Département de sciences économiques

Cahier 2001-23

Exact Nonparametric Two-Sample Homogeneity Tests for Possibly Discrete Distributions

DUFOUR, Jean-Marie
FARHAT, Abdeljelil

Université de Montréal

Département de sciences économiques
Université de Montréal
Faculté des arts et des sciences
C.P. 6128, succursale Centre-Ville
Montréal (Québec) H3C 3J7
Canada
http://www.sceco.umontreal.ca
SCECO-information@UMontreal.CA
Téléphone : (514) 343-6539
Télécopieur : (514) 343-7221

Ce cahier a également été publié par le Centre interuniversitaire de recherche en économie quantitative (CIREQ) sous le numéro 23-2001.

This working paper was also published by the Center for Interuniversity Research in Quantitative Economics (CIREQ), under number 23-2001.

ISSN 0709-9231

CAHIER 2001-23

EXACT NONPARAMETRIC TWO-SAMPLE HOMOGENEITY TESTS FOR POSSIBLY DISCRETE DISTRIBUTIONS

Jean-Marie DUFOUR[1] and Abdeljelil FARHAT[2]

[1] Canada Research Chair Holder (Econometrics), Centre de recherche et développement en économique (C.R.D.E.) and Département de sciences économiques, Université de Montréal, and Centre interuniversitaire de recherche en analyse des organisations (CIRANO)

[2] C.R.D.E., Université de Montréal, and CIRANO

October 2001

The authors thank Serge Tardif for his numerous comments. This work was supported by the Canada Research Chair Program, the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Canada Council for the Arts (Killam Fellowship), the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, and the Fonds FCAR (Government of Québec).

RÉSUMÉ

Dans ce texte, nous étudions plusieurs tests pour l'égalité de deux distributions inconnues. Deux de ces tests sont basés sur des fonctions de distribution empiriques, trois autres sur des estimateurs non paramétriques de fonctions de densité et les trois derniers sur des moments empiriques. Nous proposons de contrôler la taille des tests (sous des hypothèses non paramétriques) en employant des versions permutationnelles de ces tests conjointement avec la méthode des tests de Monte Carlo ajustée pour tenir compte de la possibilité de distributions discontinues. Nous proposons aussi une méthode pour combiner plusieurs de ces tests, le niveau de ces procédures étant aussi contrôlé par la technique des tests de Monte Carlo, laquelle possède de meilleures propriétés de puissance que les tests individuels combinés. Finalement, nous montrons dans une étude de simulation que la technique suggérée contrôle parfaitement la taille des différents tests considérés et que les nouveaux tests proposés peuvent fournir de notables améliorations de puissance.

Mots clés : méthodes non paramétriques, problème des deux échantillons, distribution discrète, distribution discontinue, test d'ajustement, test de Kolmogorov-Smirnov, Cramér-von Mises, estimateur à noyau pour une densité, test exact, test de permutations, test de Monte Carlo, bootstrap, test combiné, test induit

ABSTRACT

In this paper, we study several tests for the equality of two unknown distributions. Two are based on empirical distribution functions, three others on nonparametric probability density estimates, and the last ones on differences between sample moments. We suggest controlling the size of such tests (under nonparametric assumptions) by using permutational versions of the tests jointly with the method of Monte Carlo tests properly adjusted to deal with discrete distributions. We also propose a combined test procedure, whose level is again perfectly controlled through the Monte Carlo test technique and has better power properties than the individual tests that are combined. 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 test, bootstrap, combined test procedure, induced test

Contents

1. **Introduction** 1

2. **Test statistics** 3

3. **Exact randomized permutation tests** 5

4. **Monte Carlo standardized combined tests** 7

5. **Simulation study** 9

6. **Conclusion** 12

List of Tables

1 Continuous distributions with their means and variances 10
2 Empirical level and power for MC permutation tests of equality of two distributions 10
3 Some illustrations for empirical level for KS and CM tests of equality of two continuous distributions 10
4 Some illustrations for empirical level for KS and CM tests of 11
5 Empirical level and power for MC permutation tests of equality of two continuous distributions having same mean and same variance 13
6 Empirical level and power for MC permutation tests of equality of two continuous distributions having different means but same variance 14
7 Empirical level and power for MC permutation tests of equality of two continuous distributions having same mean but different variances 15
8 Empirical level and power for MC permutation tests of equality of two continuous distributions having different means and different variances 16
9 Empirical level and power for MC permutation tests of equality of two discrete distributions having same mean but different variances 17
10 Empirical level and power for MC permutation tests of equality of two discrete distributions having different means and same variance 18
11 Empirical level and power for MC permutation tests of equality of two discrete distributions having different means and different variances 19

1. Introduction

An important problem in statistics consists in testing whether the distributions of two random variables are identical against the alternative that they differ in some way. Specifically, consider two random samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ such that $F(x) = \mathsf{P}[X_i \leq x]$, $i = 1, \ldots, n$, and $G(x) = \mathsf{P}[Y_j \leq x]$, $j = 1, \ldots, m$. 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, many users who have to make such a confrontation resort 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 that have been suggested 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 that differ through their means, the exact and limiting distributions of the test statistics are not standard, and tables for the exact distributions are only available for a limited number of sample sizes. Thus these tests are usually performed with the help of tables based on asymptotic distributions. This leads to procedures that 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 finding test procedures with two basic features. Namely, the latter should be: (1) truly distribution-free, irrespective of whether the underlying distribution F is discrete or continuous, and (2) exact in finite samples (i.e., they must 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. For the statistics based on kernel-type density estimators, distribution-freeness 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.

