Alternative Methods For Dealing With Nonnormality And Heteroscedasticity In Paleontological Data

By: **Steven J. Hageman**

Abstract

Although numerical methods are highly useful in paleontological studies, potential problems arise with application of parametric statistical methods to paleontological data. Most common statistical tests assume data are normally distributed and that multiple populations have equal variances (homoscedasticity). Paleontological data frequently do not satisfy these assumptions, thereby affecting results of tests and potentially misleading scientific interpretations. Nonparametric tests should be used when assumptions of parametric tests are violated. Normal scores tests, which utilize expected normal deviates (rankits) substituted for original data, are the most powerful nonparametric tests. Despite their potential utility, normal scores tests have received little attention, primarily because of difficulties encountered with rankit conversion. Recent advances in microcomputer technology provide viable methods for rankit conversion, thus making normal scores tests accessible for routine application. Normal scores tests provide a practical method of dealing with nonnormality and heteroscedasticity common in paleontological data.

Hageman SJ. Alternative Methods for Dealing with Nonnormality and Heteroscedasticity in Paleontological Data. Journal of Paleontology. 1992;66(6):857-867. Publisher version of record available at: https://www.jstor.org/stable/1305944

J. Paleont., 66(6), 1992, pp. 857–867

0022-3360/92/0066-0857$03.00

ALTERNATIVE METHODS FOR DEALING WITH NONNORMALITY AND HETEROSCEDASTICITY IN PALEONTOLOGICAL DATA

STEVEN J. HAGEMAN
Department of Geology, University of Illinois, Urbana 61801

ABSTRACT—Although numerical methods are highly useful in paleontological studies, potential problems arise with application of parametric statistical methods to paleontological data. Most common statistical tests assume data are normally distributed and that multiple populations have equal variances (homoscedasticity). Paleontological data frequently do not satisfy these assumptions, thereby affecting results of tests and potentially misleading scientific interpretations. Nonparametric tests should be used when assumptions of parametric tests are violated. Normal scores tests, which utilize expected normal deviates (rankits) substituted for original data, are the most powerful nonparametric tests. Despite their potential utility, normal scores tests have received little attention, primarily because of difficulties encountered with rankit conversion.

Recent advances in microcomputer technology provide viable methods for rankit conversion, thus making normal scores tests accessible for routine application. Normal scores tests provide a practical method of dealing with nonnormality and heteroscedasticity common in paleontological data.

INTRODUCTION

NUMERICAL METHODS provide valuable tools to paleontologists; however, parametric statistical methods bear some potential pitfalls. Paleontological data frequently do not satisfy parametric assumptions of normality and homoscedasticity (equal variance). Violation of these assumptions affects results of statistical tests, which in turn influences scientific conclusions. Sample size also strongly influences the results of statistical tests. Therefore, paleontologists must exercise caution when using these techniques. The purpose of this paper is to draw attention to these well-known but often ignored problems and to review practical but historically obscure methods for dealing with nonnormality and heteroscedasticity. Morphometric data are used in examples in this paper, but principles discussed apply to a wide variety of paleontological (systematic, biostratigraphic, paleoecologic, and biogeographic) and geological data.

PARAMETRIC ASSUMPTIONS

Most common statistical tests are based on very specific assumptions: 1) objects for study are chosen at *random* (e.g., researcher does not consciously or subconsciously pick only the large specimens and ignore small ones); 2) outcome of tests are *independent* (i.e., after a measurement is taken, one cannot predict whether the next observation will be larger or smaller); 3) observations are *normally distributed* (Gaussian distribution); and 4) when two or more groups are involved, they have *equal variances* (homoscedasticity). Note that these assumptions do not apply only to sophisticated numerical methods; they are made any time the mean and standard deviation are calculated for a set of observations and employed in a test. With careful data collection [meticulous data acquisition, enhanced by digital devices (Fink, 1990)], the first two assumptions can be satisfied in most paleontological studies, but compliance with the other assumptions (normality and homoscedasticity) is dictated by the distribution of the data themselves. The latter assumptions are familiar to anyone acquainted with statistical methods (Sokal and Rohlf, 1981), but how valid are they for most paleontological data, and what are the scientific consequences of violating them? These questions are seldom addressed in the paleontological literature. The importance of sample size considerations has also been neglected (see Significance Levels and Scientific Conclusions section, and Foster and Kaesler, 1988).

The normal distribution provides a model for random error. The idea of randomness of errors is attractive, and many statistical tests have been developed based on normal distribution. Many other distributions, however, exist (e.g., beta, Cauchy, gamma, uniform, Poisson, exponential, hypergeometric) that have basis in theoretical models. Statistical tests have been developed that employ the properties of these distributions; for example, the hypergeometric distribution provides a model of sampling without replacement, which is appropriate for analyzing spatial distributions where no two organisms can occupy the same space on a grid. Ideally, one should recognize the theoretical model that describes the situation believed to exist for a given study (often the normal distribution), collect data and compare its distribution against the proper distribution, and, if the data fit the model satisfactorily, continue with the appropriate statistical test. It is important to ask why paleontological data should fit any of these models, let alone the normal distribution.

Morphological characters are often controlled by biological and physical constraints rather than random processes. Thresholds are encountered in morphological characters. It is common to find minimum sizes for characters in species, below which structures presumably are unable to perform their function, but above which the size is less constraining, resulting in a skewed but somewhat irregular distribution. Other threshold effects can also be observed. In fact, plausible biological explanations exist for many distributions observed in nature that do not fit simple mathematical models (e.g., different ontogenetic histories due to variable environmental pressures). Data resulting from these factors are often called "messy," and researchers tend to apologize for their nonnormality. However, accurately collected data are what they are. Paleontologists often deal with systems that are so complex that mathematical models have not been developed to account for them.

In the past, people have dealt with nonnormality and heteroscedasticity in a variety of ways. Some workers ascribe to the philosophy that, unless there is some a priori reason to believe otherwise, the assumption of normality is justified. Others simply ignore the problem and hope for the best; still others fervently warn that the risk that assumption violations will invalidate results is very real, and that measures should be taken regularly to guard against such mistakes.

A common method of dealing with nonnormality is to transform the data (Sokal and Rohlf, 1981; Zar, 1984). The influence of size factors on morphological data (e.g., a population that contains a range of ontogenetic growth stages) frequently results in log-normal distributions (when variance is proportional to the mean). It is therefore common practice to make logarithmic

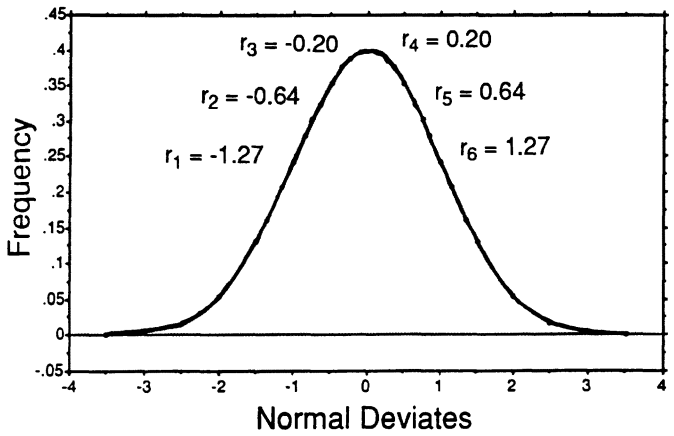

FIGURE 1—Standard normal distribution, with rankit values for $N = 6$ observations.

transformations prior to analyzing data. Other transformations such as square root, inverse, and arcsine are also commonly employed to adjust for nonnormality. Effects of transformations are not always clear, however. For example, what are the biological implications if a character is not significantly different between two species when original data are used, but is significantly different when the arcsine transformation is employed? Although there is nothing sacred about an ordinal scale, problems arise when different transformations are required for multiple populations under comparison.

