

# [www.normeddistributionfree.org](http://www.normeddistributionfree.org)

Stephen A. Kane<sup>1</sup> (Email: [skane@suffolk.edu](mailto:skane@suffolk.edu))

Assistant Professor of Finance, Frank Sawyer School of Management,  
Suffolk University, Boston, MA 02108

Aleksey V. Tazin (Email: [a\\_tazin@yahoo.com](mailto:a_tazin@yahoo.com))

Masters in Computer Science, Suffolk University, Boston, MA 02108

## Abstract

*Our web page gives techniques to assess the two sample problem (whether or not two different samples are generated by the same probability distribution or not). The web page computes normed and rank tests. All tests are distribution-free because they make no underlying distributional assumptions. The null hypothesis is that the two samples are generated by the same probability distribution. Investigators often make a trade-off between robustness and statistical power. Inasmuch as most investigations are about gathering evidence as opposed to making terminal decisions, it may make sense to run a battery of tests with different abilities as a sensitivity check. We endorse Meta statistics as an application of our techniques. We find our normed tests to be more powerful in detecting variance effects, while performing similarly for mean effects for pseudo-randomly generated normal and uniform distributions.*

---

<sup>1</sup> Corresponding author, Stephen A. Kane, Frank Sawyer School of Management,

Suffolk University, Boston, MA 02108. 617-973-5365, Fax 617-367-9307,

[skane@suffolk.edu](mailto:skane@suffolk.edu). We thank participants of Stanford's August 1996 Computational

Finance Conference, the October 2006 International Business and Economics Research

Conference in Las Vegas and UCLA's October 2006 Statistics Seminar for comments

that improved this manuscript. We are solely responsible for any errors.

Our web page gives techniques to assess the two sample problem (whether or not two different samples are generated by the same probability distribution or not). The web page computes both normed<sup>2</sup> and rank tests. All tests are distribution-free because they make no underlying distributional assumptions (Maritz 1995). The null hypothesis is that the two samples are generated by the same probability distribution. Each test uses an operationalizing assumption under the null hypothesis called exchangeability (all possible arrangements of observations of two samples that preserve the original sample size are equally likely). Our tests statistics are used to compute the relative frequency of shuffled arrangements bigger than or half those equal (within error tolerance) to the valued computed from the actual arrangement of observations (Efron 1982).

Investigators often make a trade-off between robustness and statistical power. Inasmuch as most investigations are about gathering evidence as opposed to making terminal decisions, it may make sense to run a battery of tests with different abilities as a sensitivity check (Leamer and Leonard, 1983).

By assessing the statistical power of our tests, we illustrate some of their strengths of weaknesses. Our non-parametric tests are more robust, because they do not employ all the data values in the computation of the test statistic. If there are serious data integrity concerns, then non-parametric tests may yield better statistical inferences. We find our

---

<sup>2</sup> Our norms define distance functions for empirical distributions functions when we specify right continuity. Please see Kane (1998) for details. A norm  $\|z\|$  is defined on a space  $Z$ , whenever for every  $z$  in  $Z$ , there is a non-negative number such that: (1)  $\|z\| = 0$  if and only if  $z=0$ . (2)  $\|\alpha z\| = |\alpha| \|z\|$  For all real numbers,  $\alpha$ . (3)  $\|x + y\| \leq \|x\| + \|y\|$  For all  $x$  and  $y$  in  $Z$ . Norms correspond to distance by  $\|x - z\|$ .

normed tests to be more powerful in detecting variance effects, while performing similarly for mean effects for pseudo-randomly generated normal and uniform distributions.

We have organized the paper as follows: We state the two-sample problem. We define our test statistics. We describe our test procedures. We discuss clustering concerns. We show how to use our web page and explain some of our programming decisions. We endorse Meta statistics as an application for our techniques. We illustrate the statistical power of our test procedures. Finally, we conclude.

### *Two-Sample Problem*

We suppose that there are two sets of observations. An investigator wants to make inferences about whether or not the different observations are drawn from the same underlying distribution or not. Ideally the two distributions result from a controlled experiment where investigators place subjects into classifications by employing randomization. In finance, it often happens that the subjects themselves determine their classification based on their decisions. That is, the subjects self-select their classification.<sup>3</sup> Consequently, investigators do not know whether characteristics of their subjects led to their choices, and that this and not the classification per se, is the reason for differences observed between groups (Kane 2004). Even though financial investigators might not be able to isolate the effects of a single variable, they might be able to form “match pair” samples. Investigators might use randomization to place one half of matched pair into a treatment group (to receive treatment) and the other half into a control group (to receive no treatment). An advantage of “match pair” design over

---

<sup>3</sup> For instance, marketing studies how consumers select products.

ordinary least-square regression techniques is that randomization design helps to filter out the effects of omitted variables that may affect the process under study because these variables are equally likely to appear in the treatment group or control group (Kane 2004).

The criterion of robustness emphasizes the value of methods that show insensitivity in results to small changes in underlying assumptions. Statistical tests calculate the p-value, the probability of a test statistic realization assuming the null and the operationalizing subsidiary assumptions. The p-value is a joint probability conditional on both the null and all subsidiary hypotheses (Kempthorne 1976; Kane 1995). Inferences with fewer and less stringent subsidiary assumptions are superior, because a small p-value may result from a combination of inappropriate subsidiary hypotheses with an appropriate null hypothesis (Kane 1995, 2004).

### *Test Procedures*

To assess the robustness of a testing method, investigators might devise and perform an appropriate battery of sensitivity tests (Leamer and Leonard, 1983). Inasmuch as most investigations are about gathering evidence as opposed to making terminal decisions, it may make sense to run a battery of tests with different abilities. For investigators concerned with the two sample problem, [www.normeddistributionfree.org](http://www.normeddistributionfree.org) provides seven different distribution free test statistics, namely:  $L_1, L_1^+, L_1^-, L_\infty, \text{Min}(W^+, W^-), W^+,$  and  $W^-$ .

By working with empirical distribution functions, the method makes no assumption about the underlying probability distribution from which the two samples are drawn.

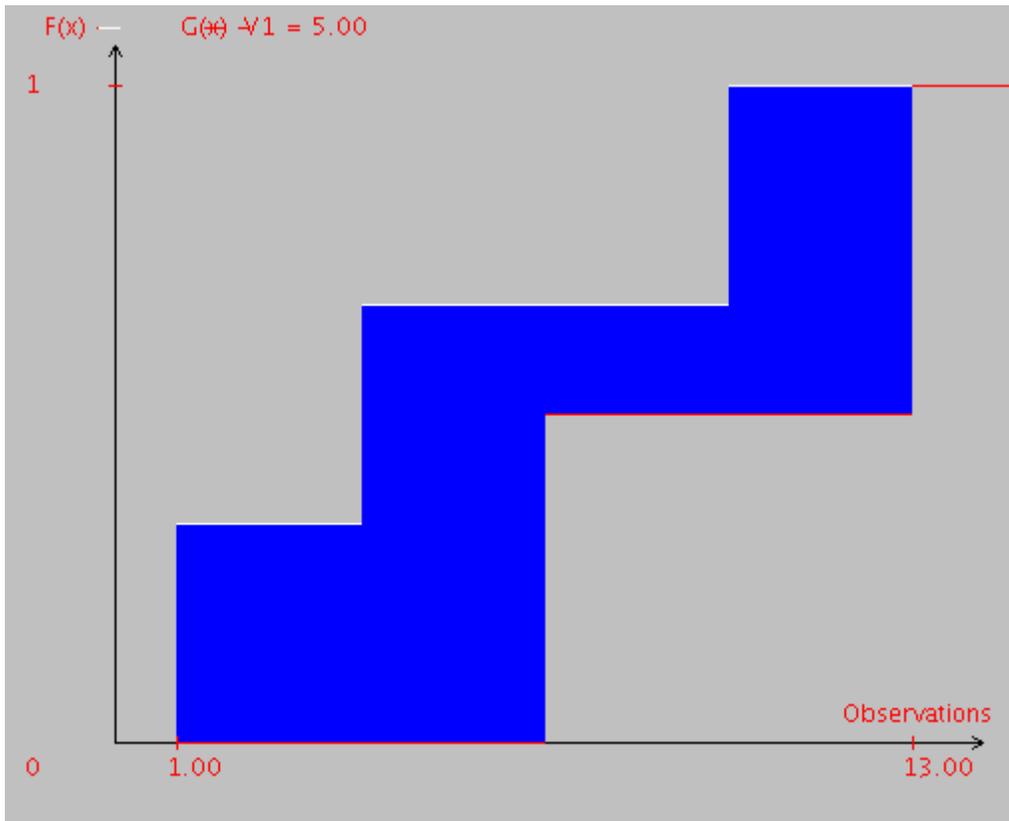
The null hypothesis is that the two samples are generated by the same probability distribution. Each test uses an operationalizing assumption under the null hypothesis called exchangeability. Under the null hypothesis, all shuffled arrangements of observations of both samples into synthetic samples which keep the same size as the original samples have an equal probability of occurring. The p-values are the relative frequency of shuffle test statistics bigger than or half of those tests that are equal (within the error tolerance) to the realized test statistic, the value of the test statistic that corresponds to the actual observations (Efron 1982).

