A SIMPLE TEST FOR DIFFERENCES IN DISTRIBUTION

ABSTRACT

We propose and discuss a simple rank-based omnibus test for differences in distribution. Simulation results suggest that the test is about equal in power to the familiar omnibus test of Kolgomorov and Smirnov. Other simulations compare the procedure to some specialized tests that focus on particular differences between distributions.

ELLEN EISEN* and ARNOLD BARNETT†

ard School of Public Health, Boston, Massachusetts.

†Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts.

1) Supported in part by U.S. Public Health Service Grant # CA 19122 from the National Cancer Institute.

2) Partially supported under Grant No. 78-NI-AX-0034, "On Another Approach to Criminal Justice Statistical Analysis," from the U.S. Department of Justice, Law Enforcement Assistance Administration, National Institute of Law Enforcement and Criminal Justice. Points of view or opinions stated in this document do not necessarily represent the official position or policies of the U.S. Department of Justice.

Acknowledgements

We are grateful to Drs. JANE DESFORGES and CYNTHIA RUTHERFORD for posing certain problems that led to his work and for their encouragement and support throughout it. We thank BYRON DAVIES and KENNETH FIELDS for their essential help with the computer work, and PROFESSOR ROY WELSCH for his advice on reliable random number generators. And we are highly appreciative of the unusually detailed and constructive referee reports that led to substantial improvements in the paper.

Introduction

Consider two large populations each member of which is associated with a particular number, and suppose these numbers are obtained for random samples from each population. Because of sampling error, one would expect some differences between the two observed distributions even if, over the entire populations, the distribution of numbers is exactly the same. Thus it becomes interesting to specify when two sample distributions differ sufficiently that they should be interpreted as reflecting genuine differences between populations rather than fluctuations associated with sampling.

Special cases of this question arise often in the analysis of both medical and crime data. The authors, for example were given certain blood-test results for Hodgkin's Disease patients who were divided into two categories: (i) early stage, for which radiotherapy is the safer treatment, (ii) later stages, for which chemotherapy is safer. We were asked whether there was any discernable relationship between the blood-test results and the extensiveness of the cancer, the hope being that these tests might provide information now obtained more dangerously through surgery.

In the context of criminal justice, one might want to investigate whether the distribution of prison sentences for a given crime is associated with, say, the race of the offender. When similar cohorts of criminals are subject to different correctional programs, one might be interested in differences in the distributions of time until first rearrest after release. In such situations any significant differences in the empirical distributions are of potential interest, not just those of mean or variance.

Introduction

In this paper we propose a rank-based "omnibus" test for differences in distribution that, like the familiar test of Kolmogorov and Smirnov, aims at sensitivity to all kinds of differences rather than those of particular form. We consider the test both conceptually simple and easy to use. Below we will try to motivate the test procedure, derive the asymptotic distribution of its key statistic, and present simulated comparisons of its power versus both the Kolgomorov-Smirnov and some specialized tests.

The simulation results identify some circumstances in which the proposed test is more powerful than its Kolgomorov-Smirnov counterpart. On balance, however, the two tests seem about equally matched and, indeed, they reached the same conclusion in the overwhelming majority of simulated cases. These results, coupled with certain advantages of our procedure that we will suggest, might lead some users to view it as a viable alternative to the Kolgomorov-Smirnov procedure for assessing differences in distribution.

I. The Test Statistic

Suppose one has two data samples A: $(x_1,\ldots,x_m)$ and B: $(y_1,\ldots,y_n)$, and wants to know whether one can reasonably assert that A and B come from the same underlying probability distribution.

Let the m + n measurements be combined into a pooled ranked sample; we assume for now that m + n is a multiple of 4. Let b_i be the number of A measurements in the i^{th} quartile of the combined sample. If the two samples do come from the same distribution each b_i should have a hypergeometric probability distribution with parameters (m, n, (m+n)/4):

$$\Pr(b_i = k) = \frac{\binom{m}{k}\binom{n}{Q-k}}{\binom{m+n}{Q}} \qquad (1)$$

where Q = (m+n)/4.