When parametric assumptions are not satisfied or when there are too few observations to satisfactorily evaluate the distribution, statisticians advocate use of nonparametric or distribution-free tests (e.g., Mann–Whitney and Kruskal–Wallace tests). Nonparametric tests do not use population parameters such as mean and standard deviation, and distribution-free tests do not make assumptions about the specific type of distribution of the data [see Bradley (1968), Marascuilo and McSweeney (1977), Conover (1980), and Zar (1984) for reviews of traditional nonparametric methods]. These tests usually deal with the rank-order of data, which allows for manipulation of data collected from imperfectly (unacceptably) calibrated measurement scales or when only count or rank data are available. Nonparametric and distribution-free tests share some disadvantages however; they are not as powerful as their parametric counterparts when distributions are approximately normal and homoscedastic (Bradley, 1968; Conover, 1980). Although the mathematics involved in nonparametric and distribution-free tests is relatively simple, bookkeeping involved in calculations is often tedious and time consuming. In addition, the wide variety of existing nonparametric tests has not been readily available in computer statistical packages.

A relatively new method for dealing with nonnormality and heteroscedasticity involves calculating standard errors and confidence limits for parameters directly from observed distributions. This procedure, known as bootstrapping (Effron and Tibshirani, 1986), is based on iterative sampling of parameters from the observed distribution. The utility of the bootstrap method has been demonstrated for paleontological studies in analyses of time-ordered evolutionary data (e.g., Gilinsky and Bambach, 1986) and morphometric analyses (Plotnick, 1989). The procedure, however, is relatively new and applications are currently being developed. As procedures become standardized, bootstrap methods will no doubt play a greater role in paleontological studies.

TABLE 1—Steps in the procedure for converting original data to rankits. Rankit values in column three are from Harter (1961, table 1).

1. Original data	2. Sorted	3. Rankits $N = 10$	4. Corrected for ties	5. Original order
6.0	2.0	−1.53875	−1.53875	0.12267
7.5	3.2	−1.00136	−1.00136	0.65606
3.2	4.7	−0.65606	−0.65606	−1.00136
4.7	5.1	−0.37576	−0.24921	−0.65606
8.1	5.1	−0.12267	−0.24921	1.00136
2.0	6.0	0.12267	0.12267	−1.53875
5.1	6.6	0.37576	0.37576	−0.24921
6.6	7.5	0.65606	0.65606	0.37576
5.1	8.1	1.00136	1.00136	−0.24921
8.7	8.7	1.53875	1.53875	1.53875

Two other, less well-known, methods for dealing with nonnormality and heteroscedasticity are: 1) empirically testing effects of violations of the assumptions of parametric tests; and 2) using normal scores tests. The rest of this paper addresses these two methods. Both use a statistic known as a rankit, of which it is helpful to have an understanding.

RANKITS

A rankit is the expected value of the Rth smallest observation in a sample of size N drawn from a standard normal distribution (Ipsen and Jerne, 1944). This can be better understood with a graphical explanation. An empirical approximation of a rankit can be obtained as follows: first, a number of observations (say six) are taken randomly from a standard normal population (a population with a mean of zero and standard deviation of one) and the observations are ranked one to six, from smallest to largest. If this procedure is repeated many times, the average of all the smallest observations provides an approximation of the rankit value corresponding to rank-order of one. From Rohlf and Sokal (1981, table 27), the rankit value corresponding to a rank-order of one from a population of six is −1.27 (Figure 1). The average of all the second smallest observations approximates the rankit corresponding to the rank-order two (−0.64), and the average of the third smallest to the rank-order three (−0.20). Note that rankit values are symmetrical about the median, with only a change in sign (Figure 1).

Rankits can be considered a normalization of ranks. In other words, the rankit for $R = 1$ (rank-order one) and $N = 6$ (number of observations) is the expected value for the smallest of six random samples from a standard normal distribution. Rankits are also known as normal-order equivalents, order-statistics, expected normal scores, and normalized ranks.

Rankits can be substituted for original data. Rankit substitution, in effect, converts a population into its equivalent standard normal distribution. The procedure for rankit substitution is illustrated in Table 1. Data are sorted from smallest to largest. Rankits for the appropriate number of observations are then arranged in order next to the sorted data. Corrections are made for tied values by averaging rankits over the spanned range. For example, in Table 1 values of 5.1 are present in the fourth and fifth rank-order position, thus the average of the fourth and fifth rankits (−0.2492) is placed in both of these positions. Rankits are substituted for original data and reordered based on the sequence of the original data. Column 5 in Table 1 represents the equivalent standard normal distribution of the original data in column 1. Rohlf and Sokal (1981, table 27), provided a table of rankit values to three digits for $N = 1$ to $N = 20$ and two digits for $N = 21$ to $N = 50$; and Harter (1961, table 1) provided a table of rankit values to five digits for $N = 1$ to $N = 100$, and includes some entries up to $N = 400$.

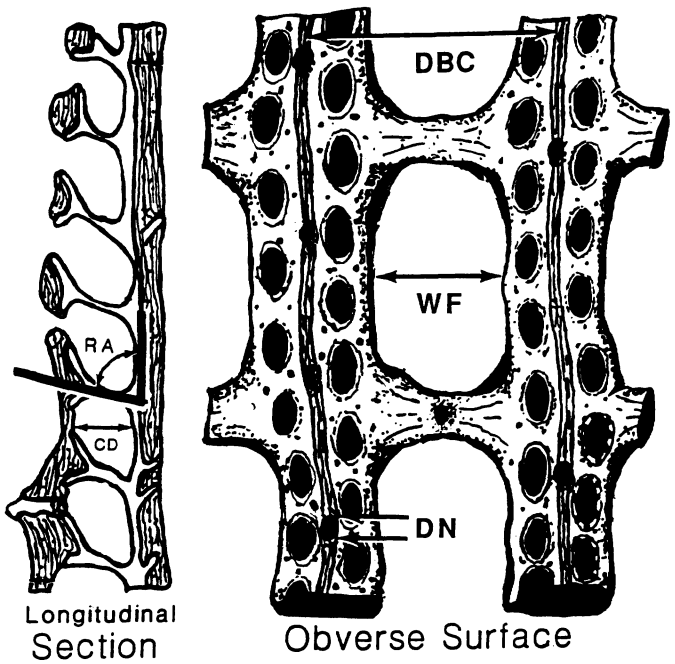

FIGURE 2—Longitudinal section and obverse surface of typical fenestrate bryozoan, illustrating location of morphometric characters used in examples (DBC = distance between branch centers, WF = width of fenestrule, DN = diameter of node, CD = autozooecial chamber depth, and RA = reverse wall budding angle).

It is important to note that rankit substitution is not analogous to a logarithmic or arcsine transformation designed to "improve" the normality of the data. It is a one-way substitution; original values cannot be obtained from rankit values.

TESTING EFFECTS OF VIOLATIONS OF ASSUMPTIONS

Rankits can be used empirically to test effects of violating assumptions of parametric tests: normality and homoscedasticity (Ghent, 1971). This is done by first performing a parametric statistical test using original data, then performing the same test using data that have been converted to rankits. If data are originally perfectly normally distributed, results of the two tests will be nearly identical. Any difference between the results is due to the effects of nonnormality or heteroscedasticity.

The following is an example of this procedure. Unpaired Student's t-tests are performed to determine if significant differences exist in five morphometric characters between two species of fenestrate bryozoans (characters illustrated in Figure 2, and defined in Table 2). It is important to note that it is known a priori that some difference exists between the two populations (see Significance Levels and Scientific Conclusions section). For character DBC (distance between branch centers) the probability that the two species have the same branch spacing is 0.0427 (4.27%), which is deemed significant using a strict $\alpha = 0.05$ significance level. A second series of t-tests are performed using rankits substituted for the original data [rankit values for $N = 24$ are obtained from Harter (1961, table 1) and corrected for ties]. With rankits, the probability that the two species have the same branch spacing is 0.0537 (5.37%). In this case, the effect of nonnormality or heteroscedasticity is to increase the significance (decrease probability of null hypothesis being true).

It is important to note that effects of nonnormality or heteroscedasticity are not predictable; sometimes significance is increased, other times it is decreased, and probabilities are altered by different magnitudes (Table 2). In one case (highlighted with an asterisk), effects of violations of assumptions of parametric tests influence scientific conclusions when a strict $\alpha = 0.05$ level of significance is used.

TABLE 2—Two-sample comparisons of means of five morphometric characters of fenestrate bryozoans (DBC = distance between branch centers, WF = width of fenestrule, DN = diameter of node, CD = autozooecial chamber depth, and RA = reverse wall budding angle). Asterisk indicates changes in scientific conclusions based on a strict 0.05 significance level.