To make the statistical techniques visual and more intuitive for users, we provide graphs of the empirical distributions functions with the respective test statistics:  $L_1$ ,  $L_1^+$ ,  $L_1^-$ , and  $L_\infty$ . We have constructed our tables and graphs so that users may copy them and insert them into Microsoft word documents.

$L_1$

The  $L_1$ -norm test statistic uses all the parameter values in its computation. It assesses the overall agreement between the two empirical distribution functions by using the area between them. Two empirical distribution functions are close in the  $L_1$ -norm sense when the area between them is small. Consequently, investigators might use this test statistic with the broadest alternative hypothesis:  $F(x) \neq G(x)$ . We choose to use the  $L_1$ -norm test statistic over  $L_2$ -norm test statistic, because it is superior numerically. For instance, the difference between two consecutive observations may be a small positive number less than one. By squaring a small difference, we make it smaller still. This exacerbates significant digit and round-off problems. With the absolute value

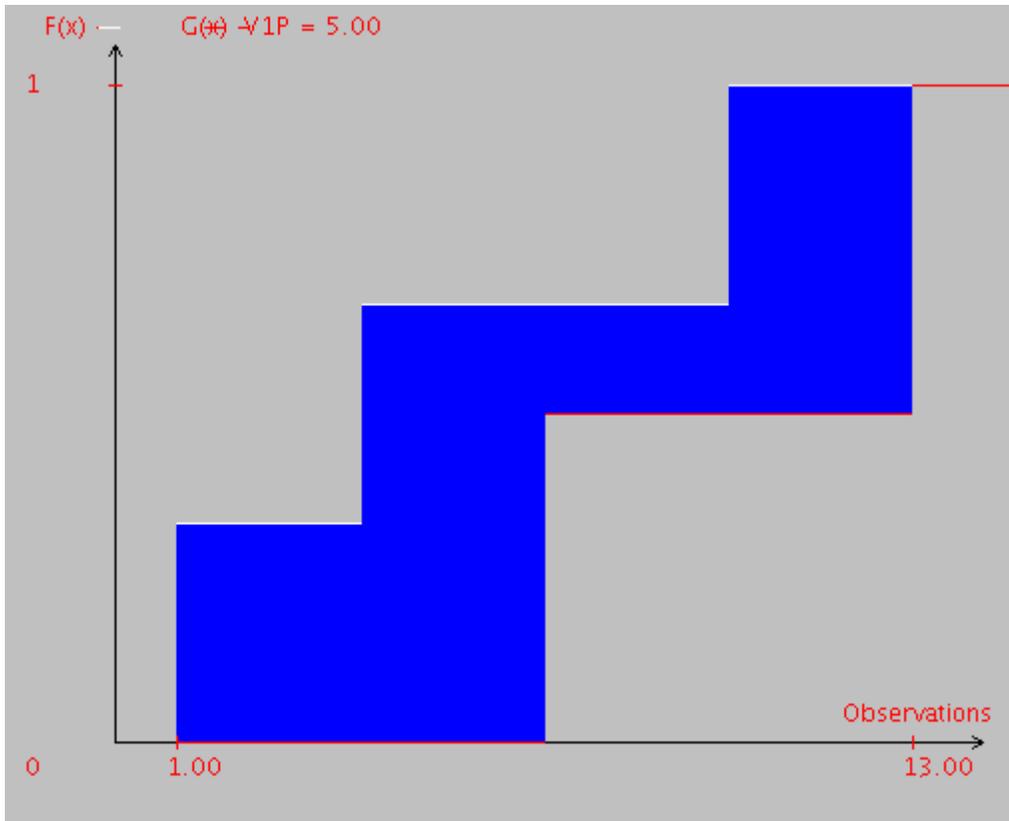
function, however, we cannot differentiate it in a neighborhood of zero. This may make estimating parameter values for other applications problematic. For example, in minimizing least squares in a regression, we take partial derivatives with respect to variables and set them equal to zero and solve for parameter estimates.



$L_1^+$

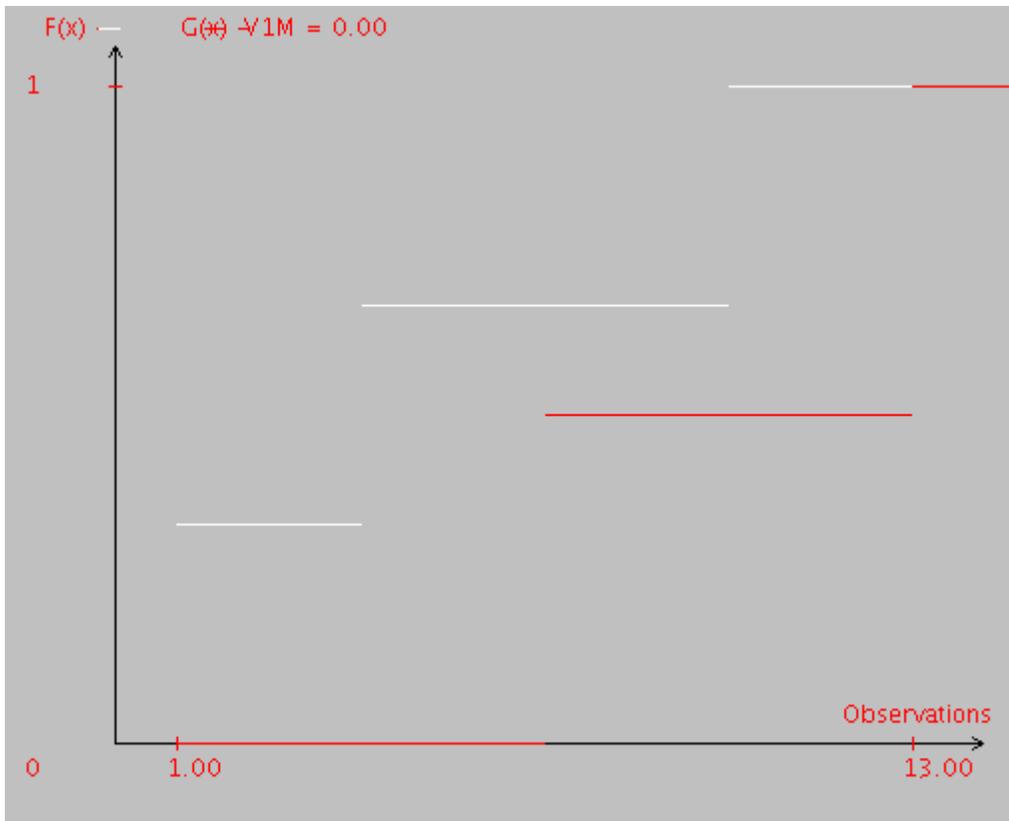
The  $L_1^+$  - norm test is parametric. The test statistic is the area where the first distribution function is above the second. Technically this test statistic is not a norm, since two empirical distribution functions could be distance zero apart, but not be equal when they differ where the second distribution is above the first. Nevertheless, it is a limit of a norm. We may multiply the area where the second distribution is above the first by an arbitrarily small positive number,  $\varepsilon > 0$ , and take the limit as  $\varepsilon \rightarrow 0$ . Investigators

might use this to test with an alternative hypothesis such as the first distribution is smaller stochastically,  $F(x) \ll G(x)$ .



