STATISTICS IN MEDICINE, VOL. 13, 747–758 (1994)

ON NON-PARAMETRIC AND GENERALIZED TESTS FOR THE TWO-SAMPLE PROBLEM WITH LOCATION AND SCALE CHANGE ALTERNATIVES

MARVIN J. PODGOR

Division of Biometry and Epidemiology, National Eye Institute, Building 31 Room 6A52, National Institutes of Health, Bethesda, Maryland 20892, U.S.A.

AND

JOSEPH L. GASTWIRTH

Department of Statistics, George Washington University, Washington, DC 20052, U.S.A.

SUMMARY

Various tests have been proposed for the two-sample problem when the alternative is more general than a simple shift in location: non-parametric tests; O'Brien's generalized t and rank sum tests; and other tests related to the t. We show that the generalized tests are directly related to non-parametric tests proposed by Lepage. As a result, we obtain a wider, more flexible class of O'Brien-type procedures which inherit the level robustness property of non-parametric tests. We have also computed the tests' empirical sizes and powers under several models. The non-parametric procedures and the related O'Brien-type tests are valid and yield good power in the settings investigated. They are preferable to the t-test and related procedures whose type I errors differ noticeably from nominal size for skewed and long-tailed distributions.

1. INTRODUCTION

The usual two-sample problem tests the null hypothesis $G(x) = F(x)$ against the alternative $G(x) = F(x - \Delta)$, that is, the treatment data have shifted to the right. Here we consider the outcome to be measured on a continuous scale (for example, intraocular pressure in mmHg, serum cholesterol in mmol/l), and we note that 'treatment' may refer to group membership (for example, diseased versus disease-free) in epidemiologic studies. When F is the normal distribution, the two-sample t-test is the most powerful test. More generally, non-parametric procedures such as the Wilcoxon rank sum test or van der Waerden's normal scores test may be used for the location shift alternative. In many biomedical situations, however, the treatment effect is more complicated than a simple location shift. For example, in HIV infected patients, protein concentration in the anterior chamber of the eye was on average higher and more widely distributed among persons with cytomegalovirus (CMV) retinitis than among those without CMV retinitis.[1] This example typifies a class of problems in which an exposure or treatment affects not only the location or variability of a distribution but can change both simultaneously. Problems of this type led Lepage[2] to consider the two-parameter alternative

$$G(x) = F\left(\frac{x - \Delta}{\sigma}\right), \tag{1}$$

CCC0277–6715/94/070747–12

where σ describes the relative change in variability, or scale. A specific form of alternative (1) relates the change in scale to the change in location through a proportionality constant l, where $\sigma = 1 + \Delta l$, so that

$$G(x) = F\left(\frac{x - \Delta}{1 + \Delta l}\right). \tag{2}$$

When $l = 0$, the usual alternative of location-shift only is obtained. When $l \to \infty$, the alternative refers to scale-change only.[3-5]

Another situation arises when only a subset of patients is responsive to treatment.[6] Among the alternatives that may be hypothesized for this setting is one based on Lehmann alternatives,

$$G(x) = (1 - p)F(x) + p[F(x)]^a, \tag{3}$$

where p is the probability that a patient responds to treatment and $a(\geqslant 1)$ can be thought of as the treatment effect (among responders).[6]

O'Brien[7] notes that the two-sample t-test is related to regressing a group indicator on the outcome variable. He considers the log odds of treatment group membership as a function of the outcome variable. When the function is linear, the simple regression is appropriate, while non-linearity of the log odds and outcome suggests use of a quadratic term in the regression model.[7,8] Although he does not specify an alternative, O'Brien observes that these regressions, or 'generalized' t-tests, should be advantageous when the treatment effect is heterogeneous, but not limited to alternative (1).[7,8] We show that the generalized tests are directly related to existing non-parametric procedures for the location-scale problem[2-5] and exploit this relationship to obtain a wider family of O'Brien-type tests. We also compare the empirical powers of the procedures.

2. BACKGROUND, NOTATION AND DEFINITIONS

O'Brien[7] has proposed regressing a group indicator (I_i) on the observation x_i and its square x_i^2. The usual F-statistic is used to test that the two regression coefficients are equal to zero. This one-step procedure was extended to a two-step procedure by Grambsch and O'Brien.[8,9] There, if a preliminary test for zero coefficient for the squared term is significant, then a test that both coefficients are simultaneously zero is performed. If the preliminary test is not significant, then I_i is regressed on x_i, and the usual test for regression is performed. Whichever path is followed, the procedure applies a 50 per cent increase to the final P-value. Because the latter regression is equivalent to the ordinary two-sample t-test, O'Brien called this procedure a generalized t-test.[7] Using the ranks of the observations, rather than the observations themselves in the regressions, yields a generalized rank sum test.[7] Some variations of O'Brien's methods have been investigated through simulations by Blair and co-workers.[10-12]

Lepage[2] proposed a test designed to have good power for alternative (1). The test is based on the sum of the squares of standardized location-change only and scale-change only linear rank tests. For alternative (2), Lepage[3] proposed a linear combination, with coefficient l, of the location-change only and scale-change only linear rank tests. Gastwirth and Podgor[5] developed maximin efficiency robust tests[13-19] based on the latter Lepage test that demonstrate good power properties when all that is known is a range of possible values for l and that the underlying distribution is one of a family of possible distributions. We show that when the non-parametric forms of O'Brien's test are based on location-change only and scale-change only scores, these new procedures have similar properties to the corresponding Lepage type tests. In particular, Lepage's 2 d.f. test[2] and O'Brien's one-step test are asymptotically equivalent.

