Evenly sensitive KS-type inference on distributions

Matt Goldman*       David M. Kaplan†

February 12, 2016

Abstract

Despite low sensitivity in the tails, the Kolmogorov–Smirnov test is widely used in empirical economics because it: 1) is distribution-free and nonparametric, 2) provides uniform confidence bands for the CDF by inversion, 3) identifies which particular part(s) of the distribution caused the test to reject, 4) supports finite-sample critical values, and 5) computes quickly. The recent Dirichlet approach of [Buja and Rolke] (2006) uses the probability integral transform to attain more even sensitivity while retaining (1)–(4), but not (5). We contribute to computation, interpretation, and power of the Dirichlet approach, including first-order stochastic dominance testing. First, we make one-sample computation nearly instant with a new formula that replaces just-in-time simulation. Second, we propose a variant of the two-sample test that also enables pre-computation of finite-sample critical values. Third, we show the Dirichlet-based CDF confidence bands to be uniform credible bands in a particular nonparametric Bayesian framework. Fourth, we interpret the global null hypotheses as families of pointwise quantile or CDF hypotheses, and we establish the one-sample and two-sample methods’ strong control of familywise error rate (FWER). Fifth, we propose stepdown and pre-test procedures to improve power while maintaining strong control of FWER. Simulations, empirical examples, and code are provided.

JEL classification: C21

Keywords: Dirichlet distribution; familywise error rate; finite-sample; first-order stochastic dominance; Kolmogorov–Smirnov; probability integral transform; uniform confidence band

*Microsoft, mattgold@microsoft.com.
†Corresponding author. Department of Economics, University of Missouri, kaplandm@missouri.edu. We thank the editor (Debopam Bhattacharya) and three anonymous referees for helpful revision ideas that substantially improved the presentation and content, including FWER and pre-testing. We thank Boaz Nadler and Amit Moscovich-Eiger for alerting us to important related works from statistics and computer science.
1 Introduction

In many economic settings, the mean does not tell the whole story. Individual quantiles add insight, but often the entire distribution is of interest. Inference based on the (weighted) $L_\infty$ distance between a continuous null hypothesis distribution and the empirical distribution has many advantages. It is nonparametric and distribution-free, includes hypothesis testing and $p$-values, generates uniform confidence bands that are equivariant to increasing transformations, can be calibrated to have exact finite-sample size (or coverage), and in many cases computes quickly. Also, unlike the Cramér–von Mises test (among others), the specific part(s) of the distribution that caused the test to reject are readily identified; this is often valuable information. The Kolmogorov–Smirnov (KS) test (Kolmogorov 1933; Smirnov 1939, 1948) and the variance-weighted KS test of Anderson and Darling (1952, Ex. 2) are the best-known examples.

However, the uneven sensitivity of the KS tests to deviations in different parts of the distribution is well known. In practice, this can mean failure to detect important features or causal effects. For example, unlike our evenly sensitive test, the KS test cannot reject equality of treatment and control group outcome distributions at a 5% level in an example from Gneezy and List (2006) revisited in Section 7.1 because the treatment effect is primarily in the lower part of the distribution where the KS is known to be insensitive. Additionally, the KS-based uniform confidence band for the distribution of city sizes in the example in Section 7.2 is far less informative in the tails than our evenly sensitive band, with little difference in the center of the distribution. The weighted KS could detect these differences in the tails, but it misses phenomena in the center of distributions, where it is known to be insensitive. In practice, we often wish to learn about features or treatment effects across the entire distribution without giving favor to particular quantiles. Additionally, an evenly sensitive method is not prone to manipulation where, after examining the data, power is

---

1Such settings include arbitrage and agricultural economics (Levy 1992), trade theory predictions for firm productivity distributions (Arnold and Hussinger 2010), firm size (Angelini and Generale 2008), welfare programs (Bitler, Gelbach, and Hoynes 2006, 2008), and other empirical examples in Section 7.
directed toward the quantiles with the largest deviations to artificially inflate the statistical significance.

In addition to uneven sensitivity across the distribution, the KS tests have uneven sensitivity to upper and lower deviations at a given point. Regardless of weighting or not, the KS uniform confidence bands for the CDF are symmetric around the empirical distribution function, \( \hat{F}(\cdot) \), so they necessarily include the value one when \( \hat{F}(x) \) is close to one, and they include zero when \( \hat{F}(x) \) nears zero. Near the tails, the consequence is dramatic and arguably undesirable. For example, if the null hypothesis distribution is Uniform\((0, 1)\), even a sample maximum of one million (which is impossible under the null) does not cause the KS tests to reject. With a sample size of \( n = 20 \), even five observations equal to one million cannot persuade either KS test to reject at a 10% level. In contrast, the evenly sensitive test that we consider always rejects if the sample maximum is greater than the upper bound of the null hypothesis distribution.

Buja and Rolke (2006) appear to have been the first to propose a Dirichlet-based approach to achieve more even sensitivity. They use the probability integral transform to reduce the problem to that of order statistics from a standard uniform distribution, which jointly follow a known ordered Dirichlet distribution in finite samples. The marginal beta distributions are used to achieve even sensitivity (i.e., equal pointwise type I error rates), while the joint Dirichlet distribution is used to achieve exact size of the overall one-sample test. In the spirit of even sensitivity, two-sided tests have equal probability of upper and lower rejections. Buja and Rolke (2006) also construct a two-sample permutation test for equality by comparing Q–Q plots for repeated permutations of the data. Overall, their proposed Dirichlet approach achieves much more even sensitivity than does the KS test, but it leaves room for improvement in computation time, interpretation, and power. We contribute to each of these areas.

Our new computational contributions make the Dirichlet approach more practical. The one-sample Buja and Rolke (2006) simulation can take over a minute for sample sizes in
the high thousands (see Table 2). We provide a new formula that obviates just-in-time simulation for almost any sample size and nominal level. Our method provides nearly instant computation of uniform confidence bands (for CDFs), hypothesis testing, and p-values. We also propose an alternative to the two-sample Buja and Rolke (2006) test that allows pre-computation of a lookup table (which we provide for commonly used values of $\alpha$), again obviating just-in-time simulation.

We also make several theoretical contributions that expand the interpretation of the Dirichlet methods and establish new properties. First, we recast the null hypothesis of distributional (in)equality as a family of pointwise quantile (in)equality hypotheses. This new multiple testing perspective enables deeper analysis of familywise error rate (FWER), defined as the probability of falsely rejecting at least one of the true individual hypotheses \cite{LehmannRomano2005}. There are two types of FWER control (Lehmann and Romano, 2005, p. 350): “strong control,” where $\text{FWER} \leq \alpha$ irrespective of which individual hypotheses are true, and “weak control,” which only guarantees $\text{FWER} \leq \alpha$ if all individual hypotheses are true.

Buja and Rolke (2006) only guarantee weak control of FWER since they only consider size control for testing the single, global hypothesis of distribution equality. We establish strong control of FWER for both one-sample and two-sample Dirichlet tests. Second, we show that the Dirichlet-based uniform confidence bands are also uniform credible bands for the nonparametric Bayesian approach of Banks (1988), who suggests a particular way of smoothing the Bayesian bootstrap of Rubin (1981). Our bands have even pointwise credibility levels, and our calibration formula removes the usual computational burden of nonparametric Bayesian methods; one can quickly plot credible bands for a variety of $\alpha$ to help visualize the posterior.

We also propose several new testing procedures that strictly improve power without compromising the FWER properties discussed above. For one-sample testing, we propose an iterative stepdown procedure to improve pointwise power, recalibrating the test each

--We thank an anonymous referee for alerting us to this distinction.
iteration based on the as yet unrejected individual hypotheses. We propose a new two-sample stepdown procedure, too, although for now its theoretical justification is only for a fixed (rather than growing) number of quantile hypotheses. For one-sided hypothesis testing, important for testing first-order stochastic dominance, we propose a pre-test to improve both pointwise and global power. The pre-test determines at which quantiles the null hypothesis inequality constraint may be binding, and the Dirichlet test is recalibrated with attention restricted to this subset, similar to the subset $\hat{B}$ in equations (13) and (14) of Linton, Song, and Whang (2010), for example.

The remainder of this section contains a literature review followed by notes on this paper’s structure and notation. The foundation for methods using $L^\infty$ distance is the works of Kolmogorov (1933), Smirnov (1939, 1948), and Anderson and Darling (1952). Using the $L^\infty$ distance to test whether the population belongs to a parametric family of distributions is beyond our paper’s scope, but see for example Lilliefors (1967) for normality and van der Vaart (1998, Thm. 19.23) more generally, as well as Buja and Rolke (2006, §5.1) and Aldor-Noiman, Brown, Buja, Rolke, and Stine (2013, §2.3). Anderson and Darling (1952) consider both Cramér–von Mises and KS tests based on a general weighted empirical process. However, for their Example 2 that assigns weight equal to the inverse pointwise standard deviation, they note that their results do not hold unless the tails are given zero weight, which undermines the goal of even sensitivity. If instead the weight is applied even in the tails, then the tails are much more sensitive, as Eicker (1979), Jaeschke (1979), and others show, which again fails to achieve the goal of even sensitivity.

Nonparametric (empirical) likelihood-based testing and confidence bands are respectively proposed by Berk and Jones (1979) and Owen (1995), based on pointwise binomial distributions. The former authors show greater asymptotic Bahadur efficiency over any weighted KS test, and the latter provides approximations for $\alpha = 0.01$ and 0.05 to avoid just-in-time simulation. For some $x$, let $k \equiv n\hat{p}$, with $\hat{p} \equiv \hat{F}(x)$ being the usual empirical distribu-

---

3We thank an anonymous referee for pointing us in this direction.
tion function, and \( p \equiv F(x) \). The pointwise nonparametric likelihood ratio is based on \( k \sim \text{Binomial}(n, p) \) and equals

\[
\frac{\hat{p}^k (1 - \hat{p})^{n-k}}{p^k (1 - p)^{n-k}}.
\]

The denominator is equivalent to a \( \beta(k + 1, n + 1 - k) \) PDF evaluated at \( p \), up to the normalizing constant; this is very similar to our \( \beta(k, n + 1 - k) \) distribution. However, including values of \( p \) that keep the PDF above some critical value leads to the narrowest possible intervals rather than equal-tailed ones. Also, determining the critical value from the distribution of the maximum of all pointwise ratios, as in (7) of Owen (1995), potentially spreads pointwise coverage probability unevenly over the distribution. In spite of these differences, the Owen (1995) bands are visually quite similar to ours. Still, our approach has advantages over Owen (1995): 1) simulation-free one-sample computation for any \( \alpha \), 2) formal results on strong control of FWER, 3) stepdown and pre-test procedures to improve power, and 4) two-sample inference.

More recently, [Barrett and Donald (2003)] discuss testing first-order stochastic dominance but primarily focus on higher-order stochastic dominance, which is beyond the scope of this paper. [Donald and Hsu (2016)] and [Linton et al. (2010)] focus on improving power for stochastic dominance tests, while [Linton, Maasoumi, and Whang (2005)] allow for dependence. [Donald and Hsu (2013)] consider estimation and inference for distributional treatment effects, using a propensity score. [Davidson and Duclos (2013)] argue for a null hypothesis of non-dominance rather than dominance; we sympathize with their arguments, but our tests are for the latter (as is more common).

Much work has been done on computation and tabulation of exact critical values and \( p \)-values, including Jaeschke (1979), Eicker (1979), and [Chicheportiche and Bouchaud (2012)] for the weighted KS, and [Wald and Wolfowitz (1939)], [Birnbaum (1952)], [Birnbaum and Tingey (1951)], and [Marsaglia, Tsang, and Wang (2003)], the latter two of which are implemented in \texttt{ks.test} in \texttt{R} ([R Core Team, 2013]), which is used in our simulation study.

As noted, Buja and Rolke (2006) first propose the Dirichlet approach, as a special case of
what they call calibration for simultaneity (CfS). They also discuss testing against families of
distributions with unknown parameters. Their one-sample Q–Q confidence band (equivalent
to a CDF band) from Section 5.1 was eventually detailed and published by Aldor-Noiman
et al. (2013).

Using the probability integral transform for goodness-of-fit testing dates back to R. A.
Fisher (1932), Karl Pearson (1933), and Neyman (1937). One important extension is that
the joint distribution of $F(X_{n:1}), \ldots, F(X_{n:n})$ 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). In the computer science literature, Moscovich-
Eiger, Nadler, and Spiegelman (2015) propose and implement an algorithm with $O(n^2)$
computational complexity for computing one-sample, one-sided $p$-values, and their Theorem
5.1 provides a formula to convert these to asymptotically valid two-sided $p$-values. Other
related computer science papers are based on Donoho and Jin (2004), who formalize John
Tukey’s multiple testing idea of “higher criticism.” Their HC* (p. 966) is similar to the
weighted KS statistic of Anderson and Darling (1952).

Concepts of multiple testing now appear in innumerable papers; Chapter 9 in Lehmann
and Romano (2005) introduces FWER, stepdown (and stepup) procedures, and other con-
cepts.

Section 2 reviews prior methods and helpful theory. Sections 3–5 contain our new results
for, respectively, computational improvements; quantile, multiple testing, and Bayesian in-
terpretations; and stepdown and pre-test procedures to improve power. Sections 6 and 7
contain simulation results and empirical examples, respectively. Appendix A has a table
summarizing all our new methods and their advantages. Proofs absent from the text are col-
lected in Appendix B. Details for methods not given in the main text are in Appendix C. In
the supplementary material, Appendix D details the steps for computation and Appendix E
contains additional simulation and empirical results.

For notation, we write $\beta(a, b)$ for a beta distribution with parameters $a$ and $b$, or a random
variable with such a distribution if the distinction is clear from context. The indicator function is $1\{\cdot\}$; $\hat{F}(\cdot)$ is the usual empirical distribution function. Acronyms repeatedly used include those for confidence interval (CI), cumulative distribution function (CDF), data generating process (DGP), familywise error rate (FWER), Kolmogorov–Smirnov (KS), probability density function (PDF), and rejection probability (RP). We use $\alpha$ for FWER and $\tilde{\alpha}$ for pointwise size, and similarly $1 - \alpha$ for uniform coverage and $1 - \tilde{\alpha}$ for pointwise coverage.

2 Setup and background

2.1 Prior results

Wilks (1962, pp. 236–238) gives many results for order statistics from a Uniform$(0,1)$ distribution in 8.7.1–8.7.6. The following reproduced results are of particular importance.

Assumption 1. For continuous CDF $F(\cdot)$, scalar observations $X_i \overset{iid}{\sim} F$, $i = 1, \ldots, n$.

Theorem 1 (Wilks 8.7.4, 8.7.1, 8.7.2). The following are true under Assumption 1. Denote the order statistics by $X_{n:1}, \ldots, X_{n:n}$. Then the spacings $F(X_{n:1}), F(X_{n:2}) - F(X_{n:1}), \ldots, F(X_{n:n}) - F(X_{n:n-1})$, are random variables having the $n$-variate Dirichlet distribution $D(1, \ldots, 1; 1)$. That is, the random variables $F(X_{n:1}), \ldots, F(X_{n:n})$ have the ordered $n$-variate Dirichlet distribution $D^*(1, \ldots, 1; 1)$. The marginal distributions are $F(X_{n:k}) \sim \beta(k, n+1-k)$ for $k = 1, \ldots, n$.