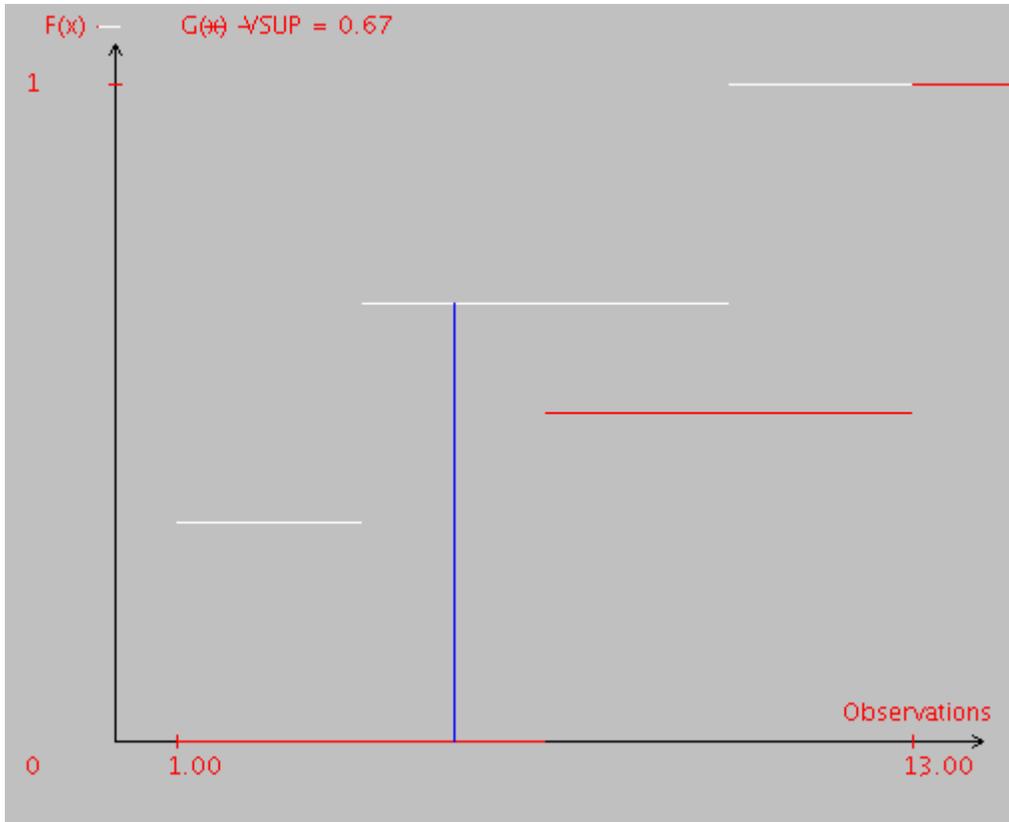
$L_1^-$

The  $L_1^-$ -norm test is parametric. The test statistic is the area where the second distribution function is above the first. Technically this test statistic is not a norm, because the two empirical distribution functions could be distance zero apart, but not be equal when they differ where the first distribution is above the second. It is a limit of a norm, however. Investigators might use this test statistic with an alternative hypothesis such as the first distribution is bigger stochastically,  $F(x) \gg G(x)$ .



$L_\infty$

The  $L_\infty$ -norm, or Kolmogorov-Smirnov, test is non-parametric, because the test statistics is not an explicit function of the parameter values, since it only uses the relative rank of the observations. The test statistic is the maximum distance between the two empirical distribution functions. It assesses with the point of worst agreement between the two empirical distribution functions. Consequently, investigators might use it to test with the broadest alternative hypothesis:  $F(x) \neq G(x)$ .



*Min (W+, W-), W+, and W-*

All three rank order tests are non-parametric, because these test statistics are only functions of the relative ranks and not the actual parameter values of the observations.

We rank all the observations in the two samples from 1 to  $j = (m + n)$  in the two samples by using  $<$  for the real numbers. For tied observations, we use the following convention:

If there are two tied observations, then we average the two integer ranks in question and assign the average to both values. Similarly if there are  $k$  ties, then we average the  $k$

ranks in question, and assign the average rank to all  $k$  observations. We let  $W^-$  equal the sum of the assigned ranks to the first data set, and  $W^+$  equal the sum of the assigned

ranks to the second. Investigators might use  $\text{Min} (W^+, W^-)$  tests with an alternative such as  $F(x) \neq G(x)$ , and the  $W^+$  test with an alternative hypothesis such as  $F(x) \ll G(x)$ ,

stochastically. Since  $\frac{j(j+1)}{2} - W^- = W^+$ , the p-value corresponding to  $W^-$  is the complement of the p-value corresponding to  $W^+$ . Investigators might use  $W^-$  in a tests with an alternative hypothesis such as  $F(x) \gg G(x)$ , stochastically.

Our previous graphs correspond to a first data set of  $\{1,4,10\}$  and a second data set of  $\{7,13\}$ . We compute our p-values by exhaustively considering all ten possible arrangements of the observations in the second data set in the table below:

	$L_1$	$L_1^+$	$L_1^-$	$L_\infty$	$W^+$	Min	$W^-$
{1,4}	7.5	0	7.5	1	3	3	12
{1,7}	5	0	5	0.67	4	4	11
{1,10}	3.5	0.5	3	0.5	5	5	10
{1,13}	4	2	2	0.5	6	6	9
{4,7}	4.5	1	3.5	0.67	5	5	10
{4,10}	3.5	3	0.5	0.5	6	6	9
{4,13}	3.5	3	0.5	0.5	7	7	8
{7,10}	4.5	3.5	1	0.67	7	7	8
{7,13}	5	5	0	0.67	8	7	7
{10,13}	7.5	7.5	0	1	9	6	6
Greater	0.2	0.1	0.8	0.2	0.1	0	0.8
Equal	0.2	0.1	0.2	0.4	0.1	0.3	0.1
P-value	0.3	0.15	0.9	0.4	0.15	0.15	0.85

### Clustering

A weakness of our methods is clustering of test statistic shuffle values near the realized test statistic value. Such clustering prevents an accurate determination of a p-value, because it is a relative frequency of shuffles. Due to the discreteness of non-parametric test statistics, clustering may be a bigger problem for:  $L_\infty$ , Min ( $W^+$ ,  $W^-$ ),  $W^+$ , and  $W^-$ . Nevertheless, the  $L_1$ ,  $L_1^+$ , and  $L_1^-$  are not immune. A lack of significant digits and even computer round-off error may induce clustering.

The web page allows users to select the fixed error tolerance, that is, the number significant digits after the decimal place to be used in computations: 0, 1, 2, 3, 4, and 5. We remind users that the number of significant digits is the lowest number of significant digits for any variable used in calculations. Furthermore, a calculated variable that is a proxy for a desired variable may have even less significant digits than its measurements.

For rounding purposes only, we carry an additional decimal digit in computing the  $L_1$ ,  $L_1^+$ , and  $L_1^-$  test statistics. We count half of the shuffled test statistics within a fixed error tolerance of the value of the realized test statistic in the relative frequency counts when computing p-values. For the above test statistics, we report the relative frequency of test statistic shuffles within the fixed error tolerance so that users may assess the severity of clustering. ‘Error -’ is the relative frequency of shuffles with test statistics below the realized test statistic, but just outside the error tolerance. ‘Error +’ is the relative frequency of shuffles with test statistics above realized test statistic, but just outside the error tolerance. ‘=’ is the relative frequency of shuffles with test statistics that are within the error tolerance of the realized test statistic

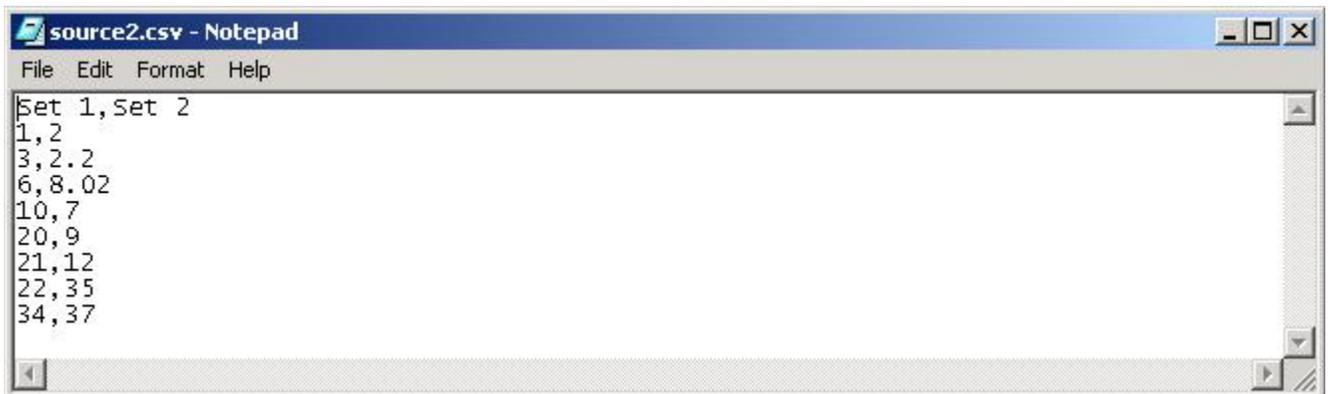
Error-: 0.0010	<b>P-value: 0.3008</b>	Error+: 0.0011
	=: 0.0011	

