Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods

Article *in* Communication in Statistics- Simulation and Computation · July 2013

DOI: 10.1080/03610918.2012.665546

1 author:

Marco Marozzi
Università Ca' Foscari Venezia
73 PUBLICATIONS **715** CITATIONS

Communications in Statistics - Simulation and Computation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lssp20

Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods

Marco Marozzi [a]

[a] Dipartimento di Economia e Statistica, Università della Calabria, Rende, (CS), Italy

Version of record first published: 10 Jan 2013.

To cite this article: Marco Marozzi (2013): Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods, Communications in Statistics - Simulation and Computation, 42:6, 1298-1317

To link to this article: http://dx.doi.org/10.1080/03610918.2012.665546

Communications in Statistics—Simulation and Computation®, 42: 1298–1317, 2013

ISSN: 0361-0918 print / 1532-4141 online
DOI: 10.1080/03610918.2012.665546

Nonparametric Simultaneous Tests for Location and Scale Testing: A Comparison of Several Methods

MARCO MAROZZI

Dipartimento di Economia e Statistica, Università della Calabria, Rende, (CS), Italy

The two-sample location-scale problem arises in many situations like climate dynamics, bioinformatics, medicine, and finance. To address this problem, the nonparametric approach is considered because in practice, the normal assumption is often not fulfilled or the observations are too few to rely on the central limit theorem, and moreover outliers, heavy tails and skewness may be possible. In these situations, a nonparametric test is generally more robust and powerful than a parametric test. Various nonparametric tests have been proposed for the two-sample location-scale problem. In particular, we consider tests due to Lepage, Cucconi, Podgor-Gastwirth, Neuhäuser, Zhang, and Murakami. So far all these tests have not been compared. Moreover, for the Neuhäuser test and the Murakami test, the power has not been studied in detail. It is the aim of the article to review and compare these tests for the jointly detection of location and scale changes by means of a very detailed simulation study. It is shown that both the Podgor–Gastwirth test and the computationally simpler Cucconi test are preferable. Two actual examples within the medical context are discussed.

Keywords Nonparametric Testing; Rank Testing; Robustness; The Location-Scale Problem.

Mathematics Subject Classification 62G09, 62G10, 62G35, 62P10.

1. Introduction

In climate dynamics, the two-sample location-scale problem arises very often when studying spatial and temporal variations in streamflow, see Zhang et al. (2009) and Yang et al. (2009), or monsoon circulation (Kwoon et al., 2007). The two-sample location-scale problem arises also in bioinformatics when comparing two groups in order to detect differentially expressed genes (Neuhäuser and Senske, 2004). In many biomedical situations, the treatment can change location and scale simultaneously, see, for example, Muccioli et al. (1996) and Rice et al. (2000). The two-sample location-scale problem arises also in detecting bull and bear stock markets (Lunde and Timmermann, 2004).

Equality of variances is a characteristic of the null hypothesis when homogeneous subjects are randomly assigned to the treatment and the placebo (Brownie et al., 1990), and then the presence of difference in variability may indicate that the treatment produces an

Received December 13, 2010; Accepted February 6, 2012
Address Correspondence to Marco Marozzi, Dipartimento di Economia e Statistica, Università della Calabria, Via Bucci 0c, 87036 Rende (CS), Italy; E-mail: marozzi@unical.it

effect. In such situations, a test for jointly detecting location and/or scale changes is more appropriate than a test for the Beherens–Fisher problem (that is to test the null hypothesis of identical locations without assuming equal variances). In practice, the normal assumption is often not fulfilled or the observations are too few to rely on the central limit theorem. This is the case of gene expression data from microarrays that are often non-normally distributed even after some preprocessing (see, among others, Giles and Kipling, 2003). Moreover, in microarrays the sample sizes are usually very small as happens also in other biomedical studies (see the review of Ludbrook and Dudley, 1998). According to many authors, these are arguments in favor of nonparametric tests that have the advantage of not assuming a specific distribution. Another argument relies on the presence of outliers which is common in microarray data, that are often non-normal even after outliers have been removed (see, among others, Magusin, 2003). Heavy-tailed data are common also in psychological studies (see the review of Micceri, 1989), in finance (Hall and Yao, 2003) and when analyzing emission level data (Freidlin et al., 2002). Also skewed data are common in many situations, for example, for ecological (Fletcher et al., 2005) and duration data (Lawless, 2003). In such situations, a nonparametric test is generally more robust and powerful than a parametric test (as the tests proposed in Marozzi, 2004a, 2004b, 2007 for the location problem).