Corresponding to (1) are the relationships:

$$E(b_i) = m/4$$

$$\sigma^2(b_i) = \frac{3mn}{16(m+n-1)}$$

Since $\sum_{i=1}^{4} b_i = m$, the different b_i's are not independent random variables.

Consider the random variables S, d_o and d_I defined by:

$$S = b_1 + b_4$$

$$d_o = b_4 - b_1$$

$$d_I = b_3 - b_2$$

If the two data samples come from the same distribution, it is clear that:

$$E(S) = m/2;\ E(d_I) = E(d_o) = 0$$

From the fact that $\sigma^2(\sum_1^4 b_i) = 0$ and the expression for $\sigma^2(b_i)$ in (2), we directly obtain the covariance of b_i and b_j $(i \neq j)$ as $-mn/(16(m+n-1))$. It follows at once that:

$$\sigma^2(S) = mn/4(m+n-1)$$
$$\sigma^2(d_o) = \sigma^2(d_I) = mn/2(m+n-1) \qquad (3)$$

We define the normalized variables $\tilde{S}$, $\tilde{d}_o$, and $\tilde{d}_I$ by

$$\tilde{S} = \frac{S-m/2}{\sigma(S)};\quad \tilde{d}_o = \frac{d_o}{\sigma(d_o)};\quad \tilde{d}_I = \frac{d_I}{\sigma(d_I)}$$

where $\sigma^2(S)$, $\sigma^2(d_o)$, $\sigma^2(d_I)$ are as specified in (3).

We further define the variable D by:

$$D = (\tilde{S})^2 + (\tilde{d}_o)^2 + (\tilde{d}_I)^2$$

The quantity D is the test statistic in our proposed procedure for investigating differences in distribution. Before discussing it further, we will obtain the asymptotic distribution for D under the null hypothesis H_o that the A and B samples arise from exactly the same distribution.

A straightforward application of Theorem 19 in Lehmann [1, P.393] shows that, as m and n increase, the jointly generalized hypergeometric variates (b_1, b_2, b_3, b_4) approach a singular multivariate normal distribution; the means, variances, and covariances of this distribution are as specified above. Since S, d_o, and d_I, are all linear combinations of the b_i's, they in turn approach a multivariate normal distribution; the fact that these variates are uncorrelated thus implies their asymptotic independence. Since $\tilde{S}^2$, $\tilde{d}_o^{\,2}$, and

$\tilde{d}_I^{\,2}$ approach independent χ^2-variates it follows that D is asymptotically χ^2-distributed with 3 degrees of freedom.

In the Appendix we present evidence in that, even for moderately large m and n, the asymptotic distribution for D is a good approximation to its actual distribution. Thus a test of H_o whose Type I error is close to α takes the form: Reject H_o if and only if $D > C_\alpha$, where C_α is the 100 $(1-\alpha)$ percentile of the χ^2-distribution with 3 degrees of freedom.

The statistic D is potentially sensitive to various differences in the two distributions from which one has empirical samples. Should the contributions of either data set gravitate towards the tails of the combined sample, a high $\tilde{S}^2$ value should reflect this; should either sample tend to "rise to the top" of the merged data, a high $\tilde{d}_o^{\,2}$ should result. Nonuniformities in the interior of the combined sample should generally show up in higher-than-usual $\tilde{d}_I^{\,2}$ values. Thus whether two distributions differ in location (e.g. mean or median), dispersion (e.g. variance) or some complex combination of both, a high D-statistic might well reflect the disparity.*

*Of course, the D statistic is insensitive to nonuniformities within the quartiles. Suppose two densitites $f_1(x)$ and $f_2(x)$ follow:

$$f_1(x) = 1/4, \text{ when } \frac{k}{8} < x < \frac{k+1}{8} \text{ for } k = 0, 2, 4, 6.$$
$$f_2(x) = 1/4, \text{ when } \frac{k}{8} < x < \frac{k+1}{8} \text{ for } k = 1, 3, 5, 7.$$

Faced with such a difference, a D-test approaches total ineffectiveness. But this situation is unusual if not pathological; the power of D in more realistic settings is considered later in the paper.