For concreteness, we suppose that we are computing with one decimal place of accuracy for our fixed error tolerance and that our procedure has computed the realized  $L_1$ -norm test statistic to be 34.59. We use the value 34.6. We count in the numerator all shuffled test statistics that lie in  $[34.55, 34.65)$  in ‘=.’ We count in the numerator all shuffled test statistics that lie in  $(34.45, 34.55)$  in ‘Error -.’ We count in the numerator all shuffled test statistics that lie in  $(34.65, 34.75)$  in ‘Error +.’

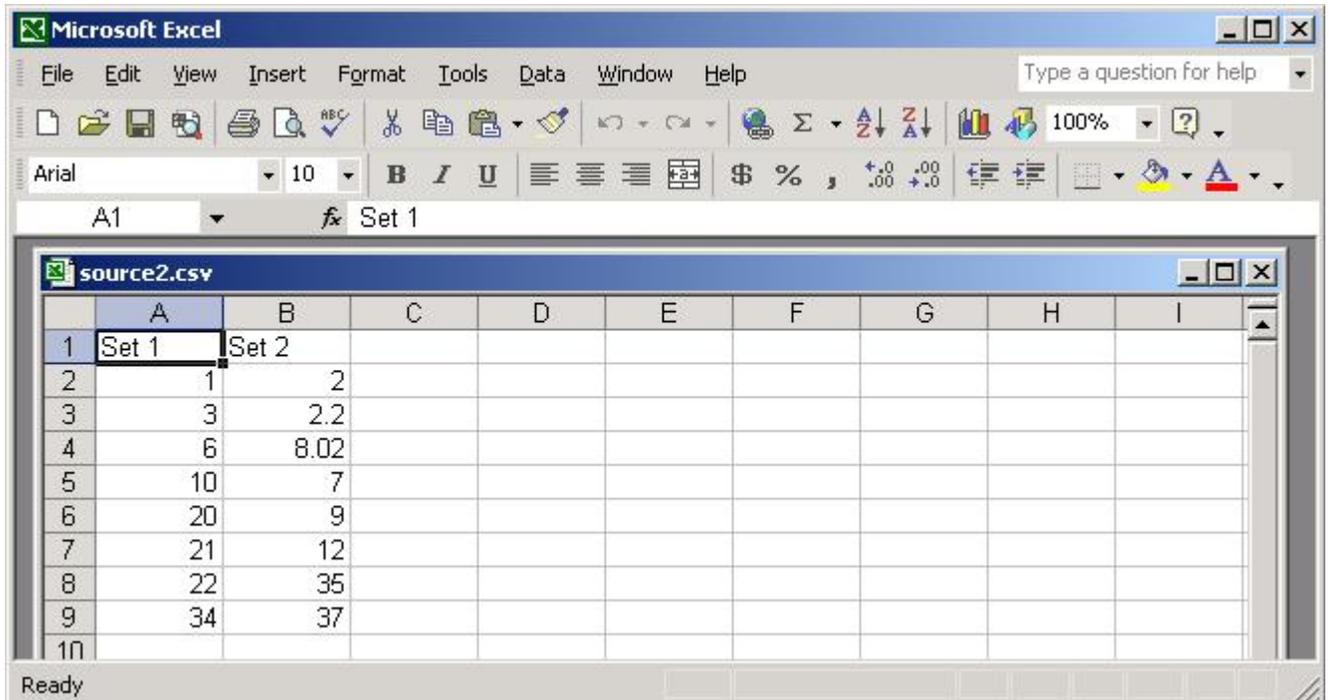
When computing with our non-parametric test statistics, we use the exact number of decimal places of accuracy after rounding by employing the convention that we round all digit 5s to the corresponding leftmost even digit. For instance, we consider 8.55 and 8.65 with one decimal place of accuracy to be both 8.6. This is one reason our procedure computes the  $L_1$ ,  $L_1^+$ , and  $L_1^-$  test statistics separately from the  $L_\infty$ ,  $\text{Min}(W^+, W^-)$ ,  $W^+$ , and  $W^-$  test statistics. The test statistics for the rank order tests are integers, but not for  $L_\infty$ . Consequently, we multiply it by  $10^5$  and truncate it by setting it equal to an integer. We then use integers to perform relative frequency counts as in the rank order tests.

### *Data Input*

We choose a data upload for our web page. Our upload works with many internet browsers, and hopefully with future changes, because data uploading is a relatively crude technique. For our data files, we chose the \*.csv (comma separated value). Many spreadsheet programs, for instance, the ubiquitous Microsoft Excel program, support this format.



We note that the two data sets do not have to be the same size. Users may create a file by opening a new spreadsheet, entering the two sets with header values Set 1 and Set 2, and saving the spreadsheet as MS-DOS CSV file.



Users may use the Browse button and locate on their hard drive where the file was saved. Then by using the Compute Test button, users can upload the file to the web page. Alternatively, users may download a sample input file to their local hard drive by right clicking the link at the bottom of the web page and edit this file in Notepad or Excel.

We dynamically define the dimensions of our arrays, since Java is an object oriented programming language. This allows us to accommodate a large number of observations in each sample without unnecessarily increasing the memory and central process unit demands for smaller samples. Unfortunately, we need to alert users that our server may not initially handle large data sets and generate an error message. Fortunately, upon resubmission of the data set, the server seems to allocate sufficient resources and perform properly.

We also provide users with the ability to download uniform and normal deviates, because users might want to use these distributions as benchmark distributions to test against other distributions. After clicking the button, an html page will appear with the iteration number and random deviate as columns. Users may copy these columns insert them into spreadsheet programs. To parse the columns in some Excel programs, users may use the paste special (Unicode) option.

From an evidentiary perspective, an investigator only needs a p-value computed with a few decimal places of accuracy. This only requires a reasonably representative sample of shuffled arrangements. Demanding the convergence of the pseudo-random p-value to the exhaustive p-value is misguided. We say a series of pseudo-random generated p-value converges to an exhaustive p-value if for every  $\varepsilon > 0$ , there exists an  $N$  such that if  $n > N$  then  $|p_{pseudo_n} - p_{exhaustive}| < \varepsilon$ . We may observe  $|p_{pseudo_n} - p_{exhaustive}|$  decreasing with  $n$  increasing on small samples. But, any finite sequence of an infinite sequence is irrelevant with respect to its limit. Nevertheless, such observations on small samples lend assurance regarding the representativeness of our pseudo-random samples in the set of shuffled arrangements, and hence, assurance concerning our estimated p-values on larger samples.

We allow the users to select the number of random shuffles: 100,000, and 500,000, and 1,000,000. When we increased the number of iterations from 100,000 to 1,000,000 the difference in p-values for the same data sets varied less than 0.001 for 100 versus 100 pseudo-random normal and uniform distribution tests. We use the algorithm of Wilf (1989) to generate the pseudo-random subsets. We employ Knuth's pseudo-random

number generator (Flannery, Press, Teukolsky, and Vetterling, 2002), because we found other generators repeated after a small number of iterations. Knuth's generator gave the best estimates the p-values from exhaustive sampling. For interested researchers, our web page performs both pseudo-random and exhaustive sampling when the total number of shuffled arrangements,  $C_m^{m+n} = \frac{(m+n)!}{(m)!(n)!}$  is less than 100,000.

### *Meta Statistics*

Meta Statistics is a natural application of our techniques. We acknowledge a 1996 discussion with George Papanicolau for this fine suggestion. Subsequent computer processing advances have made this idea practical.