	Rectifenestella tenuissima vs. *Rectifenestella tenax* $N = 24$	$N = 24$	
Character	P of Student's t	P using Rankits	Nonnorm and heterosced. effect on significance
DBC	0.0427	0.0537	Increased*
WF	0.0028	0.0040	Increased
DN	0.0002	0.0001	Decreased
CD	0.0507	0.1196	Increased
RA	0.0676	0.0652	Decreased

This procedure of directly testing effects of violations of assumptions of parametric tests, illuminates damage done by ignoring nonnormality and heteroscedasticity (Ghent, 1971). If probabilities obtained with original data are already highly significant and the effect of violations is to decrease significance (i.e., raises the probability of the effect occurring at random), then one can be confident of the original conclusions (characters DN and RA in Table 2). If, however, the effect is to increase significance of data that are only marginally significant in the first place (i.e., decreases the probability of the effect occurring at random), then caution should be applied to interpretations (character DBC in Table 2).

Why should one not simply omit the step using original data and analyze only data converted to rankits? Indeed, Fisher and Yates (1938) proposed this procedure as a method of dealing with nonnormally distributed data prior to the development of many nonparametric methods. The problem is that although test statistics obtained using rankits are close enough to provide information about general effects of violations of assumptions of parametric tests, the test statistics are not distributed exactly the same as those derived from original data (Bradley, 1968, p. 148). Statistical tests have, however, been developed that are based on the distribution of rankits, and these are called normal scores tests.

INTRODUCTION TO NORMAL SCORES TESTS

Normal scores tests are a family of distribution-free tests in which normal deviates (rankits) are substituted for original data. Many normal scores tests have been developed. Normal scores counterparts exist for all common parametric tests (Table 3).

The relative power of statistical tests (ability to reject the null hypothesis) can be compared by several methods. The asymptotic relative efficiency (A.R.E.) is one of the most common benchmarks for comparing the power of a nonparametric statistical test with its parametric counterpart. Normal scores tests are the most desirable nonparametric tests because they have asymptotic relative efficiencies of one or greater (Bradley, 1968; Marascuilo and McSweeney, 1977; Conover, 1980; Rock, 1988). This means that when data are normally distributed, normal scores tests will provide the same results as equivalent parametric tests, but when parametric assumptions of normality and homoscedasticity are violated, the power of normal scores tests increases greatly. When data are normally distributed, other parametric tests have A.R.E. values of less than one (Bradley,

TABLE 3—Common parametric tests and their normal scores and traditional nonparametric test equivalents.

Parametric tests	Normal scores tests	Traditional nonparametric tests
One-sample *t*-test	Van Eden test	Wilcoxon signed rank test
Paired *t*-test	Van Eden test	Wilcoxon signed rank test
Unpaired *t*-test	Terry-Hoeffding test	Mann–Whitney test
F-test	Klotz test	Siegel–Tukey test
One-way ANOVA	Van der Waerden test	Kruskal–Wallace test
Post-hoc comparison of means	Conover, 1980	Multiple range test
Correlation coefficients	Bradley, 1968	Spearman's rho/Kendal's tau

1968); the A.R.E. values of other nonparametric tests also increase as assumptions are violated. When data are ranked, information about the original distribution is lost; this decreases the power of nonparametric tests relative to their parametric counterparts. Substitution of rankits for ranks restores the power of nonparametric tests to their original level prior to ranking (Bradley, 1968).

Once data are converted to rankits, normal scores tests are simple to calculate. Because rankits are derived from the standard normal distribution (mean = 0, standard deviation = 1), many elements of parametric test equations are dropped or simplified from normal scores tests (e.g., sum of all rankits from any number of observations is zero). Normal scores tests are easily performed with spreadsheets such as Excel and Lotus 1–2–3 on personal computers.

Given their attractive qualities, one may question why normal scores tests have remained relatively obscure. Two reasons are: 1) the function used to generate rankits is very complex, making it difficult to calculate values for parameters not previously published in tables (see Appendix 1); and 2) although relatively simple to perform, conversion of raw data to rankits with corrections for ties (Table 3) is cumbersome and time consuming, which has made it virtually prohibitive for large data sets.

Recent advances in microcomputer technology and software provide viable methods for rankit calculation and substitution. Using a numerical approximation [see Equation (1) in Appendix 1] derived by Harter (1961), rankit values can be efficiently calculated for any R (rank-order) and N (number of observations) with *Mathematica*®, a program available for microcomputers (see Appendix 1). The program *Rankit* was written by the author to convert raw data matrices to rankit equivalents. *Rankit* reads raw data from a text file, obtains rankit values from either a look-up file (for $N < 81$) or a file of rankits created with *Mathematica*® (for $N > 80$), substitutes rankits corrected for ties, and outputs a text file of rankits that can be imported into spreadsheets for analysis.

The combination of *Mathematica*® and *Rankit* provides a viable method for rankit conversion (see Appendix 1). Large data matrices can be converted to rankits in a matter of minutes. This makes normal scores tests accessible for routine use. Data can also be converted to rankits using a function call in SPSS and Van der Waerden scores, which are a function of normal scores, and can be obtained from both SAS and SPSS. Neither of these packages, however, provides normal scores tests.

TABLE 4—Explanation of variables and test statistics.

Variables

- E_{NR_i} = rankit value for the ith smallest observation
- N = total number of observations
- R = rank order of the rth smallest observation
- n_j = number of observations in jth group
- d = difference between paired observations
- v = number of non-zero differences between paired observations
- k = number of groups
- D = difference in rankit means between ith and jth groups
- α = significance level (probability that null hypothesis is true)

Test statistics

- Z: normal distribution (Zelen and Severo, 1965, table 26.1)
- T: Student's t-distribution (Zar, 1984, table B.3)
- r: Pearon's product-moment correlation coefficient (Zar, 1984, table B.16)
- χ: χ^2 distribution (Zar, 1984, table B.1)
- S: S distribution (Klotz, 1964, table I)
- D_c: Critical difference between group means (function of t-distribution)
- B: normal scores correlation coefficient

REVIEW OF NORMAL SCORES TESTS

Normal scores tests are discussed by a number of authors (Bradley, 1968; Marascuilo and McSweeney, 1977; Conover, 1980; Rock, 1988), but few examples of practical applications are available. The following review provides examples and discussion for the most common normal scores test. The purpose of these examples is to provide models that readers can modify to suit their own needs. Data used here are from a morphometric data set compiled by Snyder (1991) and have been modified in some cases to exemplify better the nature of specific tests. Therefore, results are reported for instruction only and have no scientific value. Variables and test statistics are defined in Table 4. Step by step procedural descriptions for performing each of the normal scores tests are provided in Appendix 2 for those who are not entirely comfortable with deciphering statistical equations. One should not, however, follow this cook book approach without an understanding of the test being performed.

Before examples are presented, it is important to note that null hypotheses (e.g., no difference between the means of two populations) can never be accepted, because no two populations are identical (Foster and Kaesler, 1988). One can only reject or fail to reject a null hypothesis with a given level of confidence (probability of being in error). This principle is discussed more later.

Van Eden test.—The Van Eden normal scores test evaluates whether a significant difference exists between a series of paired observations. This test can also be applied to determine whether the mean of a single group is different than a given value (Van Eden, 1963). Its parametric analog is the paired Student's t-test, and nonparametric analog is the Wilcoxon signed rank test. The test statistic Z is approximately normally distributed:

$$Z = \frac{\sum_{R=1}^{v} E_{NR_i}}{\sqrt{\sum_{R=1}^{v} (E_{NR_i})^2}} \quad (1)$$

As an example, a Van Eden test is applied to determine whether a significant difference exists between the lengths of apertures paired across branches of the fenestrate bryozoan *Rectifenestella*

TABLE 5—Van Eden's normal scores test for differences between aperture lengths (in mm) of *Rectifenestella tenax* paired across branch (d = differences between left and right paired apertures).

