Journal of Modern Applied Statistical Methods

Volume 1 | Issue 2 Article 35

11-1-2002

Twenty Nonparametric Statistics And Their Large Sample Approximations

Gail F. Fahoome
Wayne State University

Journal of Modern Applied Statistical Methods
Fall 2002, Vol. 1, No 2, 248-268

REGULAR ARTICLES

Twenty Nonparametric Statistics And Their Large Sample Approximations

Gail Fahoome
Educational Evaluation and Research
Wayne State University

Nonparametric procedures are often more powerful than classical tests for real world data which are rarely normally distributed. However, there are difficulties in using these tests. Computational formulas are scattered throughout the literature, and there is a lack of availability of tables and critical values. The computational formulas for twenty commonly employed nonparametric tests that have large-sample approximations for the critical value are brought together. Because there is no generally agreed upon lower limit for the sample size, Monte Carlo methods were used to determine the smallest sample size that can be used with the respective large-sample approximation. The statistics reviewed include single-population tests, comparisons of two populations, comparisons of several populations, and tests of association.

Key words: nonparametric statistics, Monte Carlo methods, sample size, large sample approximation

Introduction

Classical parametric tests, such as the F and *t*, were developed in the early part of the twentieth century. These statistics require the assumption of population normality. Bradley (1968) wrote, "To the layman unable to follow the derivation but ambitious enough to read the words, it sounded as if the mathematician had esoteric *mathematical* reasons for believing in at least quasi-universal quasi-normality" (p. 8). "Indeed, in some quarters the normal distribution seems to have been regarded as embodying metaphysical and awe-inspiring properties suggestive of Divine Intervention" (p. 5).

When Micceri (1989) investigated 440 large-sample education and psychology data sets he concluded, "No distributions among those investigated passed all tests of normality, and very few seem to be even reasonably close approximations to the Gaussian" (p. 161). This is of practical importance because even though the well known Student's *t* test is preferable to nonparametric competitors when the normality assumption has been met, Blair and Higgins (1980) noted:

> Generally unrecognized, or at least not made apparent to the reader, is the fact that the *t* test's claim to power superiority rests on certain optimal power properties that are obtained under normal theory. Thus, when the shape of the sampled population(s) is unspecified, there are no mathematical or statistical imperatives to ensure the power superiority of this statistic. (p. 311)

Blair and Higgins (1980) demonstrated the power superiority of the nonparametric Wilcoxon Rank Sum test over the *t* test for a variety of nonnormal theoretical distributions. In a Monte Carlo study of Micceri's real world data sets, Sawilowsky and Blair (1992) concluded that although the *t* test is generally robust with respect to Type I errors under conditions of equal sample size, fairly large samples, and two-tailed tests, it is not powerful for skewed distributions. Under those conditions, the Wilcoxon Rank Sum test can be three to four times more powerful. See Bridge and

Gail Fahoome is a Lecturer in the College of Education. Contact her at 335 College of Education, Wayne State University, Detroit, MI 48202 for all communications regarding this paper. E-mail her at gfahoome@wayne.edu. Her areas of expertise are Monte Carlo methods with Fortran, nonparametric statistics, and structural equation modeling.

Sawilowsky (1999) and Nanna and Sawilowsky (1998) for other examples.

The prevalence of nonnormally distributed data sets in applied studies in education and related fields has its initial impact on parametric procedures with regard to Type I errors. Thus, the immediate advantage of nonparametric procedures, such as the Wilcoxon test, is that their Type I error properties are not dependent on the assumption of population normality.

A difficulty in using nonparametric tests is the availability of computational formulas and tables of critical values. For example, Siegel and Castellan (1988) noted, "Valuable as these sources are, they have typically either been highly selective in the techniques presented or have not included the tables of significance" (p. xvi). This continues to be a problem as evidenced by a survey of 20 in-print general college statistics textbooks, including seven general textbooks, eight for the social and behavioral sciences, four for business, and one for engineering. Formulas were given for only eight nonparametric statistics, and tables of critical values were given for only the following six: (a) Kolmogorov-Smirnov test, (b) Sign test, (c) Wilcoxon Signed Rank test, (d) Wilcoxon (Mann-Whitney) test, (e) Spearman's rank correlation coefficient, and (f) Kendall's rank correlation coefficient.

This situation is somewhat improved for nonparametric statistics textbooks. Eighteen nonparametric textbooks published since 1956 were also reviewed. Table 1 contains the statistical content of the eighteen textbooks. The most comprehensive texts in terms of coverage were Neave and Worthington (1988), which is currently out of print, and Deshpande Gore, and Shanubhogue (1995).

Many nonparametric tests have large sample approximations that can be used as an alternative to tabulated critical values. These approximations are useful substitutes if the sample size is sufficiently large, and hence, obviate the need for locating tables of critical values. However, there is no generally agreed upon definition of what constitutes a *large* sample size. Consider the Sign test and the Wilcoxon tests as examples. Regarding the Sign test, Hájek (1969) wrote, "The normal approximation is good for $N \geq 12$" (p. 108).

Table 1. Survey of 18 Nonparametric Books

Statistic	Number of Books That Included Tables of Critical Values
Single Population Tests	
Kolgomorov-Smirnov Test	11
Sign Test	4
Wilcoxon Signed Rank Test	14
Comparison of Two Populations	
Kolmogorov-Smirnov2-sample Test	11
Rosenbaum's Test	1
Wilcoxon (Mann-Whitney)	14
Mood Test	1
Savage Test	1
Ansari-Bradley Test	1
Comparison of Several Populations	
Kruskal-Wallis Test	10
Friedman's Test	9
Terpstra-Jonckheere Test	5
Page's Test	4
Match Test for Ordered Alternatives	1
Tests of Association	
Spearman's Rank Correlation Coefficient	12
Kendall's Rank Correlation Coefficient	10

Gibbons (1971) agreed, "Therefore, for moderate and large values of N (say at least 12) it is satisfactory to use the normal approximation to the binomial to determine the rejection region" (p. 102). Sprent (1989) and Deshpande, Gore, and Shanubhogue (1995), however, recommended N greater than 20. Siegel and Castellan (1988) suggested $N \geq 35$, but Neave and Worthington (1988) proposed $N > 50$.

The literature regarding the Wilcoxon Rank Sum test is similarly disparate. Deshpande, Gore, and Shanubhogue (1995) stated that the combined sample size should be at least 20 to use a large sample approximation of the critical value. Conover (1971) and Sprent (1989) recommended that one or both samples must exceed 20. Gibbons (1971) placed the lower limit at twelve per sample. For the Wilcoxon Signed Rank test, Deshpande, Gore, and Shanubhogue (1995) said that the approximation can be used when N is greater than 10. Gibbons (1971) recommended it when N is greater than 12, and Sprent (1989) required N to be greater than 20. The general lack of agreement may indicate that these recommendations are based on personal experience, the sample sizes commonly accommodated in tables, the author's definition of acceptable or large, or some other unstated criterion.

There are two alternatives to tables and approximations. The first is to use exact permutation methods. There is software available that will generate exact p-values for *small* data sets and Monte Carlo estimates for *larger* problems. See Ludbrook and Dudley (1998) for a brief review of the capabilities of currently available software packages for permutation tests. However, these software solutions are expensive, have different limitations in coverage of procedures, and may require considerable computing time even with fast personal computers (see, e.g., Musial, 1999; Posch & Sawilowsky, 1997). In any case, a desirable feature of nonparametric statistics is that they are easy to compute without statistical software and computers, which makes their use in the classroom or work in the field attractive.

A second alternative is the use of the rank transformation (RT) procedure developed by Conover andIman (1981). They proposed the use of this procedure as a bridge between parametric and nonparametric techniques. The RT is carried out as follows: rank the original scores, perform the classical test on the ranks, and refer to the standard table of critical values. In some cases, this procedure results in a well-known test. For example, conducting the *t* test on the ranks of original scores in a two independent samples layout is equivalent to the Wilcoxon Rank Sum test. (However, see the caution noted by Sawilowsky & Brown, 1991). In other cases, such as factorial analysis of variance (ANOVA) layouts, a new statistic emerges.

The early exuberance with this procedure was related to its simplicity and promise of increased statistical power when data sets displayed nonnormality. Iman and Conover noted the success of the RT in the two independent samples case and the one-way ANOVA layout. Nanna (1997, 2001) showed that the RT is robust and powerful as an alternative to the independent samples multivariate Hotelling's T^2.

However, Blair and Higgins (1985) demonstrated that the RT suffers power losses in the dependent samples *t* test layout as the correlation between the pretest and posttest increases. Bradstreet (1997) found the RT to perform poorly for the two samples Behrens-Fisher problem. Sawilowsky (1985), Sawilowsky, Blair, and Higgins (1989), Blair, Sawilowsky, and Higgins (1987), and Kelley and Sawilowsky (1997) showed the RT has severely inflated Type I errors and a lack of power in testing interactions in factorial ANOVA layouts. Harwell and Serlin (1997) found the RT to have inflated Type I errors in the test of $\beta = 0$ in linear regression. In the context of analysis of covariance, Headrick and Sawilowsky (1999, 2000) found the RT's Type I error rate inflates quicker than the general ANOVA case, and it demonstrated more severely depressed power properties. Recent results by Headrick (personal communications) show the RT to have poor control of Type I errors in the ordinary least squares multiple regression layout. Sawilowsky (1989) stated that the RT as a bridge has fallen down, and cannot be used to unify parametric and nonparametric methodology or as a method to avoid finding formulas and critical values for nonparametric tests.