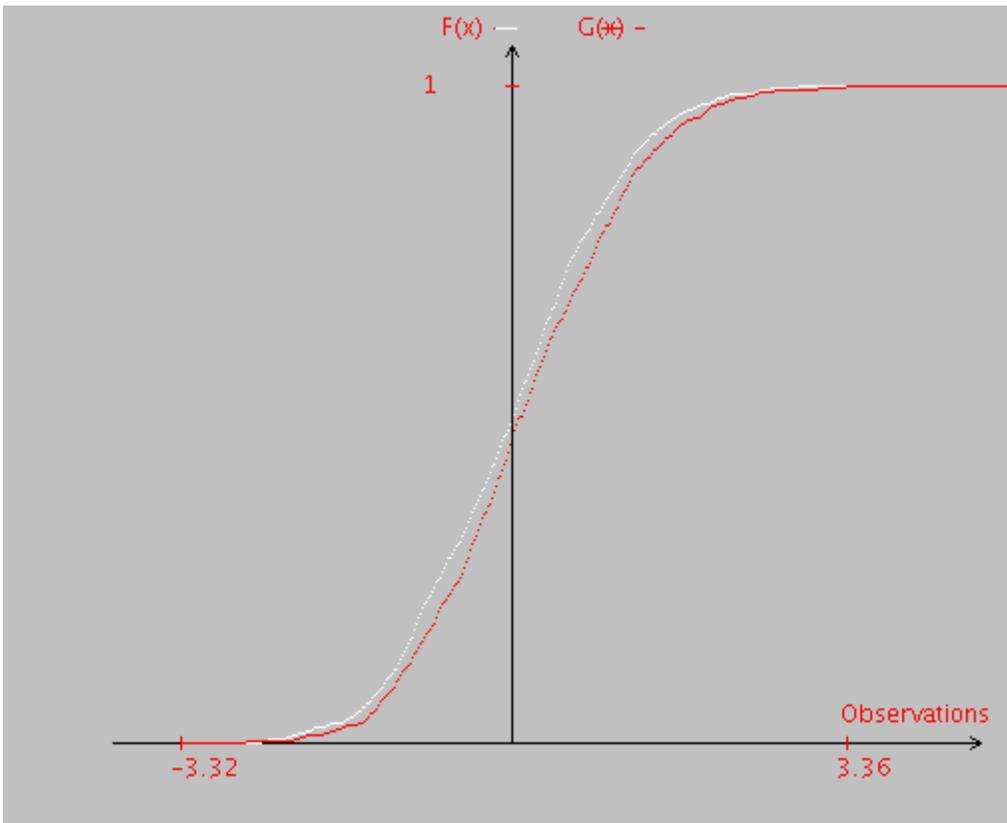
We let  $test_i$  denote a particular test of the equality of two distributions and  $H_{0i}$  denote the null hypothesis. When the number of shuffled test statistics tied with the realized test statistic is negligible, the p-value of  $test_i$  is approximately distributed uniform (0,1). Thus, we may make the operationalizing subsidiary assumption that the p-value of  $test_i$  is distributed uniform (0,1). We let the main hypothesis be  $H_0 = \bigcap_{i=1}^k H_{0i}$ , that k-pairs of observations are drawn the same underlying distribution. The alternative hypothesis is that at least one of the k-pairs has observations that are not drawn from the same distribution. Thus, we run tests of the p-values against a pseudo-randomly drawn uniform distribution to assess  $H_0$ .

## *Statistical Power*

Statistical power is the probability of rejecting a false null hypothesis given that a specific alternative hypothesis is true. The choice of which alternative hypotheses to examine depends on the applications in question. We examine the statistical power of our testing methods for mean and variance effects of the normal distribution. Then, we perform analogous assessments for our test for the uniform distribution. Relative to the ranked tests, our normed tests have more statistical power for variance effects while generally exhibiting similar power for mean effects. We use the pseudo-random number generators provided by the web page to generate our data sets.

We consider a  $N(0,1)$  versus a  $N(0.1,1)$  with data set sizes of 1,000 versus 1,000 using four decimal places of accuracy.

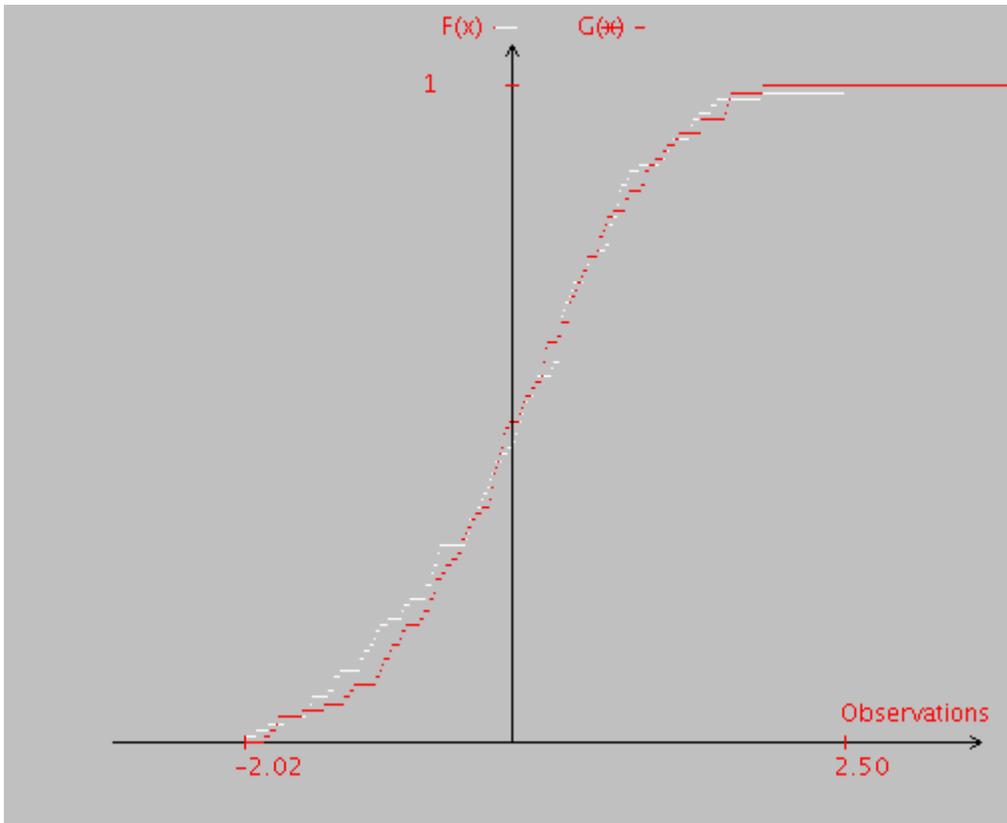
	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.0018	0.0000	0.0000	0.0000
<b>L1+</b>	0.9905	0.0021	0.0023	0.0026
<b>W+</b>	0.9982	0.0000		
<b>L1-</b>	0.0009	0.0000	0.0000	0.0000
<b>W-</b>	0.0018	0.0000		
<b>L<math>\infty</math></b>	0.0274	0.0034		
<b>Min</b>	0.9967	0.0000		



Here, both the normed and rank tests exhibit a lot of statistical power.