We also consider briefly in our simulation studies two other recently proposed tests. Conover and Salsburg[6] consider the case when only a subset of the treated group could respond to treatment, and they propose a non-parametric test for alternatives like (3). Brownie *et al.* [20] have proposed a simple modification to the standard t-test for improving power for testing location change when the treatment effect shifts the response and increases its variance. Blair and Sawilowsky[12] have compared empirical powers for the latter test, the pooled-variance t-test, and the generalized t-test incorporating a preliminary test using critical values from Blair.[10]

We assume we have two independent samples of size m and n such that $N = m + n$, where $X_1, \ldots, X_m$ have distribution $F(x)$ and $X_{m+1}, \ldots, X_{m+n}$ have distribution $G(x)$. Let I_j be a group indicator so that $I_j = 1$ when $j = 1, \ldots, m$ and $I_j = 0$ when $j = m + 1, \ldots, N$.

t: Student's t-test with $N - 2$ d.f. using a pooled estimate of variance.
One-step generalized t: O'Brien's F test with 2, $N - 3$ d.f. computed by regressing group indicator (I_j) on the observations (X_j) and their squares (X_j^2).
Rank sum: Linear rank test using Wilcoxon scores $[i/(N + 1)]$.
One-step generalized rank sum: O'Brien's[7] F test with 2, $N - 3$ d.f. computed by regressing group indicator on the ranks of the observations and the square of their ranks.
Lepage 1 d.f.: A linear combination of location-change only and scale-change only linear rank tests[3,4] with coefficient l, a proportionality constant for scale change to location change.
MERT: Gastwirth's maximin efficiency robust test.[13,14] In this study, we predominantly use a MERT based on Lepage 1 d.f. tests for given underlying distributions and values of l.[5] In one setting, we use the MERT for scale change in a gamma distribution.[15]
Lepage 2 d.f.: Sum of the squares of the standardized location-change only and scale-change only linear rank tests.[2] In one form of this test, we use scores based on the logistic distribution so that the location-change only test is Wilcoxon's test. In another form, we use scores based on the normal distribution so that the location-change only test is van der Waerden's test[21] and the scale-change only test is Klotz's test.[21]
Conover–Salsburg[6]: Linear rank test using scores $[i/(N + 1)]^4$.
Brownie et al.: t-test with $m - 1$ d.f. and the variance of the first sample used in place of the pooled variance estimate.[20]
Grambsch–O'Brien generalized t: O'Brien's procedure[7] that includes a preliminary test for zero coefficient of the squared term in the quadratic model and incorporation of the '50 per cent' rule to adjust the final P-value.[8,9] In our simulations, we use a preliminary test with level 0·10.
Grambsch–O'Brien generalized rank sum test: Similar to above, but using the ranks of the observations, rather than the observations themselves.

(Comment on terminology: What we have called the one-step generalized test was referred to as the unconditional generalized test by Blair and Morel.[11] Their term conditional generalized test corresponds to O'Brien's original procedure that incorporated a preliminary test without adjustment of the P-value.[7])

Let $X_{(i)}$ denote the ith order statistic among the N observations, and let $a(i)$ denote the score associated with $X_{(i)}$. Then, the linear rank statistic is $\sum_{i=1}^{N} a(i)I_{(i)}$. Here we use the approximate (or J-function) scores $a(i) = \phi(i/(N + 1))$, based on score function $\phi(u)$, $0 < u < 1$. When simulating, we choose the underlying distribution and can therefore use scores yielding the asymptotically most powerful rank test (AMPRT). We have computed the limiting forms of the linear rank tests (AMPRT, rank sum, Conover–Salsburg, and the component tests for Lepage's tests and the MERT). Without loss of generality, suppose we use centred (mean zero) scores $a(i)$. Then

the test statistic

$$\left(\sum_{i=1}^{N} a(i)I_{(i)}\right)\left(N(N-1)/mn\sum_{i=1}^{N} a(i)^2\right)^{1/2}$$

is compared to the critical value of the standard normal distribution.[21]

Moses has demonstrated aberrations of linear rank tests designed strictly for scale change alternatives when the distributions also differ in location.[22] In that setting, one may align the observations by an estimate of the difference in location and then compute the scale change test on the aligned observations. This estimation for finite sample sizes, however, destroys the distribution-free property of the procedure.[23,24] For alternatives (1) and (2), which we consider, we point out that the scale-change only components for the Lapage[2,3] tests and the MERT[5] are computed *without* alignment for possibly different location parameters.

3. A CLASS OF O'BRIEN-TYPE PROCEDURES AND ITS RELATIONSHIP TO LEPAGE'S PROCEDURES

O'Brien[7] applies a quadratic model to the observations themselves (generalized t) or to the ranks of the observations (generalized rank sum). One may extend his approach to scores based on the ranks. We show, in particular, that if location-change only and scale-change only scores are used, the extended procedure is asymptotically equivalent to the 2 d.f. test of Lepage.