Left	Right	d	Rank $\|d\|$	Rankit	Signed rankit	(Rankit)^2
0.0613	0.0567	0.0046	3.0	−0.729	0.729	0.531
0.0667	0.0567	0.0100	7.5	0.343	0.343	0.118
0.0667	0.0667	0.0000	•	•	•	•
0.0817	0.0650	0.0167	11.0	1.586	1.586	2.515
0.0733	0.0633	0.0100	7.5	0.343	0.343	0.118
0.0614	0.0583	0.0031	2.0	−1.062	1.062	1.128
0.0583	0.0667	−0.0084	6.0	0.000	−0.000	0.000
0.0633	0.0767	−0.0134	10.0	1.062	−1.062	1.128
0.0583	0.0633	−0.0050	4.5	−0.343	−0.343	0.118
0.0583	0.0600	−0.0017	1.0	−1.586	−1.586	2.515
0.0633	0.0500	0.0133	9.0	0.729	0.729	0.531
0.0717	0.0767	−0.0050	4.5	−0.343	−0.343	0.118
				Σ	1.458	8.820

tenax. Twelve pairs of observations are taken (Table 5) and differences between pairs calculated. Absolute values of differences are ranked (omitting observations with no differences) and rankits and their squares are obtained. Signs of rankits are changed to correspond to the signs of the original data. That is, in Table 5 rankits corresponding to all negative differences (highlighted) are given negative signs and those corresponding to positive differences (not highlighted) are given positive signs, regardless of the original rankit sign obtained from the absolute values of differences.

Using Equation (1), the test statistic Z is calculated:

$$Z = \frac{1.458}{\sqrt{8.820}} = 0.491$$

From a table of areas under the normal curve (cumulative probabilities), the normal deviate 0.491 corresponds to a two-sided probability of $P = 0.624$. Therefore, the null hypothesis (H_0: no difference between left and right paired apertures on *R. tenax*) is not rejected at the 0.05 significance level.

A Van Eden test can also be used to determine if apertural lengths of *R. tenax* differ significantly from a given value (e.g., an a priori critical size predicted from a hypothetical trophic structure). In this case, the critical value is paired with all observations (i.e., replacing "right" column in Table 5), and the test is performed in the same manner as the paired group method.

Terry–Hoeffding test.—The Terry–Hoeffding normal scores test evaluates whether two groups have the same mean (Terry, 1952). Its parametric analog is the unpaired Student's t-test, and nonparametric analog is the Mann–Whitney test. The test statistic S is simply the sum of rankits for the population with fewer observations.

$$S = \sum_{i=1}^{n} E_{NR_i} \tag{2}$$

Critical S values for $N \leq 20$ are provided in Klotz (1964, table I, p. 655) and reproduced in Bradley (1968, table VI, p. 327). Three methods have been proposed for calculating critical S values when $N > 20$. All assume that n_1/n_2 and n_2/N are not too small (but minimal sizes have not been proposed).

Method 1: t-distribution with $(N - 2)$ degrees of freedom.

$$T = \sqrt{\frac{(N-2)S^2}{\left(\frac{n_1 n_2}{N}\sum_{R=1}^{N}(E_{NR_i})^2\right) - S^2}} \tag{3}$$

Method 2: Critical S-value for a given α.

$$S_c = r_{N-2}\sqrt{\frac{n_1 n_2 \sum_{R=1}^{N}(E_{NR_i})^2}{N}} \tag{4}$$

where r is the two-tailed critical value of Pearson's product-moment correlation coefficient.

Method 3: Z is normally distributed for large N.

$$Z = \frac{S}{\sqrt{\frac{n_1 n_2}{N(N-1)}\sum_{R=1}^{N}(E_{NR_i})^2}} \tag{5}$$

As an example, a Terry–Hoeffding normal scores test is applied to determine whether a significant difference exists in branch spacing (distance between branch centers) between two fenestrate species, *Rectifenestella tenax* and *Rectifenestella tenuissima*. Twelve observations are taken for both species ($n_1 = 12$, $n_2 = 12$, $N = 24$), which are ranked as a common population (Table 6). Ranks are converted to rankits, and rankits and their

TABLE 6—Terry–Hoeffding normal scores test for differences in branch spacing (distance between branch centers, in mm) between two fenestrate species.

R. tenax				*R. tenuissima*			
Orig.	Rank	Rankit	(Rankit)^2	Orig.	Rank	Rankit	(Rankit)^2
0.447	7.0	−0.604	0.365	0.562	22.0	1.239	1.535
0.433	5.0	−0.877	0.769	0.538	18.5	0.669	0.448
0.347	1.0	−1.948	3.795	0.475	11.0	−0.156	0.024
0.548	21.0	1.041	1.084	0.542	20.0	0.877	0.769
0.425	3.5	−1.140	1.300	0.577	23.0	1.503	2.259
0.518	15.5	0.316	0.100	0.439	6.0	−0.734	0.539
0.518	15.5	0.316	0.100	0.538	18.5	0.669	0.448
0.533	17.0	0.484	0.234	0.47	9.0	−0.370	0.137
0.478	12.0	−0.052	0.003	0.425	3.5	−1.140	1.300
0.508	14.0	0.156	0.024	0.413	2.0	−1.503	2.259
0.473	10.0	−0.262	0.069	0.588	24.0	1.948	3.795
0.450	8.0	−0.484	0.234	0.497	13.0	0.052	0.003
	Σ	−3.054	8.076		Σ	3.054	13.514

TABLE 7—Klotz normal scores test for equal variances of branch spacing (in mm) between two fenestrate species.

Orig.	Orig.-$\bar{x}$	Rank	Rankit	(Rankit)^2	(Rankit)^4
		Rectifenestella tenax			
0.447	−0.026	9.0	−0.370	0.137	0.019
0.433	−0.040	6.0	−0.734	0.539	0.290
0.347	−0.126	1.0	−1.948	3.795	14.400
0.548	0.075	23.0	1.503	2.259	5.103
0.425	−0.048	5.0	−0.877	0.769	0.592
0.518	0.045	18.5	0.669	0.448	0.200
0.518	0.045	18.5	0.669	0.448	0.200
0.533	0.060	21.0	1.041	1.084	1.174
0.478	0.005	13.0	0.052	0.003	0.000
0.508	0.035	16.0	0.370	0.137	0.019
0.473	0.000	12.0	−0.052	0.003	0.000
0.450	−0.023	10.0	−0.262	0.069	0.005
$\bar{x}$ = 0.473			Σ	9.691	22.002
		Rectifenestella tenuissima			
0.562	0.057	20.0	0.877	0.769	0.592
0.538	0.033	14.5	0.209	0.044	0.002
0.475	−0.030	8.0	−0.484	0.234	0.055
0.542	0.037	17.0	0.484	0.234	0.055
0.577	0.072	22.0	1.239	1.535	2.357
0.439	−0.066	4.0	−1.041	1.084	1.174
0.538	0.033	14.5	0.209	0.044	0.002
0.470	−0.035	7.0	−0.604	0.365	0.133
0.425	−0.080	3.0	−1.239	1.535	2.357
0.413	−0.092	2.0	−1.503	2.259	5.103
0.588	0.083	24.0	1.948	3.795	14.400
0.497	−0.008	11.0	−0.156	0.024	0.001
$\bar{x}$ = 0.505			Σ	11.922	26.231

squares are summed within species. The test statistic S is obtained using Equation (2), which is simply a summation of rankits for the population with fewer observations (arbitrary in this example because $n_1 = n_2$).

$$S = 3.054$$

Because $N > 20$, the critical value of S is not available in the table published by Klotz (1964) and must be obtained by one of three methods.

Method 1, Equation (3):

$$T = \sqrt{\frac{(24-2)3.054^2}{\left[\frac{(12)(12)}{24}21.590\right] - 3.054^2}} = 1.307$$

From a table of critical t-values, a value of $t = 2.074$ is obtained for a two-sided probability of 0.05. Because 1.307 is less than 2.074, the null hypothesis (H_0: no difference in branch spacing between *R. tenax* and *R. tenuissima*) is not rejected at the 0.05 significance level. In fact, 1.321 (critical t-value for two-sided $\alpha = 0.20$) is greater than 1.307, which means H_0 is not rejected even at the 0.20 significance level.

Method 2, Equation (4):

$$S_c = 0.404\sqrt{\frac{(12)(12)(21.590)}{24}} = 4.598$$

Because 3.054 (the Terry–Hoeffding S-statistic for this example) is less than 4.598 (critical S-value for $\alpha = 0.05$) the null hypothesis is not rejected at the 0.05 significance level. In fact, the Terry–Hoeffding S is greater than 3.085, which is the critical value for a two-tailed 0.20 level of significance.

