# Supplemental material: Comparing distributions by multiple testing across quantiles or CDF values

Matt Goldman* David M. Kaplan†

February 22, 2018

**Abstract**

This supplement includes additional methods, computational details, details on two-sample quantile multiple testing, and additional simulation results.

*mattgold@microsoft.com.
†kaplandm@missouri.edu.

# B Additional methods

## B.1 One-sample methods

**Method 8.** For Task 1, modify Method 3 as follows. Define $\ell_k$ and $u_k$ as in (7). Instead of only $r_{k,i}$ corresponding to either $\ell_k$ or $u_k$, include both $r_{k,\ell,i}$ corresponding to $\ell_k$ and $r_{k,u,i}$ corresponding to $u_k$. Instead of $\hat{K}_0 = \{1, \ldots, n\}$, let $\hat{K}_0^\ell = \hat{K}_0^u = \{1, \ldots, n\}$, where $\hat{K}_0^\ell$ corresponds to the $\ell_k$ and $\hat{K}_0^u$ to the $u_k$. Replace (11) with

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

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 $H_{0\tau}\colon F^{-1}(\tau) = F_0^{-1}(\tau)$. □

**Method 9** (Pre-test only)**.** Consider the pre-test null hypotheses $H_{0\ell_k}\colon F^{-1}(\ell_k) \leq F_0^{-1}(\ell_k)$, defining $\ell_k$ 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 $\{1, \ldots, n\}$. 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 $H_{0\tau}\colon F^{-1}(\tau) \leq F_0^{-1}(\tau)$ when $\max\{X_{n:r_k} : \ell_k \leq \tau\} > F_0^{-1}(\tau)$.

To instead pre-test $H_{0u_k}\colon F^{-1}(u_k) \geq F_0^{-1}(u_k)$, reverse all inequalities and min/max, and replace $\ell_k$ 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 $\bigcap_{k=1}^{\bar{k}}$. □

## B.2 Two-sample quantile MTP and procedures to improve power

We propose a two-sample quantile MTP along with stepdown and pre-test procedures. Unlike the other methods in this paper, these are not based on finite-sample distributions of order statistics. Instead, we (slightly) extend results from Goldman and Kaplan (2018). 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 $j = 1, \ldots, M_n$, $H_{0j}\colon F_X^{-1}(t) = F_Y^{-1}(t)$ 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 (2018) use "fractional order statistics" to construct a CI for each $\Delta_j$ 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 $\Delta_j$ 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 $\Delta_j$. 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 (2018) break down, but we conjecture they still hold with $M_n = O(n^{2/5})$.

**Method 10.** 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

$$1 - \alpha = \mathrm{P}\left(\bigcap_{j\in\hat{T}_i}\big\{D_{X,j}^\ell < D_{Y,j}^u, D_{Y,j}^\ell < D_{X,j}^u\big\}\right), \tag{B.2}$$

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

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

with vector $k = (k_1, \ldots, k_{2M_n})$ containing all the $k_{X,j}^\ell$ and $k_{X,j}^u$ in ascending order so that $k_1 \leq \cdots \leq k_{2M_n}$; 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 $[Y_{n_Y:k_{Y,j}^\ell} - X_{n_X:k_{X,j}^u}, Y_{n_Y:k_{Y,j}^u} - X_{n_X:k_{X,j}^\ell}]$ 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 (B.2) 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 11.** For Task 6, using notation from Method 10, consider $H_{0j}\colon F_X^{-1}(\tau_j) \geq F_Y^{-1}(\tau_j)$. First run a pre-test of $H'_{0j}\colon F_X^{-1}(\tau_j) \leq F_Y^{-1}(\tau_j)$ using iteration $i = 0$ of Method 10 (i.e., the basic method without stepdown) with FWER level $\alpha_p = \alpha/\ln[\ln(\max\{n, 15\})]$. Then, use Method 10 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 10 and 11 have strong control of asymptotic FWER.

