Finite-sample similar nonparametric two-sample homogeneity tests for possibly discrete distributions *

Jean-Marie Dufour †
McGill University

Abdeljelil Farhat ‡
Université de Monastir

First version: April 2014
Revised: January 2016
This version: January 2016
Compiled: January 31, 2016, 22:57

*This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universitad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds de recherche sur la société et la culture (Québec).

† William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 4400 ext. 09156; FAX: (1) 514 398 4800; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

‡ Faculté des Sciences Économiques et de Gestion de Mahdia. Mailing address: Département des Méthodes Quantitatives, Faculté des Sciences Économiques et de Gestion de Mahdia, Sidi Messaoud Hiboun, Mahdia 5111, Tunisie. TEL: 216 73 683191; FAX: 216 73 683190; e-mail: abdeljelil.farhat@umontreal.ca .

ABSTRACT

In this paper, we study several distribution-free tests for the equality of two unknown distributions, without any assumption on the continuity of the relevant distributions. This is achieved by combining two techniques: first, permutational versions of the tests are considered; second, to allow for discrete unknown distributions (under which permutation tests are not typically distribution-free) as well as for the discontinuous nature of several of the test statistics considered, we use the generalized Monte Carlo test approach proposed in Dufour (2006, *Journal of Econometrics*). The tests obtained in this way are exact and similar, in then sense that the probability of rejecting the null hypothesis is precisely equal to the stated level under all distributions compatible with the null hypothesis, irrespective of the shape of the underlying distribution. The tests studied include tests based on empirical distribution functions, nonparametric probability density estimates (even if the distribution is discrete), and differences between sample moments. In view of the fact that no test dominates the others from the power viewpoint, we also propose two-stage combined test procedures, which exhibit better overall power than the individual tests on which they are based. Finally, in a simulation experiment, we show that the technique suggested provides perfect control of test size and that the new tests proposed can yield sizeable power improvements.

Key words: nonparametric methods; two-sample problem; discrete distribution; discontinuous distribution; goodness-of-fit test; Kolmogorov-Smirnov test; Cramér-von Mises; kernel density estimator; exact test; permutation test; Monte Carlo randomized test; bootstrap; combined test procedure; induced test.

Contents

List of Tables

1. Introduction

A basic problem in statistics consists in testing whether two distributions are identical; see, for example, the recent review of Thas (2010). Specifically, we consider two independent random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ such that $F(x) = \mathbb{P}[X_i \leq x]$, $i = 1, \ldots, n$, and $G(x) = \mathbb{P}[Y_j \leq x]$, $j = 1, \ldots, m$. We shall not impose here additional restrictions on the form of the cumulative distribution functions (cdf) F and G, which may be continuous or discrete. The problem consists in testing the null hypothesis

$$H_0 : F = G \tag{1.1}$$

against the alternative

$$H_1 : F \neq G. \tag{1.2}$$

H_0 is a nonparametric hypothesis, so testing H_0 requires a distribution-free procedure. Thus, users to a goodness-of-fit test, usually the two-sample Kolmogorov-Smirnov (*KS*) test [Smirnov (1939, 1948)] or the Cramér-von Mises (*CM*) test [Lehmann (1951), Rosenblatt (1952) and Fisz (1960)]. Other procedures include permutation tests based on L_1 and L_2 distances between kernel-type estimators of the relevant probability density functions (pdf) [Allen (1997)] and tests based on the difference of the means of the two samples considered [Pitman (1937), Dwass (1957), Efron and Tibshirani (1993)]. Except for the last procedure, which is designed to have power against samples which differ through their means, the exact and limiting distributions of the test statistics are not standard, and tables of the exact distributions are only available for a limited number of sample sizes. Thus these tests are usually performed using tables of asymptotic distributions. This leads to procedures which do not have the targeted size (which can easily be too small or too large) and may have low power.

In this paper, we aim at building test procedures with two basic features. Namely, the latter should be: (1) truly distribution-free, regardless whether the underlying distribution F is discrete or continuous, and (2) exact in finite samples (*i.e.*, to achieve the desired size even for small samples). In this respect, it is important to note that the finite and large-sample distributions of usual test statistics are not necessarily distribution-free under H_0. In particular, while the *KS* and *CM* statistics are distribution-free when the observations are independent and identically distributed $(i.i.d.)$ with a continuous distribution, this is not anymore the case when they follow a discrete distribution. Further, tie probabilities are unknown under the null hypothesis, because the common distribution F is unknown. For the statistics based on kernel-type density estimators, the distribution-free property does not obtain even for $i.i.d$ observations with a continuous distribution. This difficulty can be relaxed by considering a permutational version of these tests which uses the fact that all permutations of the pooled observations are equally likely when the observations are $i.i.d$ with a continuous distributions. The latter property, however, does not hold when the observations follow a discrete distribution. So none of the procedures proposed to date for testing H_0 satisfies the double requirement of yielding a test that is both distribution-free and exact. For further discussion of permutation tests, the reader may consult Dufour and Hallin (1990, 1991, 1992, 1993), Edgington (1995), Manly (1997), Good (1994), Mielke and Berry (2001), Pesarin (2001), and Pesarin and Salmaso (2010).

Given recent progress in computing power, a way to solve this difficulty consists in using

simulation-based methods, such as bootstrapping or Monte Carlo tests. The bootstrap technique however does not ensure that the level will be fully controlled in finite samples [for further discussion of bootstrapping; see Efron (1979, 2000, 2003), Hall (1992), Efron and Tibshirani (1993), Shao and Tu (1995), Davison and Hinkley (1997), Reiczigel, Zakariás and Rózsa (2005), Hall and Keilegom (2007), Li, Maasoumi and Racine (2009), Pesarin and Salmaso (2010), Thas (2010), Zhan and Hart (2012), and Duonga (2013). However, all results presented use bootstrapping or similar methods based on large numbers of permutations. The purpose of large numbers of permutations is to allow the use of asymptotic approximations. In this context, citing several examples, Zaven and Dudewicz (2000) report that the bootstrap method may not reliably control the probability of type I error even with a large sample. Indeed, the theory supporting is developed on assuming the sample size goes to infinity. For this reason, we favor Monte Carlo (MC) test methods. MC tests were introduced by Dwass (1957) and Barnard (1963). An important feature of such procedures is that exactness obtains for a given of number of MC replications, without the need to assume that the latter is large or goes to infinity. Further discussions and extensions are also available in Birnbaum (1974), Foutz (1980), Jöckel (1986), Dufour, Farhat, Gardiol and Khalaf (1998), Dufour and Kiviet (1998), Dufour and Khalaf (2001), and Dufour (2006). To perform exact tests, using this method, we will no longer need to know the distribution of the test statistic.

In this paper, we *first* show how the size of all the two-sample homogeneity tests described above can be perfectly controlled for both *continuous* and *discrete* distributions on considering their permutational distribution and using the technique of MC tests properly adjusted to deal with discrete distributions. As a result, in order to implement these tests, it is not anymore necessary to establish the distributions of the test statistics, either in finite samples or asymptotically.

Second, as a consequence of the great flexibility allowed by the MC test technique in selecting test criteria, we suggest alternative procedures that can provide power gains. These include: (i) a statistic based on the L_∞ distances between kernel-type pdf estimators; (ii) extensions of the permutational test based on the difference of two-sample means to higher order moments, such as sample variances, asymmetry (as third moments) and kurtosis sample coefficients.

Third, on observing that no single test uniformly dominates the others with respect to power, we show that different tests can be combined easily to obtain procedures with better overall power and robustness properties. The procedures proposed involve three steps: (1) in order to make the different statistics comparable, the latter are standardized using first and second moments estimated by simulation; (2) the combined test statistic is defined as the maximum of the standardized test statistics; (3) the MC test technique is used to control the size of a test based on the combined statistic. Depending of the statistics considered different combined tests can be built in this way.

Fourth, we show that the size of these combined tests can also be controlled in finite samples through the use of the MC test technique, which will automatically take account of the dependence between the test statistics as well as the discrete nature of their distribution, with a fixed (possibly very small) number of MC replications. It is of interest to note here that such control would be much more difficult, using standard distributional methods, which typically only yield finite-sample (conservative) bounds or large-sample approximations. Combined test procedures are often on the assumption of independence between the test statistics [see the review of Folks (1984)], which does not hold here, or the use of approximations based on bounds [see Miller (1981), Dufour (1989,

1990), Dufour and Torrès (1998, 2000)] or asymptotic arguments [see Westfall and Young (1993) and Pesarin (2001)]. In contrast, the method we propose here for controlling test size does not depend on the assumption that the number of observations or the number of MC replications go to infinity (as done, for example, in justifying bootstrap techniques).

Fifth, we present the results of a MC experiment which shows clearly that usual large-sample critical values do not control size, while the MC versions of the tests achieve this aim perfectly. Further, we see that the new procedures introduced, either individually or combined with other procedures, can lead to substantial power gains.