As in the previous section, let $X_{(i)}$ denote the ith order statistic among the N observations. Now let $J_1(i/(N+1))$ and $J_2(i/(N+1))$ denote two scores associated with $X_{(i)}$. We assume the scores are based on square integrable score functions $J_j(u), j = 1, 2$, where, without loss of generality $\int_0^1 J_j(u)du = 0$ (so that we use centred scores). We then have two two-sample linear rank tests

$$T_j = \sum_{i=1}^{N} J_j\left(\frac{i}{N+1}\right)I_{(i)},$$

where, as before, $I_{(i)}$ is the group indicator. By dividing the statistic by its standard deviation, we obtain the standardized statistic

$$Z_j = T_j \Big/ \sqrt{\left\{\frac{mn}{N(N-1)}\sum_{i=1}^{N}\left(J_j\left(\frac{i}{N+1}\right)\right)^2\right\}}, \quad j = 1, 2.$$

Now let

$$Q = (Z_1^2 - 2\rho_N Z_1 Z_2 + Z_2^2)(1-\rho_N^2)^{-1} \tag{4}$$

where ρ_N is the correlation of the scores,

$$\rho_N = \frac{\sum_{i=1}^{N} J_1\left(\frac{i}{N+1}\right)J_2\left(\frac{i}{N+1}\right)}{\sqrt{\left\{\sum_{i=1}^{N} J_1^2\left(\frac{i}{N+1}\right)\right\}}\sqrt{\left\{\sum_{i=1}^{N} J_2^2\left(\frac{i}{N+1}\right)\right\}}}.$$

The large sample theory is given by the following:

Theorem Suppose the multiple linear regression of $I_{(i)}$ on $J_1(i/(N+1))$ and $J_2(i/(N+1))$ yields corresponding least squares regression coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$. Then under the null hypothesis H_0: $(\beta_1, \beta_2) = (0, 0)$, Q is asymptotically distributed as a central χ^2 with 2 degrees of freedom.

The proof is outlined in the Appendix.

O'Brien's one-step procedure tests that the two regression coefficients are zero. The theorem shows that asymptotically the extended procedure can be recast as a quadratic combination of linear rank tests. In particular, O'Brien's one-step generalized rank sum test is a quadratic combination of the rank sum test and squared ranks test.[25] If the score functions are orthogonal, their correlation is zero, and the following obtains:

Corollary 1 If $J_1(\cdot)$ and $J_2(\cdot)$ are orthogonal, then $Q = Z_1^2 + Z_2^2$.

An immediate consequence is the following for particular orthogonal score functions:

Corollary 2 When $J_1(\cdot)$ is a score function for location-change only and $J_2(\cdot)$ is a score function for scale-change only, then Q is Lepage's 2 d.f. test and is asymptotically equivalent to the generalized O'Brien test regressing $I_{(i)}$ on $J_1(\cdot)$ and $J_2(\cdot)$.

The derivation of the level by Grambsch and O'Brien for their two-step test obtains asymptotically under Corollaries 1 and 2, indicating that methods from the extended class employing the '50 per cent' rule should enjoy the properties of the Grambsch–O'Brien[9] procedure.

Because the linear combination of the regression coefficients, $a\hat{\beta}_1 + b\hat{\beta}_2$, is a linear combination of the statistics T_1 and T_2 (see Appendix), it can be rewritten in the form of Lepage 1 d.f. statistics.[3-5] Therefore, the regression approach can also be used to obtain Lepage-type 1 d.f. tests.

4. SIMULATION RESULTS

Table I presents the results of applying the various tests to data from normal distributions, with and without a change in scale in addition to a shift in location. While all the tests have sizes near the nominal level, the usual t and Wilcoxon rank sum tests suffer a noticeable loss in power when both scale and location changed.[7] The power of the Brownie *et al.* test is lower than the other tests designed to detect location and scale changes. We have also included results for our extension of O'Brien's procedure by regressing on quantile normal scores[21] and their squares. From Corollary 2, we expect that the one-step procedure using these scores should have the same properties as the Lepage 2 d.f. test based on the normal distribution. The simulations bear this out. Results for the corresponding two-step procedure using the Grambsch and O'Brien 50 per cent rule are also presented and show level robustness and good power properties. We now provide more details about the test statistics and the simulation study.

Our simulations were done using double precision FORTRAN. Uniform and normal random numbers were generated with IMSL[26] routines DRNUN and DRNNOA, respectively. Other random numbers were generated with the corresponding IMSL routines or by the inverse cdf method using uniform random numbers. The alternatives in Table I are those of Table 4 of O'Brien[7] (data from which is also provided in Table 1 of Blair and Morel[11]), so that the results presented in the first four rows of Table I are similar to those of O'Brien.[7] The Lepage AMPRT is computed as the linear combination of location-only and scale-only linear rank tests using normal scores, with the proportionality constant l, where $l = (\sigma - 1)/\mu$.[5] (Thus $l = 0, 1, 2$ for the three alternatives in the table.) The MERT was computed under the assumption that the true underlying density could be normal or logistic and that l is in the interval $[0, 1]$. Correlations of the asymptotic score functions were used to compute the MERT.

