \approx 0.00019$ for $\alpha = 0.05$ and $T \approx 0.00089$ for $\alpha = 0.10$, as seen in the lookup table. For $\alpha = 0.01$, we used $M = 10^6$, $p = 0.05$, and $c = 0.05$, leading to $T \approx 0.00033$. The foregoing discussion applies equally to one-sample and two-sample inference.

C.3 Two-sample adjustment for discreteness

In the two-sample setting, the mapping from $\tilde{\alpha}$ to α is still monotonic but not continuous: it is a step function. Consequently, we suggest subtracting a small amount like 0.0001 from whichever $\tilde{\alpha}$ is found by the numerical solver. Additionally, in our lookup table of pre-computed values, we report both the smaller and larger α values at the discontinuity, to show how big the possible FWER inflation is if the simulation error is large enough that actually the next-highest α is the true FWER.

The subtraction of 0.0001 from the simulated $\tilde{\alpha}$ is because simulation error does not necessarily go to zero as the number of simulations goes to infinity, because the number of attainable α is finite. That is, the mapping from $\tilde{\alpha}$ to FWER is a step function, so if one picks the largest possible $\tilde{\alpha}$ such that FWER is below α, even an infinitessimal amount of simulation error could mean that actual FWER is above α. For example, if actual FWER equals $0.08 + 0.04\,\mathbb{1}\{\tilde{\alpha} \geq 0.03\}$, but simulated FWER is $0.08 + 0.04\,\mathbb{1}\{\tilde{\alpha} > 0.03\}$, then $\tilde{\alpha} = 0.03$ appears to control FWER below $\alpha = 0.1$ in the simulation, but actual FWER is 0.12, above α. Subtracting any small, fixed amount from the simulated $\tilde{\alpha}$ is sufficient to overcome this problem (with probability approaching one) as the number of simulation draws grows arbitrarily large.

D Additional simulations

D.1 Power compared to KS-based methods

Earlier, Figures 2 and 3 showed simulation results on the uneven sensitivity of KS-based MTPs and the (relatively) even sensitivity of the Dirichlet MTPs, in terms of pointwise type I error rates. Naturally, those differences translate into corresponding differences in pointwise power. Figures 9 and 10 show patterns similar to Figure 2: the KS-based MTP has the highest (among the three methods) pointwise power against deviations near the median of a distribution and lowest pointwise power in the tails, and the weighted KS-based MTP is usually the opposite (depending whether the null is above or below the true distribution; see below). The Dirichlet MTP has the highest pointwise power against deviations in between the middle and the tails, and it never has the lowest.

Figures 9 and 10 show examples of pointwise power for two-sided $H_{0\tau} : F^{-1}(\tau) = F_0^{-1}(\tau)$ over $\tau \in (0, 1)$. The left column graphs show $F_0\big(F^{-1}(\tau)\big)$ (dashed line). If H_0 were true, then $F_0\big(F^{-1}(\tau)\big) = \tau$ (solid line). Similar to Figure 2, the right column graphs show RPs due to each order statistic.

Figure 9 shows $X_i \overset{iid}{\sim} \mathrm{N}(0.3, 1)$ when the null is $\mathrm{N}(0, 1)$. As the left column shows, this leads to larger deviations in the middle of the distribution than in the tails. The largest peak in pointwise power is in the middle of the distribution for KS: this is where both the

deviations are largest and the KS pointwise size is largest. The Dirichlet pointwise power peaks in a similar range, but at a lower level, corresponding to its lower pointwise size in that range. The weighted KS pointwise power peaks in the lower tail, at a much lower level since the deviations are smaller.

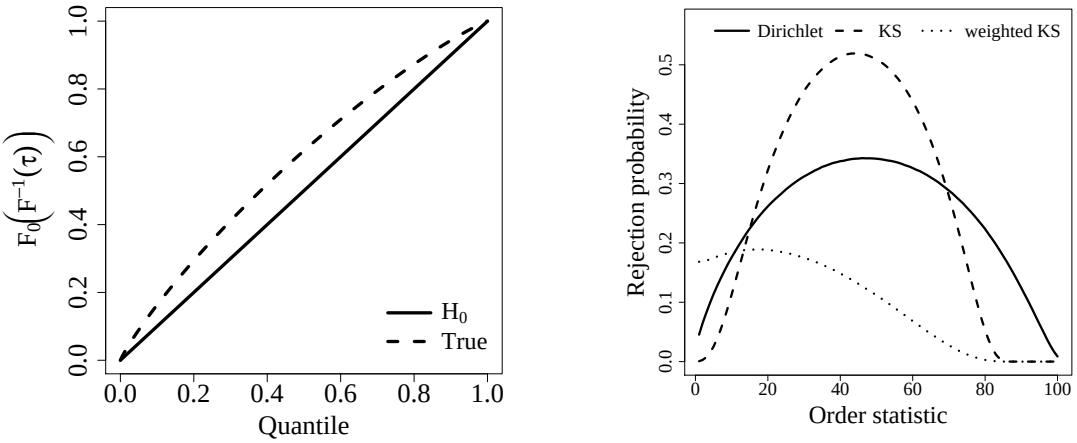

Figure 9: Simulated one-sample, two-sided RPs by order statistic when all $H_{0\tau}$ are false, $F_0 = \mathrm{N}(0,1)$, $X_i \overset{iid}{\sim} \mathrm{N}(0.3,1)$, FWER $\alpha = 0.1$, $n = 100$, 10^6 replications.