To construct a uniform confidence band for $F(\cdot)$, Aldor-Noiman et al. (2013, §2.1) apply Theorem 1. First, given a pointwise coverage level $1 - \tilde{\alpha}$, an exact, two-sided, equal-tailed CI for $F(X_{n:k})$ for any $k \in \{1, \ldots, n\}$ has for endpoints the $\tilde{\alpha}/2$ and $1 - \tilde{\alpha}/2$ quantiles of the $\beta(k, n+1-k)$ distribution. For example, a pointwise 90% CI for $F(X_{n:1})$ is between the $\tilde{\alpha}/2$ and $1 - \tilde{\alpha}/2$ quantiles of the $\beta(1, n)$ distribution. Wilks used the term “coverages” instead.
0.05-quantile and 0.95-quantile of the $\beta(1, n)$ distribution, respectively denoted $B_{0.05}^{1,n}$ and $B_{0.95}^{1,n}$. This may feel unnatural since $F(X_{n:1})$ is the random variable while $B_{0.05}^{1,n}$ and $B_{0.95}^{1,n}$ are fixed, whereas usually the CI endpoints are random while the object of interest is fixed, but nonetheless $P\left(B_{0.05}^{1,n} \leq F(X_{n:1}) \leq B_{0.95}^{1,n}\right) = 0.9$. For a more traditional interpretation, we may view $X_{n:1}$ as the lower endpoint of a one-sided 95% CI for $F^{-1}(B_{0.95}^{1,n})$, or the upper endpoint for $F^{-1}(B_{0.05}^{1,n})$. Either way, the results are exact and distribution-free.

Second, the full joint distribution of all $\{F(X_{n:k})\}_{k=1}^n$ in Theorem 1 determines the $\tilde{\alpha}$ that yields overall $1 - \alpha$ joint coverage. This is again an exact, distribution-free result.

Third, interpolating the $n$ CIs with a stair-step function leads to a uniform confidence band for $F(\cdot)$ that also has exact coverage. For simplicity, consider a one-sided band. Let event $C_k$ denote the pointwise coverage event $\hat{u}(X_{n:k}) \geq F(X_{n:k})$ for $k = 1, \ldots, n$, where $\hat{u}$ denotes an upper CI endpoint. From the first two steps, we have $P(C_k) = 1 - \tilde{\alpha}$ for all $k$ and $P(\bigcap_{k=1}^n C_k) = 1 - \alpha$. By the monotonicity of $F(\cdot)$, the events $F(X_{n:k}) \leq \hat{u}(X_{n:k})$ and $\sup_{x \leq X_{n:k}} F(x) \leq \hat{u}(X_{n:k})$ are equivalent. Thus, the events $C_k$ and

$$\bigcap_{x \in (X_{n:k-1}, X_{n:k})} F(x) \leq \hat{u}(X_{n:k})$$

are equivalent, so using upper confidence function $\hat{u}(x) \equiv \hat{u}(X_{n:k})$ over $x \in (X_{n:k-1}, X_{n:k}]$ will translate the joint coverage into uniform coverage by a stair-step interpolation. Defining $X_{n:0} \equiv -\infty$, $X_{n:n+1} \equiv \infty$, and $\hat{u}(X_{n:n+1}) \equiv 1$ completes the construction, so

$$P\left(\bigcap_{x \in \mathbb{R}} F(x) \leq \hat{u}(x)\right) = P\left(\bigcap_{k=1}^n C_k\right) = 1 - \alpha.$$

Given $n$, the ordered Dirichlet distribution in Theorem 1 can be simulated from iid Uniform$(0,1)$ draws\footnote{Alternatively, numerical integration could be used as in Moscovich-Eiger et al. (2015).} to get the mapping from $\tilde{\alpha}$ to $\alpha$, as noted by Aldor-Noiman et al. (2013, p. 253). Then, a numerical search can find the $\tilde{\alpha}$ that leads to the desired $\alpha$.

Buja and Rolke (2006) mention the straightforward extensions of one-sided bands in their Section 6.2 and $p$-values in their Section 7. For $p$-value calculation, let $\tilde{p}_k$ be the
pointwise $p$-value at $X_{n:k}$, i.e., the smallest value of $\tilde{\alpha}$ such that one of the pointwise CI endpoints coincides with the null hypothesis CDF. The $p$-value is then the simulated overall null rejection probability when $\tilde{\alpha} = \min_{k=1,\ldots,n} \tilde{p}_k$ is used[6].

For two-sample, two-sided testing and $p$-values for $H_0 : F_X(\cdot) = F_Y(\cdot)$, Buja and Rolke (2006, §5.2) propose a simulation-based permutation test for two-sample Q-Q plots. These are valid and relatively evenly sensitive, but the simulation step cannot be done ahead of time. Also, the test is only designed for weak control of FWER since the permutation test relies on the global null.

### 2.2 Characterization of the KS test’s uneven sensitivity

To further motivate the evenly sensitive Dirichlet approach to distributional inference, we quantify the unevenness of pointwise type I error rates of variance-weighted and unweighted level-$\alpha$ KS tests of $H_0 : F(\cdot) = F_0(\cdot)$. Since in practice the tests simplify to comparisons at each order statistic, we consider the rejection probability at each order statistic under the null. We are unaware of a statement of the following result in the literature, although the qualitative points are well known.

**Lemma 2.** Let Assumption 1 hold. For the unweighted KS test, the asymptotic pointwise type I error rate at extreme order statistics[7] is of order $e^{-\sqrt{n}} \to 0$, while at central order statistics it converges to a fixed value that is largest at the median. For the variance-weighted KS test, the pointwise type I error rate converges to zero at all points, but more slowly in the tails: at rate $\exp\{-\sqrt{\ln[\ln(n)]}\} \to 0$ at extreme order statistics, and no slower than $\exp\{-\ln[\ln(n)]\} \to 0$ at central order statistics.

Lemma 2 reflects the well-known insensitivity of the unweighted KS test in the tails. The lemma also shows that the weighted version actually increases tail sensitivity to an

---

[6] Buja and Rolke (2006, §7) call this a “simultaneous $p$-value” to distinguish from other types they consider.

[7] An order statistic $X_{n:r}$ is “extreme” if $r$ is fixed (or $r = n - k$ for fixed $k$) as $n \to \infty$, or “central” if $r/n \to p$ for fixed $0 < p < 1$; see for example David and Nagaraja (2003).
order of magnitude greater than that at central order statistics. These phenomena are discussed in different ways in, for example, Lockhart (1991), Jaeschke (1979), and Eicker (1979), who notes that the KS is “sensitive asymptotically only in the central range given by \( \{ u : (\log \log n)^{-1} < F(u) < 1 - (\log \log n)^{-1} \} \)” while the weighted KS is “sensitive only in the moderate tails given by, e.g., \( \{ u : n^{-1} \log n < F(u) < ((\log \log n) \log \log \log n)^{-1} \} \)” (p. 117).

Figure 3 shows that the same uneven patterns of pointwise sensitivity from Lemma 2 are strikingly manifest in finite-sample simulations (details in Section 6.1). In contrast, the Dirichlet approach ensures identical rejection probabilities at each order statistic, even in finite samples.

3 Computational improvements

For some methods in this section, further computational details are in Supplementary Appendix D. Before discussing methods, we enumerate the different tasks they should accomplish.

3.1 Objects of interest and null hypotheses

For one-sample inference, the population CDF is \( F(\cdot) \), and the sample size is \( n \). For two-sample inference, the CDFs are \( F_X(\cdot) \) and \( F_Y(\cdot) \), and the sample sizes are \( n_X \) and \( n_Y \), also defining \( n \equiv \min\{n_X, n_Y\} \). Let \( h^{-1}(\tau) \equiv \inf\{x : h(x) \geq \tau\} \), the generalized inverse.

We consider the following inferential tasks. New methods addressing each are presented throughout this paper and summarized in Table 6.

Task 1 Constructing a two-sided uniform confidence band for \( F(\cdot) \), satisfying

\[
P\left( \forall x \in \mathbb{R}, \hat{\ell}(x) \leq F(x) \leq \hat{u}(x) \right) = 1 - \alpha
\]
given \( \alpha \in (0, 1) \). This also provides a uniform confidence band for \( F^{-1}(\cdot) \):

\[
P\left( \forall x \in \mathbb{R}, \hat{\ell}(x) \leq F(x) \leq \hat{u}(x) \right) = P\left( \forall \tau \in (0, 1), \hat{u}^{-1}(\tau) \leq F^{-1}(\tau) \leq \hat{\ell}^{-1}(\tau) \right).
\]

**Task 2** Constructing a one-sided uniform confidence band satisfying

\[
P\left( \forall x \in \mathbb{R}, \hat{\ell}(x) \leq F(x) \right) = 1 - \alpha \text{ or } P\left( \forall x \in \mathbb{R}, F(x) \leq \hat{u}(x) \right) = 1 - \alpha.
\]

**Task 3** Two-sided testing of the global, one-sample null hypothesis \( H_0 : F(\cdot) = F_0(\cdot) \) against \( H_1 : F(\cdot) \neq F_0(\cdot) \). The test should control size at level \( \alpha \) and should be consistent against any fixed \( F_0(\cdot) \neq F(\cdot) \), i.e., rejecting with probability approaching one as \( n \to \infty \).

**Task 4** One-sample testing of the null of first-order stochastic dominance, i.e., one-sided testing of the global null hypothesis \( H_0 : F(\cdot) \leq F_0(\cdot) \) against \( H_1 : F(\cdot) \nleq F_0(\cdot) \), or \( H_0 : F(\cdot) \geq F_0(\cdot) \) against \( H_1 : F(\cdot) \ngeq F_0(\cdot) \), with properties as in Task 3.

**Task 5** Calculating a one-sample p-value, equal to the smallest value of \( \alpha \) such that the test in Task 3 (or Task 4) rejects.

**Task 6** Two-sided testing of the global, two-sample null hypothesis \( H_0 : F_X(\cdot) = F_Y(\cdot) \) against \( H_1 : F_X(\cdot) \neq F_Y(\cdot) \), with properties as in Task 3.

**Task 7** Two-sample testing of the null of first-order stochastic dominance, i.e., one-sided testing of the global null hypothesis \( H_0 : F_X(\cdot) \leq F_Y(\cdot) \) against \( H_1 : F_X(\cdot) \nleq F_Y(\cdot) \), or \( H_0 : F_X(\cdot) \geq F_Y(\cdot) \) against \( H_1 : F_X(\cdot) \ngeq F_Y(\cdot) \), with properties as in Task 3.

**Task 8** Calculating a two-sample p-value, equal to the smallest value of \( \alpha \) such that the test in Task 6 (or Task 7) rejects.

### 3.2 One-sample, two-sided methods

For Tasks 1, 3, and 5, the computationally intensive step in the implementation of Aldor-Noiman et al. (2013, §2.1) is determining the pointwise \( \hat{\alpha} \) given \( n \) and the desired overall \( \alpha \),
which requires repeated simulations within a numerical search. The following proposition contains a single, closed-form approximation for \( \tilde{\alpha} \) as a function of \( n \) and \( \alpha \).

**Proposition 3.** Under Assumption 1, for \( \alpha \in \{0.001, 0.01, 0.05, 0.1, 0.2, 0.5, 0.7, 0.9\} \) and \( 4 \leq n \leq 10^6 \),

\[
\tilde{\alpha} \approx \exp \left\{ -c_1(\alpha) - c_2(\alpha) \sqrt{\ln[\ln(n)]} - c_3(\alpha)[\ln(n)]^{c_4(\alpha)} \right\},
\]

where the approximation error is \( \pm 9\% \) for \( \tilde{\alpha} \) and the \( c_j(\alpha) \) functions are (rounding to three significant figures)

\[
c_1(\alpha) \approx -2.75 - 1.04\ln(\alpha), \quad c_2(\alpha) \approx 4.76 - 1.20\alpha,
\]
\[
c_3(\alpha) \approx 1.15 - 2.39\alpha, \quad c_4(\alpha) \approx -3.96 + 1.72\alpha^{0.171}.
\]

![Figure 1: Relationship between simulated \(-\ln(\tilde{\alpha})\) and \(\ln[\ln(n)]\) for values of \(\alpha\) between 0.001 (top line in each graph) and 0.90 (bottom), \(n\) up to \(10^6 \approx \exp\{\exp(2.6)\}\). Left: fit for individual \(\alpha\) separately. Right: using estimated \(c_j(\alpha)\) functions from Proposition 3.](image)

The accuracy of the Proposition 3 formulas is shown in Figure 1. The right panel shows how well the formulas in the proposition fit the simulated values. The left panel shows how slightly the fit improves when estimating the \(c_j(\alpha)\) separately for different \(\alpha\).

By monotonicity of the mapping \( \tilde{\alpha}(\alpha, n) \) in \( \alpha \), any deviation from the formulas in Proposition 3 is small for \( \alpha \) interpolated between the values mentioned explicitly in the proposition.
We conjecture that the formulas continue to be accurate for even larger \( n \) and for \( \alpha \) outside the range given; moreover, few empirical economic applications require \( n > 10^6 \), \( \alpha < 0.001 \), or \( \alpha > 0.9 \).

To characterize the approximation error in Proposition 3 another way, for example with \( n = 200 \), using \( \tilde{\alpha} \) in the range \([0.91 \times \tilde{\alpha}(0.1, 200), 1.09 \times \tilde{\alpha}(0.1, 200)]\) yields overall type I error rates in the range \((0.0923, 0.1076)\). Although not perfect, this is quite good for a worst-case bound on finite-sample type I error rates.

Although Proposition 3 is justified by simulations rather than mathematical derivation, the rate at which \( \tilde{\alpha} \to 0 \) matches that of the weighted KS test at extreme order statistics from Lemma 2, which are both slower than the rate for the weighted KS test at central order statistics. The Dirichlet rate is much slower than the unweighted KS rate at extreme order statistics, but faster than the KS at central order statistics. These theoretical comparisons match the finite-sample comparisons simulated in Figure 3.

Proposition 3 enables immediate computation of two-sided uniform confidence bands, hypothesis tests, and \( p \)-values. We now present such methods, which have exact finite-sample properties up to approximation error from Proposition 3. Notationally, \( \hat{\ell}(x) \) in Method 1 has a hat over \( \ell \) because it is a function of the data, whereas \( \ell_k \) does not have a hat because it only depends on \( k, n, \) and \( \tilde{\alpha} \), and similarly for \( \hat{u}(x) \) and \( u_k \).

**Method 1.** For Task 1, to construct a two-sided uniform confidence band for \( F(\cdot) \), given sample size \( n \) and desired confidence level \( 1 - \alpha \), first determine \( \tilde{\alpha} \) from Proposition 3. Second, for \( k = 1, \ldots, n \), construct a CI for \( F(X_{n,k}) \) from the \( \tilde{\alpha}/2 \)-quantile to the \((1 - \tilde{\alpha}/2)\)-quantile of the \( \beta(k, n + 1 - k) \) distribution; denote these lower and upper endpoints \( \ell_k \) and \( u_k \), respectively. Third, connect the CIs with a stairstep (as in Section 2.1) to create the functions \( \hat{\ell}(x) = \max\{\ell_k : X_{n,k} \leq x\} \) and \( \hat{u}(x) = \min\{u_k : X_{n,k} \geq x\} \), defining \( X_{n,n+1} = \infty \), \( X_{n;0} = -\infty \), \( u_{n+1} = 1 \), and \( \ell_0 = 0 \).