Table II presents results from exponential data and shows that the non-parametric tests tend to have better size and power characteristics than the generalized tests. The first four rows and first five columns of Table II correspond to results in Table 6 of Blair and Morel.[11] The optimal score function for exponential data is $-1 - \ln[1 - u]$, yielding the exponential scores test, and results in Table II use the approximate scores $-1 - \ln[1 - i/(N + 1)]$. Gastwirth and Mahmoud's

Table I. Type I error and power for various two-sample tests – normal data*

Statistic (two-tailed $\alpha = 0{\cdot}10$)	Null: $\mu = 0, \sigma = 1$	Location only change: $\mu = 1, \sigma = 1$	Location and scale change: $\mu = 1, \sigma = 2$	Location and scale change: $\mu = 1, \sigma = 3$
t	0·100† [0·094]‡	0·968 [0·960]	0·714 [0·728]	0·476 [0·488]
One-step generalized t	0·092 [0·100]	0·932 [0·931]	0·946 [0·960]	0·990 [0·994]
Rank sum	0·098 [0·090]	0·962 [0·954]	0·679 [0·684]	0·456 [0·463]
One-step generalized rank sum	0·099 [0·105]	0·920 [0·914]	0·942 [0·954]	0·994 [0·997]
Lepage 1 d.f. (AMPRT)	0·099	0·967	0·983	0·999
MERT – normal, logistic; $l \in [0, 1]$	0·108	0·941	0·935	0·895
Lepage 2 d.f. – logistic	0·096	0·918	0·952	0·998
Lepage 2 d.f. – normal	0·094	0·926	0·955	0·998
Conover–Salsburg	0·101	0·918	0·946	0·939
Brownie *et al.*	0·100	0·967	0·879	0·796
Grambsch–O'Brien generalized t	0·093	0·954	0·922	0·978
Grambsch–O'Brien generalized rank sum	0·097	0·944	0·918	0·990
One-step generalized normal scores	0·095	0·927	0·956	0·998
Two-step generalized normal scores (Grambsch – O'Brien 50 per cent rule)	0·093	0·943	0·931	0·996

* Proportion of times H_0 rejected in 5000 simulations. For the first sample, 25 random numbers were drawn from the standard normal distribution. For the second sample, 25 random numbers were drawn from the normal distribution with mean and variance given in the column headings

† For the null case, results of the current analysis are based on 15,000 replications

‡ Values in square brackets, given in Table 4 of O'Brien[7] and reproduced in Table 1 of Blair and Morel,[11] were based on 1000 replications

MERT[15] is formed assuming the underlying distribution is in the gamma family (ranging from the exponential to the normal distribution). Here we compute the MERT from the correlation (0·931) of the J-function scores rather than from the asymptotic correlation (0·902).[15] The MERT and Conover and Salsburg's test display good power, with the latter close to that of the exponential scores test. (That the Conover–Salsburg and exponential scores tests have similar power is expected, since the asymptotic relative efficiency of either is 0·926 when the other is optimal.[6]) As noted by Blair and Morel,[11] the one-step generalized tests have lower power. The Grambsch–O'Brien generalized tests have somewhat better power, but, in general, less than that of the non-parametric tests. In particular, the generalized t tests seem conservative, having smaller sizes than any of the other tests investigated. On the other hand, the Brownie *et al.* test is anticonservative in this situation.

We investigated the validity of the various tests under several distributions (Table III). The non-parametric tests are slightly conservative, probably reflecting the approximation to the limiting distribution for the relatively small ($N = 30$) sample size. The one-step generalized t is conservative for normal data. This was also observed by Blair and Morel,[11] who noted that, with larger sample sizes, the test approached nominal levels. Under the long-tailed Cauchy distribution, the t and both forms of the generalized t-tests were highly conservative. The Brownie *et al.*

Table II. Type I error and power for various two-sample tests – exponential data*

Statistic (two-tailed $\alpha = 0{\cdot}05$)	$\beta = 1$ (null)	$\beta = 1{\cdot}5$	$\beta = 2{\cdot}0$	$\beta = 2{\cdot}5$	$\beta = 3{\cdot}0$	$\beta = 3{\cdot}5$	$\beta = 4{\cdot}0$
t	0·044 [0·046]†	0·189 [0·192]	0·447 [0·474]	0·695 [0·696]	0·851 [0·834]	0·900	0·952
One-step generalized t	0·034 [0·038]	0·107 [0·114]	0·291 [0·306]	0·522 [0·522]	0·717 [0·695]	0·818	0·884
Rank sum	0·058 [0·051]	0·177 [0·179]	0·386 [0·416]	0·619 [0·629]	0·791 [0·764]	0·867	0·917
One-step generalized rank sum	0·052 [0·051]	0·152 [0·147]	0·328 [0·353]	0·577 [0·570]	0·751 [0·726]	0·835	0·889
Exponential scores test (AMPRT)	0·048	0·194	0·452	0·700	0·850	0·904	0·953
MERT – gamma family	0·049	0·191	0·435	0·675	0·828	0·896	0·937
Lepage 2 d.f. – logistic	0·049	0·134	0·301	0·544	0·717	0·812	0·881
Lepage 2 d.f. – normal	0·043	0·123	0·289	0·546	0·725	0·818	0·890
Conover–Salsburg	0·052	0·188	0·446	0·688	0·847	0·903	0·959
Brownie *et al.*	0·072	0·399	0·692	0·880	0·951	0·975	0·995
Grambsch–O'Brien generalized t	0·037	0·152	0·372	0·621	0·797	0·867	0·931
Grambsch–O'Brien generalized rank sum	0·051	0·164	0·352	0·595	0·768	0·850	0·904
One-step generalized normal scores	0·046	0·140	0·310	0·573	0·746	0·832	0·899
Two-step generalized normal scores (Grambsch – O'Brien 50 per cent rule)	0·047	0·155	0·340	0·582	0·758	0·842	0·898