# C Two-sample quantile MTP: difficulties

Theorem 9 does not have a corollary for interpreting Method 5 as a quantile MTP for the hypotheses $H_{0\tau}\colon F_X^{-1}(\tau) = F_Y^{-1}(\tau)$.[26] In terms of the proof of Theorem 9, the use of Lemma 2 would not be valid: rejection of $H_{0\tau}$ depends on order statistics $X_{n_X:k}$ and $Y_{n_Y:m}$ for some $k$ and $m$, but their finite-sample distributions depend on more than just $F_X^{-1}(\tau)$ and $F_Y^{-1}(\tau)$. Specifically, from Theorem 4, $F_X(X_{n_X:k}) \sim \text{Beta}(k, n_X + 1 - k)$, so the distribution of $X_{n_X:k}$ depends on all of $F_X^{-1}(\cdot)$ in finite samples.

More simply, a quantile version of Theorem 9 for Method 5 cannot be proved because it is false. A counterexample shows this. Let $X = 0.5$ be a degenerate random variable.[27] Let $\text{P}(Y = 0) = \text{P}(Y = 1) = 0.5 - e$ and $\text{P}(Y = U) = 2e$, for $U \sim \text{Unif}(0, 1)$ and small $e > 0$. Thus, among the $H_{0\tau}\colon F_X^{-1}(\tau) = F_Y^{-1}(\tau)$, only the $\tau = 0.5$ hypothesis is true. Let $n_X = 6$ and $n_Y = 12$. For $\alpha = 0.05$, Method 5 has $\tilde{\alpha} = 0.154$. Thus, $B^{\tilde{\alpha}}_{n_X,n_X} > 0.5$, $B^{1-\tilde{\alpha}}_{1,n_X} < 0.5$, $B^{\tilde{\alpha}}_{9,n_Y} > 0.5$, and $B^{1-\tilde{\alpha}}_{4,n_Y} < 0.5$, so $H_{0,\tau=0.5}$ is rejected if $Y_{n_Y:9} = 0$ or $Y_{n_Y:4} = 1$. As $e \to 0$, $\text{P}(Y_{n_Y:9} = 0)$ is the probability that no more than $n_Y - 9 = 3$ observations have $Y_i = 1$, which is the Binomial$(n_Y, 0.5)$ CDF evaluated at $n_Y - 9$, which is 0.0730. By symmetry, $\text{P}(Y_{n_Y:9} = 0) = 0.0730$, too. Thus, the quantile FWER is 0.146, slightly below $\tilde{\alpha}$ but well above $\alpha = 0.05$. If $\alpha = 0.01$, $n_X = 6$, and $n_Y = 11$, then $\tilde{\alpha} = 0.0716$, and similar calculations show FWER $= 0.0654$, again slightly below $\tilde{\alpha}$ but well above $\alpha$.

The preceding counterexample's intuition follows. For testing $H_{0\tau}\colon F_X^{-1}(\tau) = F_Y^{-1}(\tau)$ for a single $\tau$, using a Dirichlet-based method similar to ours, Goldman and Kaplan (2018, §3.3) show the importance of setting $\tilde{\alpha}$ based not only on $\alpha$, $n_X$, and $n_Y$, but also the ratio of the quantile function derivatives at $\tau$, $Q'_X(\tau)/Q'_Y(\tau)$. (This relates to the ratio of asymptotic variances of the corresponding sample quantiles.) They show that when this ratio equals one, $\tilde{\alpha}$ should be much larger than $\alpha$ to have near-exact size, but when the

[26] We thank the referees for pushing us to determine this definitively.

[27] Technically, this violates Assumption 2, but $X \sim \text{Unif}(0.5 - e, 0.5 + e)$ for arbitrarily small $e > 0$ results in FWER that is arbitrarily close to the $e \to 0$ limit. The same comment applies to the mass points in the distribution of $Y$.