**Method 2.** For Task 3, first construct the uniform confidence band from Method 1. Reject \( H_0 \) if \( \inf_{x \in \mathbb{R}} \min\left\{ \hat{u}(x) - F_0(x), F_0(x) - \hat{\ell}(x) \right\} < 0. \)
Method 3. For Task 5, to compute a two-sided \( p \)-value, first compute \( \tilde{p}_k \) for \( k = 1, \ldots, n \):

\[
\tilde{p}_k = 2 \times \min \{ F_{\beta,k}(F_0(X_{n,k})), 1 - F_{\beta,k}(F_0(X_{n,k})) \},
\]

where \( F_{\beta,k}(\cdot) \) is the \( \beta(k, n+1-k) \) CDF. The overall \( p \)-value, \( p \), satisfies \( \tilde{\alpha}(p, n) = \min_{k=1,\ldots,n} \tilde{p}_k \), using the mapping \( \tilde{\alpha}(\alpha, n) \) in Proposition 3.

\[ \square \]

3.3 One-sample, one-sided methods

With slight adjustment, the two-sided methods can be used for one-sided inference. Proposition 4 simply restates Theorem 5.1 of Moscovich-Eiger et al. (2015). Their result follows from the negative dependence (and asymptotic independence) of upper and lower one-sided rejections.

Proposition 4 (Moscovich-Eiger et al. (2015), Theorem 5.1). Let Assumption 1 hold. For Method 1 with confidence level \( 1 - \alpha_2 \), let \( 1 - \alpha_1^\ell \) be the uniform coverage probability of \( \hat{\ell}(\cdot) \) as a lower confidence bound for \( F(\cdot) \), and similarly define \( \alpha_1^u \) for \( \hat{u}(\cdot) \). Then, \( \alpha_1^\ell = \alpha_1^u \equiv \alpha_1 \), and

\[
\alpha_2 \geq 2\alpha_1 - \alpha_1^2,
\]

with equality holding asymptotically. Also, \( \alpha_2 \leq 2\alpha_1 \).

One-sided uniform confidence bands, tests, and \( p \)-values addressing Tasks 2, 4, and 5 are presented in Methods 16–18 in Appendix C. Treating \( \alpha_1 \) as an equality is conservative in finite samples, but the difference is small for conventionally used levels where \( \alpha_1^2 \) is close to zero.

3.4 Two-sample methods

With two samples, we wish to compare the population CDFs \( F_X(\cdot) \) and \( F_Y(\cdot) \) from which the samples are drawn.\(^8\) We maintain the following assumption.

\(^8\)If the samples are paired so that \( \Delta_i = Y_i - X_i \) has a meaningful interpretation, such as the treatment effect for individual \( i \), then we could perform one-sample inference to learn about the distribution of \( \Delta_i \).
**Assumption 2.** For continuous CDFs $F_X(\cdot)$ and $F_Y(\cdot)$, $X_i \overset{iid}{\sim} F_X$, $i = 1, \ldots, n_X$, and $Y_k \overset{iid}{\sim} F_Y$, $k = 1, \ldots, n_Y$. The two samples are independent: $X_i \perp \perp Y_k$ for any $i, k$.

For $H_0 : F_X(\cdot) = F_Y(\cdot)$ as in Task 6, we propose a new two-sample test (Method [4]) that achieves much more even sensitivity than the two-sample KS test. We construct two-sided bands for $F_X(\cdot)$ and $F_Y(\cdot)$ similar to Method 1, but with $\tilde{\alpha}$ approximately\[^{9}\] satisfying

$$P\left( \exists r \in \mathbb{R} \text{ s.t. } \hat{\ell}_X(r) > \hat{u}_Y(r) \text{ or } \hat{\ell}_Y(r) > \hat{u}_X(r) \mid H_0 \text{ true} \right) = \alpha. \quad (2)$$

This results in an exact test that rejects $H_0 : F_X(\cdot) = F_Y(\cdot)$ whenever there is at least one quantile at which the resulting confidence bands fail to overlap. Computationally, (2) can be evaluated using independent Dirichlet distributions for $X$ and $Y$ or using the permutation argument in Supplementary Appendix D.

The $\tilde{\alpha}$ determined by (2) depends only on $\alpha$ and the sample sizes and may thus be computed ahead of time. This is an advantage over the two-sample method in Buja and Rolke (2006) that requires repeated permutations of the observed data. With enough pre-computed values, our approach could also provide approximate $p$-values without just-in-time simulation.

For our two-sample $\tilde{\alpha}$, a relationship similar to Proposition 3 seems to exist, giving similar $\tilde{\alpha}$ at least for sample sizes $n_Y/2 < n_X < n_Y$. However, the coefficient formulas have not yet been determined, so instead a lookup table for $\tilde{\alpha}$ is provided for common sample sizes and $\alpha$.\[^{10}\]

**Method 4.** For Task 6, choose $\tilde{\alpha}$ from the provided lookup table so that (2) holds (approximately) for the desired level $\alpha$. Given $\tilde{\alpha}$, construct two-sided bands for $F_X(\cdot)$ and $F_Y(\cdot)$ as in Method 1. Reject $H_0$ if there is any $r \in \mathbb{R}$ where $\hat{\ell}_X(r) > \hat{u}_Y(r)$ or $\hat{\ell}_Y(r) > \hat{u}_X(r)$. \[\Box\]

If a needed value of $\alpha$ is not in the lookup table, then it is faster to simulate a $p$-value

\[^{9}\]Only a finite number of $\alpha$ are attainable, as discussed below.

\[^{10}\]Specifically, $n_Y \in \{10, 11, \ldots, 200, 300, \ldots, 1000, 2000, 3000, 10^4, 10^5\}$ and $n_X/n_Y \in \{0.1, 0.5, 0.9, 1.0\}$ with $\alpha \in \{0.01, 0.05, 0.10\}$. There course may be significant error if extrapolating beyond $n_X/n_Y \ll 0.1$ or $n_Y \gg 10^5$, but interpolation error is small and (in our code’s implementation) on the conservative side.
and reject $H_0$ if $p < \alpha$ than to determine $\tilde{\alpha}$.

**Method 5.** For Task 8, to compute a two-sided $p$-value, first compute $\{\tilde{p}_{X,k}\}_{k=1}^{n_X}$ and $\{\tilde{p}_{Y,k}\}_{k=1}^{n_Y}$ as follows. For $\tilde{p}_{X,k}$:

Step 1. Determine $r$ such that $Y_{n_Y:1} < X_{n_X:k} < Y_{n_Y:r+1}$, defining $Y_{n_Y:0} \equiv -\infty$ and $Y_{n_Y:n_Y+1} \equiv \infty$.

Step 2. Let $\tilde{p}^a_{X,k}$ equate the $\tilde{p}_{X,k}/2$-quantile of the $\beta(r, n_Y + 1 - r)$ distribution and the $(1 - \tilde{p}^a_{X,k}/2)$-quantile of the $\beta(k, n_X + 1 - k)$ distribution. (Note: if $r = 0$, then $\tilde{p}^a_{X,k} = 2$.)

Step 3. Let $\tilde{p}^b_{X,k}$ equate the $(1 - \tilde{p}^b_{X,k}/2)$-quantile of the $\beta(r + 1, n_Y - r)$ distribution and the $\tilde{p}^b_{X,k}/2$-quantile of the $\beta(k, n_X + 1 - k)$ distribution. (Note: if $r = n_Y$, then $\tilde{p}^b_{X,k} = 2$.)

Step 4. Let $\tilde{p}_{X,k} = \min\{1, \tilde{p}^a_{X,k}, \tilde{p}^b_{X,k}\}$.

The calculations for the $\tilde{p}_{Y,k}$ follow similarly. Writing $\tilde{\alpha}(n_X, n_Y, \alpha)$ to denote the mapping from $(n_X, n_Y, \alpha)$ to $\tilde{\alpha}$ determined by (2), the overall $p$-value solves

$$\tilde{\alpha}(n_X, n_Y, p) = \min\left\{ \min_{k=1,...,n_X} \tilde{p}_{X,k}, \min_{k=1,...,n_Y} \tilde{p}_{Y,k} \right\}.$$

Since the number of permutations of $(X_1, \ldots, X_{n_X}, Y_1, \ldots, Y_{n_Y})$ is finite, there is a finite number of attainable rejection probabilities (RPs) for any test that rejects only based on the ordering of the values, including the KS test and our Method 4. This is why (2) only holds approximately. For small sample sizes like $n_X = n_Y = 10$, the closest attainable $\alpha$ to the desired value may be a couple percentage points away, but the set of attainable $\alpha$ becomes dense on $(0, 1)$ as $n_X, n_Y \to \infty$. Moreover, Table 4 shows this issue to be much smaller for our new Dirichlet method than for KS. For similar reasons, the pointwise RP cannot be exactly equal at all points. However, it is much more even than KS and becomes increasingly even as the sample sizes grow, as our Section 6.2 simulations illustrate.

\footnote{With $n_X = n_Y = 500$, on a standard (ca. 2015) laptop, this takes around 85 seconds (or as low as 15 seconds with the parallel computing option) with our default $10^4$ Dirichlet draws; using $10^3$ draws is sufficiently accurate for exploratory analysis and takes under 10 seconds.}
The same arguments as in Section 3.3 for one-sided testing apply here; see Methods 19 and 20 in Appendix C. First-order stochastic dominance may be tested as in Barrett and Donald (2003, p. 74) and others, where the null hypothesis is weak dominance of $F_Y$ over $F_X$, $H_0: F_X(\cdot) \geq F_Y(\cdot)$.

In addition to exact size properties, our Dirichlet-based tests are consistent.

**Theorem 5.** Under Assumptions 1 and 2, the Dirichlet-based one-sample and two-sample tests for distributional equality are consistent against all fixed alternatives.

Consistency is not explicitly discussed in Buja and Rolke (2006) or Aldor-Noiman et al. (2013). In the one-sample case, Moscovich-Eiger et al. (2015, Cor. 2) strengthen the consistency result to any sequence of alternatives $G_n(\cdot)$ such that for any $\epsilon > 0$,

$$\sqrt{n/\log(n)}[\log(\log(n))]^{-0.5-\epsilon} \|G_n(\cdot) - F(\cdot)\|_\infty \rightarrow \infty.$$  

As a simple alternative to our two-sample methods, if the sample sizes $n_X$ and $n_Y$ are very different, then we may treat the empirical distribution from the larger sample as the true distribution and use one of the one-sample methods. For example, we can set $F_0(\cdot) = \hat{F}_X(\cdot)$ (or a smoothed version thereof) and test $H_{0\tau}: F_Y^{-1}(\tau) = F_0^{-1}(\tau)$ for $\tau \in (0, 1)$ using Method 7. Because our one-sample test has near-exact finite-sample size control, there is no problem with treating $n_Y$ as fixed in our asymptotic approximation. Since uniformly $\hat{F}_X(\cdot) \overset{a.s.}{\rightarrow} F_X(\cdot)$ by the Glivenko–Cantelli Theorem, if $n_X \gg n_Y$, there is little error incurred by only accounting for the sampling variation of $\hat{F}_Y(\cdot)$ and ignoring that of $\hat{F}_X(\cdot)$. Additionally, this tends to err on the conservative side (e.g., Goldman and Kaplan, 2015b, §3.3).

**Method 6.** If $n_X$ is much bigger than $n_Y$, then let $F_0(\cdot)$ be a (smoothed) estimator of $F_X(\cdot)$ and apply any of the one-sample methods.
4 New interpretations

4.1 Multiple testing and FWER

First, we define important terms related to familywise error rate (FWER) following Lehmann and Romano (2005, §9.1), and then we enumerate additional inferential tasks.

**Definition 1.** For a family of null hypotheses $H_{0h}$ indexed by $h$, and defining the set $I \equiv \{ h : H_{0h} \text{ is true} \}$, the “familywise error rate” is

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

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

**Task 9** Two-sided testing of the continuum of one-sample hypotheses $H_{0\tau} : F^{-1}(\tau) = F_0^{-1}(\tau)$ for $\tau \in (0, 1)$, with strong control of FWER.

**Task 10** Same as Task 9 but with $H_{0\tau} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ or $H_{0\tau} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$.

**Task 11** Two-sided testing of the continuum of two-sample hypotheses $H_{0r} : F_X(r) = F_Y(r)$ for $r \in \mathbb{R}$, with strong control of FWER.

**Task 12** Same as Task 11 but with $H_{0r} : F_X(r) \leq F_Y(r)$ or $H_{0r} : F_X(r) \geq F_Y(r)$.

The usual KS test accomplishes Tasks 9 and 10, albeit without even sensitivity. Moreover, any test whose inversion yields a uniform confidence band achieves strong control of FWER: given $H_{0x} : F(x) = F_0(x)$, the band covers $F(x)$ over $I \equiv \{ x : F(x) = F_0(x) \}$ with at least $1 - \alpha$ probability, so $P(\text{reject any } H_{0x}, x \in I) \leq \alpha$. We state Proposition 6 since we are unaware of such a statement in the literature in terms of strong control of FWER.

**Proposition 6.** Let Assumption 1 hold, and let $H_{0r} : F(r) = F_0(r)$ for $r \in \mathbb{R}$. The one-sided and two-sided exact KS tests have strong control of exact FWER.
Since Methods 1 and 16 provide uniform confidence bands, the corresponding Dirichlet tests have strong control of FWER. We state methods and results in terms of quantile functions for continuity with statements in Section 5, but they could also be stated in terms of CDFs to parallel the KS.

**Method 7.** For Task 9, construct a $1 - \alpha$ two-sided uniform confidence band for $F^{-1}(\cdot)$ using Method 1. Reject $H_{0r}$ when $F_0^{-1}(\tau) \notin \left[\hat{u}^{-1}(\tau), \hat{\ell}^{-1}(\tau)\right]$.

**Method 8.** For Task 10, construct a $1 - \alpha$ one-sided uniform confidence band for $F^{-1}(\cdot)$ using Method 16. Reject $H_{0r} : F^{-1}(\tau) \leq F_0^{-1}(\tau)$ when $\hat{u}^{-1}(\tau) > F_0^{-1}(\tau)$; reject $H_{0r} : F^{-1}(\tau) \geq F_0^{-1}(\tau)$ when $\hat{\ell}^{-1}(\tau) < F_0^{-1}(\tau)$.

**Proposition 7.** Under Assumption 1, Methods 7 and 8 have strong control of exact FWER, up to approximation error from Propositions 3 and 4.

For two-sample testing, the KS test’s strong control of FWER can again be established, as in Proposition 8. We are unaware of such a statement in the literature, explicitly discussing strong control of FWER in finite samples.

**Proposition 8.** Let Assumption 2 hold, and let $H_{0r} : F_X(r) = F_Y(r)$ for $r \in \mathbb{R}$. The one-sided and two-sided exact KS tests have strong control of exact FWER.

The two-sample Dirichlet test also has strong control of FWER. The key is that, given $\alpha, n_X, n_Y$, rejection of $H_{0r}$ depends only on $\hat{F}_X(r)$ and $\hat{F}_Y(r)$, whose distributions are independent (by Assumption 2) and depend only on $F_X(r)$ and $F_Y(r)$, and this extends to multiple $r$. 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 test.