II. Some Other Rank Tests

The D-test is by no means the first rank-based procedure for indentifying differences in distribution. Below we briefly review three others with which the power of the D-tests will soon be compared. We define H_o as the null hypothesis that two data samples, A and B, come from the same distribution. As before, we assume the sizes of the A and B samples are m and n respectively. We will discuss the two-sided versions of each of the tests described.

Among omnibus tests for difference in distribution, the procedure of Kolgomorov and Smirnov (hereafter K-S) is so widely known and used that it seems proper to consider it the standard approach. It begins with the preparation of a combined, ranked sample of the A and B measurements. For each K from 1 to m+n, one looks at the K lowest numbers in the combined group and counts how many of them came from the A-distribution. (Let this number be S_K). Then one calculates the quantity W_K defined by

$$W_K = \left|\frac{S_K}{m} - \frac{(K-S_K)}{n}\right|$$

Note that S_K/m and $(K-S_K)/n$ are the fractions of all A and B measurements, respectively, that fall in the lowest K places of the pooled sample. Under H_o one would expect that, except for fluctuations, W_K would be near zero. Let $u = \max_K \{W_K\}$. The K-S test is of the form: reject H_o if and only if u exceeds some threshold c. The distribution of u under H_o as a function of m and n has been extensively tabulated ([2]); one typically chooses c so as to achieve a particular significance level for the test.

A simple and familiar test for difference in location is the median test. Under it one focuses on X , the number of the m measurements from the A distribution that fall above the median of the pooled, ranked sample. X is hypergeometrically distributed under H_o with a mean of m/2. The median test rejects H_o if $|X - m/2|$ exceeds some threshold c; once again c is chosen to achieve a desired significance level.

The Siegel-Tukey test, an offshoot of the Wilcoxon rank-sum procedure, is aimed at detecting differences in dispersion. The rank one is assigned to the largest measurement in the pooled A-B sample, two to the smallest measurement, three to the second largest, etc. The test statistic X is the sum of the ranks of the m A-measurements; under H_o, the expected value of X is $m(\frac{m+n+1}{2})$. From tables for the Wilcoxon rank-sum test (e.g. [3]), one determines if $|X - m(m+n-1)/2|$ is so large that H_o should be rejected at a desired level of significance.

III. A Comparison of the D and Kolgomorov-Smirnov Omnibus Tests

We describe here a simulated comparison of the power of the D and K-S testc in various circumstances. The general procedure is to perform the two tests on H_o at the same significance level (α), and to see how often each calls for the rejection of H_o when confronted with data samples generated from different probability distributions (i.e. we are estimating β-values). Before presenting any results, we will discuss both the details and rationale of the simulation performed.

Sample Sizes: When both m and n are very large, both D and K-S should be highly effective at picking up differences in distribution. When m and/or n are small, on the other hand, neither test should be especially powerful. The the most "interesting" m and n values for comparing the power of the two tests are those that are moderately large; pursuant of this view, we focus our attention on the two cases m = n = 24 and m = 24, n = 36.

Computer Use: To generate random data samples for various A and B distribution pairs, we used subroutines for particular probability distributions in the IMSL Statistical Computing Package.* The pooling and ranking of the data and calculation of test statistics was done under a computer program we wrote and tested extensively. The work was performed at MIT on the Multics Computer System.

Distributions Used: Altogether we used 20 different pairs of A and B distributions in our comparisons; all are listed in Table 1. The distributions we

*IMSL, Subroutive Chapter G, "Generation and Testing of Random Numbers,"

chose came from the normal, exponential, uniform, beta, and linear families; the differences we explored ranged from one-dimensional (e.g. members of the same family with different means) to very extensive. Because of its importance in statistics the normal distribution received more attention than the others. Foe each pair of distributions. we generated 1000 random samples for m = n = 24 and separately for m = 24, n = 36. Thus in total we compared D and K-S 40,000 (20 x 2 x 1000) times. While we cannot claim that our comparisons exhausted all possibilities we believe they tell a great deal about how D and K-S fare in a broad range of realistic settings.