Section 2 presents the test statistics studied. In Section 3, we explain how the technique of MC randomized tests can be applied to all these statistics to control the size of the corresponding tests under nonparametric assumptions. In Section 4, we describe the method for combining several tests using simulation-based moments. Section 5 describes the results of our study, first for continuous distributions and then for discrete distributions. We conclude in Section 6.

2. Test statistics

Let $X_1, \ldots, X_n$ be a sample of independent and identically distributed observations with common cdf $F(x) = \mathbb{P}[X_i \leq x]$ and $Y_1, \ldots, Y_m$ a sample *i.i.d.* observations with cdf $G(x) = \mathbb{P}[Y_i \leq x]$. We wish to test the homogeneity hypothesis H_0 in (1.1) and, for that matter, our study will include the following test statistics. In all the tests presented below, H_0 is rejected when the test statistic is large.

The first two criteria are the *KS* and *CM* statistics. The *KS* test was introduced by Smirnov (1939, 1948) and uses the statistic

$$KS = \sup_x |F_n(x) - G_m(x)| \tag{2.1}$$

where $F_n(x)$ and $G_m(x)$ are the usual empirical distribution functions (edf) associated with the X and Y samples respectively. It is well known that *KS* is distribution-free [see Conover (1971, page 313)] under H_0 when the common distribution function F is continuous, but its exact and limiting distributions are not standard [see Massey (1951*a*, 1951*b*, 1952), Drion (1952), Gnedenko (1954), Darling (1957), Hodges (1958), Birnbaum and Hall (1960), Korolyuk (1961), Barton and Mallows (1965), Kim (1969), Steck (1969), Kim and Jennrich (1970) and Gibbons and Chakraborti (1992, Chapter 7)]. In particular, Massey (1952), Birnbaum and Hall (1960), Kim (1969) and Kim and Jennrich (1970) have supplied tables for the distribution. Further, it is important to note that *KS* is not distribution-free when F is a discrete distribution, although the critical values obtained under continuity are conservative for discrete distributions [see Goodman (1954), Noether (1963), Walsh (1963), Hájek and Šidák (1967, Section 8.2)]. Consequently, power losses may occur if the discrete nature of the distribution is not taken into account.

The two-sample *CM* statistic is defined as

$$CM = \frac{mn}{(m+n)^2}\Big\{\sum_{i=1}^{n}\big[F_n(X_i) - G_m(X_i)\big]^2 + \sum_{j=1}^{m}\big[F_n(Y_j) - G_m(Y_j)\big]^2\Big\}. \tag{2.2}$$

CM is also distribution-free under H_0 when F is continuous and, again, the exact and limiting null distributions of *CM* are not standard. Anderson (1962) and Burr (1963, 1964) provide tables for the exact distribution for small sample sizes $(n+m \leq 17)$. Otherwise, a table of the asymptotic distribution is available from Anderson and Darling (1952).

The next three statistics are based on distances (L_1, L_2 and L_∞) between kernel-based pdf estimators. If f is the pdf associated with the cdf F, Allen (1997) considered the following kernel-type density estimators:

$$f_n(x) = \frac{C_X}{n} \sum_{i=1}^{n} K[C_X(x - X_i)/h], \quad f_n(x) = \frac{C_Y}{n} \sum_{i=1}^{m} K[C_Y(x - Y_i)/h] \tag{2.3}$$

where $h = 1$,

$$C_X = n^{1/5}/(2s_X), \quad C_Y = n^{1/5}/(2s_Y), \quad K(x) = \begin{cases} \frac{1}{2}, & \text{if } |x| \leq 1, \\ 0, & \text{if } |x| > 1, \end{cases} \tag{2.4}$$

and $s_X = \left[\sum_{i=1}^{n} (X_i - \bar{X})^2/(n-1)\right]^{1/2}$ is the usual estimator of the population standard deviation [if $s_X = 0$, we set $C_X = 1$, so $f_n(x)$ simply becomes the frequency of x]. The bandwidth could, of course, be reduced $(0 < h \leq 1)$ or increased $(h > 1)$. If g is the pdf associated with the cdf G, its estimator $g_m(x)$ is defined in a way analogous to (2.3). The L_1-distance test initially proposed by Allen (1997) is based on the statistic

$$\hat{L}_1 = \sum_{i=1}^{n} |f_n(X_i) - g_m(X_i)| + \sum_{j=1}^{m} |f_n(Y_j) - g_m(Y_j)| \,. \tag{2.5}$$

Other possible criteria of this type [not considered by Allen (1997)] include the L_2 and L_∞-distance tests, which are based on the statistics:

$$\hat{L}_2 = \left\{ \sum_{i=1}^{n} \left[f_n(X_i) - g_m(X_i)\right]^2 + \sum_{j=1}^{m} \left[f_n(Y_j) - g_m(Y_j)\right]^2 \right\}^{1/2} \tag{2.6}$$

$$\hat{L}_\infty = \sup_x |f_n - g_m| = \max_{1 \leq i \leq n,\, 1 \leq j \leq m} \left\{ |f_n(X_i) - g_m(X_i)|, \, |f_n(Y_j) - g_m(Y_j)| \right\}. \tag{2.7}$$

When the distribution function F is continuous, the *KS* and *CM* statistics are distribution-free under the null hypothesis, but this is not the case (at least in finite samples) for the statistics $\hat{L}_1, \hat{L}_2$ and $\hat{L}_\infty$. When F and G are discrete, the pdf f and g are not well defined and may have to be replaced by mass functions. However, the $\hat{L}_1, \hat{L}_2$ and $\hat{L}_\infty$ statistics remain well defined and may still be used as test statistics. The main remaining problem consists in controlling the size of such tests (which will be done below). When F is discrete, none of the above statistics is distribution-free.

The next statistic to enter our study is the difference of the sample means

$$\hat{\theta}_1 = \bar{X} - \bar{Y} \,. \tag{2.8}$$

Permutation tests based on $\hat{\theta}_1$ were initially proposed by Fisher (1935) and used by Dwass (1957)

for testing the equality of means, but Efron and Tibshirani (1993, Chapter 15) suggested to extend their use, along with bootstrap tests, for testing the equality of two unknown distributions. Further, we suggest here alternative test statistics based on comparing higher-order moments. Namely, the difference between unbiased estimators of sample variances

$$\hat{\theta}_2 = \left| \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 - \frac{1}{m-1}\sum_{i=1}^{m}(Y_i-\bar{Y})^2 \right| \tag{2.9}$$

as well as statistics based on comparing sample skewness and kurtosis coefficients:

$$\hat{\theta}_3 = \left| \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)^3 - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{Y_i-\bar{Y}}{s_Y}\right)^3 \right|, \tag{2.10}$$

$$\hat{\theta}_4 = \left| \frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{s_X}\right)^4 - \frac{1}{m}\sum_{i=1}^{m}\left(\frac{Y_i-\bar{Y}}{s_Y}\right)^4 \right|, \tag{2.11}$$

where

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2 \quad \text{and} \quad s_Y^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i-\bar{Y})^2.$$

By convention, if $s_X = 0$, we set $(X_i-\bar{X})/s_X = 0$ for all i, because in such a case we have $X_1 = \cdots = X_n$, and similarly for Y if $s_Y = 0$. Skewness and kurtosis coefficients play a central role in testing normality [see Jarque and Bera (1987) and Dufour et al. (1998)]. Note also that permutation tests based on statistics of the form $\hat{\theta}_1, \ldots, \hat{\theta}_4$ are valid even if even if no division by a "standard error" is made.

3. Exact Monte Carlo permutation tests

Except for the Dwass (1957) procedure, all the tests described in the previous section involve imperfectly tabulated null distributions or are not distribution-free in finite samples. Consequently, the latter may lead to arbitrarily large size distortions. In view of obtaining distribution-free tests with known size in finite samples, we first note that truly distribution-free tests (for any given sample size) can be based on the statistics KS, CM, L_1, L_2, L_∞, $\hat{t}$, $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$ and $\hat{\theta}_4$ by considering the distribution obtained on permuting in all possible ways (with equal probabilities) the $m+n$ grouped observations $X_1, \ldots, X_n, Y_1, \ldots, Y_m$. Since these permutations are equally probable under the null hypothesis H_0, irrespective of the unknown distribution F, any test which rejects H_0 by using an exact critical value obtained from its permutational distribution [*i.e.*, its conditional distribution given the ordered statistics of the grouped observations] will have the same level conditionally (on the ordered statistics) as well as unconditionally. This property also holds under the weaker assumption where $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ are exchangeable.

