Teaching Rank-Based Tests by Emphasizing Structural Similarities to Corresponding Parametric Tests

DeWayne R. Derryberry
Sue B. Schou
Idaho State University

W. J. Conover
Texas Tech University

Journal of Statistics Education Volume 18, Number 1 (2010),
www.amstat.org/publications/jse/v18n1/derryberry.pdf

Key Words: Hypothesis Test; Nonparametric Test; Pedagogy; Skewness; Outliers.

Abstract

Students learn to examine the distributional assumptions implicit in the usual *t*-tests and associated confidence intervals, but are rarely shown what to do when those assumptions are grossly violated. Three data sets are presented. Each data set involves a different distributional anomaly and each illustrates the use of a different nonparametric test. The problems illustrated are well–known, but the formulations of the nonparametric tests given here are different from the large sample formulas usually presented. We restructure the common rank-based tests to emphasize structural similarities between large sample rank-based tests and their parametric analogs. By presenting large sample nonparametric tests as slight extensions of their parametric counterparts, it is hoped that nonparametric methods receive a wider audience.

1. Introduction

Rank-based nonparametric tests were discovered in the 1940's by Wilcoxon, who realized that outliers created problems when employing parametric tests (Salzburg, 2001). Wilcoxon provided a strong impetus for using ranks in nonparametric inference about the

same time as many other researchers including Mann and Whitney in 1947, Festinger in 1946, White in 1952, van der Reydn in 1952, as well as Kruskal and Wallis in 1952 (Conover, 1999). Though these tests generally rely on elementary statistical concepts and college algebra, most entry level students are not exposed to this class of inference.

Since there are fewer assumptions for rank based tests and they perform almost as well as, and often much better than, parametric tests, the authors wish to make a case for teaching this class of tests in conjunction with parametric tests in lower level statistics courses. In addition, these tests could easily serve as an advanced topic in Advanced Placement statistics courses. The authors will revisit Conover and Iman (1981) and offer a related approach. Both approaches offer a much improved method of teaching nonparametric inference.

Although nonparametric statistics are not currently an official part of the AP statistics curriculum, they are included as optional material in many textbooks at that level (DeVore and Peck, 2008; Moore, 2010; Utts and Heckard, 2007). Incorporating the ideas in this paper should make these optional chapters more attractive.

2. Advantages of rank-based methods

Rank-based procedures are a subset of nonparametric procedures that have three strengths: 1) as nonparametric procedures, they are preferred when certain assumptions of parametric procedures (the usual *t*- and *F*- tests) are grossly violated (example: normality assumption when the data set has outliers), 2) rank-based methods are some of the most powerful nonparametric methods, often having nearly as much power as parametric methods when the assumptions of parametric methods are met, and often having more power when the data come from non-normal populations, and 3) rank-based methods can be presented in a way that provides a natural transition for students familiar with parametric methods.

The rank-based tests discussed in this paper require fewer "shape" assumptions than their parametric alternatives. For matched-pairs data, the parametric approach involves computing the differences within pairs and performing a one-sample *t*-test. This test involves the assumption that differences within pairs are normal. The rank-based alternative discussed below, the Wilcoxon signed ranks test, assumes only that the distribution of differences within pairs be symmetric without requiring normality.

The parametric two-sample procedure, the two-sample *t*-test, assumes both samples come from populations with a normal distribution and the same variance. The Mann-Whitney test, the rank-based procedure, assumes both distributions have the same shape, but any shape. The *k*-sample parametric (one-way ANOVA) and rank-based (Kruskal-Wallis) tests have the same assumptions, respectively, as their two-sample counterparts. All *t*-tests require continuous data, while the Mann-Whitney and Kruskal-Wallis tests require only data of ordinal scale.

Conover (1999, page 269) begins his chapter on rank-based tests with a general observation:

> If data are numeric, and, furthermore are observations on random variables that have the normal distribution so that all of the assumptions of the usual parametric test are met, the loss of efficiency caused by using the methods of this chapter is surprisingly small. In those situations the relative efficiency of tests using only the ranks of the observations is frequently about .95, depending on the situation.

(Relative efficiency is the ratio of sample sizes needed for two tests to attain the same power. When relative efficiency is close to one, the two tests being compared are about equally efficient in their use of the data.) When the data are not normal, the *t*-test still performs reasonably well (Posten, 1979; Pearson and Please, 1975), but rank-based methods are often superior (Ramsey and Schafer, 2002; McDougal and Rayner, 2004). When the data contain outliers the *t*-test is suspect, while rank-based methods are unaffected (because outliers have the same rank as any other large observation). When data contain outliers, some manner of nonparametric procedure is almost mandatory.

Rank-based methods can also handle certain types of censored data. Ramsey and Schafer (2002) present a case in which students are given a task to complete and at the end of five minutes a few of the students have not completed the task. For those who did not complete the task, completion times are censored; only a lower bound for the true completion time is known. Without complete observations, the *t*-test cannot be performed, but the rank-based method need only assign the largest ranks to the censored observations.