The Form of the Test: We used D and K-S statistics to test H_o against the alternative of any difference in the distributions that spawned the A and B samples. The critical regions in the tests were chosen so as to achieve a Type I error rate (α) of .05 under both procedures. Since the null distributions of D and K-S were discrete, attaining an α of exactly .05 required randomized decision rules when "borderline values" arose for the test statistics. (Such rules, of course, would have been needed to achieve equality at any α-level chosen.)

When m = n = 24, for example, the K-S test we used for difference in distribution was:

(i) Reject H_o if $u > \frac{3}{8}$

(ii) Do not reject H_o if $u < \frac{3}{8}$

(iii) If $u = \frac{3}{8}$, select a random number x from the uniform distribution on [0,1].

If $x \leq .526$, do not reject H_o

If $x > .526$, reject H_o.

Tables in [2] on the exact null distribution of the K-S statistic show that the test rules above yield an α of exactly .050. Similar randomizations were performed for the K-S test in the 24-36 case and for the D-test in both cases. (The simulation results excerpted in the Appendix were the basis of the latter randomization.)

The results of the simulation are presented in Tables 1 and 2. We show observed β-values of D and K-S, how frequently the tests reached different conclusions, and the statistical significance of their differences in power as measured by the familiar McNemar test. The strongest difference that emerged occurred when the two distributions differed primarily in variance, ((c), (d), (i), (q)) in which case D is decisively superior to K-S. In other cases, the two tests are about equally powerful or K-S has the edge. But even when, according to the McNemar test on the differences, K-S does significantly better than D ((a), (b), (e), (k), (l), (m), (r), (s), (t)), its "margin of victory" tends to be small. In these particular cases, D and K-S reached the same conclusion 87% of the time and, when they differed, it was D and not K-S that was right about one-quarter of the time.

Over the total of 40,000 simulations, D erred 17184 times; the comparable figure for K-S was 17102, a mere 82 lower. D and K-S agreed in their conclusions 7/8 of the time. And both tests reduced their β-values by similar amounts (roughly 6%) when sample sizes increased from 24-24 to 24-36. One would hardly expect such results to be invariant across different sets of test distributions. However, coupled with the individual outcomes, they strengthen the impression that for assessing differences in distribution with no prior idea where the differences might arise, the D and K-S procedures are about equally powerful.

Table 1: Comparative Performance of KS and D Tests for Various Distribution Pairs (Sample Sizes n = m = 24; 1000 Simulations for Each Pair)

	Distribution A	Distribution B	Type II Error (β) Rates D	Type II Error (β) Rates KS	% of Time D and KS Differ	McNemar Test Significance Level
a)	N(0,1)	N(1,1)	.235	.172	.113	*
b)	N(0,1)	N(1,2)	.416	.314	.160	*
c)	N(0,1)	N(0,2.25)	.860	.919	.123	*
d)	N(0,1)	N(0,4)	.640	.815	.259	*
e)	N(0,1)	N(.5,.5)	.636	.579	.179	*
f)	N(0,1)	N(1,4)	.359	.362	.177	***
g)	N(0,1)	N(2,1)	0	0	0	***
h)	N(1,1)	e(1)	.882	.862	.108	***
i)	N(1,.09)	e(1)	.133	.290	.219	*
j)	N(1.5,1)	u(0,3)	.941	.950	.065	***
k)	e(1)	e(.5)	.625	.574	.177	*
l)	e(1)	u(0,2)	.858	.827	.131	**
m)	e(1)	u(0,3)	.570	.440	.180	*
n)	B(10,4)	u(0,1)	.017	.029	.028	**
o)	B(10,4)	B(8,6)	.006	.003	.003	***
p)	B(10,4)	B(16,6)	.909	.927	.086	***
q)	B(10,4)	L(0,1)	.187	.492	.357	*
r)	L(0,4)	U(0,1)	.679	.588	.169	*
s)	L(0,5)	U(0,4)	.159	.115	.096	*
t)	L(0,1)	L*(0,1)	.051	.032	.033	*

Notes:

$N(\mu,\sigma^2)$ means Gaussian with mean μ and variance σ^2.

u(a,b) means uniform a and b.