ratio approaches zero or infinity, $\tilde{\alpha} = \alpha$ is required. In Method 5, $\tilde{\alpha}$ is calibrated to the case when $F_X(\cdot) = F_Y(\cdot)$ and thus implicitly $Q'_X(\tau)/Q'_Y(\tau) = 1$ for all $\tau \in (0,1)$. In the counterexample, this implicit assumption is violated: $Q'_X(0.5) = 0$ and $Q'_Y(0.5) \to \infty$ as $e \to 0$, so $Q'_X(0.5)/Q'_Y(0.5) = 0$. Consequently, the rejection probability is closer to $\tilde{\alpha}$ than $\alpha$. However, $\tilde{\alpha} \to 0$ as $n_X, n_Y \to \infty$, so $\tilde{\alpha} > \alpha$ only in very small samples. This provides a helpful bound on FWER with this particular DGP (and one could thus control FWER by ensuring $\tilde{\alpha} \leq \alpha$), but it is unclear whether such a bound applies to more complex DGPs with multiple true $H_{0\tau}$.

Unlike strong control of FWER, weak control of FWER can be established for the quantile hypotheses. For two-sided hypotheses, as in Definition 2, weak control of FWER for the quantile hypotheses $H_{0\tau}$ means FWER $\leq \alpha$ if all $H_{0\tau}$ are true, i.e., if $F_X^{-1}(\tau) = F_Y^{-1}(\tau)$ for all $\tau \in [0,1]$, or more simply if $F_X^{-1}(\cdot) = F_Y^{-1}(\cdot)$. Since $F_X^{-1}(\cdot) = F_Y^{-1}(\cdot)$ is equivalent to $F_X(\cdot) = F_Y(\cdot)$, and since Method 5 has FWER $\leq \alpha$ in that case (i.e., has any rejection with less than $\alpha$ probability), then the quantile interpretation would also have FWER $\leq \alpha$ in that case.

However, a test with only weak control of FWER is in principle no more informative than a GOF test. Consider the MTP that rejects $H_{0\tau}$ for all $\tau \in (0,1)$ whenever a level-$\alpha$ GOF test rejects, and otherwise the MTP rejects none of the $H_{0\tau}$. Under $F_X^{-1}(\cdot) = F_Y^{-1}(\cdot)$, the GOF test's rejection probability is below $\alpha$, so the MTP's familywise rejection probability (and thus FWER) is also below $\alpha$, satisfying weak control of FWER. However, if enough of the $H_{0\tau}$ are false that the GOF test rejects 80% of the time (as it should), then the MTP has 80% FWER since it falsely rejects all the true $H_{0\tau}$ along with the false $H_{0\tau}$. Not only does this technically violate strong control of FWER, it makes the MTP's rejection of a particular $H_{0\tau}$ uninformative in practice: we do not know if that $H_{0\tau}$ is actually false, or if it is rejected only because some other $H_{0\tau}$ is false. Our own MTP is not nearly so egregious, with only small FWER distortion even in our highly contrived example (and with proper FWER control in many other examples we tried), but we are reluctant to endorse a method that we know lacks strong control of FWER for the foregoing reason.

If quantiles are truly desired and distributional hypotheses do not suffice, then one-sample uniform confidence bands for the two quantile functions could be combined, but the result will be conservative. The two true quantile functions have probability $1 - \alpha$ of both lying in their respective $\sqrt{1-\alpha}$ uniform confidence bands, so the quantile difference function $F_Y^{-1}(\cdot) - F_X^{-1}(\cdot)$ has at least $1 - \alpha$ probability of lying in the "difference" of the two bands. Note that our Method 5 MTP is also constructed (implicitly) using uniform confidence bands, but with lower than $1 - \alpha$ coverage, whereas $\sqrt{1-\alpha} > 1 - \alpha$, so this approach is conservative.

# D 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 $\alpha$.

## D.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 $U_i^{(m)} \overset{iid}{\sim} \text{Unif}(0,1)$ for observations $i = 1, \ldots, n$ in samples $m = 1, \ldots, M$. Given $\tilde{\alpha}$, which determines all $\ell_k$ and $u_k$, the simulated two-sided FWER (for example) is

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

While (D.1) alone is sufficient for global (GOF) $p$-value computation, we need to search for the $\tilde{\alpha}$ that leads to a specific desired $\alpha$ for the simulations informing Fact 6. Given search tolerance $T$ (see Appendix D.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 (D.1). The GOF null $H_0\colon F_X(\cdot) = F_Y(\cdot)$ 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 $H_0\colon F_X(\cdot) = F_Y(\cdot)$ and Assumptions 1 and 2; as usual, with larger sample sizes, permutations are randomly sampled rather than fully enumerated.