In Figure 9, the effect having pointwise equal-tailed (like Dirichlet) or symmetric (like KS) tests is apparent. Even though the weighted KS has greater (than Dirichlet) two-sided pointwise type I error rate in the upper tail, it has essentially zero power in the upper tail in the examples provided, whereas Dirichlet has substantial power. This is because $F_0(x) > F(x)$ in the upper tail; regardless of weighting, KS-based MTPs (or tests) are insensitive to such deviations, whereas the Dirichlet MTP is sensitive to both upper and lower deviations.

In the row of Figure 10 where $\sigma = 1.2$, the weighted KS again has pointwise power near zero even in the tails. This is an example of the same general feature seen in Figure 10: because of being pointwise symmetric instead of equal-tailed, the KS approach (whether weighted or not) has low power against a null with smaller variance than the DGP. The Dirichlet has two pointwise power peaks, reflecting the varying distance between the two curves in the corresponding left column graph. The KS has a much smaller pointwise power peak surrounding the median, where the deviations are small (and even zero right at the median) but its sensitivity is highest.

For the graph in Figure 10 with $\sigma = 0.7$, the weighted KS pointwise power has the highest peak, in the tails (and highest at the extremes) where the deviations are large and its sensitivity is large. The Dirichlet has a somewhat smaller peak, also in the tails but not at the extremes. Even smaller and closer to the middle is the KS peak. The weighted KS and KS can have very high peaks since their peak pointwise type I error rate is higher than Dirichlet's (which has no peak), but they perform poorly when their peak pointwise type I error rate coincides with low deviations from the null hypothesis. The Dirichlet is more even-keeled, yet it can still have the highest peak pointwise power of the three methods, especially if the deviations are largest in between the tails and median (where its pointwise size is largest), a case not even shown in these graphs.