If T designates a pivotal test statistic (*i.e.* its distribution does not depend on unknown parameters under the null hypothesis), we can proceed as follows to conduct a MC test. Denote by $T^{(0)}$ the test statistic computed from the observed sample. When the null hypothesis is re-

jected for large values of $T^{(0)}$, the associated critical region of size α may be expressed as $G(T^{(0)}) \leq \alpha$, where $G(x) = \mathbb{P}[T \geq x | H_0]$ is the p-value function. Generate N independent samples $(X_1^{(i)}, \ldots, X_n^{(i)}, Y_1^{(i)}, \ldots, Y_m^{(i)})$, $i = 1, , \ldots, N$, drawn from the specified null distribution F_0. This leads to N independent realizations $T^{(i)} = T\left(X_1^{(i)}, \ldots, X_n^{(i)}, Y_1^{(i)}, \ldots, Y_m^{(i)}\right)$, $i = 1, , \ldots, N$, from which we can compute an empirical p-value function:

$$\hat{p}_N(x) = \frac{N\hat{G}_N(x) + 1}{N+1} \tag{3.1}$$

where

$$\hat{G}_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(T^{(i)} \geq x), \quad \mathbf{1}[\mathbf{A}] = \begin{cases} 1, & \text{if } A \text{ holds} \\ 0, & \text{otherwise} \end{cases}.$$

The associated MC critical region is defined as

$$\hat{p}_N(T^{(0)}) \leq \alpha\,, \tag{3.2}$$

where $\hat{p}_N(T^{(0)})$ may be interpreted as an estimate of $G(T^{(0)})$. If ties among $T^{(i)}$, $i,\ j = 0, 1, \ldots, N$, have zero probability under H_0, *i.e.*

$$\mathbb{P}\left[T^{(i)} \neq T^{(j)}\right] = 0\,, \quad i,\ j = 0, 1, \ldots, N\,, \tag{3.3}$$

we have:

$$\mathbb{P}\left[\hat{p}_N(T^{(0)}) \leq \alpha\right] = \frac{I[\alpha(N+1)]}{N+1}\,,\ 0 < \alpha < 1\,, \tag{3.4}$$

under H_0, where $I[x]$ denotes the largest integer not exceeding x; see Dufour (2006). In particular, if N is chosen such that $\alpha(N+1)$ is an integer, the critical region (3.2) has size α, and thus yields an exact similar test. More generally, the level of the test is not larger than α, irrespective of the number N of replications used.

The above procedure is closely related to bootstrapping, with a fundamental difference however. Bootstrap tests are, in general, provably valid as $N \to \infty$. In contrast, we see from (3.4) that N is explicitly taken into consideration in establishing the validity of MC tests. Although the value of N has no incidence on size control, it may have an impact on power which typically increases with N.

Condition (3.1) may easily not hold for permutation distributions (which are discrete) and general statistics which not be continuous functions of the observations. Nevertheless, the technique of MC tests can be adapted to discrete distributions by appeal to the following randomized tie-breaking procedure [see Dufour (2006)]. Draw $N+1$ uniformly distributed variates $U_0, U_1, \ldots, U_N$, independently of the $T^{(i)}$'s and arrange the pairs $(T^{(i)}, U_i)$ following the lexicographic order:

$$(T^{(i)}, U_i) \leq (T^{(j)}, U_j) \Leftrightarrow \left[T^{(i)} < T^{(j)} \quad \text{or} \quad (T^{(i)} = T^{(j)} \text{ and } U_i \leq U_j)\right]. \tag{3.5}$$

Then, proceed as in the continuous case and compute

$$\widetilde{p}_N(x) = \frac{N\tilde{G}(x)+1}{N+1} \tag{3.6}$$

where

$$\begin{aligned} \tilde{G}_N(x) &= \frac{\sum_{i=1}^N \mathbf{1}[(x,U_0) \le (T^{(i)},U_i)]}{N} \\ &= 1-\frac{1}{N}\sum_{i=1}^N \mathbf{1}(T^{(i)} \le x)+\frac{1}{N}\sum_{i=1}^N \mathbf{1}(T^{(i)}=x)\ \mathbf{1}(U_i \le U_0)\,. \end{aligned} \tag{3.7}$$

The resulting critical region $\widetilde{p}_N(T^{(0)}) \le \alpha$ has level α. More precisely, for $0<\alpha<1$, we have: for

$$\mathbb{P}\big[\hat{p}_N(T^{(0)}) \le \alpha\big] \le \mathbb{P}\big[\tilde{p}_N(T^{(0)}) \le \alpha\big] = \frac{I[\alpha(N+1)]}{N+1},\ \ 0 \le \alpha \le 1. \tag{3.8}$$

The inequality in (3.8) follows on observing that $\widetilde{p}_N(x) \le \hat{p}_N(x)$ for any x, so $\hat{p}_N(T^{(0)}) \le \alpha$ entails $\hat{p}_N(T^{(0)}) \le \alpha$, and the critical region $\hat{p}_N(T^{(0)}) \le \alpha$ may be conservative at level α.

The above procedure is closely related to the parametric bootstrap, with a fundamental difference however. Bootstrap tests are, in general, provably valid as $N \to \infty$. In contrast, we see from (3.4) that N is explicitly taken into consideration in establishing the validity of MC tests. Although the value of N has no incidence on size control, it may have an impact on power which typically increases with N.

If a null hypothesis ensures that the random sample is made up of exchangeable variables and if it should be rejected for large values of the test statistic, a MC test of that hypothesis is carried out in four steps:

1. the test statistic is computed with the help of the observed sample which gives a value $T^{(0)}$, say;
2. N permutations of the sample are chosen at random and without replacement from all possible permutations;
3. the test statistic is recomputed for each of the permuted samples which gives the values $T^{(1)},\ldots,T^{(N)}$, say;
4. the null hypothesis is rejected when $\tilde{p}_N(T^{(0)}) \le \alpha$.

The fact that the procedure is randomized plays a central role in controlling the size of the test. In bootstrap-type procedures, one does as if the number of replications were infinite (even though it is not).

4. Monte Carlo combined permutation tests

Once the simulation study based on the above statistics was performed, we noticed that a group of MC randomized tests gave rise to sizable power for a first subset of alternatives but to rather poor power for a second subset. On the other hand, another group of MC randomized tests showed the opposite profile. Moreover, none of the six MC randomized tests maintained a high power against all the alternatives considered. To exploit this fact, we suggest combining statistics having different profiles in the hope of improving the power of the corresponding test over the range of all considered alternatives. Further, through the use of the MC randomized test technique, we will be able to automatically take account of the dependence between the test statistics, hence avoiding the assumption of independence often made in the literature on combining tests [see Folks (1984)] or the use of approximations based on bounds or asymptotic arguments; see Miller (1981), Hochberg and Tamhane (1987), Dufour (1989, 1990), Dufour and Torrès (1998, 2000), Westfall and Young (1993), Pesarin (2001), Dufour and Khalaf (2002), and Pesarin and Salmaso (2010).

To be more specific, we consider here tests based on the maximum of several standardized statistics. The standardization aims at ensuring comparability between the different statistics: an empirical mean is subtracted from each statistic, and the result is divided by an empirical standard error. The empirical mean and standard error are computed from the observed and simulated values of the test statistics. Formally, if $V = (T^{(1)}, \ldots, T_k)'$ denotes a vector of k selected statistics, let $V^{(0)} = (T_1^{(0)}, \ldots, T_k^{(0)})$ be its value based on the original grouped (X,Y)-sample and let $V^{(i)} = (T_1^{(i)}, \ldots, T_k^{(i)})$, $i = 1, \ldots, N$, be the values based on the N random permutations of the $(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$ sample. The standardized statistics are then:

$$\widetilde{T}_j^{(i)} = \frac{T_j^{(i)} - \overline{T}_j}{s_j}, \; j = 1, \ldots, k, \; i = 0, 1, \ldots, N, \tag{4.9}$$

where

$$\overline{T}_j = \frac{1}{N+1} \sum_{i=0}^{N} T_j^{(i)}, \quad s_j = \Big\{ \frac{1}{N} \sum_{i=0}^{N} \big(T_j^{(i)} - \overline{T}_j \big)^2 \Big\}^{1/2}, \; j = 1, \ldots, k. \tag{4.10}$$

For the observed vector of test statistics $V^{(0)}$ and each simulated vector $\big(V^{(i)}, \; i = 1, \ldots, N\big)$, we can then compute the following combined statistics:

$$\hat{Q}\big(V^{(i)}\big) = \max_{1 \le j \le k} \big\{ \widetilde{T}_j^{(i)} \big\}, \; i = 0, 1, \ldots, N. \tag{4.11}$$

The combined test based on the statistic $\hat{Q}$ rejects the null hypothesis when the maximum of the standardized statistics is "large'. In (4.11), $\hat{Q}\big(V^{(0)}\big)$ represent the statistics associated with the "actual sample" (although they also depend on randomly permuted samples thorough the empirical means and standard errors used to standardize the statistics), while $\hat{Q}\big(V^{(i)}\big)$ for $i \neq 0$ can be interpreted as values based on "simulated" (permuted) samples.