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 and Tibshirani (1993), Hall (1992), Shao and Tu (1995) and Davison and Hinkley (1997)]. For this reason, we favor Monte Carlo (MC) test methods. MC tests were introduced by Dwass (1957) and Barnard (1963). Further discussions and extensions are also available in Birnbaum (1974), Foutz (1980), Jöckel (1986), Dufour (1995), Kiviet and Dufour (1997), Dufour, Farhat, Gardiol and Khalaf (1998), Dufour and Kiviet (1998) and Dufour and Khalaf (2001)].

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.

Thirdly, 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. Note that such control would be much more difficult, using standard distributional methods, which typically only yield finite-sample (conservative) bounds or large-sample approximations. Typically combined test procedures are based 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 or large-sample arguments [see Miller (1981), Dufour (1989, 1990), Dufour and Torrès (1998, 2000), Westfall and Young (1993)]. Here, we shall control the size of the combined test through the use of the MC test technique which will automatically take account of the dependence between the test statistics.

Fourth, on observing that none of the different test statistics considered has the best power against different alternatives, we consider procedures based on combining several tests. These 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.

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 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 $F(x) = \mathsf{P}[X_i \leq x]$ and $Y_1, \ldots, Y_m$ a sample $i.i.d.$ observations with cdf $G(x) = \mathsf{P}[Y_i \leq x]$. The problem is 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*b*, 1951*a*, 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 its distribution. Further, it is important to note that KS is not distribution-free when F is a discrete distribution, although it has been noted that 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

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

CM is also distribution-free under H_0 with F 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 in the case of 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:

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

where

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

and $s_X = \left[\sum_{i=1}^n \left(X_i - \bar{X}\right)^2/(n-1)\right]^{1/2}$ 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]. 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

$$\hat{L}_1 = \sum_{i=1}^n |f_n(X_i) - g_m(X_i)| + \sum_{j=1}^m |f_n(Y_j) - g_m(Y_j)|\,. \tag{2.4}$$

The L_2-distance and L_∞-distance tests are based on the statistics

$$\hat{L}_2 = \Big\{ \sum_{i=1}^n \left[f_n(X_i) - g_m(X_i)\right]^2 + \sum_{j=1}^m \left[f_n(Y_j) - g_m(Y_j)\right]^2 \Big\}^{1/2} \tag{2.5}$$

and

$$\hat{L}_\infty = \sup_x |f_n - g_m| = \max_{1\le i\le n,\, 1\le j\le m} \left\{ |f_n(X_i) - g_m(X_i)|\,,\ |f_n(Y_j) - g_m(Y_j)| \right\} \tag{2.6}$$

respectively. 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 problem that remains 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

$$\hat{\theta}_1 = \bar{X} - \bar{Y}\,. \tag{2.7}$$

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. Contrary to Allen (1997) who also considered bootstrap tests, the statistic based on the studentized difference of sample means

$$\hat{t} = \frac{(\bar{X} - \bar{Y})/\sqrt{\frac{1}{n} + \frac{1}{m}}}{\left\{ \left[\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{j=1}^m (Y_j - \bar{Y})^2\right]/(n+m-2) \right\}^{1/2}}$$

will not be considered since our study is restricted to permutation tests and it is straightforward to

see that such tests based on $\hat{\theta}_1$ and $\hat{t}$ are equivalent [see, for instance, Lehmann (1986)]. Further, we suggest here alternative test statistics based on comparing higher-order moments. Namely, the difference between unbiased estimators of sample variances,

$$\hat{\theta}_2 = \left| \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 - \frac{1}{m-1} \sum_{i=1}^{m} \left(Y_i - \bar{Y}\right)^2 \right|, \tag{2.8}$$

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

$$\hat{\theta}_3 = \left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right)^3 - \frac{1}{m} \sum_{i=1}^{m} \left(\frac{Y_i - \bar{Y}}{s_Y} \right)^3 \right|, \tag{2.9}$$

$$\hat{\theta}_4 = \left| \frac{1}{n} \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{s_X} \right)^4 - \frac{1}{m} \sum_{i=1}^{m} \left(\frac{Y_i - \bar{Y}}{s_Y} \right)^4 \right|, \tag{2.10}$$

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 $s_Y = 0$. Note that skewness and kurtosis coefficients play a central role in testing normality [see Jarque and Bera (1987) and Dufour et al. (1998)].

3. Exact randomized 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.

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 rejected for large values of T_0, the associated critical region of size α may be expressed as $G(T_0) \leq \alpha$, where $G(x) = \mathsf{P}\left[T \geq x \,|\, H_0\right]$ 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 $T^{(i)} = T\big(X_1^{(i)}, \ldots, X_n^{(i)}, Y_1^{(i)}, \ldots, Y_m^{(i)}\big)$, $i = 1, , \ldots, N$, from which we can compute an empirical p-value function:

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

where

$$\widehat{G}_N(x) = \frac{1}{N}\sum_{i=1}^{N} 1_{[0,\infty)}(T^{(i)} - x), \quad 1_A(x) = \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases} .$$

The associated MC critical region is defined as

$$\widehat{p}_N(T_0) \leq \alpha \, , \tag{3.2}$$

where $\widehat{p}_N(T_0)$ may be interpreted as an estimate of $G(T_0)$. When T has a continuous distribution, it can be shown that [see Dufour (1995) or Dufour and Kiviet (1998)]:

$$\mathsf{P}\big[\widehat{p}_N(T_0) \leq \alpha \,|\, H_0\big] = \frac{I\left[\alpha(N+1)\right]}{N+1} \, , \; 0 \leq \alpha \leq 1 \, , \tag{3.3}$$

where $I[x]$ denotes the largest integer not exceeding x. Thus if N is chosen such that $\alpha(N + 1)$ is an integer, the critical region (3.2) has the same size as the critical region $G(T_0) \leq \alpha$. The MC test so obtained is theoretically exact, irrespective of the number N of replications used.