3. The common structure of many large sample tests

Rossman and Chance (1999) write:

> We want students to see that the reasoning and structure of statistical inference procedures are consistent, regardless of the specific technique being studied. For example, students should see the sampling distributions for several types of statistics to appreciate their similarities and understand the common reasoning process underlying the inference formulas. In addition, students can view these formulas as special cases of one basic idea. …By understanding this general structure of the formulas, students can concentrate on understanding one big idea, rather than trying to memorize a series of seemingly unrelated formulas. Students can then focus on the type and number of variables involved in order to properly decide which formula is applicable. This approach also empowers students to extend their knowledge beyond the inference procedures covered in the introductory course.

For nonparametric statisticians, most rank-based hypothesis tests are based on the permutation distribution of the ranks under the null hypothesis. This procedure is based on enumerating all possible, equally likely, arrangements of the ranks under the null

hypothesis. Although not always identified as permutation tests, many examples of this idea are both explicit and detailed in many textbooks (Randles and Wolfe, 1979; Hettmansperger, 1984; Lehmann, 1975). Simple functions of the ranks, such as the sum of the ranks for one group, are used to construct a sampling distribution from the enumerated ranks.

Large sample formulas are just approximations to permutation tests when the sample size becomes both so large that explicit enumeration becomes cumbersome and so large that the central limit theorem can be relied on to generate an approximately normal test statistic.

Many introductory statistics teachers at the high school, the community college, and even the college level are mathematicians or mathematics educators who have had little statistical training and may have heard of nonparametric tests but have never worked with these tests. Permutation tests are almost never discussed in introductory statistics textbooks. Even the basic elements needed, counting techniques, play a smaller role in each new edition of introductory textbooks (whether this trend is good or bad is another topic). On the other hand, all introductory textbooks discuss the role of sample size and the central limit theorem in developing approximately normal (or *t*) sampling distributions.

Because of this, some instructors rarely think of rank-based test statistics as large sample approximations to permutation distributions, nor do they automatically understand the motivation for the statistics traditionally used to summarize ranked data. For these colleagues it may be more natural to emphasize similarities between large sample rank-based tests and large sample parametric tests.

A general approach to rank-based methods that emphasizes structural similarities between rank-based tests and parametric tests would proceed as follows: Assign a score to the observations based on their rank, compute average scores, compute the standard deviation (or standard error) of the average scores, and form the usual *z*- or $\chi 2$ – test (or perhaps the usual *t*- and *F*- test). When sample sizes are sufficiently large, the computed test statistic can be compared to a table or the p-value can be found using a calculator or statistical software.

The Wilcoxon signed ranks test and the Mann- Whitney test can be presented in a standard form:

$$z \text{ (or } t) = \frac{\text{average score(s) - null hypothesis score}}{\text{standard devation (or standard error)}}$$

The Kruskal-Wallis test can be presented as a slight generalization of these procedures similar to the generalization of parametric *z*-tests to $\chi 2$-tests or the generalization of parametric *t*-tests to *F*-tests. For students and instructors with no training in permutation tests, but who are familiar with parametric tests, this seems more natural than the versions of these formulas found in textbooks on nonparametric statistics.

Although average scores are computed, these tests, unlike parametric tests, are not about means. Scoring the data preserves order, but loses (sometimes troublesome) distributional details. Most nonparametric tests, including all the rank-based tests we discuss, involve inference for medians and comparisons of medians.

The score assigned an observation is not always the rank of that observation. We will see with the signed ranks test (in section 5) that the scoring is slightly more complex, but still based on ranks and intuitively appealing.

The following examples will demonstrate three themes emphasized throughout the paper: 1) when the normality assumption is violated, nonparametric methods often outperform parametric methods, 2) nonparametric methods are traditionally presented in a way that does not emphasize their structural similarities to analogous parametric tests, and 3) rank-based methods, in particular, can be presented in a way that emphasizes similarities with parametric tests.

In each case two reasonable approaches to rank-based tests will be presented: 1) an approach originally suggested by Conover and Iman (1981) and 2) a revised presentation of the usual large sample rank-based test. An argument will be made that both of these approaches offer presentations of rank-based tests that can be implemented as natural extensions of what students already know and that each approach has strengths and weaknesses.

4. A rank-based two-sample test

The usual way of presenting the Mann-Whitney test is based on assigning the observations their ranks (tied observations share an average rank) and then performing a test based on the sum of the ranks for one of the groups (it does not matter which group). When the sample sizes are small, tables have been constructed to assess the size of the test statistic. When samples sizes are large (often used is a sample size of 20 or more), the test statistic is approximately normal and tabulated values are not required (Conover 1999; Daniel, 1990).

To illustrate each test procedure, we use data sets from the Data and Story Library (DASL). The first data set, "cloud seeding", involves a randomized experiment to determine whether cloud seeding increased rainfall. Each group (days when clouds were seeded and days when clouds were not seeded) has 26 observations.