It is straightforward to see that the variables

$$\hat{Q}\big(V^{(i)}\big), \; i = 0, 1, \ldots, N, \text{ are exchangeable under } H_0 . \tag{4.12}$$

Consequently, we can write: for $0 < \alpha < 1$,

$$\mathbb{P}\big[\hat{p}_N\big(\hat{Q}^{(0)}\big) \le \alpha\big] \le \mathbb{P}\big[\tilde{p}_N\big(\hat{Q}^{(0)}\big) \le \alpha\big] = \frac{I[\alpha(N+1)]}{N+1}, \tag{4.13}$$

under H_0, where $\hat{Q}^{(i)} \equiv \hat{Q}\big(V^{(i)}\big)$, $i = 0, 1, \ldots, N$, and $\hat{p}_N$, $\tilde{p}_N$ are defined as in (3.1) - (3.6) with T_i replaced by $\hat{Q}^{(i)}$.

Below we consider two special cases of such combined test statistics:

$$\hat{Q}_i \equiv \hat{Q}(V_i), \quad , \; i = 1, 2, 3, \tag{4.14}$$

where

$$V_1 = (KS, \hat{L}_\infty)', \quad V_2 = (\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)'. \quad V_3 = (KS, \hat{L}_\infty, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)'. \tag{4.15}$$

The first choice (V_1) emphasizes two overall distance measures between the two empirical distributions (KS and $\hat{L}_\infty$), (V_2) is based on the maximum distance between the four statistical moments, while the third one (V_3) also uses the differences between the first four moments of the two distributions, and should thus provide more sensitivity to differences that affect the first four moments. We will see below that no individual test has the best power against all the alternatives considered in this study.

5. Simulation study

In the simulation study, all tests [both the original tests as well as their MC randomized counterparts] were performed at the 5% level using 10000 trials. This entails that the 95% confidence interval for the nominal level is $[4.57\%, 5.43\%]$. Furthermore, they were all conducted with equal sample sizes $m = n = 22$. As mentioned earlier, each MC randomized test was carried out by picking at random N = 99 permutations of the original grouped sample and this was done by using a Fortran program based on IMSL Library random number generator. All these results were confirmed using a program developed on the software R (version 3.0.2). In other simulation studies, researchers have made use of the bootstrap where they realize generations reaching 1000 samples. In his simulation study, Allen (1997) used 2500 trials and each permutation or bootstrap test was carried out with 499 samples.

In the first part of the study, F and G are both continuous, and the following distributions were considered: normal $N(0,1)$, exponential $Exp(0, 1.5)$, gamma $\Gamma(2, 1)$, beta $B(2, 3)$, logistic $Log(-1, 1)$, lognormal $\Lambda(-1, 1.2)$ and uniform $U(0, 1)$. In this choice, care was taken to have at the same time simple parameters as well as appreciably different means and variances. Table 1 gives the list of those means and variances. Four types of setups are considered: (i) the distributions were standardized, and thus had common zero mean and unit variance; (ii) the distributions were only centered, and thus had the zero mean but different variances; (iii) the distributions were only scaled, and thus had different means and common unit variance; (iv) the distributions remained as is and thus had different means and different variances. Whatever the situation, a null hypothesis is

Table 1. Continuous distributions with their means and variances

Distribution	$N(0,1)$	$Exp(0,1.5)$	$\Gamma(2,1)$	$B(2,3)$	$Log(-1,1)$	$\Lambda(4,1.5)$	$U(0,1)$
Mean	0	1.50	2	.40	-1	0.756	.50
Variance	1	2.25	2	.04	0.55133^{-2}	1.84	1/12

Table 2. Empirical level and power for MC randomized permutation tests of equality of two distributions ($m = n = 22$, $\alpha = 0.05$)

	$F = N(0,1)$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
$N(0,1)$	4.6	5.1	5.1	5.5	4.8	4.5	4.7	4.8	4.9	5	5.1	5
$N(0.2,1)$	8.6	9.6	10.3	4.8	5.5	5.2	5.4	5.5	5.5	7.4	7.5	7.4
$N(0.3,1)$	13.2	15	16.2	4.6	5.3	5.2	5.8	5.7	6.2	10.6	10.4	10.2
$N(0.4,1)$	19.8	23.1	25	4.6	5.7	5.6	6.5	6.6	6.6	15.7	14.9	15.1
$N(0.5,1)$	28.5	33.2	37.1	3.9	5.7	5.4	7.2	7.5	7.4	22.4	22.6	22.1
$N(0.7,1)$	48.7	56.3	60.9	3.1	6.4	6.8	9.3	9.5	9.7	40.1	42	41.4
$N(0,1.2^2)$	5.7	5.4	5.2	11.8	5.2	4.9	6.9	7.4	7.5	6.7	7.2	7
$N(0,1.4^2)$	7.4	7	4.9	28.5	4.5	4.9	11	12.9	13	10.7	13.4	12.9
$N(0,1.6^2)$	10.2	9	5.1	49.6	4.2	4.3	18.9	22	20.3	16.5	26.3	23.9
$N(0,1.8^2)$	13	11.8	5.3	67.5	3.4	3.8	28.1	31.7	27.8	23.1	38.7	35.5
$N(0,2.0^2)$	17.8	16.1	5.6	80.5	2.7	3.3	40.2	44.9	37.9	31	53.4	49.4

obtained each time F and G share the same distribution and an alternative hypothesis is obtained each time F and G possess different distributions. This is explained by the fact that a linear transformation of i.i.d observations forming the two samples does not change the equality or the non-equality of the two distributions.

In the second part of the study, F and G are discrete, and five most commonly used distributions are considered: discrete uniform $[DU(n)]$ on the integers $\{1, 2, \ldots, n\}$, binomial $[Bin(n, p)]$, geometric $[Geo(p)]$, negative binomial $[Nbin(N, p)]$ and Poisson $[P(\lambda)]$. Since it is a prohibitive task to find parameters that will simultaneously give rise to either common mean and common variance, the following three situations were considered: (i) the distributions are $DU(19)$, $Bin(20, 0.5)$, $Geo(0.1)$, $Nbin(10, 0.5)$, $P(10)$ and, thus had common mean 10 and variances 33.33, 5, 90, 20 and 10 respectively; (ii) the distributions are $DU(12)$, $Bin(48, 0.5)$, $Geo(0.25)$, $Nbin\big(24, (\sqrt{3}-1)\big)$, $P(12)$, so they have means 6, 24, 4, 8.79 and 12 respectively, and common variance 12; (iii) the distributions are $DU(11)$, $Bin\big(10, 0.1\big)$, $Geo(0.3)$, $Nbin\big(10, 0.2\big)$, $P(5)$, whose means are mean 5.5, 1, 3.33, 40 and 5 respectively, variances 10, 9, 7.78, 200 and 5 respectively.

As a check on the accuracy of our study, Tables 1 and 2 of Allen (1997) were reproduced adding, however, the CM, the $\hat{L}_\infty$ and the combined MC randomized tests and by excluding the bootstrap

Table 3. Some illustrations for empirical level for KS and CM tests of equality of two continuous distributions $(\alpha = 0.05)$

	Original tests						MC randomized tests					
	$n = m = 8$		$n = m = 22$		$n = m = 50$		$n = m = 8$		$n = m = 22$		$n = m = 50$	
$F = G$	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*
N	1.9	4.9	5.0	4.9	2.3	4.9	4.7	4.8	5.0	4.8	5.0	4.7
Exp	1.8	4.7	4.8	4.8	2.2	5.0	4.8	4.5	4.9	4.8	4.7	5.2
Gam	1.9	5.3	4.9	5.0	2.1	5.1	5.3	5.0	5.0	5.1	4.9	5.2
B	1.8	5.1	5.0	5.1	2.3	4.7	5.1	5.1	4.9	5.1	4.6	4.6
Log	1.8	5.0	5.3	5.4	2.2	5.3	5.1	5.2	5.4	5.3	5.4	5.3
Ln	1.9	5.2	5.3	5.5	2.3	5.1	4.9	4.9	5.4	5.1	5.2	5.1
U	1.9	5.0	5.1	5.2	2.4	4.9	4.7	4.9	5.4	5.0	5.0	5.1

Table 4. Some illustrations for empirical level for KS and CM tests of equality of two discrete distributions $(\alpha = 0.05)$

	Original tests						MC randomized tests					
	$n = m = 8$		$n = m = 22$		$n = m = 50$		$n = m = 8$		$n = m = 22$		$n = m = 50$	
$F = G$	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*	*KS*	*CM*
UD	1.0	5.5	6.0	5.6	1.9	5.3	5.1	4.9	4.8	4.8	4.8	4.7
Bin	0.4	11.1	2.7	23.0	0.6	55.4	4.8	4.5	4.9	4.9	4.9	4.8
Geo	0.9	7.7	5.1	12.1	1.2	28.0	5.0	5.1	4.6	4.9	5.0	5.3
BinN	0.7	7.0	4.3	7.8	1.2	9.4	5.0	5.2	4.7	4.6	5.3	4.9
Poi	0.8	6.6	4.8	6.1	1.3	5.3	4.6	5.2	4.6	4.7	4.4	4.3

tests. The results appear in Table 2 and they are quite similar to those of Allen (1997).