Purpose Of The Study

As noted above, the computational formulas for many nonparametric tests are scattered throughout the literature, and tables of critical values are scarcer. Large sample approximation formulas are also scattered and appear in different forms. Most important, the advice on how large a sample must be to use the approximations is conflicting. The purpose of this study is to ameliorate these five problems.

Ascertaining the smallest sample size that can be used with a large sample approximation for the various statistics would enable researchers who do not have access to the necessary tables of critical values or statistical software to employ these tests. The first portion of this paper uses Monte Carlo methods to determine the smallest sample size that can be used with the large sample approximation while still preserving nominal alpha. The second portion of this paper provides a comprehensive review of computational formulas with worked examples for twenty nonparametric statistics. They were chosen because they are commonly employed and because large sample approximation formulas have been developed for them.

Methodology

Each of the twenty statistics was tested with normal data and Micceri's (1989; see also Sawilowsky, Blair, & Micceri, 1990) real world data sets. The real data sets represent smooth symmetric, extreme asymmetric, and multi-modal lumpy distributions. Monte Carlo methods were used in order to determine the smallest samples that can be used with large-sample approximations.

A program was written in Fortran 90 (Lahey, 1998) for each statistic. The program sampled with replacement from each of the four data sets for n = 2, 3, … N; (n_1, n_2) = (2, 2), (3,3), … (N_1,N_2), and so forth as the number of groups increased. The statistic was calculated and evaluated using the tabled values when available, and the approximation of the critical value or the transformed obtained value, as appropriate. The number of rejections was counted and the Type I error rate was computed. Nominal α was set at .05 and .01. Bradley's (1978) conservative estimates of .045 < *Type I error rate* < .055 and .009 < *Type I error rate* < .011 were used, respectively, as measures of robustness. The sample sizes were increased until the Type I error rates converged within these acceptable regions.

Limitations

In many cases there are different formulas for the large sample approximation of a statistic. Two criteria were used in choosing which formula to include: (1) consensus of authors, and (2) ease of use in computing and programming. All statistics were examined in the context of balanced layouts only.

Some statistics have different large sample approximations based on the presence of ties among the data. Ties were corrected using average ranks for rank-based tests, obviating tie correction formulae. For nonrank-based tests, simple deletion of ties results in a failure to adjust for variance. (A well-known example is the necessity of using a winsorized standard deviation – or some other modification to the estimate of population variance – in constructing a confidence interval for the trimmed mean when tied scores are deleted.) Nevertheless, many authors (e. g., Gibbons, 1976) indicated that adjustment for ties makes little difference for rank- or nonrank-based tests unless there is an extreme number of ties. The issue of correcting for ties is discussed in the section below.

Data Sets For Worked Examples In This Articl*e*

The worked examples in this study use the five data sets in Table 3 (Appendix). Some statistics converged at relatively large sample sizes. In choosing the sample size for the worked example, a compromise was made based on the amount of computation required for large samples and an unrepresentatively small but convenient sample size for presentation in this article. Therefore, a sample size of n = 15 or N = 15, as appropriate, was selected, recognizing that some statistics' large sample approximations do not converge within Bradley's (1968) limits for this sample size. The data sets were randomly selected from Micceri's (1989) multimodal lumpy data set (Table 4, Appendix). Because the samples came from the same population, the worked examples all conclude that the null hypothesis cannot be rejected.

Statistics Examined

The twenty statistics included in this article represent four layouts: (1) single population tests, (2) comparison of two populations, (3) comparison of several populations, and (4) tests of association. Single-populations tests included: (a) a goodness-of-fit test, (b) tests for location, and (c) an estimator of the median. Comparisons of two populations included: (a) tests for general differences, (b) two-sample location problems, and (c) two-sample scale problems. Comparisons of several populations included: (a) ordered alternative hypotheses, and (b) tests of homogeneity against omnibus alternatives. Tests of association focused on rank correlation coefficients.

Results

Table 2 shows the minimum sample sizes necessary to use the large sample approximation of the critical value or obtained statistic for the tests studied. The recommendations are based on results that converged when underlying assumptions are reasonably met. The minimum sample sizes are conservative, representing the largest minimum for each test. If the test has three or more samples, the largest group minimum is chosen. Consequently the large-sample approximations will work in some instances for smaller sample sizes. This is the smallest size per sample when the test involves more than one sample.

Table 2. Minimum Sample Size for Large-Sample Approximations.

Test	α= .05	α= .01
Single Population Tests		
Kolmogorov-Smirnov Goodness-of-Fit Test	$25 \le n \le 40$	$28 \le n \le 50$
Sign Test	$n > 150$	$n > 150$
Signed Rank Test	10	22
Estimator of Median for a Continuous Distribution	$n > 150$	$n > 150$
Comparison of Two Populations		
Kolmogorov-Smirnov Test	$n > 150$	$n > 150$
Rosenbaum's Test	16	20
Tukey's Test	$10 \le n \le 18$	21
Rank-Sum Test	15	29
Hodges-Lehmann Estimator	15	20
Siegel-Tukey Test	25	38
Mood Test	5	23
Savage Test	11	31
Ansari-Bradley Test	16	29
Comparison of Several Populations		
Kruskal-Wallis Test	11	22
Friedman's Test	13	23
Terpstra-Jonckheere Test	4	8
The Match Test ($k > 3$)	86	27
Page's Test $k > 4$	11	18
Tests of Association		
Spearman's Rho	12	40
Kendall's Tau	$14 \le n \le 24$	$15 \le n \le 35$

Some notes and cautionary statements are in order with regard to the entries in Table 2. The parameters for the Monte Carlo study were limited to *n* (or *N*) = 1, 2, … 150. The Kolmogorov-Smirnov goodness-of-fit test was conservative below the minimum value stated and liberal above the maximum value stated. Results for the Sign test indicated convergence for some distributions may occur close to $N = 150$. The results for the confidence interval for the Estimator of the Median suggest convergence may occur close to $N = 150$ only for normally distributed data. However, for the nonnormal data sets the Type I error rates were quite conservative (e.g., for $\alpha = .05$ the Type I error rate was only 0.01146 and for $\alpha = .01$ it was only 0.00291 for $N = 150$ and the extreme asymmetric data set).

The Kolmogorov-Smirnov two samples test was erratic, with no indication convergence would be close to 150. Results for Tukey's test were conservative for $\alpha = .05$ when the cutoff for the p-value was .05, and fell within acceptable limits for some sample sizes when .055 was used as a cutoff. The Hodges-Lehmann estimator only converged for normal data. For nonnormal data the large sample approximation was extremely conservative with $n = 10$ (e.g., for the extreme asymmetric data set the Type I error rate was only 0.0211 and 0.0028 for the .05 and .01 alpha levels, respectively) and increased in conservativeness (i.e., the Type I error rate converged to 0.0) as n increased. The Match test only converged for normally distributed data, and it was the only test where the sample size required for $\alpha = .01$ was smaller than for $\alpha = .05$.

These results relate to the large sample approximation of the critical values associated with those tests. These procedures work quite well with small sample sizes when tabled critical values are used. The difficulty, as noted above, is that tabled critical values are generally not available, or the implementation of exact procedures is still by far too time-consuming or memory intensive to compute with statistical software. For example, Bergmann, Ludbrook, and Spooren (2000), noted "What should be regarded as a large sample is quite vague …,most investigators are accustomed to using an asymptotic approximation when group sizes exceed 10" (p. 73). If they are correct with their perception of common practices using as few as $n = 11$, the results in Table 2 demonstrate that the large sample approximation of the critical value prevents the statistic from converging with nominal alpha for seventeen of the twenty procedures for $\alpha = 0.05$, and for nineteen of twenty procedures for $\alpha = 0.01$.

The vagueness of what constitutes a large sample for the purposes of using the approximation to the critical values vanishes in view of the results in Table 2. For example, with α

= 0.05, large for the Match test is greater than 85. This does not mean the test performs poorly and should be removed from the data analyst's repertoire if one has a smaller sample size; rather, it means the researcher is advised to have at least 86 per group before relying on the large sample approximation of the critical values.

Statistics, Worked Examples, Large Scale Approximations

Single Population Tests

Goodness-of-fit statistics are single-population tests of how well observed data fit expected probabilities or a theoretical probability density function. They are frequently used as a preliminary test of the distribution assumption of parametric tests. The Kolmogorov-Smirnov goodness-of-fit test was studied.

Tests for location are used to make inferences about the location of a population. The measure of location is usually the median. If the median is not known but there is reason to believe that its value is M_0, then the null hypothesis is $H_0 : M = M_0$. The tests for location studied were the Sign test, Wilcoxon's Signed Rank test, and the Estimator of the Median for a continuous distribution.

Kolmogorov-Smirnov Goodness-of-Fit Test

The Kolmogorov-Smirnov (K-S) statistic was devised by Kolmogorov in 1933 and Smirnov in 1939. It is a test of goodness-of-fit for continuous data, based on the maximum vertical deviation between the empirical distribution function, $F_N(x)$, and the hypothesized cumulative distribution function, $F_0(x)$. Small differences support the null hypothesis while large differences are evidence against the null hypothesis.