Table 1 displays both the original observations and the scores assigned to the observations. For this test, the score assigned each observation is just its rank. Note that tied observations are assigned an average rank.

Table 1: The DASL cloud seeding data: observations and the assigned scores.
* indicates ties. These data display substantial right-skewness.

Group	observation	rank	group	observation	rank
unseeded	1202.60	49	seeded	2745.60	52
unseeded	830.10	47	seeded	1697.80	51
unseeded	372.40	43	seeded	1656.00	50
unseeded	345.50	42	seeded	978.00	48
unseeded	321.20	40	seeded	703.40	46
unseeded	244.30	35	seeded	489.10	45
unseeded	163.00	31	seeded	430.00	44
unseeded	147.80	30	seeded	334.15	41
unseeded	95.00	25	seeded	302.80	39
unseeded	87.00	23	seeded	274.70	37.5*
unseeded	81.20	22	seeded	274.70	37.5*
unseeded	68.50	21	seeded	255.00	36
unseeded	47.30	20	seeded	242.50	34
unseeded	41.10	19	seeded	200.70	33
unseeded	36.60	17	seeded	198.60	32
unseeded	29.00	14	seeded	129.60	29
unseeded	28.60	13	seeded	119.00	28
unseeded	26.30	12	seeded	118.30	27
unseeded	26.10	11	seeded	115.30	26
unseeded	24.40	10	seeded	92.40	24
unseeded	21.70	9	seeded	40.60	18
unseeded	17.30	7	seeded	32.70	16
unseeded	11.50	6	seeded	31.40	15
unseeded	4.90	3.5*	seeded	17.50	8
unseeded	4.90	3.5*	seeded	7.70	5
unseeded	1.00	1	seeded	4.10	2
Column mean	165	21.31		442	31.69
Column std. dev.	278	14.36		651	14.36

The usual Mann-Whitney test involves just the sum of the ranks, *W*, from one group. It is natural, for a student familiar only with parametric tests, to immediately guess that a test could be based on comparing the average rank of each group. It may not be obvious to that same student that the sum of the ranks from one group has a one-to-one mapping to the difference in average ranks between groups. It is also easier for a student to remember a pattern when it is similar to a pattern they have already seen.

Table 2: Two approaches compared –The traditional Mann-Whitney test on the left (based on the sum of the ranks from the seeded group) and our suggested presentation (based on the average ranks for the two groups) on the right.

$W = 824, n_1 = n_2 = 26$	$\overline{R}_1 = 31.7, \ \overline{R}_2 = 21.3$
$N = n_1 + n_2 = 52$	$n_1 = n_2 = 26, \ N = n_1 + n_2 = 52$
$E(W) = n_1 \cdot (N+1)/2 = 689$	$V(R) \approx N(N+1)/12 = 229.67$
$V(W) = n_1 \cdot n_2 (N+1)/12$ $= 2985.67$	
$z \approx \dfrac{W - E(W)}{\sqrt{V(W)}} = 2.47$	$z \approx \dfrac{\overline{R}_1 - \overline{R}_2}{\sqrt{\dfrac{V(R)}{n_1} + \dfrac{V(R)}{n_2}}} = 2.47$
	p – value = 0.0136

The expression on the right side, *V(R),* is derived in the appendix. Because there are a few ties in the data, the formula is only approximately (but essentially) correct in this case.

The presentation on the left in Table 2 is identical to the presentation given in two excellent books on introductory statistics (Moore, 2010; Utts and Heckard, 2007), as well as many nonparametric textbooks (Hollander and Wolfe, 1999; Hettmansperger, 1984). The presentation on the right is new. Both sides produce identical *z*-scores, as the computations are algebraically equivalent.

The actual scoring of the data is easy enough to explain to students, but the test on the right more closely follows a pattern the students (and teachers) have already seen. There are two reasonable approaches: perform a pooled two-sample *t*-test on the ranks (Table 3), or restructure the Mann-Whitney test (Table 2) so that it is transparent that the test is a comparison of average group ranks. Both procedures produce nearly identical p-values when sample sizes are moderate or large. These procedures closely follow a pattern established with the parametric *t*-test (Table 4).

Table 3: The usual parametric *t*-test and the approach suggested by Conover and Iman (1981). Notice the test is identical to the two-sample *t*-test (pooled estimate of variance) except that ranks are used in place of the original values for the measurements.