B(c,d) means beta with parameters c and d.

e(j) means exponential with parameter j.

L(a,b) means linear distribution with density function f(x) that follows:

$$f(x) = \frac{2(x-a)}{(b-a)^2} \text{ in } (a,b), \text{ and is 0 outside } (a,b).$$

L*(1,0) means density function is 2(1-x) in (0,1) and is 0 outside (0,1).

McNemar Test:

- * means significant at .01 level
- ** means significant at .05 but not .01 level
- *** means not significant at .05 level.

Table 2: Comparative Performance of KS and D for Various Distribution Pairs with Sample Sizes m = 24, n = 36 (1000 Simulations for each pair)

	Distribution A	Distribution B	Type II Error (ß) Rates D	Type II Error (ß) Rates KS	% of Time D and KS Differ	McNemar Test Significance Level
a)	N(0,1)	N(1,1)	.156	.091	.083	*
b)	N(0,1)	N(1,1)	.279	.183	.142	*
c)	N(0,1)	N(0;2.25)	.815	.913	.146	*
d)	N(0,1)	N(0,4)	.576	.786	.284	*
e)	N(0,1)	N(.5,.5)	.554	.463	.165	*
f)	N(0,1)	N(1,4)	.219	.199	.144	***
g)	N(0,1)	N(2,1)	0	0	0	***
h)	N(1,1)	e(1)	.857	.813	.138	*
i)	N(0,.09)	e(1)	.102	.159	.119	*
j)	N(1.5,1)	U(0,3)	.927	.943	.074	***
k)	e(1)	e(.5)	.559	.455	.182	*
l)	e(1)	U(0,2)	.802	.762	.152	*
m)	e(1)	U(0,3)	.467	.320	.187	*
n)	B(10,4)	U(0,1)	.006	.011	.013	***
o)	B(10,4)	B(8,6)	0	0	0	***
p)	B(10,4)	B(16,6)	.899	.913	.100	***
q)	B(10,4)	L(0,1)	.164	.311	.215	*
r)	L(0,4)	U(0,4)	.554	.439	.171	*
s)	L(0,5)	U(0,4)	.071	.047	.044	*
t)	L(0,1)	L*(0,1)	.014	.004	.012	**
Combined results of all simulations (m = n = 24 and m = 24, n = 36)			.429	.427	.126	

At this point, an obvious question arises: if D does no better than the standard K-S procedure, what advantage is there in using it? We see three possible advantages:

i) D involves less calculation.

Once a ranked sample exists, the user of the D-test need only count the number of A-measurements in each quartile, and then do four or so simple computations to obtain a D-statistic. Except for small m and n, the significance level of the statistic can be approximated from a χ^2-table for 3 degrees of freedom. (See Appendix.) K-S, by contrast, requires substantially more counting, and in principle m + n computations to obtain the test statistic. While one can reduce such computations with graphical methods, they themselves are time-consuming when m and n are large.

ii). The reasoning behind the D-test might be more transparent to nonstatisticians.

Both K-S and D are based on pooling and ranking the data from the two populations, and on the notion that the A-measurements should fall uniformly over the combined sample. But D relies heavily on the simple notion that each quartile of the pooled data should contain roughly 1/4 of the A-measurements. The reasoning behind K-S, while hardly obscure, is perhaps less transparent to someone unfamiliar with such statistical concepts as cumulative distribution and order statistics. It would seem that conceptual simplicity, when not accompanied by loss of accuracy, is a virtue in statistics as elsewhere.

iii) D is more directly informative about how two distributions differ.

If a D-statistic is judged significant, examining which of its components is (are) particularly large can indicate how two distributions differ. If $\tilde{S}^2$ is large but $\tilde{d}_o^2$ and $\tilde{d}_I^2$ are not, for example, the distributions are probably more dissimilar in divergence than in location. In the K-S test, examining individual W_K's when the test statistic u is significant seems not as directly illuminating about where the disparity arose.