**Method 9.** For Task 11, run Method 4 and reject $H_{0r} : F_X(r) = F_Y(r)$ when $\hat{\ell}_X(r) > \hat{u}_Y(r)$ or $\hat{\ell}_Y(r) > \hat{u}_X(r)$. For Task 12, run Method 19 and reject $H_{0r} : F_X(r) \leq F_Y(r)$ when $\hat{\ell}_X(r) > \hat{u}_Y(r)$, and reject $H_{0r} : F_X(r) \geq F_Y(r)$ when $\hat{\ell}_Y(r) > \hat{u}_X(r)$.

Asymptotically, and usually not framed in terms of FWER, stronger results in more complex models exist, such as the nonparametric, uniform (over $\tau$) 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 and other functionals of potential outcome distributions in [Firpo and Galvao (2015, §4)].
Proposition 9. Under Assumption 2, Method 9 has strong control of exact FWER at level \( \alpha \).

4.2 Bayesian interpretation

The frequentist bands from Methods 1 and 16 are also valid uniform credible bands for a particular nonparametric Bayesian method.\(^\text{13}\) The KS bands share this interpretation but lack even pointwise credibility across the order statistics. Proposition 3 removes the computational burden that is often a disadvantage of nonparametric Bayesian methods.

Banks (1988) provides a histospline-smoothed variant of the Bayesian bootstrap of Rubin (1981), which in turn is the nonparametric Bayesian method of Ferguson (1973) with an improper Dirichlet process prior that takes the limit as the index measure approaches the zero function. "Nonparametric" means the distribution of the data is not specified (a priori) up to a finite-dimensional parameter vector like \((\mu, \sigma^2)\), but instead the parameter (i.e., the index measure of the Dirichlet process) is infinite-dimensional (i.e., a function). The posterior is a distribution over possible values of \(F(\cdot)\). The Rubin (1981) posterior includes only discrete distributions. Banks (1988) essentially provides a continuity correction for when \(F(\cdot)\) is continuous.

Proposition 10. For the smoothed Bayesian bootstrap of Banks (1988), Methods 1 and 16 provide uniform \(1 - \alpha\) credible bands for the unknown \(F(\cdot)\).

Figure 2 shows an example of using our method to visualize the posterior. The data are from the pension-related extract in Wooldridge (2010)\(^\text{14}\) with \(n = 226\). We use the measure of net worth in thousands of US dollars in 1989. The variable is not strictly continuous due to a small amount of bunching, but this is primarily at zero. Figure 2 shows uniform credible bands for eight credibility levels ranging from \(1 - \alpha = 0.1\) to \(0.999\). Total computation time is 0.11 seconds.

\(^{13}\)A “credible” band or interval is like a confidence band/interval, but replacing the (frequentist) coverage probability with posterior probability.

\(^{14}\)https://mitpress.mit.edu/books/econometric-analysis-cross-section-and-panel-data
5 Power improvement

5.1 Quantile framework

In Method 1, \( \ell_k \) is the \( \tilde{\alpha}/2 \)-quantile of the distribution of \( F(X_{n;k}) \), so

\[
1 - \tilde{\alpha}/2 = P(F(X_{n;k}) > \ell_k) = P(X_{n;k} > F^{-1}(\ell_k)).
\]

Thus, \( X_{n;k} \) is the upper endpoint of a one-sided \( 1 - \tilde{\alpha}/2 \) CI for the \( \ell_k \)-quantile of distribution \( F(\cdot) \). Similarly, \( X_{n;k} \) is the lower endpoint of a CI for the \( u_k \)-quantile. We may interpret Method 1 as constructing \( 2n \) one-sided quantile CIs, each with pointwise confidence level \( 1 - \tilde{\alpha}/2 \), corresponding to fixed quantiles \( \{\ell_k\}_{k=1}^n \) and \( \{u_k\}_{k=1}^n \) that are determined by \( n, \tilde{\alpha}, \) and \( k \).

5.2 One-sample stepdown procedure

For one-sample testing, we propose a new iterative stepdown procedure. Its pointwise power functions dominate those of Methods 7 and 8, and it maintains strong control of FWER.
The idea is to remove from consideration any $H_{0r}$ that were rejected by the initial test, recalibrating $\tilde{\alpha}$ as if those $H_{0r}$ are indeed false. With fewer $H_{0r}$, $\tilde{\alpha}$ can increase while maintaining strong control of FWER. If one of the initial rejections was false, we have already made an error, so it does not matter what we do next (as measured by FWER).

**Method 10.** For Task 10, let $\hat{K}_0 \equiv \{1, \ldots, n\}$, and let $r_{k,0} = k$ for $k \in \hat{K}_0$. Consider $H_{0r} : F^{-1}(\tau) \geq F_{0}^{-1}(\tau)$. Solve for the $\ell_k$ such that $\ell_k$ is the $\tilde{\alpha}$-quantile of the $\beta(k, n+1-k)$ distribution for each $k \in \hat{K}_0$ and

$$
\alpha \geq 1 - P \left( \bigcap_{k \in \hat{K}_i} X_{n;r_{k,i}} \geq F^{-1}(\ell_k) \right) = 1 - P \left( \bigcap_{k \in \hat{K}_i} F(X_{n;r_{k,i}}) \geq \ell_k \right) \tag{3}
$$

holds with equality with $i = 0$, given the distribution from Theorem 1 for $(F(X_{n;1}), \ldots, F(X_{n;n}))$. Reject $H_{0r}$ if $F_{0}^{-1}(\tau) > \min\{X_{n,k} : \ell_k \geq \tau\}$. Then, iteratively perform the following steps, starting with $i = 1$.

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

**Step 2.** Choose integers $r_{k,i} \leq r_{k,i-1}$ satisfying (3), using the distribution from Theorem 1.\(^\text{15}\)

**Step 3.** Reject any additional $H_{0r}$ 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 $H_{0r} : F^{-1}(\tau) \leq F_{0}^{-1}(\tau)$, replace $\ell_k$ with $u_k$ and $\tilde{\alpha}$ with $1 - \tilde{\alpha}$, and reverse all inequalities. \(\square\)

A two-sided stepdown procedure is given in Method 21 in Appendix C.

**Theorem 11.** Under Assumption 1, Methods 10 and 21 have strong control of exact FWER.

\(\text{15}\)This leaves open many possibilities for $r_{k,i}$; as long as the inequality is satisfied, FWER can be strongly controlled. In our code, we use a “greedy” algorithm (e.g., Sedgewick and Wayne, 2011, §4.3), starting with $r_{k,i} = r_{k,i-1}$ and iteratively decreasing (by one) whichever $r_{k,i}$ achieves the biggest increase in the corresponding pointwise rejection probability while still satisfying the FWER constraint, stopping when it is impossible to decrease any $r_{k,i}$ without violating FWER.
5.3 One-sample pre-test procedure

For one-sided testing, a pre-test can improve both pointwise and global power. The null hypothesis $H_0 : F^{-1}(\cdot) \geq F_0^{-1}(\cdot)$ is composite. The well-known “least favorable configuration” is $F(\cdot) = F_0(\cdot)$, in which case the rejection probability (RP) equals $\alpha$. For other $F(\cdot)$ satisfying $H_0$, the RP can be substantially lower, depending in part on the size of the contact set, $\{\tau : F^{-1}(\tau) = F_0^{-1}(\tau)\}$.

We implement a pre-test to determine where the null hypothesis constraint $F^{-1}(\cdot) \geq F_0^{-1}(\cdot)$ is slack, i.e., where we can reject $F^{-1}(\tau) \leq F_0^{-1}(\tau)$ in favor of $F^{-1}(\tau) > F_0^{-1}(\tau)$. Re-calibrating $\tilde{\alpha}$ using only the unrejected $\tau$ improves power.

Falsely inferring that the constraint is slack leads to distortion of the resulting test, so the probability of doing so should be small. This probability is the FWER of the pre-test. If the pre-test controls $\text{FWER} \leq \alpha_p$, then $\alpha_p \to 0$ ensures zero asymptotic size distortion.\footnote{With different notation, 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.}

Of course, in any finite sample, $\alpha_p > 0$, so it is important in practice to choose values of $\alpha_p$ that are tolerably small. Based on simulations, we suggest $\alpha_p = \alpha / \ln[\ln(\max\{n, 15\})]$, but preference may depend on the application.

To provide a concrete example, the pre-test we implement in our code is described in Method 11 (and its strong control of FWER in Proposition 12) before the overall method is described in Method 12. To pre-test the most relevant ($\ell_k$) quantiles takes some finesse beyond simply running Method 8, since then the $u_k$ would be used instead.

**Method 11 (Pre-test only).** Consider the pre-test null hypotheses $H_{0\ell_k} : F^{-1}(\ell_k) \leq F_0^{-1}(\ell_k)$, defining $\ell_k$ as in Method 10. Let $B_{k,1-\tilde{\alpha}}$ be the $(1 - \tilde{\alpha})$-quantile of the $\beta(k, n + 1 - k)$ distribution. Given $\tilde{\alpha}$, let $k = \min\{k : \ell_k \geq B_{1,1-\tilde{\alpha}}\}$ and $r_k = \max\{k' : B_{k',1-\tilde{\alpha}} \leq \ell_k\}$ (for $k \geq k$), where both $k$ and $k'$ are restricted to integers $\{1, \ldots, n\}$. Using the distribution in.
Theorem 1, calculate

\[
\alpha_p(\tilde{\alpha}, n) = 1 - P \left( \bigcap_{k=\tilde{k}}^n X_{n,r_k} \leq F^{-1}(\ell_k) \right) = 1 - P \left( \bigcap_{k=\tilde{k}}^n F(X_{n,r_k}) \leq \ell_k \right).
\]

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

To instead pre-test \(H_{0u_k}: F^{-1}(u_k) \geq F^{-1}(u_k)\), reverse all inequalities and \(\min/\max\), and replace \(\ell_k\) with \(u_k\) (also from Method 10), \(B_{k,1-\tilde{\alpha}}\) with \(B_{k,\tilde{\alpha}}\), \(\bar{k} = \max\{k : u_k \leq B_{n,\tilde{\alpha}}\}\), and \(\bigcap_{k=\tilde{k}}^n\) with \(\bigcap_{k=\bar{k}}^n\).

**Proposition 12.** Under Assumption 1, Method 11 has strong control of exact FWER.

**Method 12.** For Task 4 or Task 10, consider \(H_0: F^{-1}(\cdot) \geq F^{-1}(\cdot)\). First run a pre-test (like Method 11) of \(H_{0\tau}: F^{-1}(\tau) \leq F^{-1}(\cdot)\) for \(\tau \in (0, 1)\) that strongly controls 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 \(\ell_k\) as in Method 10. Then choose integers \(r_k \geq k\) such that

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

computed using the distribution in Theorem 1.\(^{17}\) Reject \(H_{0\tau}\) when \(\min\{X_{n,k} : \ell_k \geq \tau, k \in \hat{K}\} < F^{-1}(\tau)\), and reject the global \(H_0\) if any \(H_{0\tau}\) is rejected. For \(H_0: F^{-1}(\cdot) \leq F^{-1}(\cdot)\), reverse inequalities and replace \(\ell_k\) with \(u_k\) (from Method 10).

**Theorem 13.** Under Assumption 1, Method 12 has strong control of exact FWER at level \(\alpha + \alpha_p\).

The FWER level \(\alpha + \alpha_p\) in Theorem \(^{13}\) is very conservative. Specifically, it assumes that a false rejection of the pre-test always leads to a false rejection, whereas in reality the probability is only somewhat increased. Simulations show the FWER to be much closer to \(\alpha\) than \(\alpha + \alpha_p\), but to formalize such a statement requires considering the complex interaction between the pre-test rejection events and the overall rejection events.

\(^{17}\)Similar remarks to Footnote \(^{15}\) apply to the choice of \(r_k\) here.
5.4 Two-sample power improvement

Establishing strong control of FWER for two-sample stepdown and pre-test procedures is more difficult. We consider the following modified tasks.

**Task 13** Testing a family of $M_n = \lfloor n^{2/5} \rfloor$ two-sample quantile 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 14** Same as Task 13 but with $F_X^{-1}(t) \leq F_Y^{-1}(t)$ or $F_X^{-1}(t) \geq F_Y^{-1}(t)$.

We consider a different approach than Method 9, though still Dirichlet-based. Consider a fixed set of $M$ quantiles, $\tau_1, \ldots, \tau_M$, and let $\Delta_j \equiv F_X^{-1}(\tau_j) - F_Y^{-1}(\tau_j)$ denote $\tau_j$-quantile differences. Goldman and Kaplan (2015b) use fractional order statistics to construct CIs for the $\Delta_j$ that have joint $1 - \alpha + O(n^{-2/3})$ coverage probability.

By the same logic as the proof of Proposition 7, the corresponding test of $H_{0j} : \Delta_j = 0$ over $j = 1, \ldots, M$ has strong control of FWER. Let $\hat{\text{CI}}_j$ denote the CI for $\Delta_j$. Then, letting $I = \{j : H_{0j} \text{ is true}\}$,

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

We conjecture that the proofs would still hold if $M_n \to \infty$ at a slow enough rate, to address Tasks 13 and 14 but such modifications appear to be quite involved and are left to future work.

**Method 13.** For Tasks 13 and 14, given the desired FWER level $\alpha$ and sample sizes $n_X$ and $n_Y$, let $n = \min\{n_X, n_Y\}$ and $M_n = \lfloor n^{2/5} \rfloor$. Let $\tau_j = j/(M_n + 1)$ for $j = 1, \ldots, M_n$. Then, construct joint one-sided or two-sided CIs for the $F_X^{-1}(\tau_j) - F_Y^{-1}(\tau_j)$ using the method in Goldman and Kaplan (2015b), and reject any null hypothesis not covered by every interval. Note that the density ratios $\gamma_{x,j}$ and $\gamma_{y,j}$ in Goldman and Kaplan (2015b) are all equal to one under the null.
Conjecture 14. Under Assumption 2, Method 13 has strong control of asymptotic FWER.

We now consider refinements to Method 13. If Conjecture 14 holds, then a stepdown procedure is justified by monotonicity, as in Section 5.2. For one-sided testing, we propose a pre-test procedure similar to that in Section 5.3.

Method 14. For Tasks 13 and 14, using notation from Method 13, let \( \hat{T}_0 = \{1, \ldots, M_n\} \). For two-sided testing, given a pointwise \( \tilde{\alpha} \), let \( k_{X,j}^u \) be such that \( P(\beta(\hat{k}_{X,j}^u, n_X + 1 - \hat{k}_{X,j}^u) < \tau_j) = \tilde{\alpha}/2 \) and \( k_{X,j}^l \) such that \( P(\beta(\hat{k}_{X,j}^l, n_X + 1 - \hat{k}_{X,j}^l) > \tau_j) = \tilde{\alpha}/2 \), and similarly for \( k_{Y,j}^u \) and \( k_{Y,j}^l \) (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

\[
1 - \alpha = P \left( \bigcap_{j \in \hat{T}_i} \{ D_{X,j}^l < D_{Y,j}^u, D_{Y,j}^l < D_{X,j}^u \} \right),
\]

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

\[
(D_{X,1}, D_{X,2} - D_{X,1}, \ldots, D_{X,2M_n} - D_{X,2M_n-1}, 1 - D_{X,2M_n})
\]

\[
\sim \text{Dirichlet}(k_1, k_2 - k_1, \ldots, k_{2M_n} - k_{2M_n-1}, n_X + 1 - k_{2M_n})
\]

with vector \( k = (k_1, \ldots, k_{2M_n}) \) containing all the \( k_{X,j}^l \) 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 \( D_X \perp \perp D_Y \). For iteration \( i = 0 \), reject any \( H_{0j} \) for which \( F_0^{-1}(\tau_j) \) lies outside the joint CI. 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 \( (4) \) to construct new joint CIs.