The null hypothesis is H_0: $F_N(x) = F_0(x)$ for all x, and the alternative hypothesis is H_1: $F_N(x) \neq F_0(x)$ for at least some x where $F_0(x)$ is a completely specified continuous distribution. The empirical distribution function, $F_N(x)$, is a step function defined as:

$$F_N(x) = \frac{\text{number of sample values} \leq x}{N} \quad (1)$$

where N = sample size.

Test statistic. The test statistic, D_N, is the maximum vertical distance between the empirical distribution function and the cumulative distribution function.

$$D_N = \max\left[\max\left|F_N(x_i) - F_0(x_i)\right|, \max\left|F_N(x_{i-1}) - F_0(x_i)\right|\right] \quad (2)$$

Both vertical distances $F_N(x_i) - F_0(x_i)$ and $F_N(x_{i-1}) - F_0(x_i)$ have to be calculated in order to find the maximum deviation. The overall maximum of the two calculated deviations is defined as D_n.

For a one-tailed test against the alternatives H_1: $F_N(x) > F_0(x)$ or H_1: $F_N(x) < F_0(x)$ for at least some values of x, the test statistics are respectively:

$$D_N^+ = \max\left[F_N(x) - F_0(x)\right] \quad (3)$$

or

$$D_n^- = \max\left[F_0(x) - F_N(x)\right] \quad (4)$$

The rejection rule is to reject H_0 when $D_N \geq D_{N,\alpha}$ where $D_{N,\alpha}$ is the critical value for sample size N and level of significance α.

Large sample sizes. The null distribution of $4ND_N^{+\,2}$ (or $4ND_N^{-\,2}$) is approximately χ^2 with 2 degrees of freedom. Thus, the large sample approximation is

$$D_n^+ \approx \frac{1}{2}\sqrt{\frac{\chi^2_{\alpha,2}}{N}} \quad (5)$$

where $\chi^2_{\alpha,2}$ is the value for chi-square with 2 degrees of freedom.

Example. The K-S goodness-of-fit statistic was calculated for sample 1 (Table 3, Appendix), N = 15, against the cumulative frequency distribution of the multimodal lumpy data set. The maximum difference at step was 0.07463 and the maximum difference before step was 0.142610. Thus, the value of D_n is 0.142610. For a two-tail test, with α = .05, the large sample approximation is

$$1.3581/\sqrt{15} = 1.3581/\sqrt{15} = 0.35066.$$

Because 0.142610 < 0.35066, the null hypothesis cannot be rejected

.

The Sign Test

The Sign test is credited to Fisher as early as 1925. One of the first papers on the theory and application of the Sign test is attributed to Dixon and Mood in 1946 (Hollander & Wolfe, 1973). According to Neave and Worthington (1988), the logic of the Sign test is “almost certainly the oldest of all formal statistical tests as there is published evidence of its use long ago by J. Arbuthnott (1710)!” (p. 65).

The Sign test is a test for a population median. It can also be used with matched data as a test for equality of medians, specifically when there is only dichotomous data. (Otherwise, the Wilcoxon Signed Rank is more powerful.) The test is based on the number of values above or below the hypothesized median. Gibbons (1971) referred to the Sign test as the nonparametric counterpart of the one-sample *t* test. The Sign test tests the null hypothesis H_0: $M = M_0$, where M is the sample median and M_0 is the hypothesized population median, against the alternative hypothesis H_1: $M \neq M_0$. One-tailed test alternative hypotheses are of the form H_1: $M < M_0$ and H_1: $M > M_0$.

Procedure. Each x_i is compared with M_0. If $x_i > M_0$ then a plus symbol ‘+’ is recorded. If $x_i < M_0$ then a minus symbol ‘–’ is recorded. In this way all data are reduced to ‘+’ and ‘–’ symbols.

Test statistic. The test statistic is the number of ‘+’ symbols or the number of ‘–’ symbols. If the expectation under the alternative hypothesis is that there will be a preponderance of ‘+’ symbols, the test statistic is the number of ‘–’ symbols. Similarly, if the expectation is a preponderance of ‘–’ symbols, the test statistic is the number of ‘+’ symbols. If the test is two-tailed, use the smaller of the two. Thus, depending on the context,

$$S = \text{number of ‘+’ or ‘–’ symbols} \qquad (6)$$

Large sample sizes. The large sample approximation is given by

$$S^* = \frac{S - \frac{N}{2}}{\sqrt{\frac{N}{4}}} \qquad (7)$$

where S is the test statistic and N is the sample size. S^* is compared to the standard normal z scores for the appropriate α level.

Example. The Sign test was calculated using sample 1 (Table 3, Appendix), N = 15. The population median is 18.0. The number of minus symbols is 7 and the number of plus symbols is 8. Therefore S = 7. The large sample approximation, S^*, using formula (7) is -.258199. The null hypothesis cannot be rejected because -.258199 > -1.95996.

Wilcoxon’s Signed Rank Test

The Signed Rank test was introduced by Wilcoxon in 1945. This statistic uses the ranks of the absolute differences between x_i and M_0 along with the sign of the difference. It uses the relative magnitudes of the data. This statistic can also be used to test for symmetry and to test for equality of location for paired replicates. The null hypothesis is H_0: $M = M_0$, which is tested against the alternative H_1: $M \neq M_0$. The one-sided alternatives are H_1: $M < M_0$ and H_1: $M > M_0$.

Procedure. Compute the differences, D_i, by the formula

$$D_i = x_i - M_0 . \qquad (8)$$

Rank the absolute value of the differences in ascending order, keeping track of the individual signs.

Test statistic. The test statistic is the sum of either the positive ranks or the negative ranks. If the alternative hypothesis suggests that the sum of the positive ranks should be larger, then

$$T^- = \text{the sum of negative ranks} \qquad (9)$$

If the alternative hypothesis suggests that the sum of the negative ranks should be larger, then

$$T^+ = \text{the sum of positive ranks} \qquad (10)$$

For a two-tailed test, T is the smaller of the two rank sums. The total sum of the ranks is $\frac{N(N+1)}{2}$, which gives the following relationship:

$$T^+ = \frac{N(N+1)}{2} - T^- . \qquad (11)$$

Large sample sizes. The large sample approximation is given by

$$z = \frac{T - \frac{N(N+1)}{4}}{\sqrt{\frac{N(N+1)(2N+1)}{24}}} \qquad (12)$$

where T is the test statistic. The resulting z is compared to the standard normal z for the appropriate alpha level.

Example. The Signed Rank test was computed using the data from sample 1 (Table 3, Appendix), $N = 15$. The median of the population is 18.0. Tied differences were assigned midranks. The sum of the negative ranks was 38.5 and the sum of the positive ranks was 81.5. Therefore the Signed Rank statistic is 38.5. The large sample approximation is $\frac{-21.5}{\sqrt{310}} = \frac{-21.5}{17.6068} = -1.22112$. Because $-1.22112 > -1.95996$, the null hypothesis is not rejected.

Estimator of the Median (Continuous Distribution)

The sample median is a point estimate of the population median. This procedure provides a 1-α confidence interval for the population median. It was designed for continuous data.

Procedure. Let N be the size of the sample. Order the N observations in ascending order, $x_{(1)} \le x_{(2)} \le \ldots \le x_{(N)}$. Let $x_{(0)} = -\infty$ and $x_{(N+1)} = \infty$. These N+2 values form N+1 intervals $(x_{(0)}, x_{(1)}), (x_{(1)}, x_{(2)}), \ldots, (x_{(N-1)}, x_{(N)}), (x_{(N)}, x_{(N+1)})$. The i^{th} interval is defined as $(x_{(i-1)}, x_{(i)})$ with i = 1, 2, . . . , N, N+1. The probability that the median is in any one interval can be computed from the binomial distribution. The confidence interval for the median requires that r be found such that the sum of the probabilities of the intervals in both the lower and upper ends give the best conservative approximation of α/2, according to the following:

$$\frac{\alpha}{2} \approx \sum_{j=0}^{r} \binom{N}{j} \frac{1}{2^N} = \sum_{j=N-r}^{N} \binom{N}{j} \frac{1}{2^N}. \qquad (13)$$

Thus, ($x_{(r)}$, $x_{(r+1)}$) is the last interval in the lower end, making $x_{(r+1)}$ the lower limit of the confidence interval. By a similar process, $x_{(N-r)}$ is the upper limit of the confidence interval.

Large sample sizes. Deshpande, Gore, and Shanubhogue (1995) stated "one may use the critical points of the standard normal distribution, to choose the value of r + 1 and n – r, in the following way": r + 1 is the integer closest to

$$\frac{N}{2} - z_{\alpha/2}\left(\frac{N}{4}\right)^{\frac{1}{2}} \qquad (14)$$

where $z_{\alpha/2}$ is the upper α/2 critical value of the standard normal distribution.

Example. The data from sample 1 (Table 3, Appendix), N = 15, were used to compute the Estimator of the Median. The population median is 18.0. For the given N and α = .05, the value of r is 3. The value of r + 1 is 4, and n – r is 12. The 4th value is 13 and the 12th value is 33. Therefore the interval is (13, 33). The large sample approximation yields 7.5 – 1.95996(1.9365) = 7.5 – 3.70 = 3.80. The closest integer is r + 1 = 4, so r = 3 and N – r = 12, resulting in the same interval, (13, 33). The interval contains the population median, 18.0.