* Proportion of times H_0 rejected in 1000 simulations. For the first sample, 18 random numbers were drawn from the unit exponential distribution. For the second sample, 18 random numbers were drawn from the exponential distribution with mean β, where β is given in the column headings

† Values in square brackets are from Table 6 of Blair and Morel[11]

test was, however, anticonservative. Similar findings under other distributions were reported by Blair and Sawilowsky.[12]

Table IV presents type I error and power under the Lehmann alternative $G(x) = [F(x)]^{1+\theta}$, $\theta \geqslant 0$, where, to appropriately evaluate the t and generalized t-tests, we use the standard normal distribution for $F(x)$. We do not present results for the Conover–Salsburg and Brownie *et al.* statistics, because these tests were not designed for alternatives allowing an increase in mean accompanied by a decrease in variance, as occurs in this setting, and are expected to have low power. Savage's test is locally optimal for the Lehmann alternative, and we use the approximate scores $\ln[i/(N+1)]$. One MERT is constructed under the assumptions that either a normal or logistic distribution generates the data and l is in the interval $[-1, 0]$. This MERT has power almost coinciding with that of the Savage test. We also considered a MERT for the situation where we are less sure of the direction of change in the variance, that is, where the variance might either increase or decrease with treatment. For l in the interval $[-\frac{1}{2}, \frac{1}{2}]$, this latter MERT has somewhat lower power, close to that of the rank sum test. The t-test displays good power, while the generalized tests and the Lepage 2 d.f. tests have somewhat reduced power.

Table III. Type I error for various two-sample tests under several distributions*

Statistic (two-tailed $\alpha = 0{\cdot}05$)	Normal	Uniform	Beta (0·5, 0·5)	Cauchy
t	0·049	0·051	0·050	0·022
One-step generalized t	0·038	0·047	0·052	0·016
Rank sum	0·051	0·050	0·051	0·052
One-step generalized rank sum	0·047	0·048	0·054	0·052
MERT – normal, logistic; $l \in [0, 1]$	0·049	0·044	0·045	0·049
Lepage 2 d.f. – logistic	0·041	0·040	0·044	0·042
Lepage 2 d.f. – normal	0·037	0·035	0·039	0·038
Conover–Salsburg	0·048	0·045	0·044	0·049
Brownie *et al.*	0·051	0·045	0·045	0·222
Grambsch–O'Brien generalized t	0·044	0·050	0·054	0·018
Grambsch–O'Brien generalized rank sum	0·051	0·048	0·051	0·052
One-step generalized normal scores	0·042	0·041	0·045	0·043
Two-step generalized normal scores (Grambsch–O'Brien 50 per cent rule)	0·045	0·042	0·047	0·046

* Proportion of times H_0 rejected in 10,000 simulations. For each of the two samples, 15 random numbers were drawn from the indicated distribution

Table IV. Type I error and power for various two-sample tests – Lehmann alternatives*

Statistic (two-tailed $\alpha = 0{\cdot}05$)	$\theta = 0$ (null)	$\theta = 0{\cdot}5$	$\theta = 1{\cdot}0$	$\theta = 1{\cdot}5$	$\theta = 2{\cdot}0$	$\theta = 2{\cdot}5$	$\theta = 3{\cdot}0$
t	0·048	0·152	0·367	0·554	0·725	0·821	0·885
One-step generalized t	0·037	0·105	0·263	0·437	0·609	0·728	0·818
Rank sum	0·047	0·146	0·344	0·532	0·687	0·793	0·865
One-step generalized rank sum	0·051	0·122	0·293	0·467	0·635	0·753	0·837
Savage test	0·043	0·155	0·391	0·591	0·756	0·856	0·913
MERT – normal, logistic; $l \in [-1, 0]$	0·044	0·156	0·389	0·591	0·750	0·850	0·912
MERT – normal, logistic; $l \in [-\frac{1}{2}, \frac{1}{2}]$	0·047	0·146	0·345	0·530	0·689	0·793	0·860
Lepage 2 d.f. – logistic	0·042	0·106	0·261	0·436	0·600	0·721	0·811
Lepage 2 d.f. – normal	0·038	0·101	0·255	0·431	0·597	0·721	0·814
Grambsch–O'Brien generalized t	0·043	0·133	0·325	0·509	0·683	0·792	0·865
Grambsch–O'Brien generalized rank sum	0·048	0·134	0·321	0·505	0·667	0·783	0·852
One-step generalized normal scores	0·045	0·114	0·278	0·454	0·624	0·746	0·833
Two-step generalized normal scores (Grambsch–O'Brien 50 per cent rule)	0·042	0·125	0·308	0·488	0·653	0·769	0·842

* Proportion of times H_0 rejected in 5000 simulations. For the first sample, 15 random numbers were drawn from the standard normal distribution. For the second sample, 15 random numbers were drawn from the distribution with cdf $[\Phi(x)]^{1+\theta}$, where θ is given in the column headings