Step 3. Reject any additional \( H_{0j} \) for which \( F_0^{-1}(\tau_j) \) lies outside the joint CI.

Step 4. Increment \( i \) by one and return to Step 1.
Proposition 15. Let Assumption 2 hold. If Conjecture 14 holds, then Method 14 has strong control of asymptotic FWER.

Method 15. For Task 14, using notation from Method 13, consider \( H_0^j : F_X^{-1}(\tau_j) \geq F_Y^{-1}(\tau_j) \). First run a pre-test of \( H_0' : F_X^{-1}(\tau_j) \leq F_Y^{-1}(\tau_j) \) that strongly controls FWER at level \( \alpha_p = \alpha/\ln[\ln(\max\{n, 15\})] \), such as (we conjecture) Method 13. Then, use Method 14 starting with \( \hat{T}_0 \) containing all \( j \) such that \( H_0^j \) was not rejected by the pre-test.

\[ \square \]

Proposition 16. Let Assumption 2 hold. If Conjecture 14 holds, then Method 15 has strong control of asymptotic FWER at level \( \alpha \).

6 Simulation studies

Simulation results are provided for both one-sample and two-sample methods. Our methods (and these simulations) may be run using code available on the latter author’s website. For comparison, KS tests using \texttt{ks.test} in the \texttt{stats} package in R (R Core Team, 2013) are run, with “exact” referring to option \texttt{exact=TRUE}, as well as an exact, one-sample, weighted KS test.

Our simulations examine the following properties: 1) type I error rate of the global test, 2) distribution of pointwise type I error rates, 3) power of the global test, 4) computation time, 5) FWER, and 6) pointwise power. We show (1)–(3) to motivate adoption of the evenly sensitive Dirichlet approach generally, whereas (4)–(6) show some of our new results and benefits of our new methods. Aldor-Noiman et al. (2013) show some one-sample global power simulations, but otherwise neither they nor Buja and Rolke (2006) show simulations for (1)–(3).

6.1 One-sample simulations

In the one-sample simulations, we set \( n \in \{20, 100\} \) and \( \alpha = 0.1 \), running \( 10^6 \) replications unless otherwise noted. Since all tests considered are distribution-free under Assumption 1,
the true data generating process (DGP) for all simulations is iid Uniform(0, 1), without loss of generality.

For the weighted KS, to achieve symmetric weighting of the $k$th and $(n + 1 - k)$th order statistics (e.g., minimum and maximum), the weight for order statistic $k$ is based on $k/(n+1)$ rather than $k/n$. Asymptotic critical values from Jaeschke (1979) and Chicheportiche and Bouchaud (2012) were inaccurate, so we simulate exact critical values. However, this is time-consuming, which is a practical disadvantage.

### 6.1.1 One-sample global type I error rate

Table 1 shows type I error rates of the global tests of $H_0: F(\cdot) = F_0(\cdot)$ (Task 3). For our Method 2, this is exact by construction, up to the approximation error in Proposition 3. This error is negligible in Table 1. Figure 1 shows additional $n$ and $\alpha$.

<table>
<thead>
<tr>
<th>$\alpha$</th>
<th>$n$</th>
<th>Dirichlet</th>
<th>KS</th>
<th>KS (exact)</th>
<th>weighted KS (exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.10</td>
<td>20</td>
<td>0.101</td>
<td>0.100</td>
<td>0.100</td>
<td>0.099</td>
</tr>
<tr>
<td>0.10</td>
<td>100</td>
<td>0.101</td>
<td>0.094</td>
<td>0.100</td>
<td>0.098</td>
</tr>
<tr>
<td>0.05</td>
<td>20</td>
<td>0.050</td>
<td>0.050</td>
<td>0.050</td>
<td>0.053</td>
</tr>
<tr>
<td>0.05</td>
<td>100</td>
<td>0.050</td>
<td>0.045</td>
<td>0.050</td>
<td>0.049</td>
</tr>
</tbody>
</table>

### 6.1.2 One-sample even sensitivity

Figure 3 shows the even sensitivity of our Method 2 and the uneven sensitivity of KS tests, as discussed in Section 2.2. The figure shows the rejection probability (RP) at each order statistic under the global null $H_0: F(\cdot) = F_0(\cdot)$, where each test is calibrated to have exact size as shown in Table 1. The RPs of the weighted KS are U-shaped, much higher in the tails; the RPs of the KS follow an inverted U, much higher in the middle; and the Dirichlet RPs are identical (up to simulation error).

---

18Jaeschke (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.”
Figure 3: Simulated RP at each order statistic, under $H_0 : F(\cdot) = F_0(\cdot)$, $n = 20$ (left) and $n = 100$ (right), $\alpha = 0.1$, $10^6$ replications.

The naturally consequent differences in pointwise power among the three methods are illustrated in Supplementary Appendix E.1. One important feature missing from Figure 3 is the Dirichlet’s even sensitivity to upper and lower deviations due to its use of equal-tailed pointwise CIs. In contrast, both KS tests have symmetric pointwise CIs. This leads to very low power against deviations where $F_0(x) > F(x)$ in the upper tail, or where $F_0(x) < F(x)$ in the lower tail. For example, with $n = 20$ and $\alpha = 0.1$, even $F_0(X_{n:16}) = 1$ and $F_0(X_{n:5}) = 0$ do not cause either KS test to reject! If $F_0 = \text{Uniform}(0, 1)$, then even $X_{n:n} = 10^6$ does not cause either test to reject. In contrast, the Dirichlet approach always rejects if $F_0(X_{n:n}) = 1$ or $F_0(X_{n:1}) = 0$, regardless of $\alpha$ and $n$. For additional visual intuition, see Figure 10 (left panel).

6.1.3 One-sample global power

Aldor-Noiman et al. (2013, Table 1 and Fig. 8) show a power advantage of the Dirichlet test over the KS and Anderson–Darling (i.e., weighted Cramér–von Mises) tests for a variety of distributions. Consequently, we relegate our one-sample power results to Supplementary Appendix E.
6.1.4 One-sample computation time

Table 2 shows computation times of a level $\alpha = 0.1$ test of $H_0 : F(\cdot) = F_0(\cdot)$ (Task 3) for the Dirichlet, KS, and exact KS methods, using a standard desktop computer in the year 2015 (8GB RAM, 3.2GHz Intel i5 processor). Each data point is an average computation time over at least four repetitions. For comparison, the time to simulate $\tilde{\alpha}$ (to the same degree of precision as Proposition 3) is also shown; this is the time saved by the Proposition 3 formulas compared with just-in-time $\tilde{\alpha}$ simulation as in Buja and Rolke (2006). Since the time depends on the starting value of $\tilde{\alpha}$ in the numerical search, we report a range of values, where the lower value represents the unrealistic extreme of a correct initial $\tilde{\alpha}$ (or, $p$-value computation) and the upper value reflects a more plausible five search iterations. The upper value is comparable to Aldor-Noiman et al. (2013, p. 254), who report a runtime of 10 seconds for $n = 100$.

Table 2: Computation times (in seconds) for one-sample inference, $\alpha = 0.1$.

<table>
<thead>
<tr>
<th>$\log_{10}(n)$</th>
<th>Proposition 3</th>
<th>Buja and Rolke (2006)</th>
<th>KS</th>
<th>KS (exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>0.00</td>
<td>1.93–9.47</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>3</td>
<td>0.02</td>
<td>2.96–14.84</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>4</td>
<td>0.23</td>
<td>16.65–82.48</td>
<td>0.00</td>
<td>0.08</td>
</tr>
<tr>
<td>5</td>
<td>2.20</td>
<td>171.45–851.14</td>
<td>0.01</td>
<td>25.25</td>
</tr>
</tbody>
</table>

The KS runs in a hundredth of a second even for $n = 100\,000$. The exact KS starts to slow significantly around $n = 100\,000$, requiring over 20 seconds per test. The Dirichlet slows some, but it only takes a few seconds to compute even with $n = 100\,000$ since Proposition 3 replaces simulation; the 2.20 seconds are taken almost entirely by the 200,000 calls to the beta quantile function ($\texttt{qbeta}$). Our formula is significantly faster than just-in-time simulation as in Buja and Rolke (2006).

6.1.5 One-sample FWER

We now compare the basic one-sided Dirichlet test in Method [17] with the stepdown and pre-test procedures in Methods 10 and 12, first in terms of FWER. Table 3 shows FWER
for an example where the null distribution is Uniform(−1, 1) and $H_{0\tau} : F^{-1}(\tau) \leq F_{0}^{-1}(\tau)$ as in Task 10. The Dirichlet test in Method 17 always controls FWER, but well below the required level when $F(\cdot) \neq F_{0}(\cdot)$. The FWERs for the stepdown method and combined pre-test/stepdown method are higher but still below the nominal $\alpha = 0.1$ level, as desired. The effect of the stepdown procedure is largest when $\{\tau : F^{-1}(\tau) = F_{0}^{-1}(\tau)\} = [0, 0.5]$ or {0.5} because the initial test correctly rejects $H_{0\tau}$ for some $\tau > 0.5$. The effect of the pre-test is largest when that set is [0.5, 1] or {0.5} because the pre-test correctly detects that $F^{-1}(\tau) \leq F_{0}^{-1}(\tau)$ is not binding for some of the $\tau < 0.5$. The most important result in Table 3, though, is the strong control of FWER shown by all of our methods.

<table>
<thead>
<tr>
<th>${\tau : H_{0\tau} \text{ is true}}$</th>
<th>${\tau : F^{-1}(\tau) = F_{0}^{-1}(\tau)}$</th>
<th>Dirichlet</th>
<th>Stepdown</th>
<th>Pre+Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0, 1]</td>
<td>[0, 1]</td>
<td>0.101</td>
<td>0.101</td>
<td>0.101</td>
</tr>
<tr>
<td>[0, 0.5]</td>
<td>[0, 0.5]</td>
<td>0.048</td>
<td>0.083</td>
<td>0.082</td>
</tr>
<tr>
<td>[0, 1]</td>
<td>[0.5, 1]</td>
<td>0.068</td>
<td>0.068</td>
<td>0.079</td>
</tr>
<tr>
<td>[0, 0.5]</td>
<td>{0.5}</td>
<td>0.004</td>
<td>0.017</td>
<td>0.024</td>
</tr>
</tbody>
</table>

Table 3: FWER, nominal level $\alpha = 0.1$, $H_{0\tau} : F^{-1}(\tau) \leq F_{0}^{-1}(\tau)$, $F_{0} = \text{Uniform}(-1, 1)$, $n = 100$, 1000 replications. For $\tau$ where $F^{-1}(\tau) \neq F_{0}^{-1}(\tau)$, $F^{-1}(\tau) = 4(\tau - 0.5)$. “Dirichlet” is Method 17, “Stepdown” is Method 10, and “Pre+Step” is Methods 10 and 12 combined.

6.1.6 One-sample power improvements

Figure 4 compares the pointwise (by $\tau$) RPs of the basic Dirichlet test (Method 17), the stepdown procedure (Method 10), and a combined pre-test and stepdown procedure (Methods 10 and 12), the same methods shown in Table 3. Since all methods (correctly) have RP near zero for $\tau < 0.5$, only larger $\tau$ are shown. The DGPs (and the simulations themselves) are the same as the rows in Table 3 where $\{\tau : H_{0\tau} \text{ is true}\} = [0, 0.5]$ and $\{\tau : F^{-1}(\tau) = F_{0}^{-1}(\tau)\} = [0, 0.5]$ or {0.5}. By construction, the stepdown method rejects at least as frequently as the basic method at each quantile, and adding the pre-test also weakly increases RP.\(^{19}\) In the left panel, $F^{-1}(\tau) = F_{0}^{-1}(\tau)$ for $\tau \leq 0.5$, so the null hypoth-

\(^{19}\)Technically, since the pre-test is computed before the stepdown procedure and may lead to differently selected $r_{k,i}$, using both instead of just the stepdown does not weakly increase RP by construction, so there may exist datasets (and thus simulations) where the ordering does not hold. The figure suggests that this
esis inequality constraint is binding and the pre-test is not helpful. In the right panel, the constraint is not binding for \( \tau < 0.5 \), so the pre-test improves pointwise power beyond that of the stepdown method alone.

Figure 4: Simulated pointwise RP by quantile \((\tau)\), \( H_{0r} : F^{-1}(\tau) \leq F_0^{-1}(\tau), F_0 = \text{Uniform}(-1,1) \), \( n = 100 \), FWER level \( \alpha = 0.1 \), 1000 replications. Left: \( F^{-1}(\tau) = F_0^{-1}(\tau) \) for \( \tau \leq 0.5 \), \( F^{-1}(\tau) = 4(\tau - 0.5) \) otherwise. Right: \( F^{-1}(\tau) = 4(\tau - 0.5) \). “Dirichlet” is Method 17, “Stepdown” is Method 10, and “Pre+Step” is Methods 10 and 12 combined.

6.2 Two-sample simulations

6.2.1 Two-sample global type I error rate

Table 4 compares type I error rates for Task 6, testing \( H_0 : F_X(\cdot) = F_Y(\cdot) \), over \( 10^6 \) simulation replications. Since all tests are distribution-free, both samples are iid Uniform(0,1). The various values of \( n_X, n_Y \), and \( \alpha \) are shown in the table.

Table 4: Simulated type I error rate, two-sample, \( H_0 : F_X(\cdot) = F_Y(\cdot), 10^6 \) replications.

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( n_X )</th>
<th>( n_Y )</th>
<th>Dirichlet</th>
<th>KS</th>
<th>KS (exact)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>25</td>
<td>500</td>
<td>0.050</td>
<td>0.039</td>
<td>0.049</td>
</tr>
<tr>
<td>0.10</td>
<td>25</td>
<td>500</td>
<td>0.100</td>
<td>0.082</td>
<td>0.095</td>
</tr>
<tr>
<td>0.10</td>
<td>30</td>
<td>30</td>
<td>0.101</td>
<td>0.071</td>
<td>0.071</td>
</tr>
<tr>
<td>0.10</td>
<td>100</td>
<td>100</td>
<td>0.101</td>
<td>0.078</td>
<td>0.078</td>
</tr>
</tbody>
</table>

point is unimportant in practice.
Table 4 shows our method’s exact size. The asymptotic KS test is somewhat conservative in these cases, as is the “exact” KS test. The exact and asymptotic results being identical in some rows is initially surprising since their reported $p$-values are always different. However, for a given $(n_X, n_Y)$, there is only a finite number of different $p$-values that the KS (exact or asymptotic) can report. In these DGPs, the asymptotic and exact $p$-values are so similar (within tenths of a percentage point) that they never fall on opposite sides of $\alpha$. The discrete $p$-value distribution has a noticeable effect on KS with $n_X = n_Y = 30$ and $100$, whereas the effect on the Dirichlet test is negligible in Table 4.