The above procedure is closely related to the parametric bootstrap, with a fundamental difference however. Bootstrap tests are, in general, provably valid for $N \to \infty$. In contrast, we see from (3.3) 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.

Note that (3.3) holds for tests based on statistics with continuous distributions. In such a case, ties have non-zero probability. Nevertheless, the technique of MC tests can be adapted to discrete distributions by appeal to the following randomized tie-breaking procedure [see Dufour (1995), Dufour and Kiviet (1998), and Dufour and Khalaf (2001)]. 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:

$$(T^{(i)}, U_i) < (T^{(j)}, U_j) \Leftrightarrow \Big[T^{(i)} < T^{(j)} \quad \text{or} \quad (T^{(i)} = T^{(j)} \text{ and } U_i < U_j)\Big] .$$

Then, proceed as in the continuous case and compute

$$\widetilde{p}_N(x) = \frac{N\widetilde{G}_N(x) + 1}{N + 1} \, , \tag{3.4}$$

where

$$\widetilde{G}_N(x) = 1 - \frac{1}{N}\sum_{i=1}^{N} 1_{[0,\infty)}(x - T^{(i)}) + \frac{1}{N}\sum_{i=1}^{N} 1_{[0]}(T^{(i)} - x)\, 1_{[0,\infty)}(U_i - U_0).$$

The resulting critical region $\widetilde{p}_N(T_0) \leq \alpha$ has the same level as the region $G(T_0) \leq \alpha$, provided again $\alpha(N+1)$ is an integer. More precisely,

$$\mathsf{P}\big[\hat{p}_N(T_0) \leq \alpha \,|\, H_0\big] \leq \mathsf{P}\big[\tilde{p}_N(T_0) \leq \alpha \,|\, H_0\big] = \frac{I[\alpha(N+1)]}{N+1}, \;\; 0 \leq \alpha \leq 1. \tag{3.5}$$

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 five steps: first, the test statistic is computed with the help of the observed sample which gives a value T_0, say; second, N permutations of the sample are chosen at random and without replacement from all possible permutations; third, the test statistic is recomputed for each of the permuted samples which gives the values $T_1, \ldots, T_N$, say; fourth, if R_0 designates the rank of T_0 among the set $\{T_0, T_1, \ldots, T_N\}$ [in the case of ties, one may resort to the randomization method suggested by Dufour (1995)], the p-value associated with the MC test of the null hypothesis is given by $1 - R_0/(N+1)$; lastly, a decision is reached according to the chosen level [see Dufour (1995)]. 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.

4. Monte Carlo standardized combined tests

Once the simulation study based on the above statistics was performed, we noticed that a group of MC 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 tests showed the opposite profile. Moreover, none of the six MC 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 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), Dufour (1989, 1990), Dufour and Torrès (1998, 2000), Westfall and Young (1993)].

To be more specific, we shall consider here tests based on the maximum of several standardized statistics. The standardization aims at ensuring comparability between the different statistics and simply consists of subtracting the empirical mean from each statistic and dividing the result by the corresponding empirical standard error, where 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:

$$\widetilde{T}_j^{(i)} = \frac{T_j^{(i)} - \overline{T}_j}{s_j}, \; j = 1, \ldots, \; k, \; i = 0, 1, \ldots, N, \tag{4.6}$$

where

$$\overline{T}_j = \frac{1}{N+1} \sum_{i=0}^{N} T_j^{(i)}, \quad s_j = \Big\{ \frac{1}{N} \sum_{i=0}^{N} \big(T_j^{(i)} - \overline{T}_j\big)^2 \Big\}^{1/2}, \; j = 1, \ldots, \; k. \tag{4.7}$$

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:

$$\widehat{Q}\big(V^{(i)}\big) = \max_{1 \le j \le k} \big\{\widetilde{T}_j^{(i)}\big\}, \; i = 0, 1, \ldots, N, \tag{4.8}$$

and

$$\widehat{Q}_a\big(V^{(i)}\big) = \max_{1 \le j \le k} \big\{\big|\widetilde{T}_j^{(i)}\big|\big\}, \; i = 0, 1, \ldots, N. \tag{4.9}$$

The combined test based on the statistic $\widehat{Q}$ rejects the null hypothesis when the maximum of the standardized statistics is "large", while the one based on $\widehat{Q}_a$ does so when the absolute value of the standardized statistics is "large". In (4.8) - (4.9), $\widehat{Q}\big(V^{(0)}\big)$ and $\widehat{Q}_a\big(V^{(0)}\big)$ 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 $\widehat{Q}\big(V^{(i)}\big)$ and $\widehat{Q}_a\big(V^{(i)}\big)$ for $i \neq 0$ can be interpreted as values based on "simulated" (permuted) samples.