Method 3, Equation (5):

$$Z = \frac{3.054}{\sqrt{\frac{(12)(12)}{24(24-1)}21.590}} = 1.287$$

From a table of areas under the normal curve (cumulative probabilities) the normal deviate 1.287 corresponds to a two-sided upper tail of $P = 0.198$. Therefore, the null hypothesis is not rejected at the 0.05 significance level. Note, however, that by method three the estimated probability is less than 0.20, whereas by methods one and two the probability was greater than 0.20. This is because the normal approximation of method three improves as sample size increases. Methods one and two provide better estimations of critical values with intermediate sample sizes, as in this example. No minimal sample sizes have been proposed for method three, but the writer's experience indicates that method three provides comparable results when over 50 observations are used. The weaker normal approximation with intermediate to small sample sizes raises questions as to whether a better test statistic may also exist for the Van Eden test and Klotz test under similar circumstances.

Klotz test.—The Klotz normal scores test evaluates equality of variances between two groups (Klotz, 1962). Its parametric analog is the F-test and nonparametric analog is the Siegel–Tukey test. The Klotz test differs from other normal scores tests in that its asymptotic relative efficiency can be less than one. The A.R.E. of the Klotz test is one for a normal distribution and increases for distributions with small tails, but decreases for distributions with very large tails, such as uniform and Cauchy distributions (Klotz, 1962). In the latter case, the Siegel–Tukey test is more appropriate (Bradley, 1968). The test statistic for the Klotz test is normally distributed for large N.

$$Z \cong \frac{\sum_{i=1}^{n_i}(E_{NR_i})^2 - \frac{n_i\sum_{i=1}^{N}(E_{NR_i})^2}{N}}{\sqrt{\frac{n_1n_2}{N(N-1)}\left\{\sum_{i=1}^{N}(E_{NR_i})^4 - \frac{\left[\sum_{i=1}^{N}(E_{NR_i})^2\right]^2}{N}\right\}}} \quad (6)$$

where n_i is the number of observations in the smaller group (arbitrary in the following example because $n_1 = n_2$).

As an example, a Klotz test is applied to determine if the branch spacing of two species of fenestrate bryozoans have the same variance. Twelve observations are collected for both species and means within groups are subtracted from all observations to produce equal group means of zero (Table 7). Deviations from group means are ranked as a common population and converted to rankits. Sums for the squares of rankits and rankits raised to the fourth power are obtained for both groups. These values are used in Equation (6):

$$Z \cong \frac{11.922 - \frac{(12)(9.691+11.922)}{24}}{\sqrt{\frac{(12)(12)}{24(24-1)}\left\{(22.002+26.231) - \frac{(9.691+11.922)^2}{24}\right\}}}$$

$$= 0.407$$

TABLE 8—Van der Waerden normal scores test for differences in branch spacing (distance between branch centers, in mm) between three fenestrate species.

R. tenax $j = 1$				*R. tenuissima* $j = 2$				*R. multispinosa* $j = 3$			
Orig.	Rank	Rankit	(Rankit)²	Orig.	Rank	Rankit	(Rankit)²	Orig.	Rank	Rankit	(Rankit)²
0.447	15.0	−0.245	0.060	0.562	34.0	1.462	2.137	0.520	28.0	0.714	0.510
0.433	12.0	−0.466	0.217	0.538	30.5	0.961	0.924	0.454	18.0	−0.035	0.001
0.347	3.0	−1.462	2.137	0.475	21.0	0.174	0.030	0.354	5.0	−1.140	1.300
0.548	33.0	1.285	1.651	0.542	32.0	1.140	1.300	0.449	16.0	−0.174	0.030
0.425	10.5	−0.586	0.343	0.577	35.0	1.704	2.904	0.414	8.0	−0.806	0.650
0.518	26.5	0.586	0.343	0.439	14.0	−0.317	0.100	0.302	1.0	−2.118	4.486
0.518	26.5	0.586	0.343	0.538	30.5	0.961	0.924	0.360	6.0	−1.016	1.032
0.533	29.0	0.806	0.650	0.470	19.0	0.035	0.001	0.434	13.0	−0.390	0.152
0.478	23.0	0.317	0.100	0.425	10.5	−0.586	0.343	0.350	4.0	−1.285	1.651
0.508	25.0	0.466	0.217	0.413	7.0	−0.906	0.821	0.476	22.0	0.245	0.060
0.473	20.0	0.104	0.011	0.588	36.0	2.118	4.486	0.416	9.0	−0.714	0.510
0.450	17.0	−0.104	0.011	0.497	24.0	0.390	0.152	0.344	2.0	−1.704	2.904
	Σ	1.287	6.083		Σ	7.136	14.122		Σ	−8.423	13.286

From a cumulative normal probability table, the normal deviate 0.407 corresponds to a two-sided probability of $P = 0.684$. Therefore, the null hypothesis (H_0: no difference in variances between the two groups) is not rejected at the $\alpha = 0.05$ significance level.

Van der Waerden test.—The Van der Waerden normal scores test evaluates differences between multiple populations based on a single variable (Van der Waerden, 1952). Its parametric analog is the one-way ANOVA, and nonparametric analog is the Kruskal–Wallace test. The original Van der Waerden test was based on inverse normal scores, but McSweeney and Penfield (1969) derived an equivalent test for rankits. The test statistic is distributed as χ^2 with $(k - 1)$ degrees of freedom:

$$\chi^2_{(k-1)} \cong \frac{(N-1)}{\sum_{i=1}^{N} (E_{NR_i})^2} \sum_{j=1}^{k} \frac{\left[\sum_{i=1}^{n_j} E_{NR_i}\right]^2}{n_j} \tag{7}$$

As an example, a Van der Waerden normal scores test is applied to determine if a significant difference exists between the mean branch spacing (distance between branch centers) of three species of fenestrate bryozoans, *R. tenax, R. tenuissima,* and *R. multispinosa.* Twelve observations are taken for each species ($n_1 = 12$, $n_2 = 12$, $n_3 = 12$, $N = 36$), which are ranked as a common population (Table 8). Ranks are converted to rankits, and rankits and their squares are summed within species. The test statistic is obtained using Equation (7):

$$\chi^2_{(k-1)} = \frac{(36-1)\left[\frac{(1.287)^2}{12} + \frac{(7.136)^2}{12} + \frac{(-8.423)^2}{12}\right]}{(6.083 + 14.122 + 13.286)} = 10.758$$

The test statistic (10.758) is much greater than the critical χ^2 value for a probability of 0.05 (5.991). Therefore, the null hypothesis (H_0: no difference in mean branch spacing between the three groups) is rejected at the $\alpha = 0.05$ level of significance.

Post-hoc multiple comparison of means.—The normal scores post-hoc multiple comparison of means is used to test for significant differences between paired means in a multi-group situation (Conover, 1980). Its parametric analog is the Student–Newman–Keuls test, and nonparametric analog is the multiple range test. If a Van der Waerden test proves significant, it is desirable to test means of groups pair-wise for significance. Critical values for mean differences between groups i and j can be obtained with this equation:

$$|\bar{E}_i - \bar{E}_j| > t_\alpha \sqrt{\left[\frac{\sum_{i=1}^{N} (E_{NR_i})^2}{(N-1)}\right]\left[\frac{N-1-\chi^2}{N-k}\right]} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{8}$$

where E_i is the average rankit value of group i and t_α is the two-tailed critical t-value for a given significance level (α), with $(N - k)$ degrees of freedom, and χ^2 is the Van der Waerden test statistic obtained from Equation (7).

As an example, a post-hoc multiple comparison of means is applied to data from the previous example to determine if significant differences exist in branch spacing between pairs of three fenestrate species. First, a critical difference (D_c) is obtained using data from Table 8 and Equation (8):

$$D_c = 2.035 \sqrt{\left[\frac{6.083 + 14.122 + 13.286}{(36-1)}\right]\left[\frac{36-1-10.75}{36-3}\right]} \times \sqrt{\frac{1}{12} + \frac{1}{12}}$$
$$= 0.697$$

Pair-wise differences between means are then compared against the critical value of 0.697 (Table 9). In this example, the same

TABLE 9—Normal scores post-hoc multiple comparison of means for difference in branch spacing between three fenestrate species (data in Table 8). Asterisk denotes significance with a two-tailed probability of 0.05.