Most statistics described in the preceding sections have not been tabulated, so a study of the reliability of tabulated critical values can only be limited. In Tables 3 and 4, we present some results on this issue for the *KS* and *CM* tests. For continuous distributions, we see that the standard *KS* and *CM* tests satisfy the level constraint, although the rejection frequencies of the *KS* test are in some cases notably lower than the level. This can be explained by the fact the 0.05 level cannot be achieved by a non-randomized procedure (due to the discrete character of the distribution), so that the critical values used correspond to smaller sizes. In the case of discrete distributions, it is of interest to note that the *KS* test can be quite conservative (as predicted by theory), while the *CM* test can substantially overreject: the *CM* test is not generally conservative for discrete distributions. In all cases, irrespective of whether the distributions are continuous or discrete, the permutational MC randomized tests have rejection frequencies essentially identical to their nominal levels (as expected).

Tables 5 to 8 contain the results of our study for the case where both F and G are continuous. The following conclusions can be drawn. First, it is clear the test based on $\hat{\theta}_1$ has little power for detecting distributions that differ through other characteristics than their mean. Two distributions cannot be equal if they do not have the same mean but the converse is not true. Consequently, if the test based on $\hat{\theta}_1$ accepts the hypothesis H_0, it should not be interpreted as an acceptance of the fact that $F = G$ but rather that these distributions have equal means.

Second, $\hat{\theta}_2$ has the best power for testing the Gaussian distribution against most of the other distributions considered, but it does not perform as well in the other cases.

Third, the $\hat{L}_1$ and $\hat{L}_2$ tests behave almost identically and differ slightly from the $\hat{L}_\infty$ test. In the same way, the power of the *KS* test is not very different from that of the *CM* test.

Fourth, if we compare the powers of the tests based on edf's (*KS* and *CM*) with those based on pdf estimates ($\hat{L}_1, \hat{L}_2$ and $\hat{L}_\infty$), we notice some large power differences, but one cannot conclude that a test from one group is more powerful than all the tests in the other group. The edf tests are more powerful than those based on pdf estimates when two distributions have the same variance but different means (see Tables 2 and 6). On the other hand, if the two distributions have the same mean but different variances, the tests based on pdf estimates are the most powerful (see Tables 2 and 7).

Fifth, the combined Monte Carlo randomized tests exhibit a robust performance in the sense that their power is either the best or is only slightly lower than the one of any other test. There is no uniform dominance between the two tests based on the smaller set of statistics ($\hat{Q}_1$ and $\hat{Q}_2$) and the one based on the larger one ($\hat{Q}_3$). Not surprisingly, $\hat{Q}_3$ tends to perform better than $\hat{Q}_1$ and $\hat{Q}_2$ as it contains all statistics forming the last two.

Finally, consider the case where both F and G are discrete (Tables 9 - 11). From a qualitative viewpoint, the conclusions that emerge from these are quite similar to those reached in the continuous case: size is perfectly controlled by the MC randomized test technique, the powers of different tests can differ widely depending on the case considered, no test procedure uniformly dominates the others, and the combined test procedures exhibit a good robust overall performance.

6. Conclusion

In this paper, we first showed that finite-sample distribution-free two-sample homogeneity tests, for both continuous and discrete distributions, can be easily obtained on combining two techniques: (1) by considering permutational versions of most proposed tests for that problem; (2) by implementing the permutation procedures as Monte Carlo randomized tests with an appropriate tie-breaking technique to take account of the discreteness of the test null distributions. Second, due to the flexibility of the Monte Carlo test technique, we could easily introduce and implement several alternative procedures, including permutation tests comparing higher-order moments and procedures based on combining several test statistics. Third, in a simulation study, it was shown that the procedures proposed work as expected form the viewpoint of size control, while the new test statistics suggested yield power gains.

Table 5. Empirical level and power for MC randomized permutation tests of equality of two continuous distributions having same mean and same variance ($m = n = 22$ and $\alpha = 0.05$)

	$F = N$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
N	5.0	4.8	4.7	5.1	4.8	5.0	5.0	5.0	4.8	5.1	4.8	5.3
Exp	13.8	12.4	5.6	10.5	42.0	15.8	17.3	17.1	15.6	17.1	24.2	25.5
Γ	8.9	8.8	5.3	7.7	27.0	9.5	11.1	11.0	10.4	10.5	14.2	14.4
B	5.4	4.9	4.9	5.4	7.5	7.0	5.7	5.9	5.8	5.7	6.4	6.9
Log	5.7	5.3	5.6	5.1	5.4	6.3	5.3	5.2	5.5	5.4	5.7	6
Λ	71.8	65.2	5.7	59.0	70.2	62.7	68.6	67.6	65.7	79.4	86.2	88.9
U	6.7	5.7	4.9	6.0	6.4	16.6	6.1	6.5	6.9	6.6	11.1	11.9
	$F = Exp$											
Exp	5.3	5.3	5.5	5.2	5	5.1	5.2	5.2	5.2	5.2	5.4	5.3
Γ	5.8	5.6	4.9	6.1	8	6.4	8.8	9.1	9.3	8.3	6.5	7.4
B	10.7	9.4	5	12.6	34.9	19.8	37.4	37.5	35.2	28.6	26.2	28.5
Log	13.2	12.2	5.3	8.7	35.1	10.6	38.1	39	35.1	26.9	18.6	23.2
Λ	50	34.4	5	28.8	20.4	19.4	29.9	39.7	46.6	60.7	26.1	52.1
U	16.1	13.1	5	17.3	55.2	35	56.4	55.2	48	41	44	44.8
	$F = \Gamma$											
Γ	5.2	5.2	5.2	5.1	5	5.1	4.8	4.9	4.8	4.8	5.3	5.1
B	7	6.7	5.1	8.7	19.7	12.3	18.6	17.8	15.3	12.1	14.3	13.9
Log	9.1	8.3	5.1	6.1	22.8	7.1	22.8	22.7	18.7	14.1	11.5	12.3
Λ	45.6	31.5	5	34.2	33.4	29.6	45.7	56.4	61.2	66.7	41	61.3
U	10.5	9.1	5	12.6	35.8	25.2	33.7	31.9	25.1	20.5	28.8	28.1
	$F = B$											
B	5.1	5.3	5.4	5	5	5.1	5.1	5.2	4.8	5	5	4.9
Log	6.2	5.8	4.9	6.6	8.1	10.9	8.8	8.8	8.1	7.6	9.7	9.3
Λ	49.2	40.5	5.5	45.9	67.9	56.9	78.8	82.6	84.6	83.9	75.3	83.7
U	6	5.6	5.1	5.4	8.3	10	7.8	8	7.7	6.9	8.5	8.2
	$F = Log$											
Log	4.7	4.8	4.8	4.8	5.2	5.5	5	5	4.8	5.1	5.5	5.1
Λ	38.7	33.4	5.8	36.7	56.6	40.1	70.2	77	79.8	74.3	57.5	71.4
U	7.8	6.2	5	8	8.9	25.2	10.5	11.2	11.9	10.9	17.4	16.6
	$F = \Lambda$											
Λ	4.8	5	4.9	5.1	5	5	4.8	4.9	4.7	4.6	4.9	4.5
U	57.4	50	6	49.8	86.3	74.5	89.4	90.1	90.8	90.7	88.5	92.2

Table 6. Empirical level and power for MC randomized permutation tests of equality of two continuous distributions having different means but same variance ($m = n = 22$ and $\alpha = 0.05$)