Parametric approach:

Group	sample size	mean	standard deviation
seeded	26	442	651
unseeded	26	165	278

pooled standard deviation = 500.52 degrees of freedom = 50
t-statistic = 2.00 p-value = 0.051

Using ranks:

Group	sample size	mean	standard deviation
seeded	26	31.7	14.4
unseeded	26	21.3	14.4

pooled standard deviation = 14.4 degrees of freedom = 50
t-statistic = 2.61 p-value = 0.012

Table 4: Usual parametric procedure, the procedure suggested by Conover and Iman (1981), and the Mann-Whitney test in a format emphasizing the test can be thought of as a difference in average group ranks.

parametric test statistic	Conover and Iman rank-based test statistic	Mann-Whitney test statistic
$t_{n-2} = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$	$t_{R,n-2} = \frac{(\bar{R}_1 - \bar{R}_2) - 0}{\sqrt{\frac{s_{R,p}^2}{n_1} + \frac{s_{R,p}^2}{n_2}}}$	$z = \frac{(\bar{R}_1 - \bar{R}_2) - 0}{\sqrt{\frac{V(R)}{n_1} + \frac{V(R)}{n_2}}}$

$V(R) = N(N+1)/12$ (see Appendix, this is approximately correct when there are a limited number of ties)

The rank-based procedures produce similar p-values that are substantially smaller than that of the parametric procedure, because the data display skewness. In a case such as this, when the normality assumption is grossly violated, a nonparametric approach generally produces more powerful tests (smaller p-values when the alternative hypothesis is true.)

5. A rank-based test for matched-pairs

The parametric matched-pairs test involves finding the differences within each pair and performing a one-sample *t*-test on the resulting differences. The most common rank-

based procedure also works with these differences within pairs. These differences are the observations that are assigned a score. (The procedure discussed in this section can also be used, in a slightly different way, to perform a one sample rank-based test.)

The assignment of a score to an observation is based on the sign and magnitude of the difference. The method is rank-based because only the rank of the magnitude is considered. Denote the differences within pairs, d_i. Assign ranks to $|d_i|$, the absolute values of the differences. Recover the signs: $r_i = sign(d_i) \cdot rank(|d_i|)$; these are the signed ranks for the observations.

The DASL Fish data provide a useful illustration. The data are the 1970 and 1980 prices of a fixed unit of various types of fish. If we are interested in whether fish have risen in price at a rate different from inflation we might compute a difference = 1980 price – adjusted 1970 price, where the 1970 price has been adjusted for inflation. If the null hypothesis is true, the distribution of the paired differences is centered at zero. Similarly, the distribution of the signed ranks should have a median of zero, when the null hypothesis is true.

Table 5: DASL Fish story data. Difference = 1980 price – 2.1237(1970 price)
= 1980 price – inflation adjusted 1970 price. "SEA SCALLOPS" is an outlier in each year and after taking the difference.

Type of fish	1980 price	1970 price	difference	signed rank
COD	27.3	13.1	-0.520	-1
MENHADEN	4.5	1.8	0.677	2
OYSTERS, EASTERN	131.3	61.1	1.542	3
CLAMS, BLUE HARD	20.3	6.6	6.284	4
FLOUNDER	42.4	15.3	9.907	5
LOBSTERS	189.7	94.7	-11.414	-6
OCEAN PERCH	23	4.9	12.594	7
HADDOCK	38.7	25.8	-16.091	-8
TUNA, ALBACORE	80.1	26.7	23.397	9
SALMON, COHO	109.7	39.3	26.239	10
SHRIMP	149	47.6	47.912	11
SALMON, CHINOOK	166.3	55.4	48.647	12
CLAMS, SOFT	150.7	47.5	49.824	13
SEA SCALLOPS	404.2	135.6	116.226	14

The parametric test applied to these data produces a test statistic of 2.44, on 13 degrees of freedom, with a p-value of 0.03. Introductory textbooks are correct in suggesting parametric *t*- tests produce reliable results even when sample sizes are quite small, assuming outliers and skewness are not major problems. In the case of the fish data there is, however, an outlier (see Figure 1).

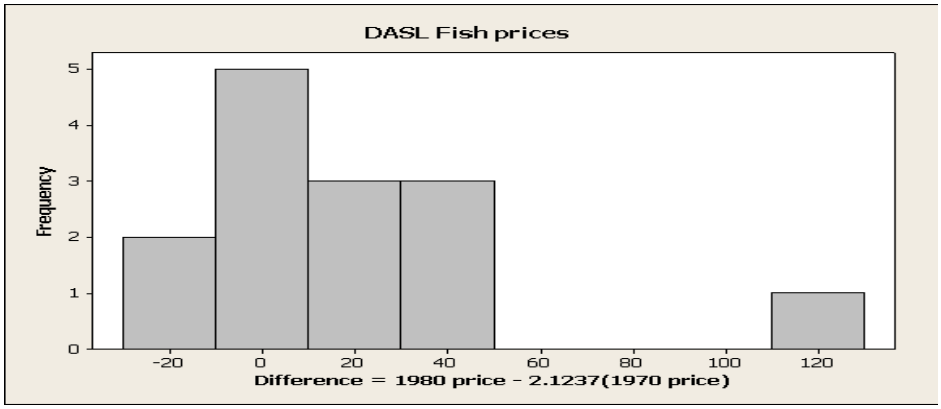

Figure 1: The fish prices data, after taking differences within paired data, have an outlier.

The parametric method is easy and familiar, but the rank-based method is preferred. Because there is an outlier in the data, to the extent that the parametric and rank-based methods produce different results, rank-based results are more reliable.