Table V. Normalized test statistics (Z-values) for data* from Conover and Salsburg[6] used in the construction of several non-parametric tests

Test	Normal score function	Logistic score function
Location-change only	0·900 ($Z_{N,l=0}$)†	0·982 ($Z_{L,l=0}$)‡
Scale-change only	2·197 ($Z_{N,l=\infty}$)§	2·263 ($Z_{L,l=\infty}$)
Location-scale change ($l = 1$)**	2·253 ($Z_{N,l=1}$)	2·307 ($Z_{L,l=1}$)

* Treated: − 1·535, − 0·547, − 0·201, − 0·201, − 0·154, − 0·095, − 0·049, 0·000, 0·000, 0·000, 0·105, 0·111, 0·201, 0·251. 0·310, 0·406, 0·511, 0·531, 0·575, 0·575, 0·773, 0·981. 1·299, 1·299, 1·322, 1·386, 1·792, 2·398 (n = 28, mean = 0·43, SD = 0·79) Controls: − 1·490, − 0·201, − 0·128, − 0·087, − 0·054, 0·000, 0·000, 0·000, 0·000, 0·000, 0·028, 0·039, 0·049, 0·061, 0·080, 0·105, 0·134, 0·193, 0·216, 0·223, 0·273, 0·288, 0·330, 0·357, 0·487, 0·541, 0·793, 1·042, 1·099, 1·609 (n = 30, mean = 0·20, SD = 0·52)

† van der Waerden's test

‡ Wilcoxon rank sum test

§ Klotz's test

** Linear combination, incorporating coefficient $l = 1$, of location-change only and scale-change only tests[3-5]

Table VI. Test statistics for Conover and Salsburg data[6]

Test	Statistic	P-value (2-tailed)
Wilcoxon rank sum*	$Z = 0{\cdot}982$	0·326
One-step generalized rank sum	$F_{2,55} = 3{\cdot}487$	0·038
MERT – normal, logistic, $l \in [0, 1]$†	$Z = 1{\cdot}770$	0·077
Lepage 2 d.f. – logistic‡	$\chi^2_2 = 6{\cdot}086$	0·048
Lepage 2 d.f. – normal§	$\chi^2_2 = 5{\cdot}636$	0·060
Conover–Salsburg	$Z = 1{\cdot}868$	0·062
Grambsch–O'Brien generalized rank sum	$F_{2,55} = 3{\cdot}487 \rightarrow 1{\cdot}5 \times (P\text{-value})$	0·057
One-step generalized normal scores	$F_{2,55} = 3{\cdot}004$	0·058
Two-step generalized normal scores (Grambsch – O'Brien 50 per cent rule)	$F_{2,55} = 3{\cdot}004 \rightarrow 1{\cdot}5 \times (P\text{-value})$	0·087
Exponential scores	$Z = 1{\cdot}629$	0·103

* $Z_{L,l=0}$ (Table V notation)

† Linear combination of $Z_{N,l=0}, Z_{L,l=0}, Z_{N,l=1}$ and $Z_{L,l=1}$, with corresponding weights 0·100, 0·454, 0·548, and 0 computed from the correlation matrix of the four component tests' J-function scores[5]

‡ $Z^2_{L,l=0} + Z^2_{L,l=\infty}$

§ $Z^2_{N,l=0} + Z^2_{N,l=\infty}$

5. EXAMPLE

Conover and Salsburg[6] present changes in pain scores, on a logarithmic scale, from baseline to four weeks post-treatment for persons with acute painful diabetic neuropathy. The data and component statistics used to construct several of the tests we considered are presented in Table V. The tests themselves are given in Table VI. Because the data are not normal, we consider only distribution-free tests. Midranks were used for tied observations. Although the rank sum test is far from significant, the other procedures indicate that the treated and control groups differ as the P-values are close to the 0·05 level. We emphasize that in practice the test procedure to be used should be specified before analysing the data. Its selection will depend on the hypothesized alternative. Indeed, if the alternative hypothesis is that the treatment changes the variability of the outcome in addition to modifying the average outcome, then procedures other than those for location-shift only should be considered. In particular, the MERT we considered applies when the alternative may range from location-shift only to location-shift accompanied by scale change

(specifically, the proportionality constant of scale change to location change ranges from 0 to 1) for either light (normal) or heavier (logistic) tailed underlying distributions. On the other hand, if one excludes the possibility of only a simple location change in favour of both location and scale changing simultaneously, the Lepage 2 d.f. tests could be used. In this example, the results using the generalized tests were similar to those of the non-parametric tests. In particular, we note the similarity of the one-step generalized normal scores test ($P = 0{\cdot}058$) and the Lepage 2 d.f. test based on normal scores ($P = 0{\cdot}060$). This is anticipated by Corollary 2, the statement of the asymptotic equivalence of these two tests.