### 6.2.2 Two-sample even sensitivity

Figure 5 is the two-sample analog of Figure 3, showing pointwise RPs of our Method 4 and the exact KS test for Task 6, two-sided testing of $H_0 : F_X(\cdot) = F_Y(\cdot)$. Both distributions are Uniform(0,1). As discussed in Section 3.4, the finite number of possible orderings of $X$ and $Y$ observations makes it impossible to perfectly equate all pointwise RPs. Nonetheless, the comparison of the Dirichlet and KS methods is striking and mirrors the one-sample setting.

![Figure 5: Simulated pointwise RP, $n_X = n_Y = 40$ (left) and $n_X = n_Y = 80$ (right), $\alpha = 0.1$, $10^6$ replications, $F_X = F_Y = \text{Uniform}(0,1)$. KS uses exact critical values.](image-url)
6.2.3 Two-sample global power

We examine power for Task 6, two-sided testing of $H_0 : F_X(\cdot) = F_Y(\cdot)$. One distribution is $N(0, 1)$, while the other is $N(\mu, \sigma^2)$; $n_X = n_Y = 30$ and 100 are considered, with $\alpha = 0.1$ and $10^6$ replications each.

Figure 6 shows power curves for our Dirichlet method (Method 4) and the exact KS. For the alternatives where $\sigma = 1$ and $\mu$ varies (left column), the Dirichlet test has better power, but the difference is small because the deviations are biggest in the middle of the distribution, where KS is most sensitive. For the alternatives where $\mu = 0$ and $\sigma$ varies (right column), the Dirichlet method has significantly better power since the deviations are largest in the tails.

6.2.4 Two-sample computation time

For the example in Table 4, using the pre-computed lookup table, the Dirichlet method runs in hundredths of a second, as do both the asymptotic KS and exact KS. Pre-computing the $\hat{\alpha}$ for the lookup table entries needed for Table 4 took between 16 minutes ($n_X = n_Y = 30$, $\alpha = 0.1$) and 130 minutes ($n_X = 25$, $n_Y = 500$, $\alpha = 0.1$) on a standard desktop computer in the year 2015. To be fair, this a different computation than the permutation test proposed in Buja and Rolke (2006). However, even for their Section 5.2 example in which $n_X = 254$, they note using only 100 permutations (p. 14) and only 50 points of evaluation “for computational expediency” (p. 15), suggesting that computation time is significantly longer than that of our lookup table method (for a comparable degree of precision).

6.2.5 Two-sample FWER

We compare the FWER of our two-sample, one-sided tests for Tasks 12 and 14 with $\alpha = 0.05$: the basic Dirichlet test in Method 9, the joint quantile difference test in Method 13, the stepdown procedure in Method 14, and the combined pre-test/stepdown procedure in Method 15. For Method 9, we forgo the adjustment of $\alpha$ based on (1) in favor of using the
Figure 6: Simulated power curves, two-sample, $\alpha = 0.1$, $10^6$ replications, $X_i \sim iid N(0,1)$, $Y_i \sim iid N(\mu,\sigma^2)$. Left column: $\sigma = 1$, $\mu$ on horizontal axis; right: $\mu = 0$, $\sigma$ on horizontal axis. Top row: $n_X = n_Y = 30$; bottom: $n_X = n_Y = 100$. 
lookup table for faster computation; the resulting size distortion in Table 5 is negligible (as expected, since $\alpha^2$ is small).

In each DGP, $F_Y(\cdot)$ is the CDF of the Uniform$(-1, 1)$ distribution, $F_X^{-1}(\tau) = 4(\tau - 0.5)$ wherever $F_X^{-1}(\tau) \neq F_Y^{-1}(\tau)$, and $n_X = n_Y = 200$; 1000 replications are run.

In Table 5, all four methods show strong control of FWER. The lone small exception is a 0.005 discrepancy for the basic test (Method 9) when the global null holds, due to not using the adjustment in (1). The basic test’s strong control of FWER reflects Proposition 9. Propositions 15 and 16 are not rigorously proved, but the stepdown and pre-test procedures’ strong control of FWER in Table 5 suggests that for these DGPs the fixed-quantile asymptotic framework provides a good finite-sample approximation. Overall, the patterns are similar to those in Table 3.

<table>
<thead>
<tr>
<th>$H_{0r}$ true</th>
<th>$F_X^{-1}(\tau) = F_Y^{-1}(\tau)$</th>
<th>Basic</th>
<th>Joint</th>
<th>Stepdown</th>
<th>Pre+Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau \in [0, 1]$</td>
<td>$\tau \in [0, 1]$</td>
<td>0.055</td>
<td>0.044</td>
<td>0.044</td>
<td>0.044</td>
</tr>
<tr>
<td>$\tau \in [0, 0.5]$</td>
<td>$\tau \in [0, 0.5]$</td>
<td>0.035</td>
<td>0.031</td>
<td>0.044</td>
<td>0.044</td>
</tr>
<tr>
<td>$\tau \in [0, 1]$</td>
<td>$\tau \in [0.5, 1]$</td>
<td>0.019</td>
<td>0.026</td>
<td>0.026</td>
<td>0.032</td>
</tr>
<tr>
<td>$\tau \in [0, 0.5]$</td>
<td>$\tau \in {0.5}$</td>
<td>0.004</td>
<td>0.000</td>
<td>0.002</td>
<td>0.006</td>
</tr>
</tbody>
</table>

Table 5: FWER, nominal level $\alpha = 0.05$, $H_{0r} : F_X^{-1}(\tau) \leq F_Y^{-1}(\tau)$, $F_Y = \text{Uniform}(-1, 1)$, $n_X = n_Y = 200$, 1000 replications. For $\tau$ where $F_X^{-1}(\tau) \neq F_Y^{-1}(\tau)$, $F_X^{-1}(\tau) = 4(\tau - 0.5)$. “Basic” is Method 9, “Joint” is Method 13, “Stepdown” is Method 14, and “Pre+Step” is Method 15.

6.2.6 Two-sample power improvement

Figure 7 compares the pointwise (by $\tau$) RPs of the methods shown in Table 5. The DGPs (and the simulations themselves) are also the same as in Table 5. As in the one-sample case, the stepdown and pre-test methods reject at least as frequently as the basic method at each quantile. The power improvements are similar to Figure 4: modest but noticeable over a range of $\tau > 0.5$.

The RPs of the joint QTE test (Method 13) are slightly larger than those of the “basic” test (Method 9) since the former explicitly focuses on fewer quantiles. One could further
increase pointwise power by examining yet fewer quantiles, but not under the mantle of “evenly sensitive” distributional inference.

7 Empirical examples

To demonstrate our new methods, we revisit the data provided by Gneezy and List (2006) and Levy (2009). Another example based on Hsu, Huang, and Tang (2007) is in Supplementary Appendix E. Results may be replicated using our code (available from latter author’s website) and data published in or along with the original papers.

7.1 Gift exchange

For a two-sample example, we examine the experimental data from Gneezy and List (2006, Tables I and V). Control group individuals work for a certain advertised hourly wage, while treatment group individuals are surprised with a larger “gift” wage upon arrival. The “gift exchange” question is whether the higher wage induces higher effort (proxied by productiv-
ity). 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.

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. Here, we examine heterogeneity in the period 1 treatment effect, testing for treatment effects across the productivity distribution. The two-sided nominal level (FWER) is $\alpha = 0.1$ to match the original one-sided 5% tests. As noted, with such small sample sizes, discreteness of the RP distribution prevents a test of exact 10% level, so exact 8.5% (library) and 9.3% (fundraising) levels are used instead.

Figure 8 shows two-sided null quantile function bands used by Method 9, for both tasks. For the library task, our test cannot reject equality of the treatment and control productivity distributions at an 8.5% level. However, there is a point near the 0.8-quantile that almost triggers a rejection. Testing first-order stochastic dominance, we cannot reject that the treatment dominates the control distribution ($p = 0.996$), while we can reject at a 10% level that the control distribution dominates the treatment distribution ($p = 0.076$). The KS cannot: its $p$-value is 0.13.

For fundraising, there are two ranges in the lower half of the distribution where a zero quantile treatment effect is rejected. These are highlighted by the filled vertical bars in Figure 8, one around the 0.2-quantile and the other around the 0.4-quantile. This suggests the gift wage most strongly affects low-productivity individuals in fundraising, in contrast to the library task where the treatment effect is (statistically) largest near the 0.8-quantile. This is consistent with the findings of Goldman and Kaplan (2015b), who test quartile treatment effects with the same data, but more informative since all quantiles of the distribution were considered, not just a few (arbitrary) quantiles. Testing first-order stochastic dominance, we cannot reject that the treatment dominates the control distribution ($p = 1$), while we can reject at a 5% level that the control distribution dominates the treatment distribution.
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): two-sided testing with exact FWER levels 8.5% (left) and 9.3% (right). Shaded (yellow) vertical rectangles in fundraising graph indicate quantiles where the bands do not overlap and zero treatment effect is rejected.
The one-sided KS $p = 0.034$: higher, but still below 0.05. For two-sided testing of zero treatment effect, the Dirichlet rejects at a 5% level while the KS cannot: the Dirichlet $p = 0.041$, while the KS $p = 0.069$.

### 7.2 City size distribution

We use our Method 1 uniform confidence band to examine the distribution of city size in the United States, as discussed by Eeckhout (2004, 2009), Levy (2009), and others. We use the population data published with Levy (2009), of 25,356 cities. Figure 9 shows the full distribution to be very close to lognormal, though exact lognormality is rejected at the 5% level. This is due to the (slight) skewness noted by Eeckhout (2004, p. 1434): in order make the lognormal CDF fit inside the uniform confidence band, it would have to become steeper below the median and less steep above the median, i.e., increased right-skewness.

![U.S. city size distribution](image)

Figure 9: Uniform 95% confidence band for U.S. city size distribution, and lognormal fit.

Figure 10 (left panel) shows the upper tail, similar to Figure 1 in Eeckhout (2009)\textsuperscript{20}. Even with large $n$, the KS insensitivity in the tails is apparent. The Dirichlet band is tight enough to exclude the lognormal (full sample) fit, while the KS is not. It is also clear that

\textsuperscript{20}The original figure shows a band obtained by inverting the Lilliefors (1967) test with the fitted lognormal as the null hypothesis, so the lower bound of the band is a constant vertical distance below the lognormal fit; this is simply a visualization of the Lilliefors (1967) test, not a confidence band for $F(\cdot)$. To compare with our Dirichlet uniform confidence band, we instead plot the KS uniform confidence band, so the lower bound is a fixed distance below the empirical distribution function.
a slightly different lognormal could fit within the Dirichlet band in the upper tail (at the expense of worse fit elsewhere), and this is informative since the band is relatively tight. The cost of the additional tail precision is shown to be quite small in the right panel of Figure 10, which shows the KS band to be slightly narrower near the median.

Figure 10: Left: close-up of upper tail in Figure 9. The “rug” at the bottom shows a vertical tick at each observed value. Right: close-up of middle of distribution in Figure 9.

8 Conclusion and extensions

We have contributed variously to the evenly sensitive Dirichlet-based approach to distributional inference and raised specific questions for further study. The impact of our contributions spans one-sample and two-sample methods for computing one-sided and two-sided hypothesis tests, p-values, and uniform confidence bands, which are summarized in Table 6. We have made computation nearly instantaneous in many cases, introduced new quantile, multiple testing, and Bayesian interpretations, established results on strong control of FWER, and improved power through stepdown and pre-test procedures. Our provided code makes these new methods readily accessible.

Many open questions remain. Discovery of two-sample $\tilde{\alpha}$ formulas analogous to those in Proposition 3 would make p-value computation instantaneous. Formal justification of the
two-sample methods that strongly control FWER may be attainable by extending the proofs in Goldman and Kaplan (2015b) to allow an increasing number of quantiles that extend into the tails asymptotically (which also means $\tilde{\alpha} \to 0$). The approach in this paper may also be extended to conditional distributional inference using a local smoothing framework like that in Goldman and Kaplan (2015a) or Shen (2016) and compared with Shen (2016) or Li, Lin, and Racine (2013), for example. Results on FWER could be extended to $k$-FWER. For two-sided uniform confidence bands (for example), instead of equal-tailed pointwise $1 - \tilde{\alpha}$ CIs, shortest length pointwise $1 - \tilde{\alpha}$ CIs could be used while maintaining even sensitivity. Last, instead of even sensitivity, the Dirichlet framework could be used to direct power to certain quantiles by allowing different $\tilde{\alpha}$ at different quantiles.

References


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.


A Table of tasks, methods, and advantages

Table 6 summarizes the new methods we introduce, ordered by which inferential task they accomplish. Advantages over existing methods are noted. All methods are implemented in the code available on the latter author’s website.

B Mathematical proofs

Lemma 2

Proof. Chicheportiche and Bouchaud (2012) show that the weighted KS critical value grows as $\sqrt{\ln[\ln(n)]}$. Imagine first the type I error rate for their pointwise CI at the sample minimum; the order of magnitude is the same at any extreme order statistic. The finite-sample distribution of the minimum of an iid sample from Uniform$(0,1)$ is $\beta(1,n)$ and has CDF $F_{\beta}(x) = 1 - (1-x)^n$ (Kumaraswamy, 1980). The implicit $\tilde{\alpha}$ is the value of $1 - F_{\beta}(x)$ when endpoint $x$ is growing with $\sqrt{\ln[\ln(n)]}/n$. Only keeping track of the order of magnitude,

$$P\left(\frac{\sqrt{n}[U_{n:1} - 1/(n+1)]}{\sqrt{1/(n+1)[1-1/(n+1)]}} > \sqrt{\ln[\ln(n)]}\right) \doteq P\left(nU_{n:1} > \sqrt{\ln[\ln(n)]}\right) = \left(1 - \sqrt{\ln[\ln(n)]}/n\right)^n$$

$$= \exp\{n \ln(1 - \sqrt{\ln[\ln(n)]}/n)\}$$

$$\doteq \exp\{-\sqrt{\ln[\ln(n)]}\},$$

the implicit order of magnitude of $\tilde{\alpha}$.