## D.2 Calibration accuracy

As introduced in Appendix D.1, to search for the $\tilde{\alpha}$ that maps to a desired $\alpha$, 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 $\alpha$, we chose to determine $\tilde{\alpha}$ such that the true FWER would be within $c\alpha$ of the desired $\alpha$ for some small $c > 0$, like $c = 0.02$ for $\alpha = 0.05$ implying FWER of $0.05 \pm 0.001$. As in Appendix D.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.

### D.3 Two-sample adjustment for discreteness

In the two-sample setting, the mapping from $\tilde{\alpha}$ to $\alpha$ 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 $\alpha$ 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 $\alpha$ 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 $\alpha$ 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 $\alpha$, even an infinitessimal amount of simulation error could mean that actual FWER is above $\alpha$. 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 $\alpha$. 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.

## E Additional simulations

### E.1 Computation time

Table 6 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 Fact 6) is also shown; this is the time saved by Fact 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 6: Computation time (seconds); one-sample, two-sided, $\alpha = 0.1$.

| $\log_{10}(n)$ | Fact 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 6, 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 Fact 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.

## E.2 Empirical-based DGP

A DGP based on the "gift wage" empirical example in Section 8.1 was constructed as follows. For both the library and fundraising tasks, using the Period 1 data, for both treatment and control groups, first a piecewise linear quantile function was interpolated between points $(\tau, F^{-1}(\tau))$ consisting of $(k/(n+1), Y_{n:k})$, $(0, 0)$, and $(1, Y_{n:n}+10)$. Second, this was modified to only include integer values (as in the data) by applying the floor function $\lfloor\cdot\rfloor$ to generate a step function $F^{-1}(\cdot)$. These are the true population quantile functions for our simulations. The sample sizes are the same as in our empirical example (library: 10 control, 9 treatment; fundraising: 10 control, 13 treatment), as is the nominal one-sided FWER level $\alpha = 0.1$. There were 10 000 simulation replications.

We compare five methods: "Basic" is our Dirichlet-based MTP in Method 5, "KS" is the KS-based MTP as in Proposition 3, "Joint" is a joint quantile difference test (iteration $i = 0$ from Method 10), "Stepdown" is Method 10, and "Pre+Step" is Method 11. Basic and KS test $H_{0r}\colon F_T(r) \geq F_C(r)$ over each integer $r$ between 0 and the maximum possible value, where subscript $T$ stands for "treatment" and $C$ for "control." In the library task, $H_{0r}$ is true for $0 \leq r \leq 27$; in the fundraising task, $H_{0r}$ is true for $42 \leq r \leq 44$. The quantile tests evaluate $H_{0\tau}$ for $\tau \in \{0.30, 0.50, 0.70\}$ for the library task and $\tau \in \{0.22, 0.41, 0.59, 0.78\}$ for fundraising; all $H_{0\tau}$ are false.

The FWER is nearly zero for both the Basic and KS MTPs: 0.002 for the library DGP and 0.001 for fundraising (for both methods). The FWER is close to zero because $H_{0r}$ is true for relatively small ranges of $r$. This is similar to the FWER in the last row of Table 4, where it is shown how FWER only gets close to the nominal level when nearly all $H_{0r}$ are true.

The Basic MTP has the best global power against $H_0\colon F_T(\cdot) \geq F_C(\cdot)$. That is, it has the highest probability of rejecting at least one $H_{0r}$. Next best are the joint quantile tests (which are all the same because the pre-test does not help for these DGPs and the stepdown cannot increase *global* power). The KS has the worst global power. Although not surprising that the Basic MTP has better global power than the KS, it is surprising that it fares better

than the joint quantile tests that focus power on a smaller number of points where all $H_{0\tau}$ are false. The simple explanation may be that these few $\tau$ do not match up with the most statistically obviously false $H_{0\tau}$. So at least here, the "evenly sensitive" approach of the Basic MTP actually leads to the best global power, too.