6. DISCUSSION

Non-parametric tests for location and scale alternatives were proposed by Lepage. Asymptotic properties of the Lepage 2 d.f. test[2] have been studied,[27-29] and Lepage[30] has provided a table of exact critical values for the small sample application of a particular form of the test. Although most investigations of the 1 d.f. Lepage test[3,4] have been asymptotic, one may compute exact tests (for example StatXact[31]) based on the linear combination of scores for small samples. Scores for the efficiency robust method, the MERT[5], may be obtained in the small sample setting from the correlation matrix of component scores. Generalized tests were proposed by O'Brien[7] and have been further investigated recently.[8-12] We have shown that the tests proposed by O'Brien are contained in a versatile class of procedures obtained by regressing group indicator on scores specifying two rank tests. This flexibility in choosing score functions enables the statistician to develop appropriate level-robust procedures with good power for the type of data at hand. Furthermore, we have demonstrated that when location-change only and scale-change only scores are used, O'Brien's one-step procedure is asymptotically equivalent to a Lepage 2 d.f. test and a linear combination of O'Brien's regression coefficients yields a Lepage 1 d.f. test.

The non-parametric tests (rank sum, Lepage 1 and 2 d.f., the MERTs, and Conover–Salsburg) are distribution-free over the class of continuous distributions. The theorem shows that the one-step generalized tests based on score functions are asymptotically distribution-free over this class. Detailed properties of the two-step procedures deserve further study. The t, generalized t, and Brownie *et al.* tests are not distribution-free, but only asymptotically distribution-free for limited classes of distributions. For example, the t-test is asymptotically distribution-free for continuous distributions with finite variance.[24] Therefore, this property does not hold for the t-test under the Cauchy distribution.

Our simulations indicate that when the underlying data are near normal, the Lepage tests, the generalized t and generalized rank sum tests, the Conover–Salsburg test, and the efficiency robust tests of Gastwirth and Podgor attain the designated nominal size and have desirable powers. When the data come from more skewed distributions or from a distribution with longer tails (the Cauchy), the t-test, the generalized t-tests, and the Brownie *et al.* test can differ noticeably from their nominal sizes. The powers of the other tests are good, but no single procedure dominates. This is not surprising, as different tests are more powerful against different alternatives. The main point is that, if one has a general idea of the types of alternative models, one can either use an existing rank test, for example Lepage's, or can readily develop an efficiency robust procedure for the problem. Our simulations imply that the non-parametric procedures and related O'Brien-type tests can be recommended. Although the Grambsch–O'Brien generalized rank sum test performs well, the adjustment to its significance level by the '50 per cent rule' seems arbitrary.[32] Additional investigation of the level of that test should yield further insight into its domain of applicability. On the other hand, the non-parametric methods are based on well-established theory and can yield powerful procedures under a variety of possible models.

APPENDIX: OUTLINE OF PROOF OF THEOREM

Obtain the least squares solutions for β_1 and β_2, the regression coefficients in the model

$$I_i = \alpha + \beta_1 J_1\left(\frac{i}{N+1}\right) + \beta_2 J_2\left(\frac{i}{N+1}\right) + e_i,$$

where we assume only that the error term is independently distributed with mean zero and constant variance. Specifically then,

$$\hat{\beta}_1 = \left(\frac{T_1}{\sum_{i=1}^{N} J_1^2\left(\frac{i}{N+1}\right)} - \frac{\sum_{i=1}^{N} J_1\left(\frac{i}{N+1}\right) J_2\left(\frac{i}{N+1}\right)}{\sum_{i=1}^{N} J_1^2\left(\frac{i}{N+1}\right) \sum_{i=1}^{N} J_2^2\left(\frac{i}{N+1}\right)} T_2 \right) (1 - \rho_N^2)^{-1}$$

and

$$V_0(\hat{\beta}_1) = \frac{mn}{N(N-1)} \left[(1 - \rho_N^2) \sum_{i=1}^{N} J_1^2\left(\frac{i}{N+1}\right) \right]^{-1}.$$

T_1 and T_2 are asymptotically normal under the null. Without loss of generality, assume the J functions are in standardized form so that $\hat{\beta}_1 \to (T_1 - \rho T_2)/(1 - \rho^2)$ and $\hat{\beta}_2 \to (T_2 - \rho T_1)/(1 - \rho^2)$. Now, $a\hat{\beta}_1 + b\hat{\beta}_2 \to [(a - b\rho)/(1 - \rho^2)\, T_1] + [(b - a\rho)/(1 - \rho^2)\, T_2]$ is of form $\sum_{i=1}^{N} [J(i/(N+1))I_{(i)}]$ for any constants a and b. Limiting normality of the linear combination results from the Chernoff–Savage Theorem,[33] so that the joint normality of $\hat{\beta}_1$ and $\hat{\beta}_2$ follows from the Cramér–Wold device. Therefore, $\hat{\beta}_1$ and $\hat{\beta}_2$ are, under the null, asymptotically jointly normal with mean (0, 0) and covariance matrix $\sum$. Then $Q = (\hat{\beta}_1, \hat{\beta}_2) \sum^{-1} (\hat{\beta}_1, \hat{\beta}_2)'$ is asymptotically χ_2^2, and with some algebra may be shown to yield the expression in equation (4).

ACKNOWLEDGEMENT

J.L.G.'s research supported in part by the National Science Foundation Grant Number SES-9209994.