The usual rank-based test, the Wilcoxon signed ranks test, involves summing either the positive or the negative signed ranks. Since there are just three negative signed ranks, summing these is easiest, so $T- = 15$. An exact p-value is found by counting the number of subsets of the ranks from 1 to 14 that give a rank sum of 15 or less. There are 136 such subsets, drawing without replacement including the null set. The one-sided p-value is found by dividing 136 by the total number of possible subsets, 2^{14} , and then multiplying by two for the two-sided p-value ($2 \cdot 136/2^{14} = 0.0166$). The one-sided p-value is in agreement with the exact tabled value of 0.008 given on page 327 of Owen (1962, using $n = 14$, $a = 15$ in his notation). The standard deviation for $T-$ is $\left[n(n+1)(2n+1)/24\right]^{1/2} = 15.93$, and the expected value for $T-$ (under the null hypothesis) is $n(n+1)/4 = 52.5$, so the z-score is (15-52.5)/15.93 = -2.35, yielding a two sided p-value = 0.0188. This is a much smaller p-value than that produced by the parametric test. (If the sum of the positive signed ranks were used, $T+ = 90$, the z-score would be +2.35).

In this form, the Wilcoxon signed ranks test, like the Mann-Whitney test, is easy to compute but unnatural for a student (or instructor) unfamiliar with permutation tests. The introductory statistics student's natural instinct, having seen the paired parametric test, would be to take differences within pairs and examine some average difference or average score. While the average score is mathematically equivalent to the sum of either the positive or negative scores (when data are ranked), the latter does not fit a pattern the teacher has spent considerable effort establishing.

As before, we present two alternatives, one directly analogous to the parametric *t*-test and suggested by Conover and Iman (1981), and the other a version of the Wilcoxon signed ranks test that emphasizes similarities to the paired *t*-test.

The parametric method and Conover and Iman's suggestion differ only in that the first approach uses the original differences while the latter uses the resulting signed ranks. The

Wilcoxon formulation is simply a form of the large sample *z*-score formula that emphasizes differences are being taken within pairs and then averaged.

Table 6: The parametric *t*-test, the approach advocated in Conover and Iman (1981) and a form of the Wilcoxon signed ranks test emphasizing similarities to the parametric approach. The value $\overline{R}_s$ is the mean of the signed ranks.

parametric approach	Conover and Iman rank-based test	Wilcoxon signed ranks
$t_{n-1} = \frac{\bar{x}_D}{s_D / \sqrt{n_D}}$	$t_{R,n-1} = \frac{\overline{R}_s}{s_{R_S} / \sqrt{n_D}}$	$z = \frac{\overline{R}_s}{s_{R_S} / \sqrt{n_D}}$

$s_{R_s} = \sqrt{V(R_s)}$ where $V(R_s) = (n_D + 1)(2 \cdot n_D + 1)/6$ (see Appendix, this is approximately correct when there are a limited number of ties).

Using the last column of Table 5 it is easy to verify that $\overline{R}_s = 5.357$, $n_D = 14$ (the number of differences) and $s_{R_D} = 6.87$. Using these values it is easy to confirm that the Wilcoxon signed ranks formula in Table 6 yields the same *z*-score as found above. Conover and Iman's approach produced the smallest p-value (0.012), generating a test statistic of 2.92 with 13 degrees of freedom.

So which answer is best? Because the sample size is small, all the p-values computed so far are approximate. Because there is an outlier, the parametric p-value is quite suspect. The best approach for these data is a rank-based permutation test. In fact, for these data the permutation test p-value is 0.0166. Both of the rank-based approaches (*z* and *t*) produced p-values much smaller than the parametric test, and reasonably close to the correct p-value. Although the tabulated result of the permutation test is best, it is not unreasonable to use rank-based tests with the normal (or *t*) approximation. However, the parametric procedure produces a poor result because of the outlier in the data.

6. Testing *k* independent samples using ranks

If students are exposed to one-way analysis of variance, the rank-based analogs are straightforward extensions of the two-sample procedures. As with the Mann-Whitney test, all the data are ranked from smallest to largest. We then want to know if the groups have similar average ranks, or if at least one group has a substantially different average rank. The Conover and Iman (1981) approach is a generalization of a *t* test and is naturally an *F* test. The second approach we advocate is a generalization of a *z* test and is, as one might suspect, a $\chi 2$ test.

Consider the DASL waste run up data. In this example there are five suppliers and the response = percentage wasted material working by hand – percentage waste using a computer layout. Is the response, some measure of excess waste, the same for all suppliers?