Table 7: FWER and global power for empirical simulation.

| | FWER | | Global power | |
|---|---|---|---|---|
| Method | Library | Fundraising | Library | Fundraising |
| Basic | 0.002 | 0.001 | 0.647 | 0.815 |
| KS | 0.002 | 0.001 | 0.477 | 0.714 |
| Joint | 0.000 | 0.000 | 0.583 | 0.758 |
| Stepdown | 0.000 | 0.000 | 0.583 | 0.758 |
| Pre+Step | 0.000 | 0.000 | 0.583 | 0.759 |

Figure 10 shows pointwise rejection probability (RP), similar to Figure 7. The joint quantile tests generally have higher pointwise RP, although the magnitude partly depends on whether $r$ is compared with $F_T^{-1}(\tau)$ (as in Figure 10) or $F_C^{-1}(\tau)$. The pre-test is useless because all $H_{0\tau}$ are false. The stepdown procedure improves pointwise RP by a few percentage points, sometimes more, sometimes less. Most notably, even though we used a nominal FWER level slightly above 10% for the KS MTP and slightly below 10% for the Dirichlet MTP, the Dirichlet MTP has significantly higher pointwise RP (i.e., power) than the KS MTP across a range of $r$.

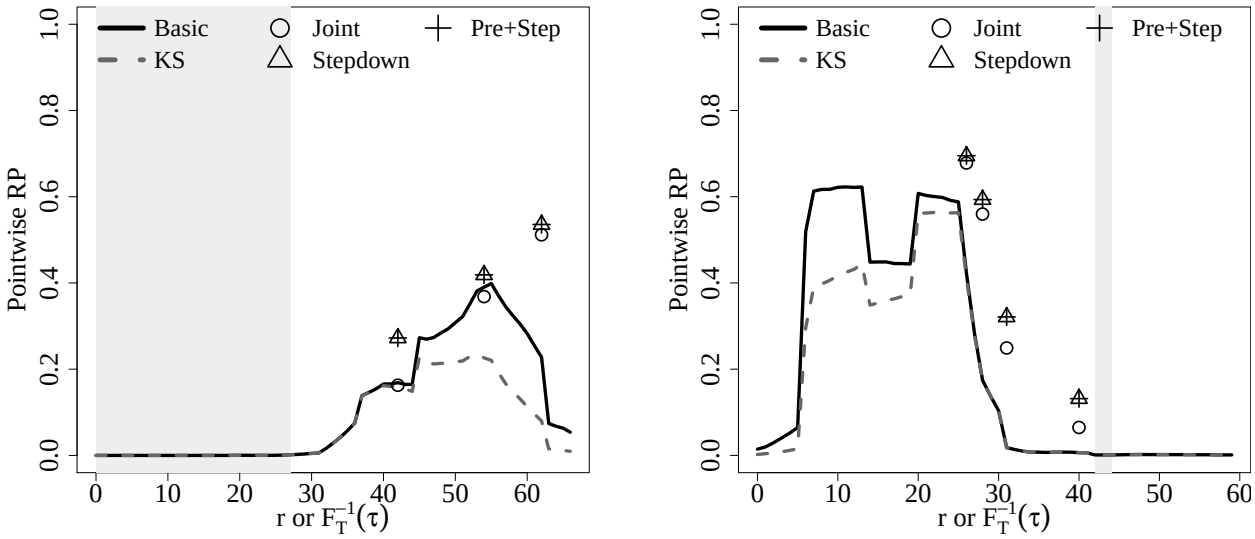


Figure 10: Simulated pointwise RP, empirical simulations, $H_{0r}\colon F_T(r) \geq F_C(r)$. For the Basic and KS MTPs, the horizontal axis shows the value of $r$; otherwise, it shows $F_T^{-1}(\tau)$. Gray shading indicates true $H_{0r}$. Left: library data entry task (units: books). Right: door-to-door fundraising task (units: dollars).