REFERENCES

1. Muccioli, C., Belfort Jr., R., Podgor, M., Sampaio, P., Hayashi, S., Neves, R., Lottenberg, C., Kim, M. K., de Smet, M. and Nussenblatt, R. 'The diagnosis of intraocular inflammation and CMV retinitis in HIV infected patients by laser flare photometry', *Investigative Ophthalmology and Visual Science*, **34**, 1110 (1993).
2. Lepage, Y. 'A combination of Wilcoxon's and Ansari–Bradley's statistics', *Biometrika*, **58**, 213–217 (1971).
3. Lepage, Y. 'Asymptotically optimum rank tests for contiguous location and scale alternatives', *Communications in Statistics*, **4**, 671–687 (1975).
4. Lepage, Y. 'Asymptotic power efficiency for a location and scale problem', *Communications in Statistics – Theory and Methods*, **A5**, 1257–1274 (1976).
5. Gastwirth, J. L. and Podgor, M. J. 'Efficiency robust rank tests for the location-scale problem', *in* Saleh, A. K. Md. E. (ed.), *Nonparametric Statistics and Related Topics*, Elsevier, Amsterdam, 1992, pp. 17–31.
6. Conover, W. J. and Salsburg, D. S. 'Locally most powerful tests for detecting treatment effects when only a subset of patients can be expected to "respond" to treatment', *Biometrics*, **44**, 189–196 (1988).
7. O'Brien, P. C. 'Comparing two samples: extensions of the *t*, rank-sum, and log-rank tests', *Journal of the American Statistical Association*, **83**, 52–61 (1988).
8. O'Brien, P. C. 'Comment on Blair and Morel', *Statistics in Medicine*, **11**, 503–505 (1992).
9. Grambsch, P. M. and O'Brien, P. C. 'The effects of transformations and preliminary tests for non-linearity in regression', *Statistics in Medicine*, **10**, 697–709 (1991).

10. Blair, R. C. 'New critical values for the generalized t and generalized rank-sum procedures', *Communications in Statistics – Simulation and Computation*, **20**, 981–994 (1991).
11. Blair, R. C. and Morel, J. G. 'On the use of the generalized t and generalized rank-sum statistics in medical research', *Statistics in Medicine*, **11**, 491–501 (1992).
12. Blair, R. C. and Sawilowsky, S. 'Comparison of two tests useful in situations where treatment is expected to increase variability relative to controls', *Statistics in Medicine* (in press).
13. Gastwirth, J. L. 'On robust procedures', *Journal of the American Statistical Association*, **61**, 929–948 (1966).
14. Gastwirth, J. L. 'The use of maximin efficiency robust tests in combining contingency tables and survival analysis', *Journal of the American Statistical Association*, **80**, 380–384 (1985).
15. Gastwirth, J. L. and Mahmoud, H. 'An efficiency robust nonparametric test for scale change for data from a gamma distribution', *Technometrics*, **28**, 81–84 (1986).
16. Lachin, J. M. and Wei, L. J. 'Estimators and tests in the analysis of multiple nonindependent 2×2 tables with partially missing observations', *Biometrics*, **44**, 513–528 (1988).
17. Burnett, R. T. and Willan, A. R. 'Linear rank tests for randomized block designs', *Communications in Statistics – Theory and Methods*, **17**, 2455–2460 (1988).
18. Burnett, R., Krewski, D. and Bleuer, S. 'Efficiency robust score tests for rodent tumorigenicity experiments', *Biometrika*, **76**, 317–324 (1989).
19. Zucker, D. M. and Lakatos, E. 'Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment', *Biometrika*, **77**, 853–864 (1990).
20. Brownie, C., Boos, D. D. and Hughes-Oliver, J. 'Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls', *Biometrics*, **46**, 259–266 (1990).
21. Randles, R. H. and Wolfe, D. A. *Introduction to the Theory of Nonparametric Statistics*, Wiley, New York, 1979.
22. Moses, L. E. 'Rank tests of dispersion', *Annals of Mathematical Statistics*, **34**, 973–983 (1963).
23. Hájek, J. 'Miscellaneous problems of rank test theory', *in* Puri, M. L. (ed.), *Nonparametric Techniques in Statistical Inference*, Cambridge University, Cambridge, 1970.
24. Lehmann, E.L. *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1975
25. Duran, B. S. and Mielke, P. W. Jr. 'Robustness of sum of squared ranks tests', *Journal of the American Statistical Association*, **63**, 338–344 (1968).
26. IMSL. *FORTRAN subroutines for statistical analysis*, IMSL, Houston, TX, 1987.
27. Duran, B. S., Tsai, W. S. and Lewis, T. O. 'A class of location-scale nonparametric tests', *Biometrika*, **63**, 173–176 (1976).
28. Lepage, Y. 'A class of nonparametric tests for location and scale parameters', *Communications in Statistics – Theory and Methods*, **A6**, 649–659 (1977).
29. Goria, M. N. 'A survey of two-sample location-scale problem, asymptotic relative efficiencies of some rank tests', *Statistica Neerlandica*, **36**, 3–13 (1982).
30. Lepage, Y. 'A table for a combined Wilcoxon Ansari–Bradley statistic', *Biometrika*, **60**, 113–116 (1973).
31. StatXact. *StatXact User Manual*, Version 2, CYTEL Software, Cambridge, MA, 1991.
32. Blair, R. C. and Morel, J. G. 'Rejoinder', *Statistics in Medicine*, **11**, 507–509 (1992).
33. Chernoff, H. and Savage, I. R. 'Asymptotic normality and efficiency of certain nonparametric test statistics', *Annals of Mathematical Statistics*, **29**, 972–994 (1958).