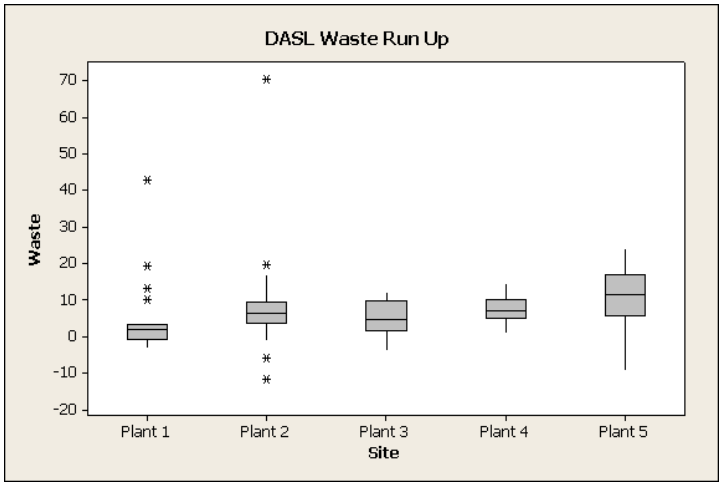

Figure 2: A data set with a large number of outliers.

The usual rank-based test, the Kruskal-Wallis test, compares the sum of the ranks and the expected sum of the ranks for each group: $H = 12/[N(N+1)] \cdot \sum [S_i - n_i(N+1)/2]^2 / n_i$, where the S_i and n_i are the sums of the ranks and sample sizes for each group, respectively. H is distributed χ^2_{k-1} when the null hypothesis is true.

The usual parametric F test statistic is 1.16 with a p-value of 0.334, but given the preponderance of outliers in the data, this result hardly seems reliable. An F test statistic on the ranked data, following Conover and Iman (1981), produces an F of 4.38 and a p-value of 0.003. The Kruskal-Wallis test produces a test statistic of 15.32 with a p-value of 0.004. Both rank-based methods produce p-values similar to each other and vastly different from the p-value produced by the (erroneously applied) parametric test.

There is a way to produce the Kruskal-Wallis test statistic that is a simple generalization of our presentation of the Mann-Whitney test statistic. First note that the average rank, under the null hypothesis, is $(N+1)/2$. Next standardize the average rank of each group: $z_i = \left[\overline{R}_i - (N+1)/2\right] / \left[V(R)/n_i\right]^{1/2}$ and $\chi^2_{k-1} = \sum_{i=1}^{k} z_i^2$ where $V(R) = N(N+1)/12$ as before (see appendix).

Table 7: The average rank for each group, the sample size, and the standardized rank based on an expected mean of $(N+1)/2 = 48$ and a variance of $V(R) \approx 760$.

Site	average rank	Sample size	standardized
Supplier 1	31.23	22	$z_1 = -2.8532$
Supplier 2	51.30	22	$z_2 = 0.5615$
Supplier 3	43.71	19	$z_3 = -0.6783$
Supplier 4	56.89	19	$z_4 = 1.4056$
Supplier 5	64.08	13	$z_5 = 2.1031$

It is easy to verify that the sum of these squared standardized values in Table 7 is 15.32, the value of the Kruskal-Wallis test statistic.

7. Which is best, the *t*-test or *z*-test?

Data exist for which parametric methods are inappropriate. Nonparametric methods should be part of a first or second course in statistics. Rank-based methods are both some of the most powerful nonparametric methods, and some of the easiest to motivate based on similarities to parametric tests. If such methods are taught at this introductory level, either approach presented here (the *t*-based approach of Conover and Iman or the *z*- based approach) is more natural than the usual presentations of the Wilcoxon signed rank test, the Mann-Whitney test, and the Kruskal-Wallis test.

The two approaches (*t*- and *z*-tests) are essentially the same since there exists a one-to-one mapping between the two procedures and both procedures are large-sample approximations (Conover and Iman, 1981.) Let ν be the degrees of freedom for a *t* test. For the Wilcoxon signed ranks test and the Mann-Whitney test $z = t_{R,\nu}[(\nu+1)/(\nu + t_{R,\nu}^2)]^{1/2}$ converts the *t*-score to a *z*-score; alternatively $t_{R,\nu} = z[\nu/(\nu+1-z^2)]^{1/2}$ converts the *z*-scores into *t*-scores. (This is easily verified with the examples given.) A similar mapping exists between the *F*- and χ2- tests (Conover and Iman, 1981.) These relationships hold exactly even when there are ties.

From the mathematical point of view, the choice is arbitrary. The choice should be based on how the nonparametric methods will fit in with the rest of a textbook, and what topics the instructor wishes to emphasize. The answer depends partly on how an instructor answers three questions:

Do I want to emphasize *t*- and *F*-procedures or *z*- and χ2- procedures?
Do I want to teach many nonparametric procedures or just rank-based procedures?
Do I want to introduce nonparametric methods embedded throughout the course, or in a single section near the end of a first course or near the beginning of a second course?