## E.3 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 11 and 12 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 11 and 12 show examples of pointwise power for two-sided $H_{0\tau}\colon F^{-1}(\tau) = F_0^{-1}(\tau)$ 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 11 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.

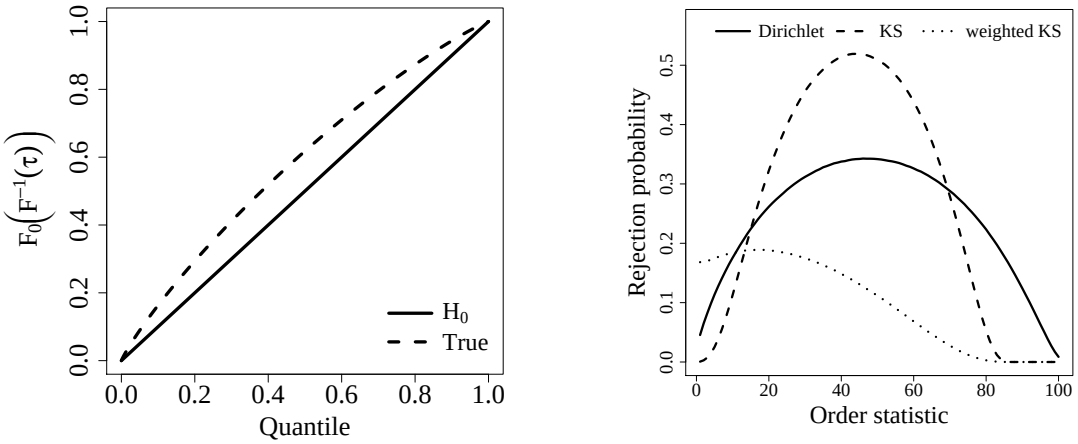


Figure 11: Simulated one-sample, two-sided RPs by order statistic when all $H_{0\tau}$ are false, $F_0 = \mathrm{N}(0,1)$, $X_i \overset{iid}{\sim} \mathrm{N}(0.3, 1)$, FWER $\alpha = 0.1$, $n = 100$, $10^6$ replications.

In Figure 11, 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 12 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 12: 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.