R. tenax vs. *R. tenuissima*	$\left\|\frac{1.287}{12} - \frac{7.136}{12}\right\| = 0.487 < 0.697$
R. tenax vs. *R. multispinosa*	$\left\|\frac{1.287}{12} - \frac{(-8.423)}{12}\right\| = 0.809 > 0.697^*$
R. tenuissima vs. *R. multispinosa*	$\left\|\frac{7.136}{12} - \frac{(-8.423)}{12}\right\| = 1.297 > 0.697^*$

TABLE 10—Normal scores test for correlation between lengths (in mm) of left and right paired apertures of the fenestrate bryozoan *Rectifenestella tenax*.

Left orig. X	Rankit X_i	(Rankit X_i)^2	Right orig. Y	Rankit Y_i	(Rankit Y_i)^2	Rankit $(X_i)(Y_i)$
0.0613	−0.537	0.288	0.0567	−0.954	0.910	0.512
0.0667	0.425	0.181	0.0567	−0.954	0.910	−0.405
0.0817	1.629	2.654	0.0650	0.312	0.097	0.508
0.0733	1.116	1.245	0.0633	0.000	0.000	0.000
0.0614	−0.312	0.097	0.0583	−0.537	0.288	0.168
0.0583	−1.179	1.390	0.0667	0.665	0.442	−0.784
0.0633	0.000	0.000	0.0767	1.373	1.885	0.000
0.0583	−1.179	1.390	0.0633	0.000	0.000	0.000
0.0583	−1.179	1.390	0.0600	−0.312	0.097	0.368
0.0633	0.000	0.000	0.0500	−1.629	2.654	0.000
0.0717	0.793	0.629	0.0767	1.373	1.885	1.089
0.0667	0.425	0.181	0.0667	0.665	0.442	0.283
	Σ	9.445		Σ	9.610	Σ 1.739

critical value is used for all comparisons because all groups have equal sample size.

The null hypothesis (H_0: no difference in branch spacing) is not rejected at the 0.05 significance level for *R. tenax* versus *R. tenuissima,* but is rejected for *R. tenuissima* versus *R. multispinosa,* and *R. tenax* versus *R. multispinosa.* This suggests that there is a difference in branch spacing between *R. multispinosa* and both other species, but that *R. tenax* cannot be distinguished from *R. tenuissima* based on branch spacing.

Normal scores test for correlation.—Bradley (1968) proposed a test for significant correlation between paired observations in a two-group situation. Its parametric analog is Pearson's product-moment correlation coefficient, and nonparametric analogs are Spearman's rho, and Kendal's tau. The normal scores correlation coefficient B is:

$$B = \sum_{i=1}^{N} (E_{NR_{Xi}})(E_{NR_{Yi}}) \tag{9}$$

The test statistic for the hypothesis that X's and Y's are uncorrelated is distributed as t with $(N - 2)$ degrees of freedom:

$$T = B \sqrt{\frac{N-2}{\left[\sum_{R=1}^{N} (E_{NR_{Xi}})^2\right]^2 - B^2}} \tag{10}$$

As an example a normal scores test for correlation is applied to determine if a significant correlation exists between the length of apertures paired across the branch of a fenestrate bryozoan, *Rectifenestella tenax.* Twelve pairs of observations are taken (Table 10), and rankits are substituted for data within groups (populations are *not* pooled for rankit conversion). The correlation coefficient $B = 1.739$ [calculated using Equation (9)] is used in Equation (10) to obtain a t-value.

$$T = 1.739 \sqrt{\frac{12-2}{9.445^2 - 1.739^2}} = 0.592$$

The two-sided critical t-value for $\alpha = 0.05$ with ten degrees of freedom is 2.228, which is much greater than the observed t-value of 0.592. Therefore, the null hypothesis (H_0: X and Y observations are uncorrelated) is not rejected at the 0.05 level of significance.

Other normal scores tests.—Variations of the normal scores tests presented here have been developed (e.g., Fraser, 1957; Capon, 1961; Bradley, 1968; McSweeney and Penfield, 1969). Presumably, new normal scores tests can be developed (adapted from existing parametric and nonparametric tests) as need arises.

SIGNIFICANCE LEVELS AND SCIENTIFIC CONCLUSIONS

Throughout discussion in this paper, null hypotheses have been rejected or not at the $\alpha = 0.05$ level of significance. This means that we are willing to accept a five percent rate of Type-I errors (rejecting a null hypothesis that is in fact true). It is important to remember that there is nothing magical about the 0.05 level of significance; in fact, the 0.05 significance level has its origins in gambling. The five percent level (20:1) was historically the odds at which bookmakers would shut down book on an event because the return on a winning bet was too small to attract interest from prospective betters on the favored side (a pay off of $1.05 for every dollar bet), and the bookmaker's commission on the transaction was too small to make it worth his time (A. W. Ghent, personal commun.).

Obviously, small differences between levels of significance near the 0.05 level are of limited meaning. This becomes especially clear when one examines the relationship between probabilities and sample sizes. The significance levels of tests generally increase as sample sizes increase. For example, based on a sample size of $N = 24$, the probability that the two fenestrate species in the Terry–Hoeffding test example have the same branch spacing (distance between branch centers) is greater than 0.20. When 12 different observations are collected for both species ($N = 24$), the probability that there is no difference in branch spacing is also greater than 0.20 (data not presented here). When these two data sets are combined ($N = 48$), the probability that the two species have the same branch spacing drops to between $0.10 < P < 0.05$. Undoubtedly, if more observations were added to the analysis, the probability would drop below the $\alpha = 0.05$ significance level. The point being that no two populations are exactly the same, and given a large enough sample size and degree of measurement precision, statistically significant differences can always be found (Foster and Kaesler, 1988).

Given the last statement, one may question the value of statistical tests. If, however, researchers weigh sample size and significance levels into final interpretations, relative levels of significance can provide a great deal of information. For example, the conclusion that one can draw from the results of the Van der Waerden test in this paper is that the relative difference in branch spacing between *R. tenuissima* and *R. multispinosa* is much greater than between *R. tenax* and *R. tenuissima.* It is, however, improper to conclude that there is no difference in

branch spacing between *R. tenax* and *R. tenuissima,* even though the null hypothesis was not rejected in this instance.

Principles discussed in this section are well known, and there are many excellent reviews of hypothesis testing and significance levels (e.g., Sokal and Rohlf, 1981, section 7.8; Foster and Kaesler, 1988, section 2.2.4). Regrettably, however, scientific conclusions are frequently made without consideration for effects of sample size.

SUMMARY

Normal scores tests provide practical methods for dealing with nonnormality and heteroscedasticity common in paleontological data. Despite their potential utility, normal scores tests have not received widespread acceptance as a common statistical method, primarily because the required rankit substitution is cumbersome and time consuming. Recent advances in microcomputer technology and software provide a viable method for converting raw data to rankits. This makes normal scores tests, which are the most powerful nonparametric tests, accessible for routine use.

ACKNOWLEDGMENTS

I thank D. B. Blake, A. W. Ghent, and J. F. Pachut for their comments and suggestions; I also thank A. F. Budd and R. J. Kaesler, whose comments greatly improved this paper. M. S. Wilkerson adapted the numerical approximation for rankit calculation for *Mathematica*®, which is provided in Appendix 1, and provided helpful suggestions in the preparation of this manuscript. This work was supported by PRF Grant 22927-BLA and NSF Grant BSR 89-03506 to D. B. Blake. Acknowledgment is made to the Donors of The Petroleum Research Fund, administered by the American Chemical Society, for partial support of this research.

REFERENCES

BRADLEY, J. V. 1968. Distribution-free Statistical Tests, Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 388 p.

CAPON, J. 1961. Asymptotic efficiency of certain locally most powerful rank tests. Annals of Mathematical Statistics, 29:972–994.

CONOVER, W. J. 1980. Practical Nonparametric Statistics, 2nd ed. John Wiley and Sons, New York, 493 p.

EFFRON, B., AND R. TIBSHIRANI. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science, 1:54–77.