The main advantages of the *z*-based approach are that the *z*-scores are identical to the rank-based tests, the generalization to k-samples does not require knowledge of *F*-tests

(the Kruskal –Wallis test can be taught without teaching one-way ANOVA!), and the emphasis is on *z*-scores where the variance is known. This approach is ideally suited when a collection of nonparametric methods (including methods like the signs test, which are not rank-based) are presented as a group after discussion of correlation, ANOVA, and $\chi 2$. For a class in which $\chi 2$ is emphasized, this approach is ideal because the methods make connections between *z* and $\chi 2$; this form of the Kruskal-Wallis test reinforces and builds upon $\chi 2$ ideas.

The main advantage of the *t*-based approach is that the different rank-based tests can be presented side by side with their companion *t* –tests in different sections as each procedure is introduced. In such a course each rank-based test is presented as soon as the need arises. In a course that spends a lot of time on *t*- and *F*- tests, and little time on $\chi 2$ tests, this approach is far more natural, especially when the Kruskal-Wallis test is included. For an example of a textbook that has already adopted some elements of this approach see Iman (1994).

There is an additional advantage to the *t*-based approach. There are other methods of scoring data based on ranks (Lehmann, 1975; Randles and Wolfe, 1979). The *t*-based procedure handles these other cases just as easily as the cases presented. The *z*-based method would require the derivation of a new variance following the manner of reasoning found in the appendix. This variance might not always have a concise closed form.

8. Summary

Nonparametric methods are important and should be part of a first or second course in statistics. As usually presented, they are best viewed as permutation tests, a topic rarely discussed in introductory statistics. However, as presented in this article, an important class of nonparametric tests, rank-based tests, are straightforward extensions of parametric methods widely taught in introductory statistics.

Skewed data (cloud seeding example) and data with outliers (waste run up example) present difficulties for parametric methods that are easily addressed by switching to rank-based methods. Even for small data sets (fish prices example), the large sample approximations to rank-based tests are reliable to the same degree that parametric methods are reliable for small samples.

As useful as rank-based methods are, the presentation of the large sample formulas frequently focus on a permutation point of view, which is unfamiliar to many students (and teachers) of introductory statistics. The method of assigning scores is itself easy to explain, but once these scores are assigned the resulting formulas do not develop in the way these students would naturally expect. It seems more natural, especially for this particular audience, to work with averages and standard deviations (standard errors) of those averages as we typically do with all the parametric *z*, *t* and *F* tests.

In fact, there are two simple solutions. The first is to just treat the resulting scores like any other data, producing *t* and *F* tests nearly identical to their parametric cousins. The second approach is to derive the traditional tests in a manner that emphasizes structural similarities to corresponding parametric tests. Both approaches offer a way to present nonparametric tests that can be easily motivated and fit with the mainstream of introductory inferential statistics.

Appendix: derivation of new versions of rank-based tests

One sample case:

For sufficiently large sample sizes, the average of the signed ranks is approximately normal. The test is straightforward once the variance of the signed ranks is known.

The signed ranks, R_s, take on the integers $1...n_D$ with a sign associated with each ranked value. *Under the null hypothesis*, the average of the signed ranks is zero, $E(R_s) = 0$.

$$V(R_s) = \sum_{k=1}^{n_D} [\pm R_k - E(R_k)]^2 / n_D = \frac{1}{n_D} \sum_{k=1}^{n_D} (\pm R_k)^2 = \frac{1}{n_D} \sum_{k=1}^{n_D} k^2 = \frac{(n_D)(n_D+1)(2n_D+1)}{6n_D}$$

which simplifies to the formula given in the paper.

When there are ties the variance is still $\frac{1}{n_D} \sum_{k=1}^{n_D} |R_k|^2$, but the closed form formula is now $\frac{(n_D+1)(2n_D+1)}{6} - \frac{1}{12n_D} \sum_{k=1}^{g} (t_k - 1)t_k(t_k + 1)$ where g is the number of groups with ties and t_k is the number of ties in group k (this formula is analogous to Hollander and Wolfe, 1999, page 38).

K sample case (including two sample case):

The following "derivation" is more intuitive than rigorous. Nevertheless, the resulting formulas produce the correct *z*-score for the Mann-Whitney test and the correct chi-square test statistic for the Kruskal-Wallis test.

Assume all the observations really come from a single distribution (this implies both that the null hypothesis is true and that groups share a common shape and spread) and are independent. Suppose further the sample size is sufficiently large that average ranks are approximately normally distributed. The observations are assigned ranks. Under the null hypothesis the ranks can now be thought of as a single group of independent observations with a common shape, center and spread. The ranks, *R*, take on the integers $1...N$.

The mean of the ranks is obviously *(N+1)/2*, and the pooled variance is

$$V(R) = \frac{\sum_{j=1}^{N} [R_j - (N+1)/2]^2}{N-1} = \frac{\sum_{j=1}^{N} [j - (N+1)/2]^2}{N-1} = \frac{N(N+1)}{12}$$

In this case the denominator is $N - 1$ because of the covariance of the ranks among themselves, as the ranks must average *(N+1)/*2. (In the derivation of the one sample variance on signed ranks the signed ranks have zero covariance with each other.)