The most familiar nonparametric test for the two-sample location-scale problem is the test of Lepage (1971) which is a combination of the Wilcoxon test for detecting location changes and the Ansari–Bradley test for detecting scale changes. In this article, we also consider several other nonparametric tests due to Podgor and Gastwirth (1994), Neuhäuser (2000), Murakami (2007), Cucconi (1968), and Zhang (2006). It should be noted that Lepage (1971) proposed two other combinations of the standardized Wilcoxon and Ansari–Bradley statistics, which are not considered here because Marozzi (2008) found that the corresponding tests are less powerful than the L test. Podgor and Gastwirth (1994) extended results of O'Brien (1988) and proposed a class of nonparametric tests for the location-scale problem. This approach can be recast as a quadratic combination of a rank test for location and a rank test for scale which generates Lepage type tests. Marozzi (2009) compared several Podgor and Gastwirth tests and found the one based on the Wilcoxon test and the Mood squared rank test to be generally preferable. For this reason, this test is considered here. Each of Neuhäuser (2000) and Murakami (2007) proposed a modification of the Lepage test and concluded that their modified tests should be preferred to the classical one. Nevertheless in our opinion their simulation studies were too limited to draw a clear conclusion. Cucconi (1968) proposed a rank test that is earlier than the Lepage one. This rank test is of interest because contrary to the other location-scale tests it is not a combination of a test for location and a test for scale. It is based on squared ranks and squared contrary-ranks. Even if this test is not much familiar, a detailed simulation study by Marozzi (2009) showed that its size is very close to α and is more powerful than the Lepage test. Zhang (2006) proposed a new approach for constructing nonparametric tests for the general two-sample problem which generates as particular cases traditional tests like the Kolmogorov–Smirnov, Cramer–Von Mises, and Anderson–Darling tests, as well as new powerful tests based on the likelihood ratio. Zhang recommended the test that is the analog of the Cramer–Von Mises test because it has the simplest form within the proposed class. This test is considered in our article. It should be emphasized that so far, all these tests have not been compared. It is the aim of the article to review and compare these tests for the jointly detection of location and scale changes by means of a very detailed simulation study. In section 2, we describe the tests. Size and power of the tests are studied in section

3. Section 4 illustrates two practical applications and section 5 concludes the article with some remarks.

2. Description of the Tests

Let m and n be the sizes of two random samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, $N = m + n$. We assume that the elements within each sample are independent and identically distributed, and we assume independence between the two samples. Let F_1 and F_2 be the continuous distribution functions underlying the two samples. The null hypothesis to be tested is that both samples come from the same population

$$H_0 : F_1(t) = F_2(t) \text{ for all } t \in (-\infty, \infty).$$

We wish to test H_0 against the location-scale alternative

$$H_1 : F_2(t) = F_1\left(\frac{t-\mu}{\sigma}\right)$$

with

$$\mu \neq 0 \quad \text{or} \quad \sigma \neq 1, \quad \mu \in (-\infty, \infty), \quad \sigma \in (0, \infty). \tag{1}$$

Note that $\mu = \mu_1 - \mu_2$ is the possible location change between F_1 and F_2, and that $\sigma = \sigma_1/\sigma_2$ is the possible scale change where $\mu_1(\mu_2)$ and $\sigma_1(\sigma_2)$ denote the location and the scale parameters of F_1 and F_2.

2.1. *The Lepage L Test for the Location-Scale Problem*

The test of Lepage (1971) is a combination of the Wilcoxon test for location and the Ansari–Bradley test for scale. The Wilcoxon statistic is defined as

$$W = \sum_{j=1}^{N} j V_j,$$

where $V_j = 1$ when the jth smallest of the N observations is an X and $V_j = 0$ otherwise. The Ansari–Bradley statistic is defined as

$$AB = \frac{1}{2}m(N+1) - \sum_{j=1}^{N} \left| j - \frac{1}{2}(N+1) \right| V_j.$$

The Lepage statistic is defined as

$$L = \frac{(W - E_0(W))^2}{VAR_0(W)} + \frac{(AB - E_0(AB))^2}{VAR_0(AB)}.$$

Under H_0, we have

$$E_0(W) = m(N+1)/2,$$
$$VAR_0(W) = mn(N+1)/12,$$

$$E_0(AB) = \begin{cases} m(N+2)/4 & \text{if } N \text{ is even} \\ m(N+1)^2/(4N) & \text{if } N \text{ is odd} \end{cases},$$

$$VAR_0(AB) = \begin{cases} mn(N+2)(N-2)/(48(N-1)) & \text{if } N \text{ is even} \\ mn(N+1)(3+N^2)/(48N^2) & \text{if } N \text{ is odd} \end{cases}.$$

For the formulae to be used in the presence of ties see Hollander and Wolfe (1999, p. 109 and 146). It should be noted that Lepage (1971) proposed two other combinations of the standardized Wilcoxon and Ansari–Bradley statistics, which are not considered here because Marozzi (2008) found that the corresponding tests are less powerful than the L test.