Two Sample Tests

The two-sample layout consists of independent random samples drawn from two populations. This study examined two sample tests for general differences, two sample location tests, and two sample scale tests.

When differences between two samples are not expected to be predominantly differences in location or differences in scale, a test for general differences is appropriate. Generally differences in variability are related to differences in location. Two tests for differences were considered, the Kolmogorov-Smirnov test for general differences and Rosenbaum's test.

Two sample location problems involve tests for a difference in location between two samples when the populations are assumed to be similar in shape. The idea is that $F_1(x) = F_2(x+\theta)$ or $F_1(x) = F_2(x-\theta)$ where θ is the distance between the population medians. Tukey's quick test, the Wilcoxon (Mann-Whitney) statistic, and the

Hodges-Lehmann estimator of the difference in location for two populations were considered.

In two sample scale tests, the population distributions are usually assumed to have the same location with different spreads. However, Neave and Worthington (1988) cautioned that tests for difference in scale could be severely impaired if there is a difference in location as well. The following nonparametric tests for scale were studied: the Siegel-Tukey test, the Mood test, the Savage test for positive random variables, and the Ansari-Bradley test.

Kolmogorov-Smirnov Test for General Differences

The Kolmogorov-Smirnov test compares the cumulative distribution frequencies of the two samples to test for general differences between the populations. The sample cdf "is an approximation of the true cdf of the corresponding population – though, admittedly, a rather crude one if the sample size is small" (Neave & Worthington, 1988, p. 149). This property was used in the goodness-of-fit test above. Large differences in the sample cdfs can indicate a difference in the population cdfs, which could be due to differences in location, spread, or more general differences in the distributions. The null hypothesis is $H_0 : F_1(x) = F_2(x)$ for all *x*. The alternative hypothesis is $H_1 : F_1(x) \neq F_2(x)$ for some *x*.

Procedure. The combined observations are ordered from smallest to largest, keeping track of the sample membership. Above each score, write the cdf of sample 1, and below each score write the cdf of sample 2. Because the samples are of equal sizes, it is only necessary to use the numerator of the cdf. For example, the cdf(x_i) = $\frac{i}{n}$. Then, write *i* above x_i for sample 1. Find the largest difference between the cdf for sample 1 and the cdf for sample 2.

Test statistic. The test statistic is *D**. $D^* = n_1 n_2 D$, and $D^* = n^2 D$ for equal sample sizes. The above procedure yields *nD*. Thus

$$D^* = n(nD)\ . \qquad (15)$$

The greatest difference found by the procedure is multiplied by the sample size.

Large sample sizes. The distribution is approximately χ^2 with 2 degrees of freedom as sample size increases, as it is for the goodness-of-fit test. The large sample approximation for *D* is

$$D = \frac{1}{2}\sqrt{\frac{\chi^2_{\alpha,2}(n_1 + n_2)}{n_1 n_2}} \qquad (16)$$

where $\chi^2_{\alpha,2}$ is the value for chi-square with 2 degrees of freedom for the appropriate alpha level, and n_1 and n_2 are the two sample sizes. The resulting *D* is used in formula (15).

Example. This example used the data from sample 1 and sample 5 (Table 3, Appendix), $n_1 = n_2 = 15$. The greatest difference (*nD*) between the cdfs of the two samples is $nD = 3$. Therefore $D^* = 15(3) = 45$. The large sample approximation is $15^2(1.3581)\sqrt{\frac{30}{225}} = 225(1.3581)(.365148) = 111.579301$. Because $45 < 111.579301$, the null hypothesis cannot be rejected.

Rosenbaum's Test

Rosenbaum's test, which was developed in 1965, is useful in situations where an increase in the measure of location implies an increase in variation. It is a quick and easy test based on the number of observations in one sample greater than the largest observation in the other sample. The null hypothesis is that both populations have the same location and spread against the alternative, that both populations differ in location and spread.

Procedure. The largest observation in each sample is identified. If the largest overall observation is from sample 1, then count the number of observations from sample 1 greater than the largest observation from sample 2. If the largest overall observation is from sample 2, then count the number of observations from sample 2 greater than the largest observation from sample 1.

Test statistic. The test statistic is the number of extreme observations. *R* is the number of observations from sample 1 greater than the largest observation in sample 2, or the number of observations from sample 2 greater than the largest observation in sample 1.

Large sample sizes. As sample sizes increase, $\frac{n_1}{N} \to p$ and the probability that the number of extreme values equals h approaches p^h.

Example. Rosenbaum's statistic was calculated using samples 1 and 5 (Table 3, Appendix), $n_1 = n_2 = 15$. The maximum value from sample 1 is 39, and from sample 2 it is 33. There are three values from sample 1 greater than 33: 34, 36, and 39. Hence, $R = 3$. The large sample approximation is $(.5)^3 = 0.125$. Because 0.125 > .05, the null hypothesis cannot be rejected.

Tukey's Quick Test

Tukey published a quick and easy test for the two-sample location layout in 1959. It is easy to calculate and in most cases does not require the use of tables. The most common one-tailed critical values are 6 ($\alpha = .05$) and 9 ($\alpha = .01$). These critical values can be used for most sample sizes. The statistic is the sum of extreme runs in the ordered combined samples. When a difference in location exists, more observations from sample 1 will be expected at one end and more observations from sample 2 will be expected at the other end.

Procedure. The combined samples can be ordered, but it is only necessary to order the largest and smallest observations. If both the maximum and minimum values come from the same sample the test is finished, the value of $T_y = 0$, and the null hypothesis is not rejected.

For the one-tailed test, the run on the lower end should come from the sample expected to have the lower median, and the run on the upper end should come from the sample expected to have the larger median. For a two-tailed test, it is possible to proceed with the test as long as the maximum and minimum observations come from different samples.

Test statistic. T_y is defined as follows for the alternative hypothesis, H_1: $M_1 > M_2$. T_y is the number of observations from sample 2 less than the smallest observation of sample 1, plus the number of observations from sample 1 greater than the largest observation from sample 2. For the alternative H_1: $M_2 > M_1$ the samples are reversed. For the two-tailed hypothesis H_1: $M_1 \neq \mathrm{M}_2$, both possibilities are considered.

Critical values. As stated above, generally, the critical value for $\alpha = .05$ is 6, and is 9 for $\alpha = .01$. There are tables available. As long as the ratio of n_x to n_y is within 1 to 1.5, these critical values work well. There are corrections available when the ratio exceeds 1.5. For a two-tailed test the critical values are 7 ($\alpha = .05$) and 10 ($\alpha = .01$).

Large sample sizes. The null distribution is based on the order of the elements of both samples at the extreme ends. It does not depend on the order of the elements in the middle. Neave and Worthington (1988, p. 125) gave the following formula:

$$\text{Prob}(T_y \geq h) = \frac{pq(q^h - p^h)}{q-p} \tag{17}$$

for $h \geq 2$. When the sample sizes are equal, p = q = .5. Then the probability of $T_y \geq h$ is $h2^{-(h+1)}$. For a two-tailed test the probability is doubled.

Example. The Tukey test was calculated using the data in sample 1 and sample 5 (Table 3, Appendix), $n_1 = n_2 = 15$. The maximum value (39) is from sample 1 and the minimum (2) is from sample 5, so the test may proceed. The value of $T_y = 1 + 3 = 4$. For a two-tailed test with $\alpha = .05$, the large sample approximation is $2(4)(2^{-5}) = 0.25$. Because 0.25 > .05, the null hypothesis cannot be rejected.

Wilcoxon (Mann-Whitney) Test

In 1945, Wilcoxon introduced the Rank Sum test, and in 1947 Mann and Whitney presented a different version of the same test. The Wilcoxon statistic is easily converted to the Mann-Whitney U statistic. The hypotheses of the test are $H_0 : F_1(x) = F_2(x) \text{ for all } x$ against the two-tailed alternative, $H_0 : F_1(x) \neq F_2(x)$. The one-tailed alternative is $H_1 : F_1(x) = F_2(x+\theta)$.

Procedure. For the Wilcoxon test, the combined samples are ordered, keeping track of sample membership. The ranks of the sample that is expected, under the alternative hypothesis, to have the smallest sum, are added. The Mann-Whitney test is conducted as follows. Put all the observations in order, noting sample membership. Count how many of the observations of one sample exceed each observation in the first sample. The sum of these counts is the test statistic, U.

Test statistic. For the Wilcoxon test,

$$S_n = \sum_{j=1}^{n} R_j \tag{18}$$

where R_j are the ranks of sample n and S_n is the sum of the ranks of the sample expected to have the smaller sum.

For the Mann-Whitney test, calculate the U statistic for the sample expected to have the smaller sum under the alternative hypothesis.

U_{n2} = the sum of the observations in n_1 exceeding each observation in n_2. (19)

U_{n1} = the sum of the observations in n_2 exceeding each observation in n_1. (20)

There is a linear relation between S_n and U_n. It is expressed as

$$U_{n_1} = S_{n_1} - \frac{1}{2} n_1 (n_1 + 1) \tag{21}$$

and similarly,

$$U_{n_2} = S_{n_2} - \frac{1}{2} n_2 (n_2 + 1) \tag{22}$$

where

$$U_{n_1} = n_1 n_2 - U_{n_2} \quad . \tag{23}$$

In a two-tailed test, use the smallest U statistic to test for significance.