FINK, W. L. 1990. Data acquisition for morphometric analysis in systematic biology, p. 9–19. *In* F. J. Rohlf and F. L. Bookstein (eds.), Proceedings of the Michigan Morphometrics Workshop, Special Publication 2. The University of Michigan Museum of Zoology, Ann Arbor, Michigan.

FISHER, R. A., AND F. YATES. 1938. Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd, London, 78 p.

FOSTER, D. W., AND R. L. KAESLER. 1988. Shape analysis: ideas from the Ostracoda, p. 53–69. *In* M. L. McKinney (ed.), Heterochrony in Evolution: A Multidisciplinary Approach. Plenum Press, New York.

FRASER, D. A. S. 1957. Most powerful rank-type tests. Annals of Mathematical Statistics, 28:1040–1043.

GHENT, A. W. 1971. Theory and application of several tests of sequential nonrandomness. The Biologist, 53:169–214.

GILINSKY, N. L., AND R. K. BAMBACH. 1986. Bootstrapping a regression equation: some empirical results. Paleobiology, 12:251–268.

HARTER, H. L. 1961. Expected values of normal order statistics. Biometrika, 48:151–165.

IPSEN, J., AND N. K. JERNE. 1944. Graphical evaluation of the distribution of small experimental series. Acta Pathologica et Microbiologica Scandinavia, 21:343–361.

KLOTZ, J. 1962. Nonparametric tests for scale. Annals of Mathematical Statistics, 33:498–512.

——. 1964. On the normal scores two-sample rank test. American Statistical Association Journal, 59:653–664.

MARASCUILO, L. A., AND M. MCSWEENEY. 1977. Nonparametric and Distribution Free Methods for the Social Sciences. Brooks/Cole Publishing Company, Monterey, California, 556 p.

MCSWEENEY, M., AND D. PENFIELD. 1969. The normal scores test for the c-sample problem. The British Journal of Mathematical and Statistical Psychology, 22:177–192.

PLOTNICK, R. E. 1989. Application of bootstrap methods to reduced major axis line fitting. Systematic Zoology, 38:144–153.

ROCK, N. M. S. 1988. Lecture Notes in Earth Sciences, 18 Numerical Geology. Springer-Verlag, New York, 427 p.

ROHLF, F. J., AND R. R. SOKAL. 1981. Statistical Tables, 2nd ed. W. H. Freeman and Company, New York, 219 p.

SNYDER, E. M. 1991. Revised taxonomic procedures and paleoecological applications for some North American Mississippian Fenestellidae and Polyporidae (Bryozoa). Palaeontographica Americana 57, 351 p.

SOKAL, R. R., AND F. J. ROHLF. 1981. Biometry, 2nd ed. W. H. Freeman and Company, New York, 859 p.

TERRY, M. E. 1952. Some rank order tests which are most powerful against specific parametric alternatives. Annals of Mathematical Statistics, 23:346–366.

VAN DER WAERDEN, B. L. 1952. Order tests for the two-sample problem and their power. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 55:453–458.

VAN EDEN, C. 1963. The relation between Pitman's asymptotic relative efficiency of two tests and the correlation coefficient between their test statistics. Annals of Mathematical Statistics, 34:1442–1451.

ZAR, J. 1984. Biostatistical Analysis. Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 718 p.

ZELEN, M., AND SEVERO, N. C. 1965. Probability functions, p. 925–926. *In* M. Abramowitz and I. A. Stegun (eds.), Handbook of Mathematical Functions. Dover Publications Inc., New York.

ACCEPTED 22 NOVEMBER 1991

APPENDIX 1

The equation for calculating rankit values:

$$E(\mu_{r|n}) = \frac{n!}{(n-r)!(r-1)!} \times \int_{-\infty}^{+\infty} \mu[1/2 - \Phi(\mu)]^{r-1}[1/2 + \Phi(\mu)]^{n-r}\phi(\mu)\, d\mu \quad (1)$$

where $\phi(\mu) = (2\pi)^{-1/2}e^{-1/2\mu^2}$ and $\Phi(\mu) = \int_0^{\mu} \phi(\mu)\, d\mu$

can be solved via numerical integration as outlined by Harter (1961). The method determines the solution for each n and r by writing Equation (1) in terms of logarithms:

$$\log_e I(n, r, \mu) = \log_e n! - \log_e(n-r)! - \log_e(r-1)! + \log_e\mu + (r-1)\log_e(1/2 - \Phi(\mu)) + (n-r)\log_e(1/2 + \Phi(\mu)) + \log_e\phi(\mu) \quad (2)$$

where $\phi(\mu) = (2\pi)^{-1/2}e^{-1/2\mu^2}$ and $\phi(\mu) = \int_0^{\mu} \phi(\mu)\, d\mu$

Because $I(n, r, m) \approx 0$ when $|\mu| > 7.60$, $E(\mu_{r|n})$, Equation (1) can be solved by summing the exponential of the right hand side of Equation (2) for μ from -7.60 to 7.60 in steps of δ, and then multiplying the result by δ (Harter, 1961). This generates values of $E(\mu_{r|n})$ which are accurate to within a unit in the fifth decimal place.

The procedure employed here, unlike that of Harter (1961), does not necessitate the input of tables into a mainframe computer in order to solve the problem. *Mathematica*®, a program available for *Macintosh*® personal computers, can provide a numerical solution for this complex equation by utilizing a programming language with built-in mathematical functions. The *Mathematica*® function used to solve Equation (1) is:

```
NormStat [n_, r_]:
        = Block [{φ, Φ, first, second, answer, μ, t},
      φ = (2•Pi)^(−1/2)•Exp[(−1/2)•μ^2];
      Φ = Integrate [φ, {μ, 0, t}];
      Φ = Φ/.t −> μ;
  first = ((1/2) − Φ);
 second = ((1/2) + Φ);
 answer = Log[n!] − Log[(n − r)!] + Log[μ]
          + (r − 1)•Log[first] + (n − r)•Log[second] + Log[φ];
          Sum[Exp[answer], {μ, −7.60, 7.60, 0.05}]•0.05]
```

A copy of *Rankit* can be obtained by sending a Macintosh-formatted or IBM-formatted 5.25″ or 3.5″ disk to S. J. Hageman. Information about *Mathematica*® can be obtained from: Wolfram Research, Inc., P.O. Box 6059, Champaign, IL 61826-6059.

APPENDIX 2

This appendix provides step by step procedural descriptions for normal scores tests discussed above. All descriptions include steps of ranking data and conversion to rankits. These steps are provided here to demonstrate how the test would be performed by hand, where rankit values are obtained from tables (e.g., Harter, 1961, table 1). When the computer program *Rankit* is used, it performs these steps and provides rankit values corrected for ties.

Van Eden's test.—See text Equation (1) and Table 5.

1. Paired data are collected and arranged in two columns aligned by pairs.
2. The "Right" column is subtracted from the "Left," and results are placed in a new column "*d*."
3. Negative differences in column "*d*" are noted and highlighted.
4. Entries in column "*d*" are ranked by their absolute values and results are placed in a new column "Rank $|d|$." Note that entries with no difference (0.0000 in column "*d*") are dropped from the analysis at this point.
5. Rankit values are determined for the entries in column "Rank $|d|$" and placed in a new column "Rankit."
6. The signs of the entries in the "Rankit" column are changed to correspond to the sign of entries in column "*d*" (see step 3). These values are placed in a new column "Signed Rankit."
7. Entries in "Signed Rankit" column are squared and placed in a new column "(Rankit)^2."
8. Sums are obtained for the "Signed Rankit" and "(Rankit)^2" columns.
9. The test statistic Z is calculated by taking the sum of the "Signed Rankit" column and dividing it by the square root of the sum of "(Rankit)^2" column [see text Equation (1)].
10. The test statistic Z is equivalent to a normal deviate for which a probability can be obtained from a table of the probability distribution of a normal curve (e.g., Zelen and Severo, 1965, table 26.1).

Terry–Hoeffding test.—See text Equation (2) and Table 6.