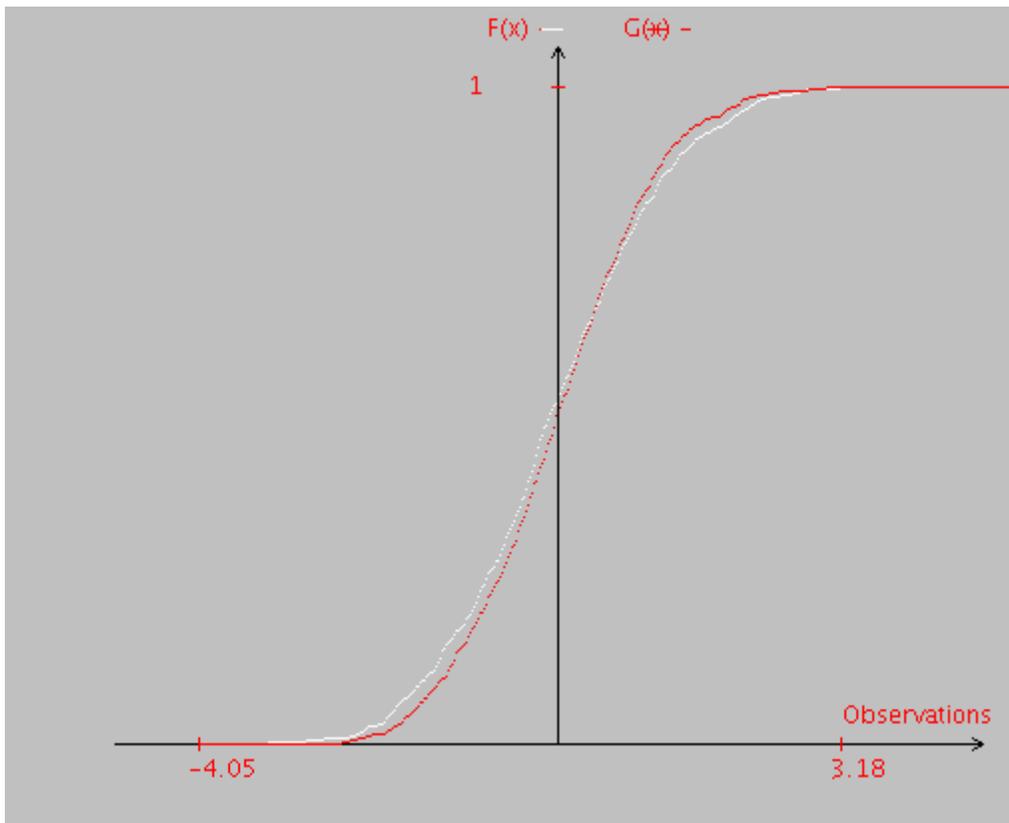
When we drop the data set sizes to 100 versus 100, the statistical power lessens substantially.

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.9491	0.0004	0.0003	0.0005
<b>L1+</b>	0.7569	0.0007	0.0010	0.0010
<b>W+</b>	0.5921	0.0008		
<b>L1-</b>	0.4726	0.0005	0.0006	0.0004
<b>W-</b>	0.4079	0.0008		
<b>L<math>\infty</math></b>	0.8603	0.0896		
<b>Min</b>	0.1876	0.0018		



We examine a  $N(0,1.1)$  versus  $N(0,1)$  with 100,000 iterations, four decimal places of accuracy and a data set sizes of 1,000 versus 1,000.

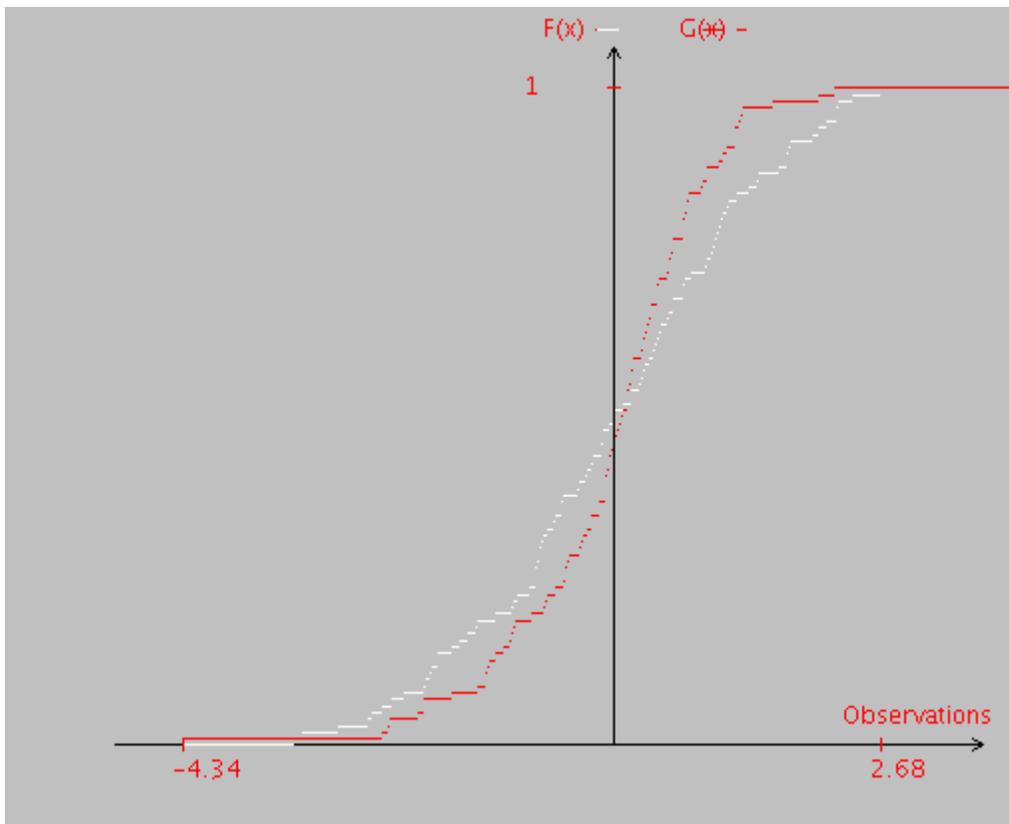
	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.0170	0.0001	0.0001	0.0001
<b>L1+</b>	0.3008	0.0011	0.0010	0.0011
<b>W+</b>	0.8425	0.0000		
<b>L1-</b>	0.0528	0.0003	0.0003	0.0002
<b>W-</b>	0.1575	0.0000		
<b>L<math>\infty</math></b>	0.2083	0.0181		
<b>Min</b>	0.6832	0.0000		



The  $L_1$ -norm test statistic test exhibits good statistical power, but the rank tests do not. This is no surprise, because the ranks are affected symmetrically.

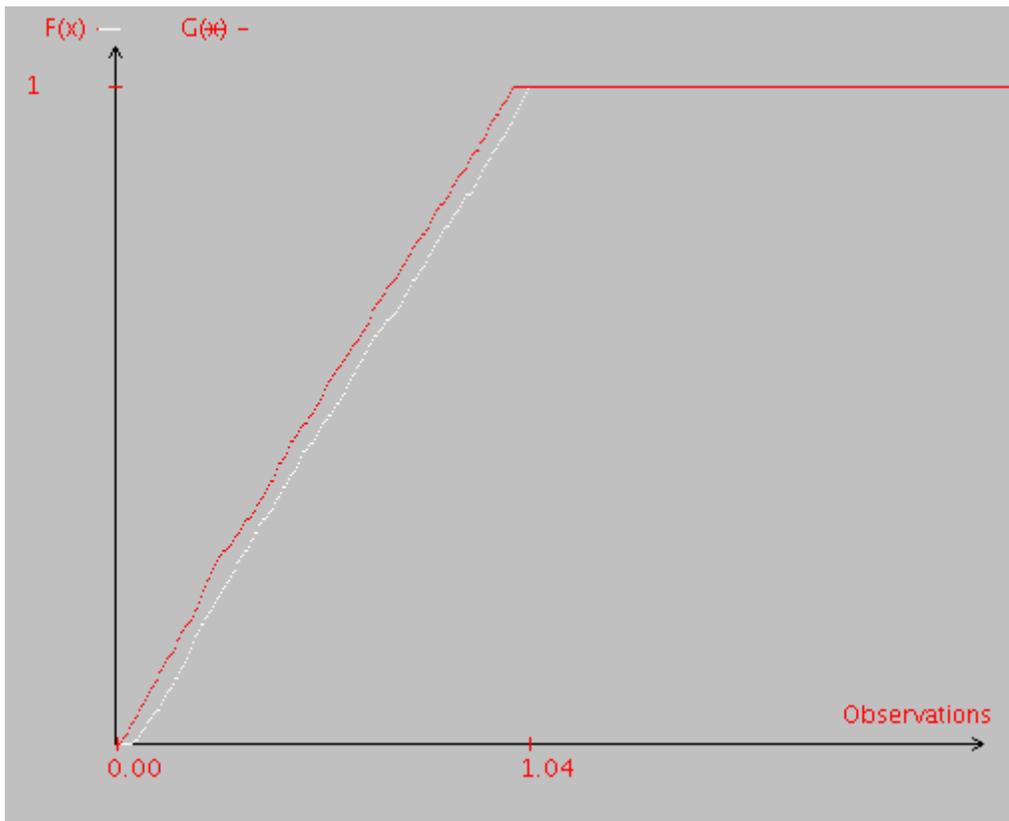
The statistical power diminishes substantially when we reduce the data set sizes to 100 versus 1000. Consequently, we present  $N(0,1.3)$  versus  $N(0,1)$ . We note that the  $\text{Min}(W_+, W_-)$  has a low p-value, because the values of  $W_+$  and  $W_-$  are close due to the symmetrical affect on the ranks for this pseudo-random realization.