It is straightforward to see that the variables

$$\widehat{Q}\big(V^{(i)}\big), \; i = 0, 1, \ldots, N, \text{ are exchangeable under } H_0, \tag{4.10}$$

and similarly for $\widehat{Q}_a\big(V^{(i)}\big)$, $i = 0, 1, \ldots, N$. Consequently, we can write:

$$\mathsf{P}\big[\hat{p}_N\big(\widehat{Q}^{(0)}\big) \le \alpha\big] \le \mathsf{P}\big[\tilde{p}_N\big(\widehat{Q}^{(0)}\big) \le \alpha \,|\, H_0\big] = \frac{I[\alpha(N+1)]}{N+1}, \;\; 0 \le \alpha \le 1, \tag{4.11}$$

where $\widehat{Q}^{(i)} \equiv \widehat{Q}\big(V^{(i)}\big)$, $i = 0, 1, \ldots, N$, and $\hat{p}_N$, $\tilde{p}_N$ are defined as in (3.1) - (3.4) with T_i replaced by $\widehat{Q}^{(i)}$. Of course, the same holds for tests based on $\widehat{Q}_a^{(i)} \equiv \widehat{Q}_a\big(V^{(i)}\big)$, $i = 0, 1, \ldots, N$.

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

$$\widehat{Q}_i \equiv \widehat{Q}(V_i), \quad \widehat{Q}_{ai} \equiv \widehat{Q}_a(V_i), \; i = 1, 2, \tag{4.12}$$

where

$$V_1 = (KS, \hat{L}_\infty)', \quad V_2 = (KS, \hat{L}_\infty, \hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)'. \tag{4.13}$$

The first choice (V_1) emphasizes two overall distance measures between the two empirical distributions (KS and $\hat{L}_\infty$), while the second one (V_2) also uses the differences between the first four moments of the two distributions, and thus should 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 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 test was carried out by picking at random N = 99 permutations of the original grouped sample and this was done by using the IMSL (1987) Program Library random number generator. In his simulation study, Allen (1997) used 2500 trials and each permutation or bootstrap test was carried out with 499 samples.

For the first part of the study where F and G are both continuous, 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 $\Lambda(4,\ 1.5)$ 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 situations were 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 obtained each time F and G share the same distribution from the list and an alternative hypothesis is obtained each time F and G possess different distributions from that list.

For the second part of the study where F and G are discrete, the five most commonly used distributions were retained: 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 were $DU(19)$, $Bin(20,\ 0.5)$, $Geo(0.1)$, $Nbin(8,\ 0.2)$, $P(10)$ and, thus had common mean 10 and variance 30, 5, 90, 2.5 and 10 respectively; (ii) the distributions were $DU(10)$, $Bin(33,\ 0.5)$, $Geo\big((\sqrt{34}\ -\ 1)/16.5\big)$, $Nbin\big(3,\ (\sqrt{108}\ -\ 3)/16.5\big)$, $P(8.25)$ and, thus had mean 5.5, 16.5, 3.42, 2.23 and 8.25 respectively but common variance 8.25; (iii) the distributions were $DU(10)$, $Bin\big(10,\ 0.1\big)$, $Geo(0.3)$, $Nbin\big(10,\ 0.2\big)$, $P(5)$ and, thus had mean 5.5, 1, 3.33, 50 and 5 respectively and variance 8.25, 0.9, 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 tests and by excluding the bootstrap 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 well tabulated, so a study of the reliability of tabulated critical values can only be limited. In Tables 3 and 4, we present some

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	168.17	.50
Variance	1	2.25	2	.04	0.55133^{-2}	240055	1/12

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

	$F = N(0,1)$												
G	KS	CM	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
$N(0,1)$	5.0	4.8	4.7	5.1	4.8	5.0	5.0	5.0	4.8	5.1	4.8	5.1	4.7
$N(0.2,\ 1)$	5.7	5.9	5.7	5.2	4.9	5.1	5.9	5.8	5.8	5.7	5.5	5.8	5.5
$N(0.3,\ 1)$	8.0	8.9	9.6	4.3	5.4	5.2	6.6	6.5	6.2	7.3	7.4	7.3	7.4
$N(0.4,\ 1)$	13.1	15.0	15.7	4.9	5.3	5.4	10.0	9.7	9.0	11.8	10.5	11.6	10.5
$N(0.5,\ 1)$	28.2	33.0	36.0	3.7	6.3	5.8	19.3	18.8	17.0	24.3	22.8	24.2	22.8
$N(0.7,1)$	49.2	56.8	61.9	3.2	6.5	6.7	35.2	34.2	31.7	43.5	43.1	43.4	43.1
$N(0,\ 1.2^2)$	5.6	5.3	4.5	12.1	4.8	5.0	10.0	10.3	10.6	8.7	7.7	8.7	7.7
$N(0,\ 1.4^2)$	7.1	6.7	4.8	28.5	4.2	5.0	22.6	23.1	23.4	18.7	15.3	18.6	15.3
$N(0,\ 1.6^2)$	9.6	8.6	4.7	48.7	3.4	4.4	41.5	42.0	42.1	35.0	28.8	34.9	28.8
$N(0,\ 1.8^2)$	13.4	11.8	4.9	67.9	3.2	3.7	60.4	61.2	60.4	52.8	43.8	52.7	43.8
$N(0,\ 2.0^2)$	17.5	15.9	5.2	80.4	2.7	3.4	74.7	75.5	75.0	67.8	59.8	67.8	59.8

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

	Original tests						MC 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 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

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 earlier theoretical results), 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 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.