	$F = N$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
N	5.1	4.8	4.8	4.8	5	4.8	5	5.1	5	5.2	4.9	5
Exp	89.3	88.3	92.1	5.1	19.1	13.2	34.8	31.8	23.6	84.7	84.4	86.3
Γ	98.7	99.6	99.9	1.4	14.2	10.6	36.8	34.3	24.4	97.5	99.2	99.1
B	100	100	100	0.1	7.4	12.6	31.5	32.1	30.3	99.9	100	100
Log	35.8	41	42.6	4.4	7.2	7.9	9.4	9.7	9.5	28.5	28.7	28.8
Λ	86.5	70.9	49.6	35.4	55.5	49.1	71.2	74.3	72.2	90.5	76	88.5
U	99.5	99.9	100	0.2	6.8	18.4	21.7	22	21.4	98.8	99.8	99.7
	$F = Exp$											
Exp	4.9	4.8	4.9	5	4.9	5	5	4.9	5	4.9	4.8	5.1
Γ	37.3	42.4	29.9	4.6	10.9	9.5	15.6	16.6	16.7	31.4	22.2	27.4
B	90.6	94	86.2	4.6	50.5	40.4	67.7	68.9	67.3	87.4	81.2	85.9
Log	100	100	100	1.9	35.4	18.2	69.3	68.4	52.7	100	99.8	99.9
Λ	54	58	43.6	13.1	23.1	20.2	36.5	44.7	47.3	52.7	42.9	50.8
U	67.3	72.8	64.7	11.8	64.5	50.2	70.3	70.2	65.8	68.6	69.1	70.2
	$F = \Gamma$											
Γ	5.4	5.4	5.4	5	5.2	5.2	5.2	5	5	5.2	5	4.9
B	47.7	53.8	48.1	5.5	28.5	20.6	32.1	32	28.4	43	40.3	41.9
Log	100	100	100	0.5	31	20.2	70.4	72.4	65.8	100	100	100
Λ	97.4	98.7	83.3	7.5	40.8	36.4	77.4	84.7	87.8	96.6	84.5	95.6
U	24.7	25.1	19.9	11.4	41.3	31.1	40.6	39.4	32.7	30.1	35.9	35.1
	$F = B$											
B	5	5.3	5	4.9	5.1	5	5.3	5.2	5	5.1	5.2	5.4
Log	100	100	100	0	16.8	34.3	64.8	67.4	67.7	100	100	100
Λ	100	100	96.8	10.8	79	69.7	99.1	99.4	99.4	100	98.8	100
U	8.6	10.3	12.9	4.7	7	8.3	6.6	6.5	6	8.1	9.6	9.3
	$F = Log$											
Log	4.6	4.8	5	5	5.1	4.8	4.6	4.5	4.7	4.6	4.8	4.8
Λ	100	99.8	98.5	20.4	59.5	49	88.6	87.8	81.2	99.9	98.7	99.9
U	100	100	100	0	14	41.9	53.5	56.2	53.9	100	100	100
	$F = \Lambda$											
Λ	4.9	4.5	4.7	4.8	5	5	4.8	4.9	5	4.8	4.8	4.9
U	99.1	99.5	92.9	33.6	87.5	75.7	98.6	98.7	98.5	99.1	98.3	99.2

Table 7. Empirical level and power for MC randomized permutation tests of equality of two continuous distributions having same mean but different variances ($m = n = 22$ and $\alpha = 0.05$)

	$F = N$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
N	5.2	5.2	5	5.3	4.9	5.1	5	4.9	4.9	4.7	5.1	5
Exp	19.2	18	6	18.2	38.1	6.2	49.6	49.3	39.6	31.3	21.4	26.1
Γ	13.2	12.4	5.8	19.3	22.6	5	31.3	31.6	26.1	20.5	14.6	16.8
B	74.7	76	5.6	100	0.4	0.2	99.9	99.8	93.1	90.1	97.6	96.9
Log	10.2	9.2	5.2	56.4	2.6	2.4	19.1	23	20.8	16.6	27.1	24.4
Λ	33	29.3	6.9	23.8	67	39.8	68.3	72	74.3	68	54.6	65
U	52.8	48.9	5.5	99.9	0.3	0	98.5	97.9	80.6	73.8	94.3	91.5
	$F = Exp$											
Exp	5.1	5	5.4	5.4	5.2	5.3	5	5	5.2	5.1	5.5	5.3
Γ	6	6	5	5.2	7.8	6.1	8.7	8.6	9	7.9	6.7	7.5
B	95.9	96.5	9.1	100	1.6	0.6	100	100	98.8	99.2	90.8	99.2
Log	14.1	12.8	5.2	18.3	35.4	15	41.3	45	43.2	34.1	24.4	30.3
Λ	60.9	44.7	5.4	35.5	21.4	21.1	35.8	47.1	52.7	71.4	31.8	63.9
U	90.3	88.8	9	99.9	6.7	1.2	99.9	99.9	96.2	97.2	89.4	97.6
	$F = \Gamma$											
Γ	4.7	4.6	4.8	5	4.9	4.9	5	5.1	5	4.8	5	5
B	93.7	93.8	7.1	100	0.4	0.2	100	100	98.5	98.5	94.5	98.8
Log	9.1	8.8	5.2	17.9	22.8	9.8	26	28.1	26.1	19.2	16.8	18.4
Λ	49.2	34.3	5.1	36.9	32.6	29.5	47.7	58.6	63.7	69.7	42.6	64.1
U	83.2	81.3	7	100	1.6	0.3	100	99.9	94.5	94.6	93.4	96.8
	$F = B$											
B	4.8	5	5	5.2	5.4	4.8	5.2	5.1	5	5	4.9	5.1
Log	91.6	93.4	5.2	100	0.3	0	100	100	99.3	98.6	94.8	98.6
Λ	95.5	95.5	16.6	96.4	15.5	7.8	97.8	98.1	92.8	97.8	76.2	97.3
U	12.5	11.3	5.3	57.2	9.3	11.4	26.1	29.5	27	23.2	38	36.9
	$F = Log$											
Log	4.9	5	5.1	5	4.8	4.5	5.2	5	5.2	5	4.7	4.9
Λ	51.1	43.8	5.9	56.2	54.3	43.5	81	87	87	84.8	71.1	83.2
U	83.2	81.1	6	100	0.1	0	100	100	96.2	94.7	94.2	96.1
	$F = \Lambda$											
Λ	5	5.2	4.9	5.1	5.1	5.3	5.3	4.9	5.2	5.1	4.9	5.1
U	77.1	80	14.8	86	48.3	13.8	93.1	92.6	74.6	83.1	71.4	83.3

Table 8. Empirical level and power for MC randomized permutation tests of equality of two continuous distributions having different means and different variances ($m = n = 22$ and $\alpha = 0.05$)

	$F = N$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
N	5	4.9	4.9	5.1	5	5.4	5.1	5.2	5.6	5.3	5.3	5.2
Exp	94.9	97.2	99.2	0.4	16.4	4.4	39.8	31.8	16.4	91.4	97	96.4
Γ	99.8	100	100	0	18.4	5.4	53.5	52	37.9	99.6	100	100
B	92.4	90.6	43.3	99.8	0.8	0.1	100	100	95.1	96.5	97.4	98.2
Log	63.3	67.5	59.9	46.8	4.6	3.3	33.2	37.3	33	59.1	57.2	58.8
Λ	88.1	74.7	68	17.6	43.4	41.1	55.7	54.5	52.4	87.1	74.2	84.7
U	89.3	85.1	58.5	98.4	0.8	0.3	99.4	99.2	87	92.4	95.5	95.9
	$F = Exp$											
Exp	5	4.8	5.4	5.2	5	5	4.8	4.9	4.8	4.8	5.1	4.8
Γ	31.1	35.1	22.9	5.8	9.6	8.3	13.8	14.3	14.6	26.2	17.2	22.4
B	97.6	96.2	99.5	81.6	12.1	0.2	99.9	100	98.1	98.7	98.6	98.7
Log	100	100	99.8	5.7	38.8	22.5	73.5	75.3	61.9	99.9	99.5	99.9
Λ	61	65.8	52.3	14.9	24.3	21.4	43.3	51.8	54.6	60.9	50.8	58.6
U	90.6	84.8	96.6	81.7	22.2	0.3	99.5	99.5	92	92.9	93.4	93.8
	$F = \Gamma$											
Γ	5.2	5.2	4.8	5.1	4.9	4.7	4.7	4.7	4.6	5	4.9	4.8
B	100	100	100	46.9	6.1	0	100	100	99.9	100	100	100
Log	100	100	100	3	31.3	21.8	75	78.1	72	100	100	100
Λ	98	99.1	84.9	8.3	41.1	37	78.8	86.1	89.2	97.1	85.8	96.5
U	100	100	100	47.1	12.9	0.1	100	100	98.6	100	100	100
	$F = B$											
B	5	4.7	4.7	5.1	5	5	5.1	5	5.1	4.9	5	4.9
Log	100	99.9	92.8	96.9	2.4	0.7	100	100	99.9	100	99.3	100
Λ	21.5	17.1	26.3	84.8	32.3	4	80	80.9	56	44.3	42	43.5
U	25.8	25.8	24.1	52.7	9.6	9	27.3	31.3	30	30.5	39.3	37.8
	$F = Log$											
Log	5	4.8	5	5	5.4	5.1	5.2	5.2	5.3	5.1	5	5
Λ	99.9	99.6	96.6	38.3	65	50.5	93.2	93.1	89.2	99.9	98.4	99.8
U	99.9	99.8	95.1	95.1	1.5	1.4	100	100	99.7	100	99.4	99.9
	$F = \Lambda$											
Λ	5.5	5.8	5.3	5.3	5.2	5	5.4	5.4	5.8	5.7	5	5.4
U	15.6	12.7	9.3	64.3	53.2	7.3	75.3	70.4	42.3	32.3	46	37.6