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.0402	0.0001	0.0001	0.0001
<b>L1+</b>	0.1615	0.0002	0.0002	0.0003
<b>W+</b>	0.4738	0.0010		
<b>L1-</b>	0.2044	0.0002	0.0002	0.0002
<b>W-</b>	0.5267	0.0010		
<b>L<math>\infty</math></b>	0.2463	0.0705		
<b>Min</b>	0.0540	0.0020		



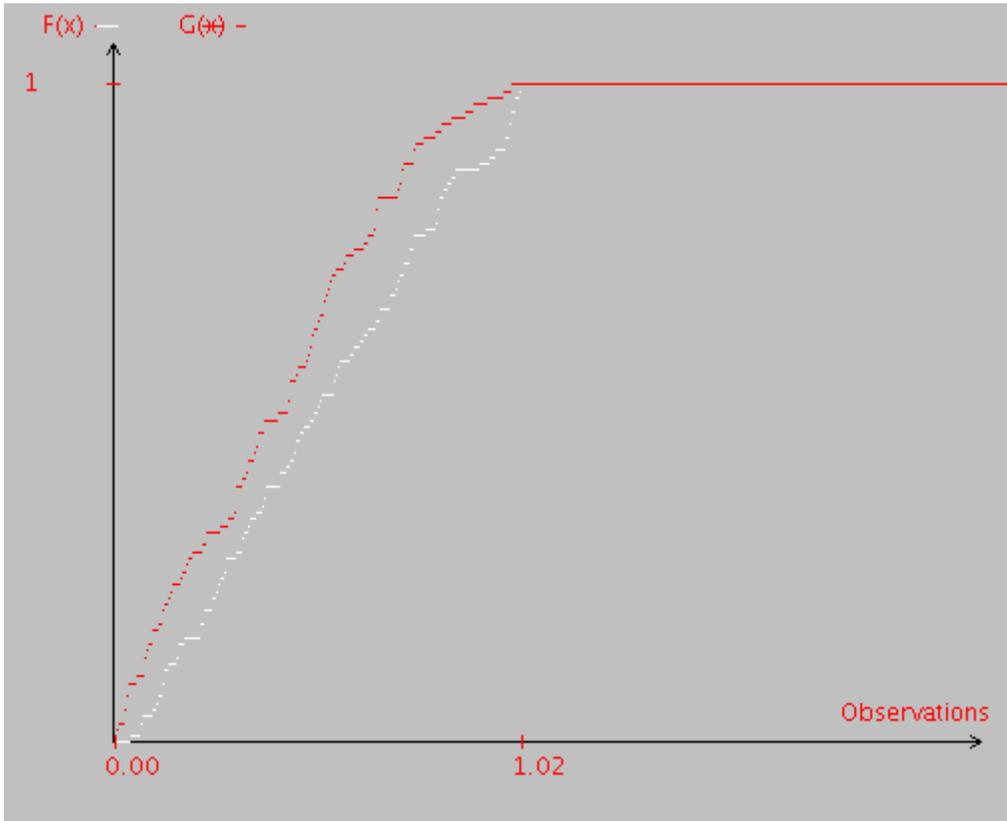
We consider 100,000 iterates, 1,000 versus 1,000, four decimal places of accuracy with a  $U(0,1) + 0.0398$  versus  $U(0,1)$ . We use 0.0398 because it is the area under a  $N(0,1)$  density between 0 and 0.1 with four decimal places of accuracy.

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.0004	0.0000	0.0000	0.0000
<b>L1+</b>	0.0002	0.0000	0.0000	0.0000
<b>W+</b>	0.0002	0.0000		
<b>L1-</b>	0.9896	0.0208	0.0000	0.0288
<b>W-</b>	0.9998	0.0000		
<b>L<math>\infty</math></b>	0.0340	0.0043		
<b>Min</b>	0.9996	0.0000		



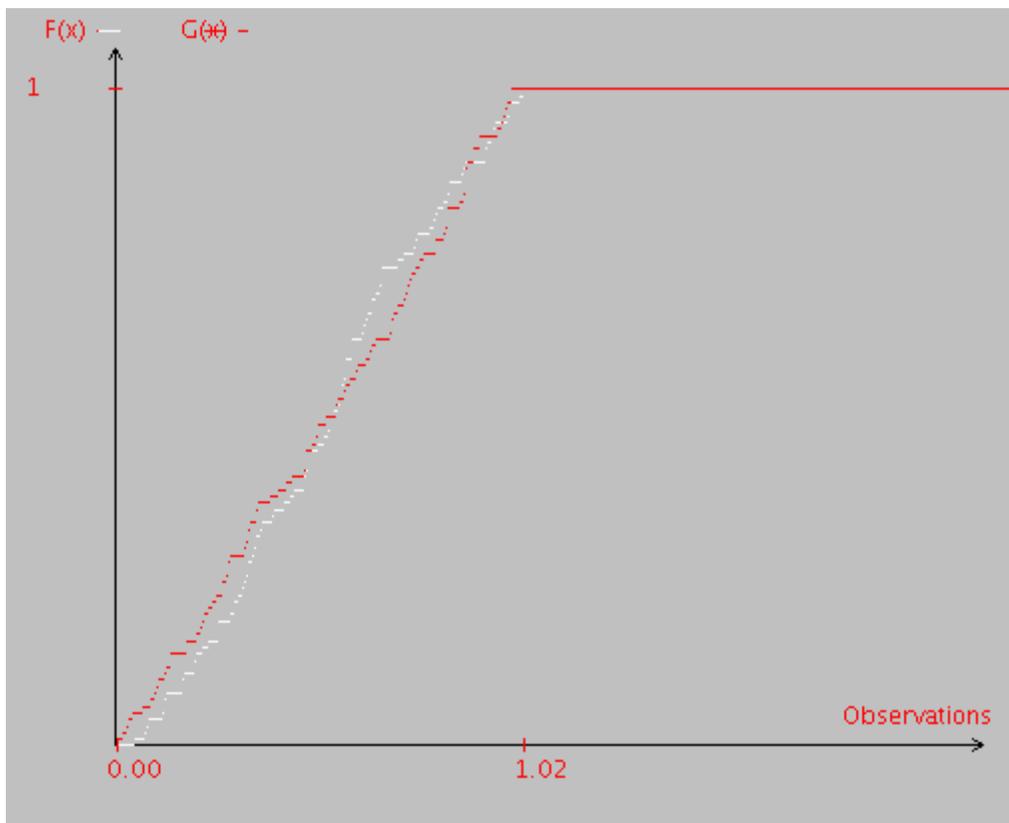
We next examine two different data sets of size 100 versus 100. We see that the two pseudo random sample realizations are quite different. This may make our resulting p-values dependent on the “luck of the draw.” Consequently, it is difficult to assess the statistical power of our test procedures for 100 versus 100. Thus, we recommend using a larger number of pseudo random deviates when performing benchmark testing.

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.0053	0.0000	0.0000	0.0000
<b>L1+</b>	0.0028	0.0000	0.0000	0.0000
<b>W+</b>	0.0035	0.0000		
<b>L1-</b>	0.9917	0.0166	0.0000	0.0137
<b>W-</b>	0.9930	0.0001		
<b>L<math>\infty</math></b>	0.0446	0.0174		
<b>Min</b>	0.9930	0.0001		