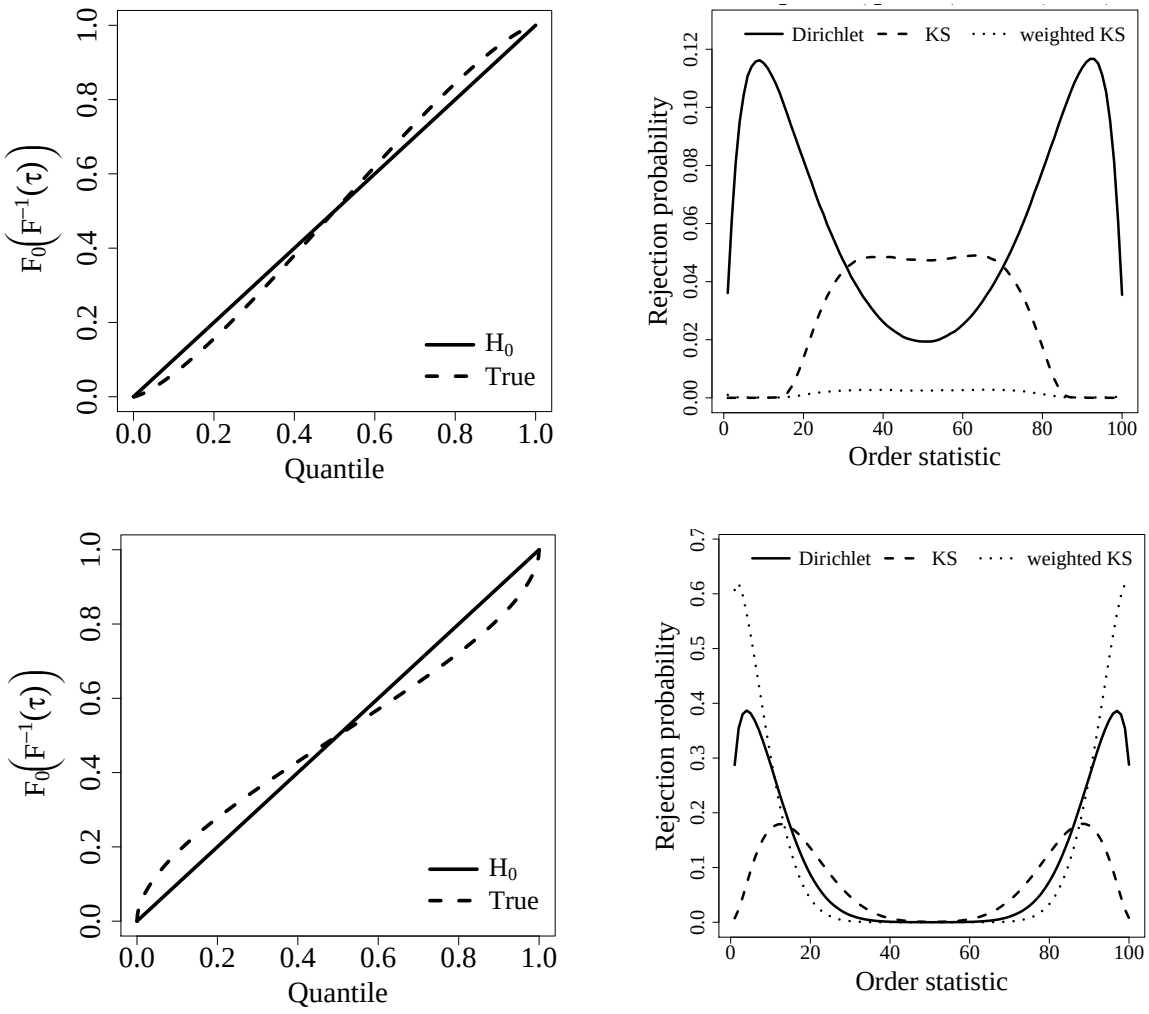


Figure 12: Simulated one-sample, two-sided RPs by order statistic when all $H_{0\tau}$ are false (except $\tau = 0.5$), $F_0 = \mathrm{N}(0,1)$, FWER $\alpha = 0.1$, $n = 100$, $10^6$ replications; $X_i \overset{iid}{\sim} \mathrm{N}(0,\sigma^2)$ with $\sigma = 1.2$ (top) or $\sigma = 0.7$ (bottom).

For the graph in Figure 12 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.

Table 8: Simulated global power, one-sample, two-sided, $\alpha = 0.1$, $n = 100$.

| $\mu$ | $\sigma$ | 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 |

**Note:** $H_0\colon F(\cdot) = F_0(\cdot)$, $F_0 = \mathrm{N}(0,1)$, $X_i \overset{iid}{\sim} \mathrm{N}\big(\mu, \sigma^2\big)$, $10^6$ replications. RPs are shown as percentages. All methods have exact size.

Table 8 shows global power for one-sample, two-sided GOF tests of $H_0\colon F(\cdot) = F_0(\cdot)$ 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 $H_0\colon F_0(X_i) \overset{iid}{\sim} \mathrm{Unif}(0,1)$, or $H_0\colon F_0\big(F^{-1}(\tau)\big) = \tau$.[28] For pure location shifts with $\mu \neq 0$ and $\sigma = 1$, the deviations (of $F_0\big(F^{-1}(\tau)\big)$ from $\tau$) 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 $F_0\big(F^{-1}(\tau)\big)$ 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.

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.

# References

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

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

URL https://doi.org/10.1080/00031305.2013.847865

Buja, A., Rolke, W., 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.

Goldman, M., Kaplan, D. M., 2018. Nonparametric inference on conditional quantile differences and linear combinations, using $L$-statistics. Econometrics Journal XXX (XX), XXX–XXX.
URL https://doi.org/10.1111/ectj.12095

Stigler, S. M., 1977. Fractional order statistics, with applications. Journal of the American Statistical Association 72 (359), 544–550.
URL https://www.jstor.org/stable/2286215