Thirdly, 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, 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 MC 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 test based on the smaller set of statistics ($\widehat{Q}_1$) and the one based on the larger one ($\widehat{Q}_2$). Not surprisingly, $\widehat{Q}_2$ tends to perform better than $\widehat{Q}_1$ when the distributions compared have different variances (because $\hat{\theta}_2$ is used by $\widehat{Q}_2$ but not by $\widehat{Q}_1$). The combined statistics $\widehat{Q}$ and $\widehat{Q}_a$ exhibit very similar behaviors, so there appears to be little ground for preferring one over the other.

Let us now consider the case where both F and G are discrete. The results of our simulation are presented in Tables 9 to 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 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 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. Thirdly, 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 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
N	5.0	4.8	4.7	5.1	4.8	5.0	5.0	5.0	4.8	5.1	4.8	5.1	4.7
Exp	13.8	12.4	5.6	10.5	42.0	15.8	17.3	17.1	15.6	17.1	24.2	17.1	24.1
Γ	8.9	8.8	5.3	7.7	27.0	9.5	11.1	11.0	10.4	10.5	14.2	10.4	14.2
B	5.4	4.9	4.9	5.4	7.5	7.0	5.7	5.9	5.8	5.7	6.4	5.5	6.4
Log	5.7	5.3	5.6	5.1	5.4	6.3	5.3	5.2	5.5	5.4	5.7	5.3	5.7
Λ	71.8	65.2	5.7	59.0	70.2	62.7	68.6	67.6	65.7	79.4	88.9	79.7	88.3
U	6.7	5.7	4.9	6.0	6.4	16.6	6.1	6.5	6.9	6.6	11.1	6.6	11.1
	$F = Exp$												
Exp	4.8	5.0	5.2	4.9	4.7	5.0	5.2	5.2	5.2	5.0	4.8	4.9	4.9
Γ	6.0	6.0	4.7	5.6	8.3	6.5	6.0	5.9	5.9	6.1	6.8	6.2	6.9
B	11.1	9.1	4.8	12.9	35.9	20.8	15.9	16.0	15.8	15.2	24.6	15.1	24.6
Log	13.4	12.7	5.4	8.5	36.4	11.7	14.5	14.3	12.9	15.3	19.6	15.2	19.5
Λ	85.4	73.1	5.4	50.3	30.6	31.5	54.9	54.6	54.5	86.7	86.8	86.9	86.7
U	17.1	13.7	5.5	17.2	54.8	35.3	22.5	22.7	24.1	23.4	41.5	23.3	41.5
	$F = \Gamma$												
Γ	4.5	4.8	4.7	5.4	5.0	4.8	4.9	4.9	5.0	4.8	4.9	4.9	4.9
B	7.3	6.6	5.0	8.8	20.3	12.5	9.8	9.9	9.6	9.2	13.8	9.3	13.8
Log	9.7	8.8	5.5	6.3	23.4	7.4	10.0	10.1	9.3	10.1	11.7	10.0	11.6
Λ	80.5	67.6	5.3	54.2	41.8	42.5	60.9	60.5	60.6	84.1	86.9	84.3	86.7
U	10.5	9.5	5.1	12.6	35.9	25.1	14.6	14.8	15.2	14.0	26.2	13.8	26.2
	$F = B$												
B	5.0	5.0	5.0	4.9	5.1	4.9	5.2	4.9	5.0	4.8	5.2	4.8	5.2
Log	6.1	5.6	4.7	6.0	8.3	10.9	6.8	6.7	6.9	6.6	9.1	6.6	9.0
Λ	77.2	70.4	5.5	62.5	74.9	70.3	71.8	71.2	69.6	84.0	93.2	84.2	92.8
U	5.5	5.2	5.1	5.4	8.1	10.5	5.6	5.5	5.8	5.6	8.5	5.6	8.5
	$F = Log$												
Log	5.0	5.0	5.1	5.0	5.0	5.1	4.6	4.6	4.8	4.8	4.9	4.9	4.8
Λ	66.9	59.8	5.8	54.9	60.2	52.6	64.2	63.3	61.2	75.1	83.9	75.4	83.4
U	8.0	6.7	5.2	8.3	8.9	25.6	8.7	9.0	9.6	9.4	16.6	9.3	16.6
	$F = \Lambda$												
Λ	5.0	4.9	5.1	5.6	5.0	4.9	5.4	5.4	5.3	5.4	5.2	5.4	5.2
U	82.2	77.4	5.8	65.5	90.3	83.5	76.6	75.6	73.4	88.0	96.5	88.6	96.1