Large sample sizes. The large-sample approximation using the Wilcoxon statistic, S_{n1} is:

$$z = \frac{S_{n_1} - \frac{n_1(n_1 + n_2 + 1)}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} . \tag{24}$$

The large-sample approximation with the U statistic is

$$z = \frac{U + \frac{1}{2} - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}} . \tag{25}$$

In either case, reject H_0 if $z < -z_\alpha$ (or $z < -z_{\alpha/2}$ for a two-tailed test).

Example. The Wilcoxon Rank Sum (Mann-Whitney) statistic was calculated with data from sample 1 and sample 5 (Table 3, Appendix), $n_1 = n_2 = 15$. The combined samples were ranked, using midranks in place of the ranks of tied observations. The rank sum for sample 1 was 258.5 and for sample 5, 206.5. Hence S = 206.5. Calculating the U statistic, U= 206.5 – 0.5(15)(16) = 86.5. The large sample approximation for U is $\frac{86.5 + .5 - .5(15^2)}{\sqrt{\frac{15^2(31)}{12}}} = \frac{-25.5}{24.1091} = -1.05769$. Because $-1.05769 > -1.95996$, the null hypothesis cannot be rejected.

Hodges-Lehmann Estimator of the Difference in Location

It is often useful to estimate the difference in location between two populations. Suppose two populations are assumed to have similar shapes, but differ in locations. The objective is to develop a confidence interval that will have the probability of 1-α that the difference lies within the interval.

Procedure. All the pairwise differences are computed, x_i–y_j . For sample sizes of n_1 and n_2, there are n_1n_2 differences. The differences are put in ascending order. The task is to find two integers l and u such that the probability that the difference lies between l and u is equal to 1–α. These limits are chosen symmetrically. The appropriate lower tail critical value is found for the Mann-Whitney U statistic. This value is the upper limit of the lower end of the differences. Therefore, l is the next consecutive integer. The upper limit of the confidence interval is the u^{th} difference from the upper end, found by $u = n_1n_2 - l+1$. The interval (l, u) is the confidence interval for the difference in location for the two populations.

Large sample sizes. Approximate *l* and *u* by

$$l = \left[\frac{n_1 n_2}{2} - z_{\alpha/2}\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} - \frac{1}{2} \right] \quad (26)$$

and

$$u = \left[\frac{n_1 n_2}{2} + z_{\alpha/2}\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} - \frac{1}{2} \right] \quad (27)$$

“where the square brackets denote integer nearest to the quantity within, and $z_{\alpha/2}$ is the suitable upper critical point of the standard normal distribution” (Deshpande, et al., 1995, p. 45, formulas rewritten for consistency of notation with this article).

Example. The Hodges-Lehmann estimate of the difference in location was computed using samples 1 and 5 (Table 3, Appendix), $n_1 = n_2 = 15$. All possible differences were computed and ranked. Using the large sample approximation formula (26), l = 112.5–1.95596(24.109)–0.5 = 64.844. Thus, l = 65 and the lower bound is the 65th difference, which is -4. The upper bound is the 65th difference from the upper end, or the 225 –65+1=161st value, 14. The confidence interval is (-4, 14).

Siegel-Tukey Test

The Siegel-Tukey test was developed in 1960. It is similar in procedure to the Wilcoxon Rank Sum test for difference in location. It is based on the logic that if two samples come from populations with the same median, the one with the greater variability will have more extreme scores. An advantage of the Siegel-Tukey statistic is that it uses the Wilcoxon table of critical values or can be transformed into a *U* statistic for use with the Mann-Whitney *U* table of critical values.

The hypotheses for a two-tailed test are H_0: There is no difference in spread between the two populations, which is tested against the alternative H_1: There is some difference in spread between the two populations.

Procedure. The two combined samples are ordered, keeping track of sample membership. The ranking proceeds as follows: the lowest observation is ranked 1, the highest is ranked 2, and the next highest 3. Then the second lowest is ranked 4 and the subsequent observation ranked 5. The ranking continues to alternate from lowest to highest, ranking two scores at each end. If there is an odd number of scores, the middle score is discarded and the sample size reduced accordingly. Below is an illustration of the ranking procedure:

1 4 5 8 9 … N … 7 6 3 2
where $N = n_1 + n_2$.

Test statistic. The sum of ranks is calculated for one sample. The rank sum can be used with a table of critical values or it can be transformed into a *U* statistic by one of the following formulas:

$$U^* = R_{n_1} - \frac{1}{2} n_1 (n_1 + 1) \quad (28)$$

or

$$U^* = R_{n_2} - \frac{1}{2} n_2 (n_2 + 1)\,. \quad (29)$$

Large sample sizes. The large-sample approximations are the same for the Siegel-Tukey test as for the Wilcoxon Rank Sum or the Mann-Whitney *U* statistic, formulas (24) and (25).

Example. The Siegel-Tukey statistic was calculated using sample 1 and sample 5 (Table 3, Appendix), $n_1 = n_2 = 15$. The samples were combined and ranked according to the method described. Then, tied ranks were averaged. The sum of ranks was 220.5 for sample 1, and 244.5 for sample 5. The *U* statistic is 220.5 – .5(15)(16) = 100.5. The large sample approximation is

$$z = \frac{100.5 + .5 - .5(15^2)}{\sqrt{\frac{15^2(31)}{12}}} = \frac{-11.5}{24.109127} = -0.476998.$$

Because –0.476998> –1.95996, the null hypothesis cannot be rejected.

The Mood Test

In 1954, the Mood test was developed based on the sum of squared deviations of one sample’s ranks from the average combined ranks. The null hypothesis is that there is no difference in spread against the alternative hypothesis that there is some difference.

Procedure. Let sample 1 be $x_1, x_2, \ldots, x_{n1}$ and let sample 2 be $y_1, y_2, \ldots, y_{n2}$. Arrange the combined samples in ascending order and rank the observations from 1 to n_1+ n_2. Let R_i be the rank of x_i. Let $N = n + n_2$. If N is odd, the middle rank is ignored to preserve symmetry.

Test statistic. The test statistic is

$$M = \sum_{i=1}^{n_1} \left(R_i - \frac{n_1 + n_2 + 1}{2} \right)^2 . \qquad (30)$$

Large sample sizes. The large sample approximation is

$$z = \frac{M - \dfrac{n_1(N^2-1)}{12}}{\sqrt{\dfrac{n_1 n_2 (N+1)(N^2-4)}{180}}} \qquad (31)$$

where N = n_1 + n_2 and M is the test statistic.

Example. The Mood statistic was calculated using sample 1 and sample 5 (Table 3, Appendix), n_1 = n_2 = 15. The combined samples are ranked, with midranks assigned to the ranks of tied observations. The mean of the ranks is 15.5, and the sum of squared deviations of the ranks from the mean for sample 1 was calculated, yielding M=1257. The large sample approximation is $\dfrac{1257 - 1123.75}{\sqrt{34720}} = \dfrac{133.25}{186.333} =$ 0.71512. Because 0.71512 < 1.95596, the null hypothesis cannot be rejected.

The Savage Test for Positive Random Variables

Unlike the Siegel-Tukey test and the Mood test, the Savage test does not assume that location remains the same. It is assumed that differences in scale cause a difference in location. The samples are assumed to be drawn from continuous distributions.

The null hypothesis is that there is no difference in spread, which is tested against the two-tailed alternative that there is a difference in variability.

Procedure. Let sample 1 be $x_1, x_2, \ldots, x_{n1}$ and let sample 2 be $y_1, y_2, \ldots, y_{n2}$. The combined samples are ordered, keeping track of sample membership. Let R_i be the rank for x_i. The test statistic is computed for either sample.

Test statistic. The test statistic is

$$S = \sum_{i=1}^{n_1} a(R_i) \qquad (32)$$

where

$$a(i) = \sum_{j=N+1-i}^{N} \frac{1}{j} \qquad (33)$$

such that

$$a(1) = \frac{1}{N},\ a(2) = \frac{1}{N-1} + \frac{1}{N},\ \ldots,\ a(N) = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{N-1} + \frac{1}{N}.$$

Large sample sizes. For large sample sizes the following normal approximation may be used.

$$S^* = \frac{S - n_2}{\sqrt{\dfrac{n_1 n_2}{N-1}\left(1 - \dfrac{1}{N}\displaystyle\sum_{j=1}^{N} \frac{1}{j}\right)}} . \qquad (34)$$

S^* is compared to the critical z value from the standard normal distribution.

Example. The Savage statistic was calculated using samples 1 and 5 (Table 3, Appendix), n_1 = n_2 = 15. Using sample 1, S = 18.3114. The large sample approximation is $\dfrac{18.3114 - 15}{\sqrt{7.7586(.86683)}} = \dfrac{3.114}{2.59334} =$ 1.27689. Because 1.27689 < 1.95596, the null hypothesis cannot be rejected.

Ansari-Bradley Test

This is a rank test for spread when the population medians are the same. The null hypothesis is that the two populations have the same spread, which is tested against the alternative that the variability of the two populations differs.