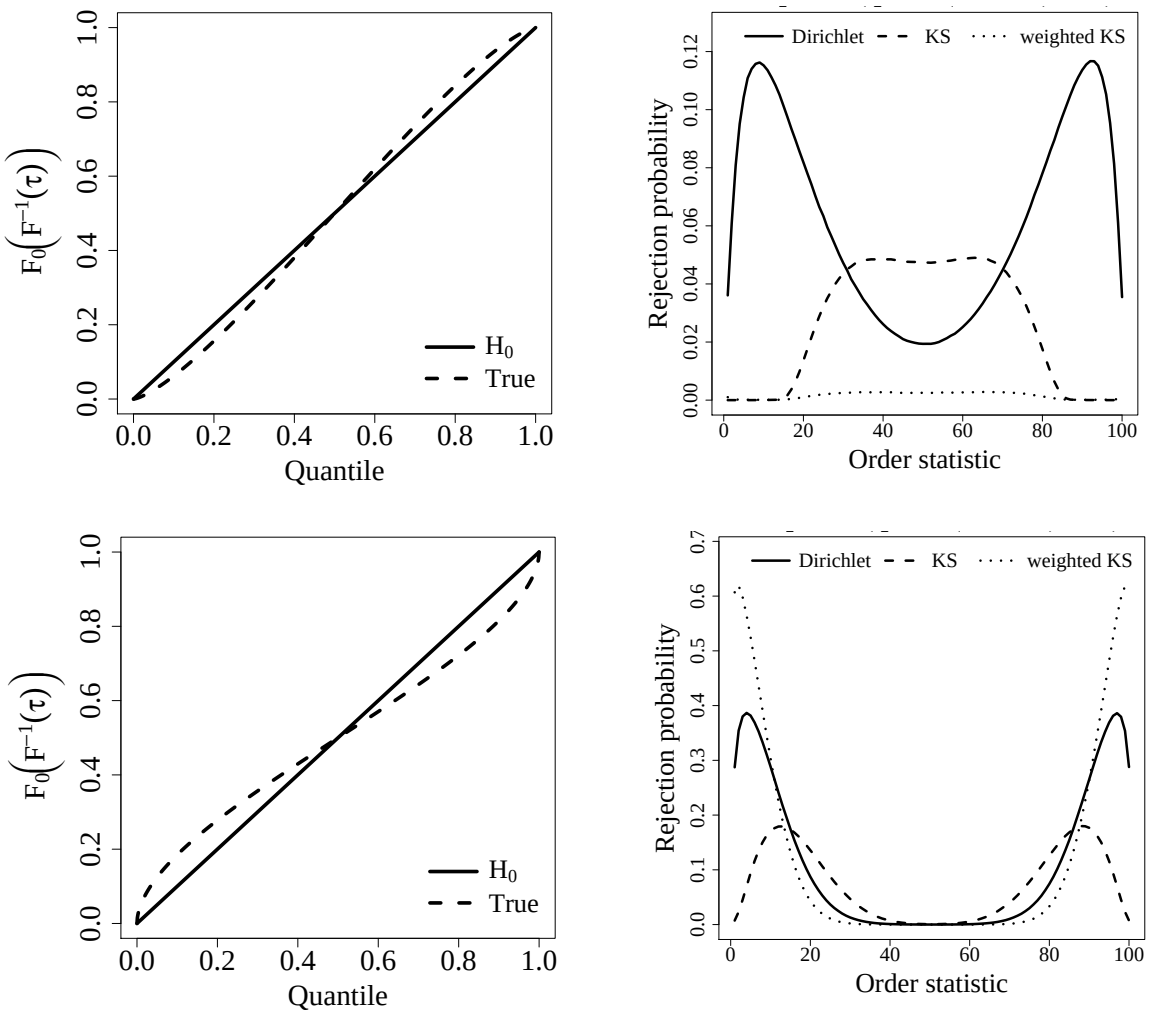

Figure 10: Simulated one-sample, two-sided RPs by order statistic when all $H_{0\tau}$ are false (except $\tau = 0.5$), $F_0 = \mathrm{N}(0,1)$, FWER $\alpha = 0.1$, $n = 100$, 10^6 replications; $X_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$ with $\sigma = 1.2$ (top) or $\sigma = 0.7$ (bottom).

Table 7 shows global power for one-sample, two-sided GOF tests of $H_0 : F(\cdot) = F_0(\cdot)$ with $F_0 = \mathrm{N}(0,1)$ and $X_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$. For the Dirichlet, KS, and weighted KS tests alike, this is equivalent to testing $H_0 : F_0(X_i) \overset{iid}{\sim} \mathrm{Unif}(0,1)$, or $H_0 : F_0\big(F^{-1}(\tau)\big) = \tau$.[16] For pure location shifts with $\mu \neq 0$ and $\sigma = 1$, the deviations (of $F_0\big(F^{-1}(\tau)\big)$ from τ) are largest near the middle of the distribution, where KS has the largest pointwise power. The weighted KS is not very sensitive to such deviations, so it has the worst power by far. The Dirichlet power is below KS, but only by a couple percentage points. With $\mu = 0$ and $\sigma = 0.7$, the largest vertical deviations of $F_0\big(F^{-1}(\tau)\big)$ are in the tails (i.e., near zero and one). Consequently, the weighted KS has the best power. The KS test has significantly lower power, but the Dirichlet is close to the weighted KS. With $\mu = 0$ and $\sigma = 0.8$, Dirichlet power is again between weighted KS (best) and KS (worst). When $\mu = 0$ and $\sigma = 1.2$, the deviations of

[16] When the population CDF is $F(\cdot)$, then $F_0(X_i) = F_0(F^{-1}(F(X_i))) = F_0(F^{-1}(U_i))$, $U_i \overset{iid}{\sim} \mathrm{Unif}(0,1)$.

$F_0\big(F^{-1}(\tau)\big)$ are no longer largest at the extremes. This poses a problem for the weighted KS, and its power is even lower than its size. Even though there is zero deviation at $\tau = 0.5$, KS has better power than weighted KS in this case because it has better pointwise power around the upper and lower quartiles. The Dirichlet pointwise power is even higher in those regions, so its global power is far above either KS or weighted KS.

Table 7: Simulated one-sample, two-sided, global power against $H_0 : F(\cdot) = F_0(\cdot)$, $F_0 = \mathrm{N}(0, 1)$, $\alpha = 0.1$ (all methods have exact size here), $n = 100$, 10^6 replications, $X_i \overset{iid}{\sim} \mathrm{N}(\mu, \sigma^2)$. RPs are shown as percentages.

μ	σ	Dirichlet	KS	weighted KS
0.3	1.0	80.5	82.4	62.4
0.2	1.0	49.4	52.2	33.9
0.0	0.7	92.0	65.6	98.6
0.0	0.8	50.1	26.7	76.6
0.0	1.2	64.2	25.5	2.8

Additionally, Table 1 and Figure 8 in Aldor-Noiman et al. (2013) show a power advantage of the Dirichlet GOF test over the KS and Anderson–Darling (i.e., weighted Cramér–von Mises) tests for a variety of distributions.