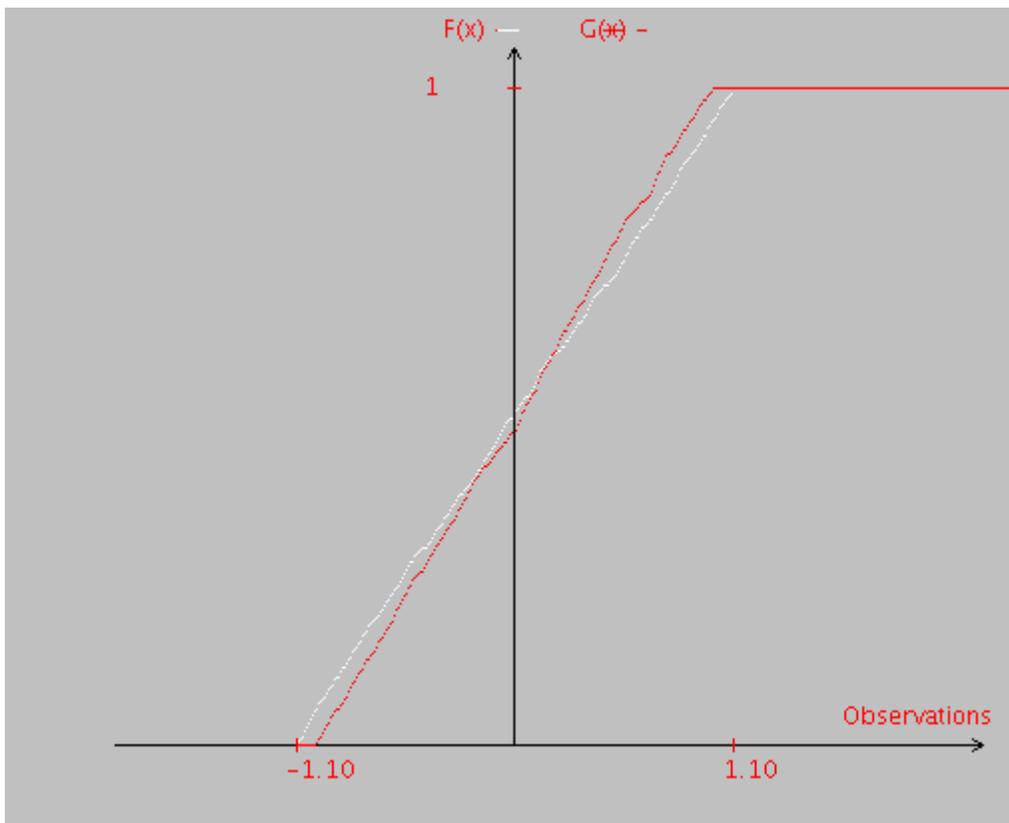
We give the second realization.

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.4747	0.0023	0.0024	0.0025
<b>L1+</b>	0.3591	0.0016	0.0018	0.0018
<b>W+</b>	0.4614	0.0010		
<b>L1-</b>	0.4571	0.0016	0.0014	0.0014
<b>W-</b>	0.5380	0.0011		
<b>L<math>\infty</math></b>	0.5252	0.1142		
<b>Min</b>	0.0809	0.0021		



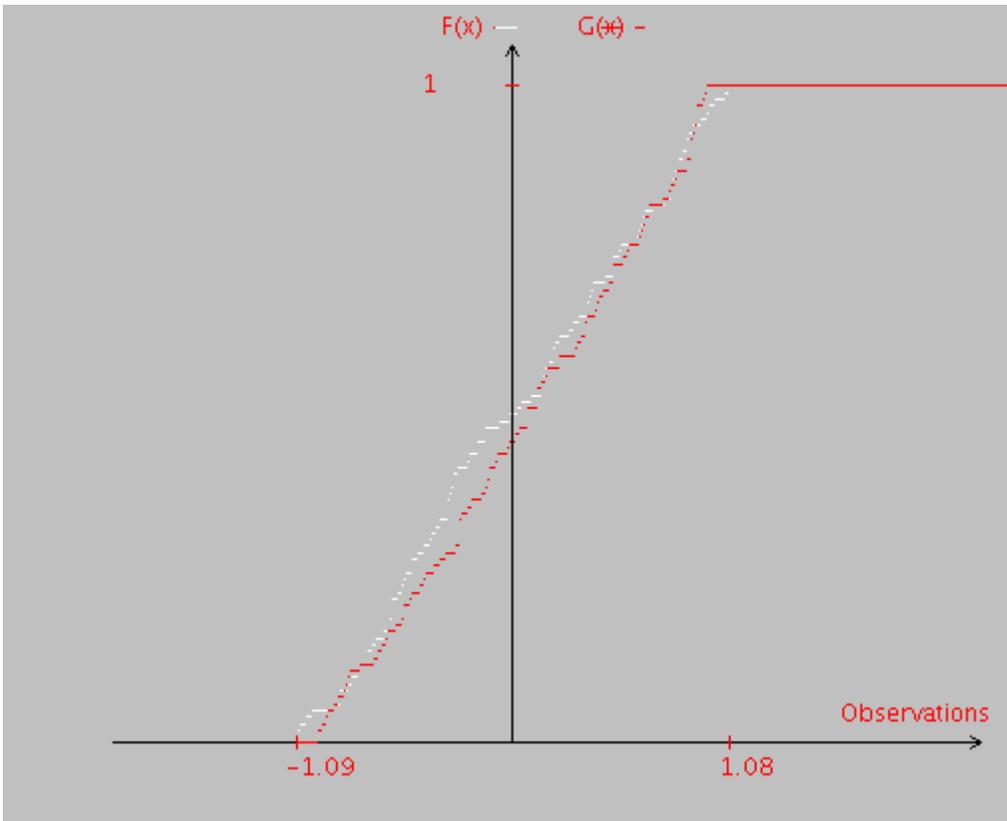
For 100,000 iterates, four decimal places of accuracy, and 1000 versus 1000, we investigate  $U(-1.1,1.1)$  versus  $U(-1.0,1.0)$ .

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.0085	0.0001	0.0001	0.0001
<b>L1+</b>	0.1045	0.0008	0.0006	0.0007
<b>W+</b>	0.5195	0.0000		
<b>L1-</b>	0.0970	0.0009	0.0008	0.0007
<b>W-</b>	0.4805	0.0000		
<b>L<math>\infty</math></b>	0.0503	0.0060		
<b>Min</b>	0.0426	0.0000		



Finally, we consider a data set of 100 versus 100. Again, we lack statistical power and our p-values may depend on the “luck of the draw.”

	<b>P-value</b>	<b>Equal</b>	<b>Error-</b>	<b>Error+</b>
<b>L1</b>	0.7065	0.0012	0.0014	0.0013
<b>L1+</b>	0.8347	0.0023	0.0023	0.0022
<b>W+</b>	0.7562	0.0007		
<b>L1-</b>	0.2875	0.0005	0.0005	0.0005
<b>W-</b>	0.2442	0.0009		
<b>L<math>\infty</math></b>	0.4171	0.1005		
<b>Min</b>	0.5106	0.0014		



## *Conclusions*

The two sample problem is broad and across many disciplines. We look forward to the surge of other minds in finding and implementing other applications for their research problems. Thus, we made our techniques easy to use and available on the World Wide Web.

We argued the importance of robustness. We explained our statistical procedures and some of our programming choices. We discussed some of the strengths and weaknesses of our different test statistics. We endorsed Meta statistics as a natural application of our techniques. We recommended using a large number of pseudo-random deviates to generate benchmark distributions used for testing, to reduce the likelihood that a low p-value being the result of the “luck of the draw.” We found our normed tests to be more powerful in detecting variance effects, while performing similarly to rank tests for mean effects for pseudo-randomly generated normal and uniform distributions.

## References

- Efron, Bradley, 1982. The jackknife, the bootstrap, and other resampling plans.  
Philadelphia PA: Society for Industrial and Applied Mathematics.
- Press, William, Saul Teukolsky, William Vetterling, and Brian Flannery, 2002.  
*Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge  
University Press, New York.
- Kane, Stephen, 1995, "APT: Testability with an Undetermined Number of Factors,"  
*Advances in Investment Analysis and Portfolio Management*, Vol. 3, pp. 249-  
258.
- Kane, Stephen, 1998, "Normed Distribution-Free Testing for Identically Distributed  
Residuals," *Advances in Financial Planning and Forecasting*, Vol. 8, pp. 95-  
111.
- Kane, Stephen, 2004, "Scientific Methods in Finance," *International Review of Financial  
Analysis*, Vol. 13, pp. 105-118.
- Kempthorne, Oscar, 1976, "Of What Use are Tests of Significance and Hypothesis?"  
*Communications in Statistics – Theory and Method*, A5(8): pages 763-77.
- Leamer, Edward and Herman Leonard, 1983, "Reporting the Fragility of Regression  
Estimates," *Review of Economics and Statistics*, Vol. 65, pp. 306-17.
- Maritz, Johannes, 1995. *Distribution-Free Statistical Methods*, Chapman & Hall,  
London, second edition.
- Wilf, Herbert, 1989. *Combinatorial Algorithms : An Update*. C B M S - N S F Regional  
Conference Series in Applied Mathematics.