At a central order statistic, the weighted KS asymptotic distribution is normal. Again only tracking order of magnitude, for some $0 < p < 1$ and $Z \sim N(0,1)$,

$$P\left(\frac{\sqrt{n}(U_{n:p(n+1)} - p)}{\sqrt{p(1-p)}} > \sqrt{\ln(n)}\right) \doteq P\left(Z > \sqrt{\ln(n)}\right)$$
<table>
<thead>
<tr>
<th>Task</th>
<th>Method(s)</th>
<th>Advantage over KS</th>
<th>Advantage over Buja and Rolke (2006)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (1-sample, 2-sided uniform band)</td>
<td>1</td>
<td>Even sensitivity</td>
<td>Instant computation</td>
</tr>
<tr>
<td>2 (1-sample, 1-sided uniform band)</td>
<td>16</td>
<td>Even sensitivity</td>
<td>Instant computation</td>
</tr>
<tr>
<td>3 (1-sample, 2-sided global test)</td>
<td>2</td>
<td>Even sensitivity</td>
<td>Instant computation</td>
</tr>
<tr>
<td>4 (1-sample, 1-sided global test)</td>
<td>17, 12</td>
<td>Even sensitivity</td>
<td>Instant computation</td>
</tr>
<tr>
<td>5 (1-sample p-value)</td>
<td>3, 18</td>
<td>Even sensitivity</td>
<td>Instant computation</td>
</tr>
<tr>
<td>6 (2-sample, 2-sided global test)</td>
<td>4</td>
<td>Even sensitivity</td>
<td>Faster computation</td>
</tr>
<tr>
<td>7 (2-sample, 1-sided global test)</td>
<td>19</td>
<td>Even sensitivity</td>
<td>Now possible</td>
</tr>
<tr>
<td>8 (2-sample p-value)</td>
<td>5, 20</td>
<td>Even sensitivity</td>
<td>Faster computation</td>
</tr>
<tr>
<td>9 (1-sample, 2-sided multiple testing)</td>
<td>7, 21</td>
<td>Even sensitivity; power improvement in Method 21</td>
<td>Now possible</td>
</tr>
<tr>
<td>10 (1-sample, 1-sided multiple testing)</td>
<td>8, 10, 12</td>
<td>Even sensitivity; power improvement in Methods 10 and 12</td>
<td>Now possible</td>
</tr>
<tr>
<td>11, 12 (2-sample multiple testing)</td>
<td>9</td>
<td>Even sensitivity</td>
<td>Now possible</td>
</tr>
<tr>
<td>13 (2-sample, 2-sided multiple testing)</td>
<td>14</td>
<td>Even sensitivity; power improvement</td>
<td>Now possible</td>
</tr>
<tr>
<td>14 (2-sample, 1-sided multiple testing)</td>
<td>14, 15</td>
<td>Even sensitivity; power improvement</td>
<td>Now possible</td>
</tr>
<tr>
<td>$n_X \gg n_Y$</td>
<td>6</td>
<td>See Tasks 3, 4, 5, 9, 10 above</td>
<td>See Tasks 3, 4, 5, 9, 10 above</td>
</tr>
</tbody>
</table>

Table 6: Summary of inferential tasks, showing applicable new methods and their advantages over prior methods. “Now possible” means no such prior method existed. Task “$n_X \gg n_Y$” refers to any two-sample task where one sample size is much larger than the other.
\[
\leq \frac{\exp\{-\sqrt{\ln[\ln(n)]^2/2}\}}{\sqrt{\ln[\ln(n)]\sqrt{2\pi}}}
= \exp\{-\ln[\ln(n)] - \ln\left(\sqrt{\ln[\ln(n)]}\right)\}.
\]

For the unweighted KS, the critical value is fixed. Since we only track order of magnitude, we let it equal one. For the first order statistic,

\[
P\left(\sqrt{n}(U_{n,1} - 1/(n+1)) > 1\right) = P(U_{n,1} > 1/\sqrt{n}) = (1 - 1/\sqrt{n})^n \approx e^{-\sqrt{n}} \to 0.
\]

In contrast, at a central order statistic, again with 0 < p < 1 and Z ∼ N(0,1),

\[
P\left(\sqrt{n}(U_{n,p(n+1)} - p) > 1\right) = P(Z > 1/\sqrt{p(1-p)}),
\]

which is a fixed, strictly positive value. This value is largest when p(1 − p) is largest, which occurs at p = 1/2. \(\square\)

**Proposition 3**

*Proof.* Figure 1 (right panel) verifies the formulas by simulation. Monotonicity in n ensures that interpolation error is negligible. \(\square\)

**Proposition 4**

*Proof.* See Moscovich-Eiger et al. (2015, §A.5). \(\square\)

**Theorem**

*Proof.* Consider the one-sample case, and fix α. If \(H_0\) is false, then \(\exists x\) s.t. \(p = F(x) = F_0(x) - 2\epsilon\); let \(x = 0\) and \(\epsilon > 0\) without loss of generality. By the continuity in Assumption 1, there also exists \(\delta > 0\) s.t. \(F(\delta) = F_0(0) - \epsilon = p + \epsilon\). Let \(r \equiv \lfloor n(p + \epsilon)\rfloor + 1\), so \(X_{n,r} = \hat{F}^{-1}(p + \epsilon)\).

Asymptotically, if 1) \(P(X_{n,r} > 0) \to 1\) and 2) the upper endpoint of the pointwise CI at \(X_{n,r}\) converges to \(p + \epsilon\), then the test is consistent: with probability approaching one, \(F_0(X_{n,r})\) exceeds the upper CI endpoint at \(X_{n,r}\). Since \(X_{n,r} \xrightarrow{p} F^{-1}(p + \epsilon) > F^{-1}(p) = 0\) implies (1), it remains to show (2).

For (2), we invoke Lemma 7(iii) of Goldman and Kaplan (2015a). Applied to the present special case of \(U_{n,r} \equiv F(X_{n,r}) \sim \beta(r, n + 1 - r)\), whose \(1 - \bar{a}/2\) quantile is the upper CI.
To establish first-order consistency, we need only find \( a_n \) such that \( \sqrt{a_n/n} \to 0 \) and \( \alpha^2 n / 2 \gtrsim \sqrt{\ln[\ln(n)]} \) (strictly), using the rate \( \tilde{\alpha} \propto \exp\{-\sqrt{\ln[\ln(n)]}\} \) from Proposition 3. These conditions are satisfied by \( a_n = \ln(n) \), for example.

For the two-sample case, similarly, let \( x = 0, \epsilon > 0, \delta > 0, F_X(0) = p, F_Y(0) = p + 2\epsilon, F_X(\delta) = p + \epsilon \). Let \( r = \lfloor n_X(p+\epsilon) \rfloor + 1 \) and \( k = \lfloor n_Yp \rfloor + 1 \). To satisfy (1), \( X_n: r p \to F_X^{-1}(p+\epsilon) > 0 \) and \( Y_n: k p \to F_Y^{-1}(p) < 0 \). To satisfy (2), the upper CI endpoint at \( X_n: r \) again converges asymptotically to \( p + \epsilon \), and similarly the lower CI endpoint at \( Y_n: k \) converges to \( p + 2\epsilon \).

**Proposition 6**

**Proof.** Let

\[
D_n^x \equiv \sqrt{n} \left| \hat{F}(x) - F(x) \right|, \quad D_n \equiv \sup_{x \in \mathbb{R}} D_n^x,
\]

and let critical value \( c_n(\alpha) \) satisfy

\[
P(D_n > c_n(\alpha)) = \alpha \text{ in finite samples}.
\]

Let

\[
D_n^{x,0} \equiv \sqrt{n} \left| \hat{F}(x) - F_0(x) \right|, \quad I \equiv \{ x : H_{0x} \text{ is true} \}, \quad D_n^I \equiv \sup_{x \in I} D_n^{x,0} = \sup_{x \in I} D_n^x.
\]

Then, since \( I \subseteq \mathbb{R} \),

\[
\text{FWER} \equiv P(D_n^I > c_n(\alpha)) \leq P(D_n > c_n(\alpha)) = \alpha.
\]

The one-sided case follows the same argument (after modifying \( D_n^x \) and \( D_n^{x,0} \)), with the additional inequality that if \( H_{0x} : F(x) \leq F_0(x) \) is true, then \( \hat{F}(x) - F_0(x) \leq \hat{F}(x) - F(x) \).

**Proposition 7**

**Proof.** Since Method 7 is based on a band with (approximate) \( 1 - \alpha \) uniform coverage probability regardless of \( I \) or \( F(\cdot) \), and since \( I \subseteq (0,1) \),

\[
\alpha \approx P \left( \bigcup_{\tau \in (0,1)} F^{-1}(\tau) \notin \left[ \hat{u}^{-1}(\tau), \hat{\ell}^{-1}(\tau) \right] \right)
\]

\[\text{The critical value is distribution-free, depending only on } n \text{ and } \alpha. \text{ The following arguments could also be applied asymptotically to the } F\text{-Brownian bridge limit of the empirical process.}\]
\[ \geq P\left( \bigcup_{\tau \in I} F^{-1}(\tau) \notin \left[ \hat{u}^{-1}(\tau), \hat{\ell}^{-1}(\tau) \right] \right) \]
\[ = P\left( \bigcup_{\tau \in I} \{ \text{reject } H_{0\tau} \} \right) \equiv \text{FWER}. \]

The same argument applies to Method 8, replacing \( \hat{\ell}^{-1}(\tau) = \infty \) or \( \hat{u}^{-1}(\tau) = -\infty \).

**Proposition 8**

**Proof.** Consider any given \( n_X, n_Y, \) and \( \alpha \). Let

\[ D_{n_X,n_Y}^r \equiv \sqrt{\frac{n_Xn_Y}{n_X + n_Y}} \left\{ \hat{F}_X(r) - \hat{F}_Y(r) \right\}, \quad D_{n_X,n_Y} \equiv \sup_{\tau \in \mathbb{R}} D_{n_X,n_Y}^r, \quad K_{n_X,n_Y} \overset{d}{=} D_{n_X,n_Y} \mid F_X(\cdot) = F_Y(\cdot). \]

Let critical value \( c_{n_X,n_Y}(\alpha) \) satisfy \( P(K_{n_X,n_Y} > c_{n_X,n_Y}(\alpha)) \leq \alpha \) in finite samples. Letting

\[ D_{n_X,n_Y}^l \equiv \sup_{\tau \in I} D_{n_X,n_Y}^r, \quad I \equiv \{ \tau : H_{0\tau} \text{ is true} \}, \]

a familywise error is committed if and only if \( D_{n_X,n_Y}^l > c_{n_X,n_Y}(\alpha) \).

Pointwise, \( n_X \hat{F}_X(r) \sim \text{Binomial}(n_X, F_X(r)) \), \( n_Y \hat{F}_Y(r) \sim \text{Binomial}(n_Y, F_Y(r)) \), and \( \hat{F}_X(\cdot) \perp \perp \hat{F}_Y(\cdot) \) by Assumption 2, so the distribution of \( D_{n_X,n_Y}^r \) depends only on \( F_X(\cdot) \) and \( F_Y(\cdot) \). More generally, the distribution of

\[ \left( n_X[\hat{F}_X(r_1) - \hat{F}_Y(r_1)], \ldots, n_X[\hat{F}_X(r_m) - \hat{F}_Y(r_{m-1})] \right) \]

is multinomial with parameters \( n_X \) and \( (F_X(r_1), F_X(r_2) - F_X(r_1), \ldots, F_X(r_m) - F_X(r_{m-1})) \), and the joint distribution of the \( D_{n_X,n_Y}^r \) over any set \( r \in S \) depends only on \( F_X(\cdot) \) and \( F_Y(\cdot) \) for \( r \in S \). This implies that the distribution of \( D_{n_X,n_Y}^l \) is the same as if \( F_X(\cdot) = F_Y(\cdot) \) for all \( r \in \mathbb{R} \), even outside \( I \), so \( D_{n_X,n_Y}^l \) is stochastically dominated by \( K_{n_X,n_Y} \) since \( I \subseteq \mathbb{R} \). FWER is thus

\[ P(D_{n_X,n_Y}^l > c_{n_X,n_Y}(\alpha)) \leq P(K_{n_X,n_Y} > c_{n_X,n_Y}(\alpha)) \leq \alpha. \]

As with the proof of Proposition 6, the one-sided case follows the same argument, except with different definitions and an additional inequality.

**Proposition 9**

**Proof.** Consider any given \( n_X, n_Y, \) and \( \alpha \). Define \( \phi(r) \equiv 1\{ \text{reject } H_{0r} \} \), the pointwise rejection function for the Dirichlet test. From the construction of Methods 4 and 9, we
can write $\phi(r) = h(\hat{F}_X(r), \hat{F}_Y(r); n_X, n_Y, \alpha)$ for some function $h(\cdot)$: rejection of $H_{0r}$ is determined only by the random variables $\hat{F}_X(r)$ and $\hat{F}_Y(r)$, given $n_X$, $n_Y$, and $\alpha$. Thus, we may apply the same argument as in the proof of Proposition 8, although the specific expressions differ.

As before, let $I \equiv \{ r : H_{0r} \text{ is true} \}$. A familywise error is committed when $R_I \equiv \sup_{r \in I} \phi(r) = 1$. Define $R^g_I$ to be a random variable with the distribution that $R_I$ would have if the global (hence “g”) null $F_X(\cdot) = F_Y(\cdot)$ were true, and define $R^g_R$ to be the distribution of $\sup_{r \in \mathbb{R}} \phi(r)$ if $F_X(\cdot) = F_Y(\cdot)$.22

As in the proof of Proposition 8, the joint distribution of $\{\phi(r)\}_{r \in I}$ depends only on $n_X$, $n_Y$, $\alpha$, and $\{F_X(r), F_Y(r)\}_{r \in I}$. Thus, the distribution of $R_I$ is invariant to the behavior of $F_X(r)$ and $F_Y(r)$ for $r \notin I$, so $R_I \overset{d}{=} R^g_I$. Since $I \subseteq \mathbb{R}$, $R^g_I \leq R^g_R$ in any realization, so the distribution of $R^g_I$ is stochastically dominated by that of $R^g_R$. Consequently,

$$P(R_I = 1) = P(R^g_I = 1) \leq P(R^g_R = 1) \leq \alpha,$$

where the last inequality uses the global test’s (Method 4) size control at level $\alpha$.

As with the proof of Proposition 8, the one-sided case follows the same argument, except with different definitions and an additional inequality.

**Proposition 10**

*Proof.* The frequentist sampling distribution in Wilks (1962) matches the Banks (1988) posterior, although the interpretation differs. Let $X_{n:0} = a$ and $X_{n:n+1} = b$; the values of $a$ and $b$ do not matter for our comparison. The Banks (1988) posterior first draws

$$(p_0, \ldots, p_n) \sim \text{Dir}(1, \ldots, 1)$$

and then uniformly spreads the probability $p_k$ over the interval $(X_{n:k}, X_{n:k+1}]$. The Banks (1988) posterior joint distribution of

$$(F(X_{n:1}) - F(X_{n:0}), \ldots, F(X_{n:n}) - F(X_{n:n-1}))$$

is the same as that of $(p_0, \ldots, p_{n-1})$, which matches the Wilks (1962) Dirichlet sampling distribution from Theorem 1. Thus, the Dirichlet method can be interpreted as constructing $n$ intervals for the $F(X_{n:k})$ with joint $1 - \alpha$ credibility. The stair-step interpolation does not increase the uniform credibility level for the same reasons as in Section 2.1. \qed

---

22Since the distribution of $R^g_I$ (but not $R^g_R$) depends on $F_X(\cdot)$, assume it refers to counterfactually changing $F_Y(\cdot)$ to match $F_X(\cdot)$, although the opposite is also valid.
Theorem 11

Proof. Consider the one-sided case with $H_0: F^{-1}(\tau) \geq F_0^{-1}(\tau)$ for $\tau \in (0, 1)$, using the quantile perspective from Section 4.1; the other cases are analogous.

As in the analysis of Method 17, we focus on the $n$ hypotheses $F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k)$ for $k = 1, \ldots, n$ since rejections at other $\tau$ are simply by logical implication of the monotonicity of $F_0^{-1}(\cdot)$. For example, for the initial test, $X_{n,k} < F_0^{-1}(\ell_k) \implies$ reject $H_0$ for all $\tau \leq \ell_k$ such that $X_{n,k} < F_0^{-1}(\tau) \implies X_{n,k} < F_0^{-1}(\ell_k)$.

Let $K \equiv \{k : F^{-1}(\ell_k) \geq F_0^{-1}(\ell_k)\}$, the (true) set of true hypotheses. For $k \in K$, let $r_k$ satisfy $r_k \leq k$ and

$$\alpha = 1 - P\left(\bigcap_{k \in K} \{X_{n:r_k} \geq F_0^{-1}(\ell_k)\}\right).$$

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

The proof is by induction. Consider any dataset where

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

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 (5), this implies FWER $\leq \alpha$.