Table 9. Empirical level and power for MC randomized permutation tests of equality of two discrete distributions having same mean but different variances ($m = n = 22$ and $\alpha = 0.05$)

	$F = UD$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
UD	5.2	5.4	5.3	5	5.1	5	5.2	5.1	5	5	4.9	5
Bin	57.1	66.6	5.3	99.6	6.7	12.6	92.8	94	85.4	80.5	97	96.4
Geo	23.7	25.7	11	32.4	38.8	13	56.8	55.4	43.2	36.1	28.8	31.2
$BinN$	13.5	11.2	4.8	35.7	18.5	27.8	26.9	29.2	27	23.1	36.1	34.9
Poi	30.7	30.9	5.2	88.2	11.8	22.6	62.7	67.5	59.3	52.9	75.9	74.3
	$F = Bin$											
Bin	5.3	5	5.4	4.8	5.3	5	4.8	4.8	5	4.8	4.9	5
Geo	91.7	94.6	15.3	99.4	14.4	2.2	98.7	99.2	95.5	97	91.5	97.7
$BinN$	18.6	21.4	5.5	79	4.9	3.2	41.8	47.1	39.4	32.6	49.6	45.7
Poi	7.4	8	5.2	29.2	4.8	4.8	11.9	13.3	13.2	10.6	14.3	13.4
	$F = Geo$											
Geo	5.2	5.3	5.3	4.6	4.9	5.3	5.2	5.2	5.3	5.6	5.2	5.3
$BinN$	53.4	51.2	12.1	67	10.1	5.4	53.3	58.9	52.9	60.3	39.7	54.9
Poi	77.6	79.4	13.8	94.8	12.5	3.4	87	90.1	82	86.5	76.3	86.3
	$F = BinN$											
$BinN$	5.3	5.2	5	4.8	4.8	5.1	4.7	4.7	5	5.2	5.1	4.8
Poi	9	8.3	5.4	27.5	4.7	4.7	11.4	13	13	11.6	13.1	12.8

Table 10. Empirical level and power for MC randomized permutation tests of equality of two discrete distributions having different means and same variance ($m = n = 22$ and $\alpha = 0.05$)

	$F = UD$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
UD	5	4.9	5.2	4.9	5.1	5.2	4.7	4.7	5	4.9	4.8	4.8
Bin	100	100	100	0	6.3	72.2	90.7	91.9	93.6	100	100	100
Geo	73.9	84.8	77.1	10.6	67.2	52.4	76.1	75.9	71.9	75	77.1	77.8
$BinN$	50.2	62	72	2.7	4	8.9	6.1	5.5	5.8	40.2	53.1	49.9
Poi	99.6	99.9	100	0.2	4.7	13	17.8	16.8	15.6	98.7	99.9	99.8
	$F = Bin$											
Bin	4.7	4.9	5.2	5	5	4.7	4.9	4.9	4.7	4.6	4.8	4.8
Geo	100	100	100	0.1	59.9	87.2	99.8	99.8	99.9	100	100	100
$BinN$	100	100	100	0	17	57.2	88.1	89.6	90.9	100	100	100
Poi	100	100	100	0	14.2	40.5	77.4	79.8	81.4	100	100	100
	$F = Geo$											
Geo	4.7	4.7	5.1	4.8	5	5.2	4.9	5	5.1	5.1	5.1	5
$BinN$	99.9	100	99.6	3.2	50	47.6	82.7	84.6	84.5	99.8	99.2	99.8
Poi	100	100	100	1.8	58.8	62.5	94.7	95	95.1	100	100	100
	$F = BinN$											
$BinN$	5.1	5	5.2	5.1	5.1	5.4	5.1	5.3	5.3	5.1	5.2	5.4
Poi	74.4	82.2	83.8	2.6	9.8	10.4	15.3	15.3	14.9	65.9	68	67.6

Table 11. Empirical level and power for MC randomized permutation tests of equality of two discrete distributions having different means and different variances ($m = n = 22$ and $\alpha = 0.05$)

	$F = UD$											
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\hat{Q}_1$	$\hat{Q}_2$	$\hat{Q}_3$
UD	5	5.4	5.5	5.1	5	4.8	4.8	4.8	4.7	4.5	5.2	5
Bin	100	100	100	80.6	18.4	8.6	100	99.9	96.6	100	100	100
Geo	82.7	92.6	89.5	14.7	69.4	56.2	82.4	82.8	80.6	84.1	88.1	87.7
$BinN$	100	100	100	0.7	10.9	11.6	100	100	100	100	100	100
Poi	21.3	20.1	9.7	51.2	15.3	25.3	32.5	36.3	33.9	30.3	43.9	42.8
	$F = Bin$											
Bin	5.1	4.9	5.1	5.1	5	5.2	4.8	5	4.8	5	5	5.1
Geo	45	54.2	58.1	69.4	3.2	0.6	59.9	61.1	36.4	42.8	46	42.7
$BinN$	100	100	100	0.3	9.2	21.5	100	100	100	100	100	100
Poi	100	100	100	7.6	14.2	10.7	99.5	99.6	98.8	100	100	100
	$F = Geo$											
Geo	4.9	5.1	5.2	5.1	5	5.1	5.4	5.4	5.2	5.2	4.9	5.1
$BinN$	100	100	100	0.2	39.1	46.2	100	100	100	100	100	100
Poi	96.8	97.6	88.4	13.1	47	36.8	71.4	72.7	70.9	95.2	84.7	93.3
	$F = BinN$											
$BinN$	5.3	5.3	5	5.4	5.3	5	5.5	5.5	5.2	5.4	5.1	5.3
Poi	100	100	100	1.1	8.2	11.7	100	100	100	100	100	100

References

Allen, D. L. (1997), 'Hypothesis testing using an L_1-distance bootstrap', *The American Statistician* **51**, 145–150.

Anderson, T. W. (1962), 'On the distribution of the two-sample Cramér-von Mises criterion', *Annals of Mathematical Statistics* **33**, 1148–1159.

Anderson, T. W. and Darling, D. A. (1952), 'Asymptotic theory of certain 'goodness of fit' criteria based on processes', *Annals of Mathematical Statistics* **23**, 193–212.

Barnard, G. A. (1963), 'Comment on 'The spectral analysis of point processes' by M. S. Bartlett', *Journal of the Royal Statistical Society, Series B* **25**, 294.

Barton, D. E. and Mallows, C. L. (1965), 'Some aspects of the random sequence', *Annals of Mathematical Statistics* **36**, 236–260.

Birnbaum, Z. W. (1974), Computers and unconventional test-statistics, *in* F. Proschan and R. J. Serfling, eds, 'Reliability and Biometry', SIAM, Philadelphia, PA, pp. 441–458.

Birnbaum, Z. W. and Hall, R. A. (1960), 'Small sample distributions for multi-sample statistics of the Smirnov type', *Annals of Mathematical Statistics* **31**, 710–720.

Burr, E. J. (1963), 'Distribution of the two-sample Cramér-von Mises criterion for small equal samples', *Annals of Mathematical Statistics* **34**, 95–101.

Burr, E. J. (1964), 'Small samples distributions of the two-sample Cramér-von Mises' W^2 and Watson's U^2', *Annals of Mathematical Statistics* **35**, 1091–1098.

Conover, W. J. (1971), *Practical Nonparametric Statistics*, John Wiley & Sons, New York.

Darling, D. A. (1957), 'The Kolmogorov-Smirnov, Cramér-von Mises tests', *Annals of Mathematical Statistics* **28**, 223–238.

Davison, A. and Hinkley, D. (1997), *Bootstrap Methods and Their Application*, Cambridge University Press, Cambridge (UK).

Drion, E. F. (1952), 'Some distribution free tests for the difference between two empirical cumulative distributions', *Annals of Mathematical Statistics* **23**, 563–564.

Dufour, J.-M. (1989), 'Nonlinear hypotheses, inequality restrictions, and non-nested hypotheses: Exact simultaneous tests in linear regressions', *Econometrica* **57**, 335–355.

Dufour, J.-M. (1990), 'Exact tests and confidence sets in linear regressions with autocorrelated errors', *Econometrica* **58**, 475–494.

Dufour, J.-M. (2006), 'Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics', *Journal of Econometrics* **133**, 443–477.

Dufour, J.-M., Farhat, A., Gardiol, L. and Khalaf, L. (1998), 'Simulation-based finite sample normality tests in linear regressions', *The Econometrics Journal* **1**, 154–173.

Dufour, J.-M. and Hallin, M. (1990), 'An exponential bound for the permutational distribution of a first-order autocorrelation coefficient', *Statistique et analyse des données* **15**, 45–56.

Dufour, J.-M. and Hallin, M. (1991), 'Nonuniform bounds for nonparametric *t* tests', *Econometric Theory* **7**, 253–263.