Table 6. Empirical level and power for MC 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
N	5.0	4.8	4.7	5.1	4.8	5.0	5.0	5.0	4.8	5.1	4.8	5.1	4.7
Exp	90.1	88.9	92.5	5.0	19.3	13.8	38.0	38.6	47.7	83.8	86.6	83.9	86.4
Γ	98.6	99.4	99.9	1.4	14.5	10.2	77.2	77.6	82.2	97.0	98.7	97.1	98.8
B	100	100	100	0.1	7.2	12.4	99.1	99.1	99.4	100	100	100	100
Log	36.1	40.4	42.6	3.8	7.2	8.0	23.6	22.8	21.0	30.4	29.5	30.2	29.4
Λ	88.4	75.7	22.8	55.8	66.0	62.9	68.5	67.8	67.8	90.1	94.0	90.3	93.5
U	99.6	99.9	100	0.1	6.6	18.2	95.2	95.3	96.2	98.9	99.8	98.9	99.8
	$F = Exp$												
Exp	4.8	5.0	5.2	4.9	4.7	5.0	5.2	5.2	5.2	5.0	4.8	4.9	4.9
Γ	36.9	42.1	30.0	4.5	11.4	9.6	17.9	17.3	15.7	30.8	27.1	30.8	27.0
B	90.3	94.0	86.5	4.4	51.4	41.3	78.8	78.5	75.9	87.9	85.5	87.9	85.5
Log	100	100	100	1.9	35.1	18.1	84.3	85.6	92.9	99.9	100	100	100
Λ	93.4	96.0	76.7	18.7	33.7	30.6	60.7	59.4	55.3	90.5	88.9	90.6	88.7
U	67.6	72.7	63.4	12.0	64.2	49.3	61.4	60.8	57.5	66.6	69.1	66.6	69.1
	$F = \Gamma$												
Γ	4.5	4.8	4.7	5.4	5.0	4.8	4.9	4.9	5.0	4.8	4.9	4.9	4.9
B	48.5	53.7	48.0	5.5	28.6	20.7	40.5	40.2	36.6	45.1	42.6	45.1	42.6
Log	100	100	100	0.5	32.0	19.8	98.5	98.8	99.6	100	100	100	100
Λ	100	100	93.0	9.6	51.7	47.8	82.6	81.7	80.5	99.9	99.8	99.9	99.8
U	24.7	24.5	19.4	11.7	41.5	30.5	28.0	27.5	24.7	26.9	34.9	26.7	34.9
	$F = B$												
B	5.0	5.0	5.0	4.9	5.1	4.9	5.2	4.9	5.0	4.8	5.2	4.8	5.2
Log	100	100	100	0.1	17.0	33.8	100	100	100	100	100	100	100
Λ	100	100	98.2	14.2	86.9	79.2	95.7	95.7	96.7	100	100	100	100
U	8.7	10.0	12.6	4.6	6.8	8.2	6.9	7.0	6.8	8.1	9.5	8.0	9.5
	$F = Log$												
Log	5.0	5.0	5.1	5.0	5.0	5.1	4.6	4.6	4.8	4.8	4.9	4.9	4.8
Λ	100	99.5	93.3	41.8	66.4	57.6	76.0	75.7	79.9	99.9	99.9	99.9	99.9
U	100	100	100	0.0	14.9	43.4	99.8	99.8	99.9	100	100	100	100
	$F = \Lambda$												
Λ	5.0	4.9	5.1	5.6	5.0	4.9	5.4	5.4	5.3	5.4	5.2	5.4	5.2
U	100	100	96.4	45.9	91.5	81.0	91.4	91.3	92.9	100	100	100	100

Table 7. Empirical level and power for MC 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
N	5.0	4.8	4.7	5.1	4.8	5.0	5.0	5.0	4.8	5.1	4.8	5.1	4.7
Exp	19.2	18.4	6.4	18.0	37.6	6.8	18.0	18.7	19.1	20.6	22.3	20.4	22.0
Γ	12.4	11.8	5.4	19.4	22.9	5.2	19.1	19.4	19.9	17.8	16.5	17.6	16.4
B	75.1	76.4	5.5	100	0.4	0.1	100	100	100	100	100	100	100
Log	11.1	9.9	5.7	57.0	2.6	2.3	50.9	51.6	51.1	43.2	34.4	43.1	34.2
Λ	100	100	24.1	99.9	18.3	8.5	100	100	100	100	100	100	100
U	51.9	49.1	5.3	99.8	0.2	0.0	99.8	99.8	99.6	99.5	98.7	99.5	98.7
	$F = Exp$												
Exp	4.8	5.0	5.2	4.9	4.7	5.0	5.2	5.2	5.2	5.0	4.8	4.9	4.9
Γ	5.9	6.1	4.8	4.9	7.7	6.0	5.3	5.2	5.3	5.8	6.2	5.8	6.2
B	96.2	96.3	8.3	100	1.2	0.5	100	100	100	100	100	100	100
Log	13.9	12.4	5.4	17.9	36.3	15.7	24.0	24.0	22.9	22.4	26.5	22.4	26.5
Λ	100	100	24.5	99.8	1.4	8.3	100	100	100	100	100	100	100
U	89.5	88.0	8.2	99.9	6.2	1.1	100	100	100	100	100	100	100
	$F = \Gamma$												
Γ	4.5	4.8	4.7	5.4	5.0	4.8	4.9	4.9	5.0	4.8	4.9	4.9	4.9
B	93.5	93.5	6.9	100	0.5	0.2	100	100	100	100	100	100	100
Log	9.5	8.8	5.4	18.2	23.0	9.7	21.0	21.1	20.6	18.3	18.9	18.2	19.1
Λ	100	100	24.0	99.8	2.1	8.1	100	100	100	100	100	100	100
U	82.8	81.2	6.8	99.9	2.1	0.2	100	100	100	100	100	100	100
	$F = B$												
B	5.0	5.0	5.0	4.9	5.1	4.9	5.2	4.9	5.0	4.8	5.2	4.8	5.2
Log	92.4	93.7	5.9	100	0.2	0.0	100	100	100	100	100	100	100
Λ	100	100	25.3	99.9	8.5	11.0	100	100	100	100	100	100	100
U	12.5	10.9	4.9	56.8	9.3	12.3	32.7	35.6	41.8	31.1	37.6	30.8	37.6
	$F = Log$												
Log	5.0	5.0	5.1	5.0	5.0	5.1	4.6	4.6	4.8	4.8	4.9	4.9	4.8
Λ	100	100	24.2	99.9	23.2	7.0	100	100	100	100	100	100	100
U	81.4	80.1	5.4	100	0.1	0.0	100	100	100	100	100	100	100
	$F = \Lambda$												
Λ	5.0	4.9	5.1	5.6	5.0	4.9	5.4	5.4	5.3	5.4	5.2	5.4	5.2
U	100	100	24.9	99.8	14.8	14.8	100	100	100	100	100	100	100