Procedure. Order the combined samples, keeping track of sample membership. Rank the smallest and largest observation 1. Rank the second lowest and second highest 2. If the combined sample size, N, is odd, the middle score will be ranked $\dfrac{N+1}{2}$ and if N is even the middle two ranks will be $\frac{N}{2}$. The pattern will be either 1, 2, 3, . . . , $\frac{N+1}{2}$, . . . , 3, 2, 1 (N odd), or 1, 2, 3, . . ., $\frac{N}{2}$, $\frac{N}{2}$, . . . , 3, 2, 1 (N even).

Test statistic. The test statistic, W, is the sum of the ranks of sample 1.

$$W = \sum_{i=1}^{n_1} R_i \qquad (35)$$

where R_i is the rank of the i^{th} observation of a sample.

Large sample sizes. There are two formulas. If N is even, use

$$W^* = \frac{W - \frac{n_1(n_1+n_2+2)}{4}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+2)(n_1+n_2-2)}{48(n_1+n_2-1)}}} \qquad (36)$$

and if N is odd, use

$$W^* = \frac{W - \frac{n_1(n_1+n_2+1)^2}{4(n_1+n_2)}}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)[3+(n_1+n_2)^2]}{48(n_1+n_2)^2}}} . \qquad (37)$$

Reject the null hypothesis if $W^* \geq z_{\alpha/2}$.

Example. The Ansari-Bradley statistic was calculated using samples 1 and 5 (Table 3, Appendix), $n_1 = n_2 = 15$. The combined samples were ranked using the method described, and the ranks of tied observations were assigned average ranks. The two-tailed statistic, W, is 126.5, the rank sum of sample 5. The large sample approximation is $\frac{126.5-120}{\sqrt{144.8276}} = \frac{6.5}{12.034}$ = 0.54. Because 0.54 < 1.95596, the null hypothesis cannot be rejected.

Comparisons Of Several Populations

This section considered tests against an omnibus alternative and tests involving an ordered hypothesis. The omnibus tests were the Kruskal-Wallis test and Friedman's test. The tests for ordered alternatives are the Terpstra-Jonckheere test, Page's test, and the Match test.

The Kruskal-Wallis statistic is a test for independent samples. It is analogous to the one-way analysis of variance. Friedman's test is an omnibus test for k related samples, and is analogous to a two-way analysis of variance.

Comparisons of several populations with ordered alternative hypotheses are extensions of a one-sided test. When an omnibus alternative states only that there is some difference between the populations, an ordered alternative specifies the order of differences. Three tests for an ordered alternative were included: the Terpstra-Jonckheere Test, Page's Test, and the Match Test.

Kruskal-Wallis Test

The Kruskal-Wallis test was derived from the F test in 1952. It is an extension of the Wilcoxon (Mann–Whitney) test. The null hypothesis is that the k populations have the same median. The alternative hypothesis is that at least one sample is from a distribution with a different median.

Procedure. Rank all the observations in the combined samples, keeping track of the sample membership. Compute the rank sums of each sample. Let R_i equal the sum of the ranks of the i^{th} sample of sample size n_i. The logic of the test is that the ranks should be randomly distributed among the k samples.

Test statistic. The formula is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1) \qquad (38)$$

where N is the total sample size, n_i is the size of the ith group, k is the number of groups, and R_i is the rank-sum of the ith group. Reject H_0 when H ≥ critical value.

Large sample sizes. For large sample sizes, the null distribution is approximated by the χ^2 distribution with k – 1 degrees of freedom. Thus, the rejection rule is to reject H_0 if $H \geq \chi^2_{\alpha,k-1}$ where $\chi^2_{\alpha,k-1}$ is the value of χ^2 at nominal α with k – 1 degrees of freedom.

Example. The Kruskal-Wallis statistic was calculated using samples 1–5 (Table 3, Appendix), $n_1 = n_2 = n_3 = n_4 = n_5 = 15$. The combined samples were ranked, and tied ranks were assigned midranks. The rank sums were: R_1 = 638, R_2 = 595, R_3 = 441.5, R_4 = 656.5, and R_5 = 519. The sum of R_i^2 = 1,656,344.5, i = 1, 2, 3, 4, 5.

$$H = \frac{12}{75(76)}\left(\frac{1{,}656{,}344.5}{15}\right) - 3(76) =$$
$$0.00211(110{,}422.97 - 228 = 4.47$$

Thus, H = 4.47. The large sample approximation with 5 – 1 = 4 degrees of freedom at α = .05 is χ^2 = 9.488. Because 4.47 < 9.488, the null hypothesis cannot be rejected.

Friedman's Test

The Friedman test was developed as a test for k related samples in 1937. The null hypothesis is that the samples come from the same population. The alternative hypothesis is that at least one of the samples comes from a different population. Under the truth of the null hypothesis, this test only requires exchangeability (or, if variances differ, compound symmetry) and the ability to rank the data. The data are arranged in k columns and n rows, where each row contains k related observations.

Procedure. Rank the observations for each row from 1 to k. For each of the k columns, the ranks are added and averaged, and the mean is designated $\overline{R}_j$. The overall mean of the ranks is $\overline{R} = \frac{1}{2}(k+1)$. The sum of the squares of the deviations of mean of the ranks of the columns from the overall mean rank is computed. The test statistic is a multiple of this sum.

Test statistic. The test statistic for Friedman's test is M, which is a multiple of S, as follows:

$$S = \sum_{j=1}^{k} (\overline{R}_j - \overline{R})^2 \qquad (39)$$

$$M = \frac{12n}{k(k+1)} S \qquad (40)$$

where n is the number of rows, and k is the number of columns. An alternate formula that does not use S is as follows.

$$M = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1) \qquad (41)$$

where n is the number of rows, k is the number of columns, and R_j is the rank sum for the jth column, j = 1, 2, 3, . . . , k.

Large sample sizes. For large sample sizes, the critical values can be approximated by χ^2 with k – 1 degrees of freedom.

Example. Friedman's statistic was calculated with samples 1 – 5 (Table 3, Appendix), $n_1 = n_2 = n_3 = n_4 = n_5 = 15$. The rows were ranked, with the ranks of tied observations replaced with midranks. The column sums are: R_1 = 48.5, R_2 = 47, R_3 = 33, R_4 = 52.5, and R_5 = 44. The sum of the squared rank sums is 10,342.5. $M = \frac{12}{15 \cdot 5 \cdot 6}(10{,}342.5) - 3 \cdot 15 \cdot 6$ =0.0267(10,342.5) –270 = 5.8. The large sample approximation is χ^2 with 5 – 1 = 4 degrees of freedom and α = .05, which is 9.488. Because 5.8 < 9.488, the null hypothesis cannot be rejected.

Terpstra-Jonckheere Test

This is a test for more than two independent samples. It was first developed by Terpstra in 1952 and later independently developed by Jonckheere in 1954. The null hypothesis is that the medians of the samples are equal, which is tested against the alternative that the medians are either decreasing or increasing. This test is based on the Mann-Whitney U statistic, where U is calculated for each pair of samples and the U statistics are added.

Suppose the null hypothesis is H_0: $F_1(x) \geq F_2(x) \geq F_3(x) \geq \ldots \geq F_k(x)$ and the alternative hypothesis is H_0: $F_1(x) < F_2(x) < F_3(x) < \ldots < F_k(x)$ for i = 1, 2, . . . k. The U statistic is calculated for each of the $\frac{k(k-1)}{2}$ pairs, which are ordered so that the smallest U is calculated.

Test statistic. The test statistic is the sum of the U statistics.

$$W = U_{k,1} + U_{k,2} + \ldots + U_{3,1} + U_{3,2} + U_{2,1} \qquad (42)$$

where $U_{i,j}$ is the number of pairs when the observation from sample j is less than the observation from sample i.

Large sample sizes. The null distribution of W approaches normality as the sample size increases. The mean of the distribution is

$$\mu = \frac{(N^2 - \sum n_i^2)}{4} \tag{43}$$

and the standard deviation is

$$\sigma = \sqrt{\frac{N^2(2N+3) - \sum n_i^2(2n_i+3)}{72}} \tag{44}$$

The critical value for large samples is given by

$$W \le \mu - z\sigma - \frac{1}{2} \tag{45}$$

where z is the standard normal value, and $\frac{1}{2}$ is a continuity correction.

Example. The Terpstra-Jonckheere statistic was calculated with samples 1 – 5 (Table 3, Appendix), $n_1 = n_2 = n_3 = n_4 = n_5 = 15$. This was done as a one-tailed test with α = .05. The U statistics for each sample were calculated. $U_{5,1}$ = 135, $U_{5,2}$ = 124, $U_{5,3}$ = 91, $U_{5,4}$ = 136, $U_{4,1}$ = 103, $U_{4,2}$ = 97, $U_{4,3}$ = 71, $U_{3,1}$ = 145, $U_{3,2}$ = 142, and $U_{2,1}$ = 121, for a total W = 1,165. The large sample approximation was calculated with μ = 1125 and σ = 106.94625. The approximation is 1125 – 1.6449(106.9463) - .5 = 948.584. Because 1165 > 948.584 the null hypothesis cannot be rejected.

Page's Test

Page's test for an ordered hypothesis for k > 2 related samples was developed in 1963. It takes the form of a randomized block design with k columns and n rows. The null hypothesis is $H_0 : M_1 = M_2 = \ldots = M_k$ and the alternative hypothesis is $H_1 : M_1 < M_2 < \ldots < M_k$ for i = 1, 2, . . . k. For this test, the alternative must be of this form. The samples need to be reordered if necessary.