IV. Comparison of D with Some Specialized Tests

We have also compared the power of D to those of two specialized tests: the median test for difference in location and the Siegel-Tukey test for difference in dispersion. Our results are summarized in Table 3; they are based on the same randomly-generated data that led to the previous tables.

With the lone exception of distribution pair (m), there were no situations in which the median test did substantially better than D. The Siegel-Tukey test, on the other hand, was clearly superior to the D-test when one distribution essentially "surrounded" the other (i.e. (c), (d), (i), (q)). But except in this general setting, the Siegel-Tukey test was strikingly ineffective in picking up differences within the distribution pairs we studied.

Especially since other specialized tests exist for differences in distribution, we should avoid extreme statements based on Table 3. Yet certain remarks are suggested by the results. If one has strong prior knowledge on how two distributions would differ if in fact they do, a test that focuses on this discrepancy is probably preferable to an omnibus test like D. But if one has weak or even moderately strong prior feelings one might do well to use an omnibus test, for specialized procedures seem to lose power rapidly as one departs from their bailiwicks.

Table 3: Comparative Performances of D, Median, and Siegal-Tukey Tests for Various Distribution Pairs

(Combined Results m = n = 24 and m = 24, n = 36; 2000 Simulations for each pair)

	Distribution		Type II Error (β) Rates		
	A	B	D	Median	Siegal-Tukey
a)	N(0,1)	N(1,1)	.196	.184	.961
b)	N(0,1)	N(1,2)	.347	.347	.876
c)	N(0,1)	N(0,2.25)	.838	(.946)	.670
d)	N(0,1)	N(0,4)	.608	(.931)	.299
e)	N(0,1)	N(.5,.5)	.595	.641	.737
f)	N(0,1)	N(1,4)	.289	.499	.454
g)	N(0,1)	N(2,1)	0	0	.943
h)	N(1,1)	e(1)	.870	.856	.863
i)	N(1.09)	e(1)	.118	.673	.032
j)	N(1.5,1)	U(0,3)	.934	(.951)	.949
k)	e(1)	e(.5)	.592	.574	.917
l)	e(1)	U(0,2)	.830	.805	.818
m)	e(1)	U(0,3)	.519	.423	.943
n)	B(10,4)	U(0,1)	.012	.330	.040
o)	B(10,4)	B(8,6)	.003	.007	.988
p)	B(10,4)	B(16,6)	.904	.933	.866
q)	B(10,4)	L(0,1)	.176	.924	.025
r)	L(0,4)	U(0,4)	.617	.603	.872
s)	L(0,5)	U(0,4)	.115	.198	.977
t)	L(0,1)	L*(0,1)	.033	.034	.982

Note: ( ) around β-value for median test means A and B distributions had the same median; thus the test is inherently insensitive to their difference.

V. When m + n is not a Multiple of Four

The discussion about D so far has assumed that m + n = 4R, where R is some positive integer. We propose below some minor modifications of the procedure in cases where this assumption is not satisfied.

i) If m + n = 4R + 1, form the ranked, pooled sample, delete the median measurement, and then calculate the b_i's in the usual way.

ii) If m + n = 4R + 2, no measurements are deleted but the pooled sample is divided into its lowest R measurements, the next lowest R + 1, the next R + 1, and the highest R. The b_i's for the four groups continue to record the number of A-measurements in each.

iii) When m + n = 4R + 3, deleted the median measurement and then proceed as in the 4R + 2 case above.

The quantities S, d_o and d_I retain their usual definitions in terms of the b_i's.