Table 8. Empirical level and power for MC 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
N	5.0	4.8	4.7	5.1	4.8	5.0	5.0	5.0	4.8	5.1	4.8	5.1	4.7
Exp	94.9	96.8	99.0	0.5	15.5	4.0	48.1	47.6	53.6	90.0	95.8	90.0	95.6
Γ	99.9	100	100	0.1	17.9	5.0	90.7	90.7	92.9	99.5	100	99.5	100
B	92.6	90.4	43.4	99.8	0.7	0.1	100	100	100	100	100	100	100
Log	62.9	67.2	59.0	46.5	4.2	3.2	76.4	75.9	74.2	75.2	68.8	75.0	68.7
Λ	100	100	100	15.6	69.9	17.6	100	100	100	100	100	100	100
U	89.5	85.4	59.8	98.3	0.6	0.2	99.9	99.9	99.9	99.8	99.6	99.8	99.6
	$F = Exp$												
Exp	4.8	5.0	5.2	4.9	4.7	5.0	5.2	5.2	5.2	5.0	4.8	4.9	4.9
Γ	31.0	35.5	23.1	5.5	10.0	8.7	13.8	13.2	12.3	25.2	22.0	25.1	21.8
B	97.8	96.5	99.5	81.5	11.8	0.2	100	100	99.9	99.8	99.7	99.8	99.7
Log	100	100	99.8	5.5	38.8	23.3	87.2	88.2	94.6	99.9	99.9	99.9	99.9
Λ	100	100	100	17.3	20.4	8.9	100	100	100	100	100	100	100
U	90.7	85.4	96.7	81.2	21.9	0.3	99.3	99.2	98.8	98.0	97.0	98.0	96.8
	$F = \Gamma$												
Γ	4.5	4.8	4.7	5.4	5.0	4.8	4.9	4.9	5.0	4.8	4.9	4.9	4.9
B	100	100	100	46.6	6.2	0.1	100	100	100	100	100	100	100
Log	100	100	100	3.0	32.4	21.6	98.4	98.6	99.5	100	100	100	100
Λ	100	100	100	17.8	32.1	12.1	100	100	100	100	100	100	100
U	100	100	100	47.6	13.1	0.1	100	100	100	100	100	100	100
	$F = B$												
B	5.0	5.0	5.0	4.9	5.1	4.9	5.2	4.9	5.0	4.8	5.2	4.8	5.2
Log	100	100	92.5	97.0	2.3	0.4	100	100	100	100	100	100	100
Λ	100	100	100	16.6	60.8	20.0	100	100	100	100	100	100	100
U	27.4	26.4	24.3	52.8	9.4	9.3	45.0	46.8	48.0	42.6	40.9	42.2	40.9
	$F = Log$												
Log	5.0	5.0	5.1	5.0	5.0	5.1	4.6	4.6	4.8	4.8	4.9	4.9	4.8
Λ	100	100	100	15.7	68.5	13.8	100	100	100	100	100	100	100
U	100	99.9	95.4	95.0	1.3	1.5	100	100	100	100	100	100	100
	$F = \Lambda$												
Λ	5.0	4.9	5.1	5.6	5.0	4.9	5.4	5.4	5.3	5.4	5.2	5.4	5.2
U	100	100	100	16.7	69.7	23.7	100	100	100	100	100	100	100

Table 9. Empirical level and power for MC 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
UD	4.9	4.9	4.8	4.8	4.7	5.0	4.5	4.7	4.8	4.6	4.9	4.6	4.9
Bin	51.5	45.1	4.8	99.2	7.6	14.4	97.1	97.1	96.7	95.2	96.2	95.2	96.3
Geo	16.4	17.8	6.0	36.4	34.7	8.4	21.8	22.9	26.0	22.1	23.2	22.1	22.7
$BinN$	83.1	71.5	5.6	99.9	25.3	21.5	99.9	99.9	99.9	99.7	99.8	99.7	99.8
Poi	25.1	20.5	5.0	82.8	12.3	24.2	69.0	71.2	71.7	65.3	70.1	65.3	70.1
	$F = Bin$												
Bin	5.1	5.0	4.9	5.2	4.9	5.0	5.1	4.9	4.9	5.2	5.2	5.1	5.2
Geo	82.4	82.2	8.5	99.6	13.9	1.3	99.7	99.7	99.7	99.4	99.0	99.4	98.9
$BinN$	10.2	7.4	5.0	33.8	16.6	9.8	28.3	28.7	26.9	24.3	25.3	24.0	25.4
Poi	8.4	8.0	5.0	29.0	5.5	4.5	21.9	22.1	21.7	18.5	16.7	18.3	16.7
	$F = Geo$												
Geo	5.0	4.9	4.8	4.7	4.8	4.8	4.4	4.6	4.5	4.8	4.8	4.8	4.8
$BinN$	97.1	91.7	8.5	99.9	1.2	2.6	100	100	100	100	99.9	100	99.9
Poi	61.0	57.3	7.1	94.7	11.0	2.1	92.9	93.6	93.5	91.7	87.8	91.6	87.4
	$F = BinN$												
$BinN$	4.7	4.9	4.9	4.8	4.7	4.9	4.6	4.5	4.5	4.8	4.7	4.9	4.7
Poi	23.5	21.4	5.4	78.1	9.3	9.2	62.9	64.4	65.0	58.4	60.3	58.3	60.5