Procedure. The data are ranked from 1 to k for each row, creating a table of the ranks. The ranks of each of the k columns are totaled. If the null hypothesis is true, the ranks should be evenly distributed over the columns, whereas if the alternative is true, the ranks sums should increase with the column index.

Test statistic. Each column rank-sum is multiplied by the column index. The test statistic is

$$L = \sum_{i=1}^{k} iR_i \tag{46}$$

where i is the column index, i = 1, 2, 3, . . . , k, and R_i is the rank sum for the i^{th} column.

Large sample sizes. The mean of L is

$$\mu = \frac{nk(k+1)^2}{4} \tag{47}$$

and the standard deviation is

$$\sigma = \sqrt{\frac{nk^2(k+1)(k^2-1)}{144}}\,. \tag{48}$$

For a given α, the approximate critical region is

$$L \ge \mu + z\sigma + \frac{1}{2}\,. \tag{49}$$

Example. Page's statistic was calculated with samples 1 – 5 (Table 3, Appendix), $n_1 = n_2 = n_3 = n_4 = n_5 = 15$. This was done as a one-tailed test with α = .05. The rows are ranked with midranks assigned to tied ranks. The column sums are: R_1 = 48.5, R_2 = 47, R_3 = 33, R_4 = 52.5, and R_5 = 44. The statistic, L, is the sum of iR_i^2 = 671.5, where i = 1, 2, 3, 4, 5. The large sample approximation was calculated with μ = 675 and σ = 19.3649. The approximation is 675 + 1.64485(19.3649) + .5 = 707.352. Because 671.5 < 707.352, the null hypothesis cannot be rejected.

The Match Test for Ordered Alternatives

The Match test is a test for k > 2 related samples with an ordered alternative hypothesis. The Match test was developed by Neave and Worthington (1988). It is very similar in concept to Page's test, but instead of using rank-sums, it uses the number of matches of the ranks with the expected ranks plus half the near matches. The null hypothesis is H_0: M_1= M_2= ... = M_k and the alternative hypothesis is H_0: M_1< M_2< ...< M_k for i = 1, 2, . . . k.

Procedure. A table of ranks is compiled with the observations in each row ranked from 1 to k. Tied observations are assigned average ranks. Each rank, r_i, is compared with the expected rank, i, the column index. If the rank equals the column index, it is a match. Count the number of matches. Every non-match such that $0.5 \le |r_i - i| \le 1.5$ is counted as a near match.

Test statistic. The test statistic is

$$L_2 = L_1 + \frac{1}{2}(\text{number of near matches}) \qquad (50)$$

where L_1 is the number of matches.

Large sample sizes. The null distribution approaches a normal distribution for large sample size. The mean and standard deviation for L_2 are as follows:

$$\mu = n\left(2 - \frac{1}{k}\right) \qquad (51)$$

and

$$\sigma = \sqrt{\frac{n}{k}\left(\frac{3(k-2)}{2}\right) + \frac{1}{k(k-1)}}\,. \qquad (52)$$

For a given level of significance α the critical value approximation is

$$L_2 \ge \mu + z\sigma + \frac{1}{2} \qquad (53)$$

where z is the upper-tail critical value from the standard normal distribution and $\frac{1}{2}$ is a continuity correction.

Example. The Match statistic was calculated with samples 1 – 5 (Table 3, Appendix), $n_1 = n_2 = n_3 = n_4 = n_5 = 15$. This was done as a one-tailed test with α = .05. The rows are ranked, with midranks assigned for tied observations. The number of matches for the five columns are 3, 3, 2, 2, and 1, for L_1 = 11. The number of near matches were 1, 6, 8, 8, and 4, for L_2 = 27. The statistic, L = 11 + .5(27) = 24.5. For the large sample approximation, μ = 27 and σ = 3.68103. The approximation is 27 + 1.6449(3.68103) + .5 = 33.5549. Because 24.5 < 33.5549, the null hypothesis cannot be rejected.

Rank Correlation Tests

The rank correlation is a measure of the association of a pair of variables. Spearman's rank correlation coefficient (rho) and Kendall's rank correlation coefficient (tau) were studied.

Spearman's Rank Correlation Coefficient

Spearman's rank correlation (rho) was published in 1904. Let X and Y be the two variables of interests. Each observed pair is denoted (x_i, y_i). The paired ranks are denoted (r_i, s_i), where r_i is the rank of x_i and s_i is the rank of y_i. The null hypothesis for a two-tailed test is $H_0 : \rho = 0$, which is tested against the alternative $H_1 : \rho \ne 0$. The alternative hypotheses for a one-tailed test are $H_1 : \rho > 0$ or $H_1 : \rho < 0$.

Procedure. Rank both X and Y scores while keeping track of the original pairs. Form the rank pairs (r_i, s_i) which correspond to the original pair, (x_i, y_i). Calculate the sum of the squared differences between r_i and s_i.

Test statistic. If there are no ties, the formula is

$$\rho = 1 - \frac{6T}{n(n^2-1)} \qquad (54)$$

where

$$T = \sum (r_i - s_i)^2\,. \qquad (55)$$

Large sample sizes. For large n the distribution of ρ is approximately normal. The critical values can be found by $z = \rho\sqrt{n-1}$. The rejection rule for a two-tailed test is to reject H_0 if $z > z_{\alpha/2}$ or $z < - z_{\alpha/2}$ where $z_{\alpha/2}$ is the critical value for the given level of significance.

Example. Spearman's rho was calculated using sample 1 and sample 5 (Table 3, Appendix), n = 15. The sum of the squared rank differences for the two samples is T = 839. Rho is $1 - \frac{6(839)}{15(224)} = 1 - \frac{5034}{3360}$ =1–1.498 = –0.498. So z =

$-0.498\sqrt{14} = -1.864$. Because $-1.864 > -1.956$, the null hypothesis cannot be rejected.

Kendall's Rank Correlation Coefficient

Kendall's rank correlation coefficient (tau) is similar to Spearman's rho. The underlying concept is the tendency for concordance, which means that if $x_i > x_j$ then $y_i > y_j$. Concordance implies that the differences $x_i - x_j$ and y_i - y_j have the same sign, either "+" or "–". Discordant pairs have opposite signs, that is, $x_i > x_j$ but $y_i < y_j$, or the opposite, $x_i < x_j$ but $y_i > y_j$.

Procedure. Arrange the pairs in ascending order of *X*. Count the number of y_i smaller than y_1. This is the number of disconcordant pairs (N_D) for x_1. Repeat the process for each x_i, counting the number of $y_j < y_i$, where $j = i + 1, i + 2, i + 3, \ldots, n$.

Test statistic. Because the total number of pairs is $\frac{1}{2}n(n-1)$, $N_c = \frac{1}{2}n(n-1) - N_D$. The tau statistic ($\tau$) is defined as

$$\tau = \frac{N_C - N_D}{\frac{1}{2}n(n-1)}. \qquad (56)$$

This formula can be simplified by substituting $N_c = \frac{1}{2}n(n-1) - N_D$ into the formula so that

$$\tau = 1 - \frac{4N_D}{n(n-1)}. \qquad (57)$$

Large sample sizes. For large sample sizes, the formula is

$$z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \qquad (58)$$

where *z* is compared to the *z* score from the standard normal distribution for the appropriate alpha level.

Example. Kendall's tau was calculated using sample 1 and sample 5 (Table 3, Appendix), *n* = 15. The number of discordant pairs for each pair, (x_1, x_5), were 12, 8, 8, 5, 9, 5, 6, 3, 5, 3, 0, 3, 0, 1, and 0. The total number of discordant pairs, N_D is 68. Tau is $1 - \frac{4 \cdot 68}{15 \cdot 14} = 1 - \frac{272}{210} = -0.295$.

Thus $z = \frac{3(-.295)\sqrt{(15)(14)}}{\sqrt{2(35)}} = \frac{-12.835}{8.366} = -1.534$.

Because $-1.534 > -1.95596$, the null hypothesis cannot be rejected.

References[1]

*Anderson, D.R., Sweeney, D.J., & Williams, T. A. (1999). *Statistics for business and economics* (7th ed.). Cincinnati: South-Western College Publishing Co.

*Berenson, M. L., Levine,D. M., & Rindskopf, D. (1988). *Applied statistics: A first course.* Englewood Cliffs, NJ: Prentice – Hall, Inc.

Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *The American Statistician, 54*, 72-77.

Blair, R. C., & Higgins, J. J. (1985). A comparison of the power of the paired samples rank transformation to that of Wilcoxon's signed rank statistic. *Journal of Educational Statistics*, *10*, 368-383.

Blair, R. C., & Higgins, J. J. (1980). A comparison of the power of Wilcoxon's rank-sum statistic to that of Student's *t* statistic under various nonnormal distributions. *Journal of Educational Statistics, 5,* 309-335.

Blair, R. C., Sawilowsky, S. S., & Higgins, J. J. (1987). Limitations of the rank transformation in factorial ANOVA. *Communications in Statistics*, *16*, 1133-1145.

Bradley,J.V. (1968). *Distribution-free statistical tests*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31*, 144-152.