Dufour, J.-M. and Hallin, M. (1992), 'Improved Berry-Esseen-Chebyshev bounds with statistical applications', *Econometric Theory* **8**, 223–240.

Dufour, J.-M. and Hallin, M. (1993), 'Improved Eaton bounds for linear combinations of bounded random variables, with statistical applications', *Journal of the American Statistical Association* **88**, 1026–1033.

Dufour, J.-M. and Khalaf, L. (2001), Monte Carlo test methods in econometrics, *in* B. Baltagi, ed., 'Companion to Theoretical Econometrics', Blackwell Companions to Contemporary Economics, Basil Blackwell, Oxford, U.K., chapter 23, pp. 494–519.

Dufour, J.-M. and Khalaf, L. (2002), 'Exact tests for contemporaneous correlation of disturbances in seemingly unrelated regressions', *Journal of Econometrics* **106**(1), 143–170.

Dufour, J.-M. and Kiviet, J. F. (1998), 'Exact inference methods for first-order autoregressive distributed lag models', *Econometrica* **66**, 79–104.

Dufour, J.-M. and Torrès, O. (1998), Union-intersection and sample-split methods in econometrics with applications to SURE and MA models, *in* D. E. A. Giles and A. Ullah, eds, 'Handbook of Applied Economic Statistics', Marcel Dekker, New York, pp. 465–505.

Dufour, J.-M. and Torrès, O. (2000), 'Markovian processes, two-sided autoregressions and exact inference for stationary and nonstationary autoregressive processes', *Journal of Econometrics* **99**, 255–289.

Duonga, T. (2013), 'Local significant differences from nonparametric two-sample tests', *Journal of Nonparametric Statistics* **25**(3), 635–645.

Dwass, M. (1957), 'Modified randomization tests for nonparametric hypotheses', *Annals of Mathematical Statistics* **28**, 181–187.

Edgington, E. S. (1995), *Randomization Tests, 3rd Edition*, Marcel Dekker, New York.

Efron, B. (1979), 'Bootstrap methods: Another look at the jacknife', *The Annals of Statistics* **7**, 1–26.

Efron, B. (2000), 'The bootstrap and modern statistics', *Journal of the American Statistical Association* **95**, 1293–1296.

Efron, B. (2003), 'Second thoughts on the bootstrap', *Statistical Science* **18**(2), 135–140.

Efron, B. and Tibshirani, R. J. (1993), *An Introduction to the Bootstrap*, Vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, New York.

Fisher, R. A. (1935), *The Design of Experiments*, Oliver and Boyd, London.

Fisz, M. (1960), 'On a result by M. Rosenblatt concerning the Mises-Smirnov test', *Annals of Mathematical Statistics* **31**, 427–429.

Folks, J. L. (1984), Combination of independent tests, *in* P. R. Krishnaiah and P. K. Sen, eds, 'Handbook of Statistics 4: Nonparametric Methods', North-Holland, Amsterdam, pp. 113–121.

Foutz, R. V. (1980), 'A method for constructing exact tests from test statistics that have unknown null distributions', *Journal of Statistical Computation and Simulation* **10**, 187–193.

Gibbons, J. D. and Chakraborti, S. (1992), *Nonparametric Statistical Inference, Third Edition, Revised and Expanded*, Marcel Dekker, New York.

Gnedenko, B. V. (1954), 'Tests of homogeneity of probability distributions in two independent samples (in Russian)', *Doklady Akademii Nauk SSSR* **80**, 525–528.

Good, P. (1994), *Permutation Tests. A Practical Guide to Resampling Methods for Testing Hypotheses*, second edn, Springer-Verlag, New York.

Goodman, L. A. (1954), 'Kolmogorov-Smirnov tests for psychological research', *Psychological Bulletin* **51**, 160–168.

Hájek, J. and Šidák, Z. (1967), *Theory of Rank Tests*, Academic Press, New York.

Hall, P. (1992), *The Bootstrap and Edgeworth Expansion*, Springer-Verlag, New York.

Hall, P. and Keilegom, I. V. (2007), 'Two-sample tests in functional data analysis from discrete data', *Statistica Sinica* **17**, 1511–1531.

Hochberg, Y. and Tamhane, A. C. (1987), *Multiple Comparison Procedures*, John Wiley & Sons, New York.

Hodges, Jr., J. L. (1958), 'The significance probability of Smirnov two-sample test', *Arkivfoer Matematik, Astronomi och Fysik* **3**, 469–486.

Jarque, C. M. and Bera, A. K. (1987), 'A test for normality of observations and regression residuals', *International Statistical Review* **55**, 163–172.

Jöckel, K.-H. (1986), 'Finite sample properties and asymptotic efficiency of Monte Carlo tests', *The Annals of Statistics* **14**, 336–347.

Kim, P. J. (1969), 'On the exact and approximate sampling distributions of the two sample Kolmogorov-Smirnov crirerion $D_{mn},\ m \leq n$', *Journal of the American Statistical Association* **64**, 1625–1637.

Kim, P. J. and Jennrich, R. I. (1970), Tables of the exact distribution of the two sample Kolmogorov-Smirnov crirerion $D_{mn},\ m \leq n$, *in* H. L. Harter and D. B. Owen, eds, 'Selected Tables in Methematical Statistics', Vol. 1, American Mathematical Society, Providence, Rhode Island, pp. 79–170.

Korolyuk, V. S. (1961), 'On the discrepancy of empiric distributions for the case of two independent samples', *Selected Translations in Mathematical Statistics and Probability* **1**, 105–121.

Lehmann, E. L. (1951), 'Consistency and unbiasedness of certain nonparametric tests', *Annals of Mathematical Statistics* **22**, 165–179.

Li, Q., Maasoumi, E. and Racine, J. S. (2009), 'A nonparametric test for equality of distributions with mixed categorical and continuous data', *Journal of Econometrics* **148**(4), 186–200.

Manly, B. F. J. (1997), *Randomization, Bootstrap and Monte Carlo Methods in Biology; Permutation Test*, second edn, Chapman & Hall, London.

Massey, F. J. (1951*a*), 'The distribution between of the maximum deviation between two sample cumulative step functions', *Annals of Mathematical Statistics* **22**, 125–128.

Massey, F. J. (1951*b*), 'A note on a two sample test', *Annals of Mathematical Statistics* **22**, 304–306.

Massey, F. J. (1952), 'Distribution table for the deviation between two sample cumulatives', *Annals of Mathematical Statistics* **23**, 435–441.

Mielke, Jr., P. W. and Berry, K. J. (2001), *Permutation Methods: A Distance Function Approach*, Springer-Verlag, New York.

Miller, Jr., R. G. (1981), *Simultaneous Statistical Inference*, second edn, Springer-Verlag, New York.

Noether, G. E. (1963), 'Note on the Kolmogorov statistic in the discrete case', *Metrika* **7**, 115–116.

Pesarin, F. (2001), *Multivariate Permutation Tests with Applications in Biostatistics*, John Wiley & Sons, New York.

Pesarin, F. and Salmaso, L. (2010), *Permutation Tests for Complex Data: Theory, Applications and Software*, John Wiley & Sons, New York.

Pitman, E. J. G. (1937), 'Significance tests which may be applied to samples from any populations', *Journal of the Royal Statistical Society, Series A* **4**, 119–130.

Reiczigel, J., Zakariás, I. and Rózsa, L. (2005), 'A bootstrap test of stochastic equality of two populations', *The American Statistician* **59**(2), 156–161.

Rosenblatt, M. (1952), 'Limit theorems associated with variants of the von Mises statistic', *Annals of Mathematical Statistics* **23**, 617–623.

Shao, S. and Tu, D. (1995), *The Jackknife and Bootstrap*, Springer-Verlag, New York.

Smirnov, N. V. (1939), 'Sur les écarts de la courbe de distribution empirique (Russian/French summary)', *Matematičeskiĭ Sbornik N.S.* **6**, 3–26.

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

Steck, G. P. (1969), 'The Smirnov two sample tests as rank tests', *Annals of Mathematical Statistics* **40**, 1449–1466.

Thas, O. (2010), *Comparing Distributions*, Springer Statistics, Springer-Verlag, New York.

Walsh, J. E. (1963), 'Bounded probability properties for Kolmogorov-Smirnov and similar statistics for discrete data', *Annals of the Institute of Statistical Mathematics* **15**, 153–158.

Westfall, P. H. and Young, S. S. (1993), *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*, John Wiley & Sons, New York.

Zaven, A. K. and Dudewicz, E. J. (2000), *Fitting Statistical Distributions: The Generalized Lambda Distr bution and Genera zed Bootstrap Methods*, CRC Press, Boca Raton, Florida.

Zhan, D. and Hart, D. (2012), 'Testing equality of a large number of densities', *Biometrika* **99**(1), 1–17.