Table 10. Empirical level and power for MC 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
UD	4.8	4.8	4.8	4.8	4.7	4.7	4.7	4.7	4.9	4.7	5.0	4.7	5.0
Bin	100	100	100	0.0	8.1	55.6	100	100	100	100	100	100	100
Geo	67.5	76.9	64.2	12.0	67.1	51.2	60.5	59.7	57.6	65.7	69.3	65.4	69.3
$BinN$	20.4	13.5	5.4	47.7	40.8	38.1	36.3	37.7	39.2	35.2	54.7	35.0	54.7
Poi	65.3	72.7	86.9	2.0	3.2	9.7	46.9	46.9	47.8	58.6	68.4	58.4	68.4
	$F = Bin$												
Bin	5.3	5.3	5.3	4.8	4.8	5.0	5.1	5.2	5.1	5.2	5.2	5.3	5.2
Geo	100	100	100	0.3	63.8	79.9	100	100	100	100	100	100	100
$BinN$	100	100	100	0.0	40.0	61.7	100	100	100	100	100	100	100
Poi	100	100	100	0.0	15.7	32.4	100	100	100	100	100	100	100
	$F = Geo$												
Geo	5.4	5.1	4.9	4.9	5.1	5.0	4.9	4.9	4.8	5.1	5.1	5.1	5.2
$BinN$	95.9	92.7	72.1	21.6	14.2	20.6	69.8	71.8	76.3	94.0	89.8	93.9	89.8
Poi	99.9	100	99.7	3.5	55.8	49.9	99.6	99.6	99.3	99.9	99.8	99.9	99.8
	$F = BinN$												
$BinN$	5.0	5.0	4.9	5.0	5.2	5.3	4.9	5.0	4.9	5.1	5.1	5.0	5.1
Poi	90.9	92.4	94.0	10.3	29.3	19.8	84.8	84.5	83.7	89.9	89.0	90.0	89.0

Table 11. Empirical level and power for MC 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	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}_4$	$\hat{L}_1$	$\hat{L}_2$	$\hat{L}_\infty$	$\widehat{Q}_1$	$\widehat{Q}_2$	$\widehat{Q}_{a1}$	$\widehat{Q}_{a2}$
UD	4.8	4.8	4.8	4.8	4.7	4.7	4.7	4.7	4.9	4.7	5.0	4.7	5.0
Bin	100	100	100	61.9	19.6	11.1	99.8	99.9	99.9	100	100	100	100
Geo	69.8	79.1	67.2	13.0	67.0	52.5	62.7	61.9	59.7	68.0	71.5	67.9	71.5
$BinN$	100	100	100	0.2	27.4	33.3	100	100	100	100	100	100	100
Poi	15.5	13.1	9.4	32.8	14.4	26.2	25.7	26.5	26.0	24.7	32.6	24.4	32.5
	$F = Bin$												
Bin	5.1	4.8	5.2	4.9	4.7	4.9	4.6	5.1	5.2	5.2	4.7	5.1	4.8
Geo	94.0	84.4	99.7	14.8	3.9	1.0	65.2	66.3	70.7	93.0	96.5	92.9	96.3
$BinN$	100	100	100	0.0	7.8	81.1	100	100	100	100	100	100	100
Poi	100	100	100	7.4	13.7	9.9	100	100	100	100	100	100	100
	$F = Geo$												
Geo	5.0	5.3	4.7	4.7	4.9	5.1	4.5	4.6	4.6	4.7	4.8	4.7	4.7
$BinN$	100	100	100	8.3	33.1	76.1	100	100	100	100	100	100	100
Poi	76.6	74.6	57.2	12.2	40.5	22.6	54.8	55.3	56.3	71.9	66.2	71.9	66.2
	$F = BinN$												
$BinN$	4.9	5.0	5.3	5.1	4.6	5.0	5.0	4.9	4.7	4.9	4.8	5.1	4.8
Poi	100	100	100	0.2	15.6	49.4	100	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. (1995), Monte Carlo tests with nuisance parameters: A general approach to finite-sample inference and nonstandard asymptotics in econometrics, Technical report, C.R.D.E., Université de Montréal.

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 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 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.

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

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.

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.

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.

Kiviet, J. and Dufour, J.-M. (1997), 'Exact tests in single equation autoregressive distributed lag models', *Journal of Econometrics* **80**, 325–353.

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.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses, 2nd edition*, John Wiley & Sons, New York.

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.

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.

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.

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.

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.