When there are ties the variance is still $\frac{\sum_{j=1}^{N}\left[R_j-(N+1)/2\right]^2}{N-1}$, but the closed form formula is now $\frac{N(N+1)}{12}-\frac{1}{12(N-1)}\sum_{k=1}^{g}(t_k-1)t_k(t_k+1)$ where g is the number of groups with ties and t_k is the number of ties in group k (this formula is analogous to Hollander and Wolfe, 1999, page 109).

Suppose the observations are arbitrarily broken up into k subgroups (assuming each subgroup is still sufficiently large that average ranks are approximately normal) and an average rank is computed for each subgroup. Let subgroup i have n_i observations and an average rank denoted $\overline{R}_i$. Given the initial suppositions (common mean and spread), $E(\overline{R}_i)=(N+1)/2$ and $V(\overline{R}_i)=V(R)/n_i$ so $z_i=[\overline{R}_i-(N+1)/2]/\sqrt{V(R)/n_i}$ are the standardized average ranks. Because the average ranks are approximately normal, z_i are approximately standard normal.

These standardized average ranks can be combined using the usual argument that sums of squared (approximately) standard normal random variables are (approximately) $\chi 2$, but with only k-1 degrees of freedom because of the covariance among the average ranks. When k = 2 this produces the Mann-Whitney test (in this case, equivalence to the formula found on the right side of Table 2 involves tedious algebra).

Acknowledgements

The authors thank two anonymous referees and the editor for many valuable comments on a previous version of this paper, both with regard to numerous details and overall focus.

References

Conover, W. J. (1999), *Practical Nonparametric Statistics* (3rd ed.), New York: Wiley.

Conover W. J., and Iman, R. L. (1981), "Rank Transformations as a Bridge Between Parametric and Nonparametric Statistics," *The American Statistician*, 35, 124-129.

Daniel, W. D. (1990), *Applied Nonparametric Statistic* (2nd ed.), Boston: PWS-KENT.

DASL, The Data and Story Library, http://lib.stat.cmu.edu/DASL/Stories/CloudSeeding.html

DeVore, J. L., and Peck, R. (2008), *Statistics: The Exploration and Analysis of Data* (6th ed.), Pacific Grove, CA: Brooks/Cole.

De Veaux, R. D., Velleman, P. F., and Bock, D. E. (2009), *Intro Stats* (3rd ed.), Reading, MA: Addison-Wesley.

Eddington, E. S. (1995), *Randomization Tests* (3rd ed.), New York: Dekker.

Hettmansperger, T. P. (1984), *Statistical Inference Based on Ranks*, New York: Wiley.

Hollander, M., and Wolfe, D. A. (1999), *Nonparametric Statistical Methods* (2nd ed.), New York: Wiley.

Iman, R. L. (1994), *A Data – Based Approach to Statistics*, Pacific Grove, CA: Duxbury.

Lehmann, E. L. (1975), *Nonparametrics: Statistical Methods Based on Ranks*. San Francisco: Holden-Day.

McDougall, M. K., and Rayner, G. D. (2004), "Robustness to non-normality of various tests for the one-sample location problem," *Journal of Applied Mathematics and Decision Sciences*, 8, 235-246.

Moore, David S. (2010), *The Basic Practice of Statistics* (5th ed.), New York: Freeman.

Owen, D. B. (1962), *Handbook of Statistical Tables*, Reading, MA: Addison-Wesley.

Pearson, E. S., and Please, N. W. (1975), "Relation between the shape of population distribution and the robustness of four simple test statistics," *Biometrika*, 62, 223-242.

Posten, Harry O. (1979), "The robustness of the one-sample t-test over the Pearson system," *Journal of Statistical Computation and Simulation*, 9, 133-150.

Ramsey, F. L., and Schafer, D. W. (2002), *The Statistical Sleuth* (2nd ed.), Pacific Grove, CA: Duxbury.

Randles, R. H., and Wolfe, D. A. (1979), *Introduction to the Theory of Nonparametric Tests*, New York: Wiley.

Rossman, A. J., and Chance, B. L. (1999), "Teaching the Reasoning of Statistical Inference: A "Top Ten" List," *The College Mathematics Journal*, Vol. 30, No. 4, 297-305.

Salzburg, D. (2001), *The Lady Tasting Tea*, New York: W. H. Freeman and Company.

Utts, J. M., and Heckard, R. F. (2007), *Mind on Statistics* (3rd ed.), Belmont, CA: Brooks/Cole.

DeWayne R. Derryberry
Department of Mathematics
921 S. 8th Ave., STOP 8085
Pocatello, ID 83209
Email: derrdewa@isu.edu
Phone: 253-278-3447
Fax number: 208-282-2636

Sue B. Schou
College of Business
921 S. 8th Ave., STOP 8020
Pocatello, ID 83205
Email: schosue@isu.edu

W. J. Conover
Information Systems and Quantitative Science
Texas Tech University
Box 42101
Lubbock, TX 79409
Email: jay.conover@ttu.edu