Bradstreet, T. E. (1997). A Monte Carlo study of Type I error rates for the two-sample Behrens-Fisher problem with and without rank transformation. *Computational Statistics and Data Analysis*, *25*, 167-179.

Bridge, P. K., & Sawilowsky, S. S. (1999). Increasing physician's awareness of the impact of statistical tests on research outcomes: Investigating the comparative power of the Wilcoxon Rank-Sum test and independent samples t test to violations from normality. *Journal of Clinical Epidemiology*, *52*, 229-235.

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

*Daly, F., Hand, D.J., Jones, M.C., Lunn, A..D., & McConway, K.J. (1995). *Elements of statistics.* Workingham, England:Addison-Wesley.

*Daniel, W.W. (1978). *Applied nonparametric statistics.* Boston: Houghton Mifflin Co.

Deshpande, J.V., Gore,A..P.,& Shanubhogue, A.. (1995). *Statistical analysis of nonnormal data.* New York: John Wiley & Sons, Inc.

*Ferguson, G. A. (1971). *Statistical analysis in psychology and education* (3rd ed.). New York: McGraw-Hill book Company.

*Ferguson, G. A. (1981). *Statistical analysis in psychology and education*_(5th ed.). New York: McGraw-Hill Book Company.

*Gravetter, F. J., & Wallnau, L. B. (1985). *Statistics for the behavioral sciences.* St. Paul: West Publishing Co.

Gibbons, J. D. (1971). *Nonparametric statistical inference.* New York: McGraw-Hill Book Company.

Hájek, J. (1969). *A course in nonparametric statistics.* San Francisco: Holden-Day.

Harwell, M., & Serlin, R. C. (1997). An empirical study of five multivariate tests for the single-factor repeated measures model. *Communications in Statistics*, *26*, 605-618.

*Hays, W. L. (1994). Statistics (5th ed.). Fort Worth: Harcourt Brace College Publishers.

Headrick, T. C., & Sawilowsky, S. S. (2000). Type I error and power of the RT ANCOVA. American Educational Research Association, SIG/ Educational Statisticians. New Orleans, LA

Headrick, T. C., & Sawilowsky, S. S. (1999). Type I error and power of the rank transform in factorial ANCOVA. Statistics Symposium on Selected Topics in Nonparametric Statistics. Gainesville, FL.

*Hildebrand, D. (1986). *Statistical thinking for behavioral scientists.* Boston: Duxbury Press.

Hollander, M. & Wolfe, D. (1973). *Nonparametrical statistical methods.* New York: John Wiley & Sons.

*Jarrett, J. & Kraft, A. (1989). *Statistical analysis for decision making.* Boston: Allyn and Bacon.

Jonckheere, A. R. (1954). A distribution-free *k*-sample test against ordered alternatives. *Biometrika*, 41, 133-143.

Kelley, D. L., & Sawilowsky, S. S. (1997). Nonparametric alternatives to the F statistic in analysis of variance. *Journal of Statistical Computation and Simulation*, *58*, 343-359.

*Knoke, D. and Bohrnstedt, G. W. (1991). *Basic social statistics*. New York: F. E. Peacock Publishers, Inc.

*Kraft, C. H. & van Eeden, C. (1968). *A nonparametric introduction to statistics*. New York: Macmillan Co.

*Krauth, J. (1988). *Distribution-free statistics: An application-oriented approach*. Amsterdam: Elsevier.

*Kurtz, N. R. (1983). *Introduction to social statistics*. New York: McGraw-Hill Book Co.

Lahey (1998). *Essential Lahey Fortran 90*. Incline Village, NY: Lahey Computer Systems, Inc.

*Lehmann, E. L. & D'Abrera, H.J.M. (1975). *Nonparametric statistical methods based on ranks*. New York: McGraw-Hill International Book Company.

Ludbrook, J. & Dudley, H. (1998). Why permutation tests are superior to *t* and *F* tests in biomedical research. *The American Statistician*, *52*, 127-132.

*Manoukian, E. B. (1986). *Mathematical nonparametric statistics*. New York: Gordon & Breach Science Publications.

*McClave, J. T., Dietrich II, F. H. (1988). *Statistics* (4th ed.). San Francisco: Dellen Publishing Company.

*Mendenhall, W. & Reinmuth, J. E. (1978). *Statistics for management and economics* (3rd ed.). North Scituate, MA: Duxbury Press.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156-166.

*Montgomery, D.C., & Runger, G. C. (1994). *Applied statistics and probability for engineers*. New York: John Wiley and Sons, Inc.

Musial, J., III. (1999). Comparing exact tests and asymptotic tests with colorectal cancer variables within the National Health and Nutrition Examination Survey III. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Nanna, M. J. (1997). Robustness and comparative power properties of Hotelling's T^2 versus the rank transformation test using real pre-test/post-test likert scale data. Unpublished doctoral dissertation, Wayne State University, Detroit, MI.

Nanna, M. J. (2001, in press). Hotelling's T^2 vs the rank transform with real Likert data. *Journal of Modern Applied Statistical Methods, 1.*

Nanna, M. J., & Sawilowsky, S. S. (1998). Analysis of Likert scale data in disability and medical rehabilitation evaluation. *Psychological Methods, 3,* 55-67.

Neave, H. R., & Worthington, P. L. (1988). *Distribution-free tests.* London: Unwin Hyman Ltd.

*Newmark, J. (1988). *Statistics and probability in modern life*_(4th ed.). New York: Saunders College Publishing.

Posch, M.A., & Sawilowsky, S. (1997). A comparison of exact tests for the analysis of sparse contingency tables. Joint Statistical Meetings, American Statistical Association.

*Rosenberg, K.M.(1990). *Statistics for beavioral scientists*. Dubuque, IA: Wm. C. Brown Pub.

*Runyon, R. P. (1977). *Nonparametric statistics: A contemporary approach.* Reading MA: Addison-Wesley Publishing Co.

Sawilowsky, S. S. (1985). Robust and power analysis for the 2x2x2 ANOVA, rank transformation, random normal scores, and expected normal scores transformation tests. Unpublished doctoral dissertation, University of South Florida, Tampa, FL.

Sawilowsky, S. S. (1989). Rank transformation: the bridge is falling down. American Educational Research Association, SIG/ Educational Statisticians, San Francisco, CA.

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and Type II error properties of the *t* test to departures from population normality. *Psychological Bulletin, 111,* 352-360.

Sawilowsky, S. S., & Brown, M. T. (1991). On using the t test on ranks as an alternative to the Wilcoxon test. *Perceptual and Motor Skills*, *72*, 860-862.

Sawilowsky, S. S., Blair, R. C., & Higgins, J. J. (1989). An investigation of the type I error and power properties of the rank transform procedure in factorial ANOVA. *Journal of Educational Statistics*, *14*, 255-267.

Sawilowsky, S. S., Blair, R. C., & Micceri, T. (1990). REALPOPS.LIB: A PC FORTRAN library of eight real distributions in psychology and education. *Psychometrika*, *55*, 729.

Siegel, S. & Castellan, Jr., N. J. (1988). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, Inc.

*Snedecor, G. W. & Cochran, W. G. (1967). *Statistical methods*. Ames, IA: Iowa State University Press.

Sprent, P. (1989). *Applied nonparametric statistical methods.* London: Chapman and Hall.

*Triola, M. (1995). *Elementary statistics* (6th ed.). Reading MA: Addison – Wesley Publishing Company.

*Wilcox, R. R. (1996). *Statistics for the social sciences.*_San Diego:
Academic Press.

*Zikmund, W. G. (1991). *Business research methods* (3rd ed.). Chicago: The Dryden Press.

[1] Entries with the "*" refer to the textbook survey results compiled in Table 1, but not cited in this article.

Appendix

Table 3. Samples Randomly Selected from Multimodal Lumpy Data Set (Micceri, 1989)

Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
20	11	9	34	10
33	34	14	10	2
4	23	33	38	32
34	37	5	41	4
13	11	8	4	33
6	24	14	26	19
29	5	20	10	11
17	9	18	21	21
39	11	8	13	9
26	33	22	15	31
13	32	11	35	12
9	18	33	43	20
33	27	20	13	33
16	21	7	20	15
36	8	7	13	15

Table 4. Multimodal Lumpy Set (Micceri, 1989).

Score	cum freq	cdf	score	cum freq	cdf
0	5	0.01071	22	269	0.57602
1	13	0.02784	23	279	0.59743
2	21	0.04497	24	282	0.60385
3	24	0.05139	25	287	0.61456
4	32	0.06852	26	297	0.63597
5	38	0.08137	27	306	0.65525
6	41	0.08779	28	309	0.66167
7	50	0.10707	29	319	0.68308
8	62	0.13276	30	325	0.69593
9	80	0.17131	31	336	0.71949
10	91	0.19486	32	351	0.75161
11	114	0.24411	33	364	0.77944
12	136	0.29122	34	379	0.81156
13	160	0.34261	35	389	0.83298
14	180	0.38544	36	401	0.85867
15	195	0.41756	37	418	0.89507
16	213	0.45610	38	428	0.91649
17	225	0.48180	39	434	0.92934
18	234	0.50107	40	445	0.95289
19	244	0.52248	41	454	0.97216
20	254	0.54390	42	460	0.98501
21	261	0.55889	43	467	1.00000