Proposition 12

Proof. The test is based on a one-sided uniform confidence band, so it strongly controls FWER by the same argument as in the proof of Proposition 7.

Theorem 13

Proof. The stated FWER bound is very 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 8, i.e., the test without using a pre-test, never
falsely rejects when the pre-test falsely rejects. Then, the FWER is \( \alpha + \alpha_p \). □

**Proposition 15**

**Proof.** The proof is similar to that of Theorem 11, relying primarily on the monotonicity property and induction. Let \( T \equiv \{ j : H_{0j} \text{ is true} \} \), the (true) set of true hypotheses. Let \( \alpha^* \) solve (4) when \( \hat{T}_i = T \). The stepdown procedure ensures \( \hat{T}_0 \supset \hat{T}_1 \supset \cdots \), which in turn implies \( \tilde{\alpha}_0 < \tilde{\alpha}_1 < \cdots \), where \( \tilde{\alpha}_i \) is the solution to (4) with \( \hat{T}_i \). Since the function \( k_{X,j}^\ell(\tilde{\alpha}) \) is increasing in \( \tilde{\alpha} \) while \( k_{X,j}^u(\tilde{\alpha}) \) is decreasing in \( \tilde{\alpha} \), and similarly for \( Y \), rejection of \( H_{0j} \) is “monotonic” over iterations, for all \( j \): \( H_{0j} \) is rejected in iteration \( i \) if (but not only if) it is rejected in \( i - 1 \).

Proposition 15 assumes that the joint CIs calibrated by (4) using the true \( \hat{T} = T \) have asymptotically exact coverage probability \( 1 - \alpha \). That is, there is (approximately) \( 1 - \alpha \) probability of a dataset where

\[
1 = \prod_{j \in T} \{ X_{nX:k_{X,j}^\ell(\tilde{\alpha}^*)} \leq Y_{nY:k_{X,j}^u(\tilde{\alpha}^*)} \} \times \{ X_{nX:k_{X,j}^u(\tilde{\alpha}^*)} \geq Y_{nY:k_{X,j}^\ell(\tilde{\alpha}^*)} \}.
\]

In iteration \( i \), if \( \hat{T}_i \supseteq T \), then none of the true hypotheses are rejected since \( \tilde{\alpha}_i \leq \tilde{\alpha}^* \), which implies \( k_{X,j}^u(\tilde{\alpha}_i) \geq k_{X,j}^u(\tilde{\alpha}^*) \), \( k_{Y,j}^u(\tilde{\alpha}_i) \geq k_{Y,j}^u(\tilde{\alpha}^*) \), \( k_{X,j}^\ell(\tilde{\alpha}_i) \leq k_{X,j}^\ell(\tilde{\alpha}^*) \), and \( k_{Y,j}^\ell(\tilde{\alpha}_i) \leq k_{Y,j}^\ell(\tilde{\alpha}^*) \). Consequently, \( \hat{T}_{i+1} \supseteq T \). Since \( \hat{T}_0 \supseteq T \), the stepdown procedure does not reject any true hypothesis in such a dataset, for any \( T \), so FWER is strongly controlled. □

**Proposition 16**

**Proof.** The proof is the same as that of Theorem 13. □

### C Additional methods

**Method 16.** For Task 2 with confidence level \( 1 - \alpha \), solve for \( a_2 \) in \( \alpha = a_2 + (a_2^2 + a_4^3)/2 + 5a_4^2/8 + 7a_5^2/8 \), construct a two-sided \( 1 - 2a_2 \) uniform confidence band using Method 1, and take \( \ell(\cdot) \) or \( u(\cdot) \). More directly, letting \( B_{k,\tilde{\alpha}} \) denote the \( \tilde{\alpha} \)-quantile of the \( \beta(k, n + 1 - k) \) distribution, let \( \ell_k \) solve

\[
\ell_k = B_{k,\tilde{\alpha}}, \quad k = 1, \ldots, n,
\]

\[
1 - \alpha = P \left( \bigcap_{k=1}^{n} X_{n;k} \geq F^{-1}(\ell_k) \right) = P \left( \bigcap_{k=1}^{n} F(X_{n;k}) \geq \ell_k \right).
\]
for some $\tilde{\alpha}$, using the distribution in Theorem 1. Analogously, the $u_k$ would directly solve

$$u_k = B_{k,1-\tilde{\alpha}}, \quad k = 1, \ldots, n,$$

$$1 - \alpha = P\left(\bigcap_{k=1}^{n} X_{n,k} \leq F^{-1}(u_k)\right) = P\left(\bigcap_{k=1}^{n} F(X_{n,k}) \leq u_k\right).$$

**Method 17.** For Task 4, first use Method 16 to construct a one-sided uniform confidence band. Reject $H_0$ if $F_0(x)$ is outside the band at any $x$.

**Method 18.** For Task 5, to compute a one-sided $p$-value, first compute the two-sided $p$-value, $p_2$, using Method 3. Then, solve for the one-sided $p$-value in (1), $p_2 = 2p_1 - p_1^2$.

**Method 19.** For Task 7 with level $\alpha$, solve for $a_2$ in $\alpha = a_2 + (a_2^2 + a_3^2)/2 + 5a_2^4/8 + 7a_3^2/8$ and run Method 4 with level $2a_2$. Reject $H_0 : F_X(\cdot) \leq F_Y(\cdot)$ if there exists $r \in \mathbb{R}$ where $\hat{\ell}_X(r) > \hat{u}_Y(r)$; reject $H_0 : F_X(\cdot) \geq F_Y(\cdot)$ if there exists $r \in \mathbb{R}$ where $\hat{u}_X(r) < \hat{\ell}_Y(r)$.

**Method 20.** For Task 8, to compute a one-sided $p$-value, first compute the two-sided $p$-value, $p_2$, using Method 5. Then, solve for the one-sided $p$-value in (1), $p_2 = 2p_1 - p_1^2$.

**Method 21.** For Task 9, modify Method 10 as follows. Define $\ell_k$ and $u_k$ as in Method 16. 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 (3) with

$$\alpha \geq 1 - P\left(\bigcap_{k \in \hat{K}_0^\ell} \{X_{n,r_{k,\ell,i}} \geq F^{-1}(\ell_k)\} \cap \bigcap_{k \in \hat{K}_0^u} \{X_{n,r_{k,u,i}} \leq F^{-1}(u_k)\}\right)$$

$$\geq 1 - P\left(\bigcap_{k \in \hat{K}_0^\ell} \{F(X_{n,r_{k,\ell,i}}) \geq \ell_k\} \cap \bigcap_{k \in \hat{K}_0^u} \{F(X_{n,r_{k,u,i}}) \leq u_k\}\right),$$

(6)

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 10; either implies rejection of $H_{0,\tau} : F^{-1}(\tau) = F_0^{-1}(\tau)$.
Supplemental material:
Evenly sensitive KS-type inference on distributions

Matt Goldman*    David M. Kaplan†
February 12, 2016

Abstract

This supplement includes details on computation and additional simulation and empirical results.

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

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

\[
(U_{n:1}, U_{n:2} - U_{n:1}, U_{n:3} - U_{n:2}, \ldots, U_{n:n} - U_{n:n-1}, 1 - U_{n:n}) \sim \text{Dirichlet}(1, \ldots, 1).
\]

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 \( \ell_k \) and \( u_k \), the simulated two-sided type I error rate (for example) is

\[
\hat{\alpha} = 1 - \frac{1}{M} \sum_{m=1}^{M} 1\{\ell_1 < U_{n:1}^{(m)} < u_1\} \times \cdots \times 1\{\ell_n < U_{n:n}^{(m)} < u_n\}.
\] (7)

While (7) alone is sufficient for \( p \)-value computation, we need to search for the \( \tilde{\alpha} \) that leads to a specific desired \( \alpha \) for the simulations informing Proposition 3. 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 (7). The 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. 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. Under \( H_0 : F_X(\cdot) = F_Y(\cdot) \) and Assumption 2, the distribution of \( (X_1, \ldots, X_{n_X}, Y_1, \ldots, Y_{n_Y}) \) is the same as that of any permutation of the vector, satisfying the “randomization hypothesis” in Definition 15.2.1 of Lehmann and Romano (2005), for example. One possibility is to construct a test based on permutations of the observed data, following Theorem 15.2.1 of Lehmann and Romano (2005); this is the approach of Buja and Rolke (2006). Alternatively,
we use the fact that each of the \( \binom{n_X + n_Y}{n_X} \) orderings is equally likely under \( H_0 \), following Theorem 15.2.2 of Lehmann and Romano (2005). This result depends on \( H_0 \) being satisfied but not on any specific \( F_X(\cdot) \) or \( F_Y(\cdot) \): the argument is distribution-free, so our \( \tilde{\alpha} \) is only a function of \( \alpha, n_X, \) and \( n_Y \) and can be computed ahead of time. 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 type I error rate would be within \( c \alpha \) of the desired \( \alpha \) for some small \( c > 0 \), like \( c = 0.02 \) for \( \alpha = 0.05 \) implying type I error rate 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 rejections 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 \( \hat{\alpha} \) yields a true type I error rate above \( \alpha (1 + c) \). If the true type I error rate is \( \alpha (1 + c) \), then the total number of simulated rejections follows a Binomial\((M, \alpha (1 + c))\) distribution, so \( \hat{\alpha} \sim N(\alpha (1 + c), \alpha (1 + c) [1 - \alpha (1 + c)] / M) \), and we choose \( T \) and \( M \) to equate \( T \) with the \( p \)-quantile of this distribution:

\[
\alpha + T = \alpha (1 + c) + \Phi^{-1}(p) \sqrt{\alpha (1 + c) (1 - \alpha (1 + c)) / \sqrt{M}},
\]

\[
T = c \alpha - \Phi^{-1}(1 - p) \sqrt{\alpha (1 + c) (1 - \alpha (1 + c)) / \sqrt{M}},
\]

\[
M = \left( \frac{\Phi^{-1}(1 - p) \sqrt{\alpha (1 + c) (1 - \alpha (1 + c))}}{c \alpha - T} \right)^2.
\]

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 is all the same whether considering one-sample or two-sample inference.

### E Additional simulations and example

#### E.1 One-sample power simulations

Differences in pointwise size translate into differences in power. Following the same patterns as in Figure 3, the KS test has the highest pointwise power against deviations near the
median of a distribution and lowest pointwise power in the tails, and the weighted KS is
the opposite. The Dirichlet method 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 testing of \( H_0 : F(\cdot) = F_0(\cdot) \). The left column graphs show \( F_0(F^{-1}(\tau)) \) (dashed line). If \( H_0 \) were true, then \( F_0(F^{-1}(\tau)) = \tau \) (solid line).

For the location shifts in Figure 11, \( X_i \overset{iid}{\sim} N(0.2, 1) \) or \( N(0.3, 1) \). As the left column shows, this leads to larger deviations in the middle of the distributions than in the tails. The largest peak in pointwise power is in the middle of the distribution for the 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.

In Figure 11, the effect having pointwise equal-tailed (like Dirichlet) or symmetric (like KS) CIs is apparent. Even though the weighted KS has greater 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 tests are insensitive to such deviations, whereas the Dirichlet test 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. The Dirichlet has two pointwise power peaks, in between the lower tail and the middle and in between the middle and the upper tail, 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 graphs in Figure 12 with \( \sigma < 1 \), 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. Yet smaller and closer to the middle is the KS peak. The weighted KS and KS can have very high peaks since their peak pointwise size is higher than Dirichlet’s (which has no peak), but they perform poorly when their peak pointwise size 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 7 shows global power for one-sample, two-sided tests with \( F_0 = N(0, 1) \) and \( X_i \overset{iid}{\sim} \)
Figure 11: Simulated one-sample, two-sided, pointwise power against location shift, $F_0 = N(0, 1)$, $\alpha = 0.1$, $n = 100$, $10^6$ replications. Top: $X_i \sim N(0.3, 1)$; bottom: $X_i \sim N(0.2, 1)$. 
Figure 12: Simulated one-sample, two-sided, pointwise power against scale difference, $F_0 = N(0,1)$, $\alpha = 0.1$, $n = 100$, $10^6$ replications; $X_i \overset{iid}{\sim} N(0,\sigma^2)$ with $\sigma = 1.2$ (top), $\sigma = 0.8$ (middle), or $\sigma = 0.7$ (bottom).
$N(\mu, \sigma^2)$. For location shifts with $\mu \neq 0$ and $\sigma = 1$, the deviations are largest near the middle of the distribution, where KS has the largest pointwise power. The weighted KS is not sensitive to such deviations, so it has the worst power by far. The Dirichlet power is below KS, but very close. With $\mu = 0$ and $\sigma = 0.7$, the largest vertical deviations are at the extreme order statistics. Consequently, the weighted KS has the best power. The KS test has significantly lower power, but the Dirichlet is quite close to the weighted KS. When $\mu = 0$ and $\sigma = 1.2$, the deviations 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 no deviation at the median, KS has better power than weighted KS in this case because it has better pointwise power in between (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, $F_0 = N(0, 1)$, $\alpha = 0.1$, $n = 100$, $10^6$ replications, $X_i \sim N(\mu, \sigma^2)$. RPs are shown as percentages.

<table>
<thead>
<tr>
<th>$\mu$</th>
<th>$\sigma$</th>
<th>Dirichlet</th>
<th>KS</th>
<th>weighted KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>0.7</td>
<td>92.0</td>
<td>65.6</td>
<td>98.6</td>
</tr>
<tr>
<td>0.0</td>
<td>0.8</td>
<td>50.1</td>
<td>26.7</td>
<td>76.6</td>
</tr>
<tr>
<td>0.0</td>
<td>1.2</td>
<td>64.2</td>
<td>25.5</td>
<td>2.8</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>49.4</td>
<td>52.2</td>
<td>33.9</td>
</tr>
<tr>
<td>0.3</td>
<td>1.0</td>
<td>80.5</td>
<td>82.4</td>
<td>62.4</td>
</tr>
</tbody>
</table>

E.2 Empirical example: minimax play in professional tennis

We revisit the analysis of minimax play at professional tennis tournaments in Hsu et al. (2007), who provide their data in a supplemental file. Their approach is the same as Walker and Wooders (2001), who describe both the game theoretic setup and statistical approach in detail. The main idea is that the server may choose to serve to the right (R) or left (L), and under the minimax-optimal mixed strategy, the equilibrium probability of winning the point should be equal for R or L. A Pearson goodness-of-fit test is run on data from multiple points within a single match, generating a sample of $p$-values from multiple matches.

Figure 13 shows 90% uniform confidence bands constructed by our Method 1, related to the left column of Figure 1 in Hsu et al. (2007), which compares the distribution of $p$-values to the uniform distribution that would obtain under the null hypothesis. In this case, the

---

23They actually break down each match into four separate “point games,” which may have different winning probabilities: player A serving on the deuce side, player A serving on the ad side, and the same for player B.
Figure 13: Dirichlet and KS 90% uniform confidence bands for $p$-values from Hsu et al. (2007).
null hypothesis is not rejected for any of the groups (men, women, juniors, or all combined) by either test. However, the visually apparent precision difference in the tails suggests we may feel more secure in the non-rejection of the Dirichlet, i.e., not worry as much about type II error.