1. Unpaired data are collected for two populations (not necessarily same number of observations per population).
2. Original data are ranked as a whole (both populations are pooled together) and results are placed in a new column "Rank."
3. Rank values are converted to their rankit equivalents, which are placed in a new column "Rankit."
4. Entries in the "Rankit" column are squared and the results are placed in a new column "(Rankit)^2."
5. The sums of the "Rankit" and "(Rankit)^2" columns are calculated.
6. The test statistic S [see text Equation (2)] is simply the sum of the "Rankit" column for the population with fewer number of observations (arbitrary if population sizes are equal).
7. Several methods exist for determining critical S-values.
   - 7a. Critical S-values can be obtained directly from a table (Klotz, 1964, table 1) when the pooled population size is less than 21. If the number of observations is greater than 20, one of the following methods can be employed.
   - 7b. Critical S-values can be obtained using the Student's t-distribution [see text Equation (3)].
     - 7b1. The test statistic S is squared and multiplied by the total number of observations minus two.
     - 7b2. The sum of the "(Rankit)^2" for population one is added to the sum of the "(Rankit)^2" for population two. This is multiplied by the number of observations in population one and by the number of observations in population two. This value is then divided by the total number of observations.
     - 7b3. The square of the test statistic S is then subtracted from the value obtained in step 7b2.
     - 7b4. A t-value is then calculated by dividing the value obtained in step 7b1 by the value obtained in step 7b3, and then taking the square root of the results.
     - 7b5. The t-value obtained in the previous step is then compared to critical t-values from a table (e.g., Zar, 1984, table B.3). If the t-value is greater than the critical t-value at a given confidence level (with the total number of observations minus two as the degrees of freedom), then the null hypothesis is rejected.
   - 7c. Critical S-values can be obtained for a given confidence level using the Pearson's product-moment correlation coefficient [see text Equation (4)].
     - 7c1. Step number 7b2 is performed and the square root of the value is obtained.
     - 7c2. The value obtained in the previous step is then multiplied by the two-tailed critical value of the Pearson's product-moment correlation coefficient for a given level of confidence, which can be obtained from a table (e.g., Zar, 1984, table B.16).
     - 7c3. If the test statistic S (step 6) is greater than the critical S_c (step 7c2), then the null hypothesis is rejected.
   - 7d. For a large number of observations, a critical S-value can be obtained from the normal distribution [see text Equation (5)].
     - 7d1. Step number 7b2 is performed and the result is divided by the total number of observations in both populations minus one.
     - 7d2. The critical Z-value is obtained by dividing the test statistic S by the square root of the value obtained in the previous step.
     - 7d3. The test statistic Z is equivalent to a normal deviate for which a probability can be obtained from a table of the probability distribution of a normal curve (e.g., Zelen and Severo, 1965, table 26.1).

Klotz test.—See text Equation (6) and Table 7.

1. Data are collected for two populations, and means are calculated within each population.
2. The within group means are subtracted from each observation and results are placed in a new column "Orig. $-\bar{x}$"
3. Entries in the "Orig. $-\bar{x}$" columns are ranked, with the populations pooled and results are placed in a new column "Rank."
4. Ranks are converted to their equivalent rankits and placed in a new column "Rankit."
5. Rankits are squared and raised to the fourth power to form two new columns "(Rankit)^2" and "(Rankit)^4," respectively.
6. Entries in the "(Rankit)^2" and "(Rankit)^4" columns are summed within each group.
7. The test statistic Z is calculated [see text Equation (6)].
   - 7a. The sum of the "(Rankit)^2" column for population one is added to the sum of the "(Rankit)^2" column for population two.
   - 7b. The results of step 7a are multiplied by the number of observations from the population with fewer observations and divided by the total number of observations in both populations.
   - 7c. The results from step 7b are subtracted from the results of step 7a [this forms numerator of text Equation (6)].
   - 7d. The results from step 7a are squared, and divided by the total number of observations in both populations.
   - 7e. The sum of the "(Rankit)^4" column for population one is added to the sum of the "(Rankit)^4" column for population two.

7f. The results of step 7d are subtracted from the results of step 7e. The result is then multiplied by the number of observations in population one and multiplied by the number of observations in population two. This result is then divided by the total number of observations in both populations, and then again divided by the total number of observations in both populations minus one.
7g. The test statistic Z is then obtained by dividing the result of step 7c by the square root of the result of step 7f.
7h. The test statistic Z is equivalent to a normal deviate for which a probability can be obtained from a table of the probability distribution of a normal curve (e.g., Zelen and Severo, 1965, table 26.1).

Van der Waerden test.—See text Equation (7) and Table 8.

1. Data are collected for three or more populations and placed in a column "Orig."
2. Data are pooled and ranked as if they came from a single population. The rank order values are placed in a new column "Rank."
3. Rankit equivalents are determined for the ranked data and placed in a new column "Rankit."
4. Rankit values are squared and placed in a new column "(Rankit)^2."
5. Entries in the "Rankit" and "(Rankit)^2" columns are summed within each group.
6. The test statistic, which is distributed as Chi squared, is calculated [see text Equation (7)].
   6a. The sum of the "Rankit" column for population one (step 5) is squared and the result is divided by the number of observations for population one. This step is repeated for each population (e.g., "Rankit" sum for population two divided by the number of observations in population two, and for population three . . .).
   6b. The results for each population from step 6a are added together.
   6c. The sums of all the "(Rankit)^2" columns (step 5) are added together to obtain a total sum of rankit squares.
   6d. The test statistic is obtained by multiplying the results of step 6b by the total number of observations minus one and then dividing by the results of step 6c.
   6e. The test statistic is then evaluated as χ^2 with the number of populations minus one as the degrees of freedom. This can be obtained from a table (e.g., Zar, 1984, table B.1).

Post-hoc multiple comparison of means.—See text Equation (8) and Tables 8 and 9.

1. A critical difference (D_c) is obtained for comparison of mean values between populations, two at a time. Only one D_c is required when all populations have the same number of observations. The base part of D_c is calculated as follows.
2. The two-tailed critical t-value for a given level of significance (based on the number of observations minus the number of groups as the degrees of freedom) is obtained from a table (e.g., Zar, 1984, table B.3).
3. The sums of all the "(Rankit)^2" columns (step 5 from Klotz test) are added together to obtain a total sum of rankit squares. This result is divided by the total number of observations minus one.
4. The χ^2 test statistic, obtained from the Klotz test (step 7e), is subtracted from the total number of observations minus one. This result is divided by the total number of observations minus the number of groups.
5. The results of step 3 are multiplied by the results of step 4.
6. The basic part of the critical difference value D_c is obtained by multiplying the critical t-value (step 2) by the square root of the results of step 5.
7. For any given pair of observations the critical difference D_c is obtained by adding the inverse of the number of observations in the first group to the inverse of the number of observations in the second group, taking the square root of the sum, and multiplying it by the results in step 6.
8. If the absolute value of the mean of group one minus the mean of group two is greater than the critical difference D_c (step 7), the difference is significant for the confidence level chosen in step 2.
9. If populations have different numbers of observations, steps 7 and 8 are repeated for each pair-wise comparison of group means. If populations have equal numbers of observations, then the critical difference value D_c is calculated once in step 7, and step 8 is repeated for each pair-wise comparison of group means.

Normal scores test for correlation.—See text Equations (9) and (10) and Table 10.

1. Paired data are collected and placed in columns designated "X" and "Y."
2. Data are ranked within each group (not pooled).
3. Rankit values are obtained for entries within each group and placed in new columns "Rankit X_i" and "Rankit Y_i."
4. Entries in the "Rankit X_i" and "Rankit Y_i" columns are squared and results are placed in new columns "(Rankit X_i)^2" and "(Rankit Y_i)^2," respectively.
5. Paired entries in the "Rankit X_i" and "Rankit Y_i" columns are multiplied and results are placed in a new column "Rankit $(X_i)(Y_i)$."
6. The normal scores correlation coefficient B is obtained by summing the entries in the "Rankit $(X_i)(Y_i)$" column [text Equation (9)].
7. The normal scores correlation coefficient B is tested for significance using text Equation (10).
   7a. The result of step 6 is squared and subtracted from the square of the sum of the "(Rankit X_i)^2" column.
   7b. The number of pairs of observations minus two is divided by the results obtained in step 7a.
   7c. The test statistic T is equal to the result of step 6 multiplied by the square root of the results of step 7b.
   7d. The two-tailed critical t-value for a given level of significance (based on the number of pairs of observations minus two as the degrees of freedom) is obtained from a table (e.g., Zar, 1984, table B.3).