Under the rules above, d_o and d_I continue under H_o to have means of zero. But the mean of S and the variances of all three quantities are different in the cases above than when m + n = 4R. Their new values under H_o, which are obtained straightforwardly, are listed below.

m + n = 4R + 1: N = m + n in all results below.)

$$\mu(S) = \frac{m(N-1)}{2N}$$

$$\sigma^2(d_o) = \sigma^2(d_I) = \frac{mn}{2N}$$

$$\sigma^2(S) = \frac{mn(N+1)}{4N^2}$$

m + n = 4R + 2:

$$\mu(S) = \frac{m(N-2)}{2N}$$

$$\sigma^2(d_o) = \frac{mn(3N-4)}{2N(3N+2)}$$

$$\sigma^2(d_I) = \frac{mn(N+2)}{2N(N-1)}$$

$$\sigma^2(S) = \frac{mn(N^2-4)}{4N^2(N-1)}$$

m + n = 4R + 3:

$$\mu(S) = \frac{m(N-3)}{2N}$$

$$\sigma^2(d_o) = \frac{m(N-3)[3nN(N+1) - 2m(N-1)]}{2N^2(N-1)(3N+2)}$$

$$\sigma^2(d_I) = \frac{m(N+1)[Nn(3N-1) + m(N-1)]}{2N^2(N-1)(3N-1)}$$

$$\sigma^2(S) = \frac{mn(N^2-9)}{4N^2(N-1)}$$

With the expressions above, the standardized variables $\tilde{S}$, $\tilde{d}_o$, and $\tilde{d}_I$ can be calculated directly. The statistic $D = \tilde{S}^2 + \tilde{d}_o^{\,2} + \tilde{d}_I^{\,2}$ continues to have an asymptotic distribution that is χ^2 with 3 degrees of freedom.

References

1. E.H. Lehmann, Nonparametrics. Statistical Methods Based on Ranks, Holden-Day, San Francisco, 1975.

2. P.J. Kim and R.I. Jennrich, "Tables of the Exact Sampling Distribution of the Two-Sample Kolmogorov-Smirnov Criterion," in Selected Tables in Mathematical Statistics, I, IMS, Markham (Chicago), 1970, pp. 79-170.

3. F. Wilcoxon, S.K. Katti, R. Wilcox, "Critical Values and Probability Levels for the Wilcoxon Rank Sum Test and the Wilcoxon Signed Rank Test," in Selected Tables in Mathematical Statistics, I, IMS, Markham (Chicago), 1970, pp. 177-260.

Appendix: The Null Distribution of the D Statistics for Moderately Large m and n Values

It is of interest how rapidly the probability distribution of D under H_o approaches its asymptotic form of χ^2 with 3 d.f. Simulation-based evidence presented below suggests that, even for moderately large m and n values, the actual distribution of D is rather well approximated by the asymptotic distribution. This circumstance makes somewhat superfluous the construction of elaborate tables that detail the significance levels of outcomes of D-tests.

For the special cases m = n = 24 and m = 24, n = 36, we randomly generated A and B data samples from the same (unit normal) probability distribution. (The samples came from the IMSL Computer Package on MIT's Multics Computer System). Then a computer program caluculated the D-value in each of 10,000 trials for the case m = n = 24, and in 10,000 separate trials for m = 24, n = 36. We observed how many of the obtained D-values fell below 6.25, 7.81, and 11.34, the 90th, 95th, and 99th percentiles of the χ^2 distribution with 3 d.f. These percentiles, of course, correspond to the 10%, 5% and 1% significance levels of a test of H_o and thus warrant particular attention. The results of the simulation appear in the chart below.

D-Statistics in 10,000 Simulations in Which H_o is Correct

	Number of Outcomes Below 6.25	Number of Outcomes Below 7.81	Number of Outcomes Below 11.34
Expected under asymptotic distribution of D	9000	9500	9900
Actual for case m=n=24	9038	9445	9921
Actual for case m=24, n=36	8986	9543	9919

In the chart, most of the observed fractions of D-values below the listed levels did not differ significantly from those predicted from D's asymptotic distribution. Even where the differences were statistically significant, they were still only a fraction of a percentage point. Taken together, the results clearly suggest that using a χ^2 table to approximate the significance level of a D-test outcome is not a procedure prone to serious inaccuracy, even at the upper tail of the null distribution where such inaccuracy might be most feared.

END