2.2. *The Podgor-Gastwirth PG Test for the Location-Scale Problem*

Let I_i, $i = 1, \ldots, N$ be a group indicator so that $I_i = 1$ when the ith element of the pooled sample is an X, $I_i = 0$ otherwise. The O'Brien (1988) and Podgor and Gastwirth (1994) PG test statistic is the F statistic with 2 and $N - 3$ df computed by regressing group indicators I_i on the ranks S_i and the squared ranks S_i^2 of the observations in the pooled sample

$$PG = \frac{(\boldsymbol{b}^T \boldsymbol{S}^T I - m^2/N)/2}{(m - \boldsymbol{b}^T \boldsymbol{S}^T \boldsymbol{I})/(N-3)},$$

where T denotes the transpose operator, $\boldsymbol{b}$ is the 3×1 column vector of the OLS estimate of the intercept term and the regression coefficients, $\boldsymbol{S}$ is a $N \times 3$ matrix with the first column of 1s, the second column of $S_1, \ldots, S_N$ and the third column of $S_1^2, \ldots, S_N^2$, $\boldsymbol{I}$ is the $N \times 1$ column of the group indicators $I_1, \ldots, I_N$. Podgor and Gastwirth (1994) showed that asymptotically the PG test can be recast as a quadratic combination of the Wilcoxon rank test for location and the Mood squared rank test for scale.

2.3. *The Neuhäuser NEU Test for the Location-Scale Problem*

Neuhäuser (2000) proposed a modification of the Lepage test by replacing the W statistic with the Baumgartner et al. (1998) B statistic. His modified Lepage test statistic is defined as

$$NEU = \frac{(B - E_0(B))^2}{VAR_0(B)} + \frac{(AB - E_0(AB))^2}{VAR_0(AB)},$$

where $B = (B_X + B_Y)/2$ and

$$B_X = \frac{1}{m}\sum_{i=1}^{m}\left(G_i - \frac{N}{m}i\right)^2 \Bigg/ \left(\frac{i}{m+1}\left(1 - \frac{i}{m+1}\right)\frac{nN}{m}\right),$$

$$B_Y = \frac{1}{n}\sum_{j=1}^{n}\left(H_j - \frac{N}{n}j\right)^2 \Bigg/ \left(\frac{j}{n+1}\left(1 - \frac{j}{n+1}\right)\frac{mN}{n}\right),$$

G_i, $i = 1, \ldots, m$ and H_j, $j = 1, \ldots, n$ are the ranks in increasing order of each X_i and Y_j in the pooled sample. Neuhäuser (2000) proposed this modification of the Lepage test because he noted that when there is a constant difference in means, but an increasing difference in variability the loss of power is more severe for the W test than for the B test. After performing a simulation study which considers normal, exponential and uniform

distributions with $(m,n) = (10,10)$ Neuhäuser concluded that his test is more powerful than the L test. Nevertheless, it should be noted that the difference in power is very small and that the simulation study is limited to only one sample size setting and only three distributions.

2.4. *The Murakami MUR Test for the Location-Scale Problem*

Murakami (2007) proposed a modification of the *NEU* test, based on the following statistic

$$MUR = A^T K^{-1} A,$$

where

$$A = \begin{pmatrix} \tilde{B} - E_0(\tilde{B}) \\ M - E_0(M) \end{pmatrix}, \quad K = \begin{pmatrix} VAR_0(\tilde{B}) & COV_0(\tilde{B}, M) \\ COV_0(\tilde{B}, M) & VAR_0(M) \end{pmatrix}, \quad COV_0(\tilde{B}, M)$$

is the covariance under H_0 between $\tilde{B}$ and M, M is the Mood statistic and $\tilde{B}$ is a modification of B which uses the exact mean and variance of R_i and H_j:

$$\tilde{B} = (\tilde{B}_X + \tilde{B}_Y)/2$$

where

$$\tilde{B}_X = \frac{1}{m}\sum_{i=1}^{m}\left(G_i - \frac{N+1}{m+1}i\right)^2 \Bigg/ \left(\frac{i}{m+1}\left(1 - \frac{i}{m+1}\right)\frac{n\,(N+1)}{m+2}\right),$$

$$\tilde{B}_Y = \frac{1}{n}\sum_{j=1}^{n}\left(H_j - \frac{N+1}{n+1}j\right)^2 \Bigg/ \left(\frac{j}{n+1}\left(1 - \frac{j}{n+1}\right)\frac{m\,(N+1)}{n+2}\right).$$

After performing a simulation study which considers normal, exponential and Gumbel distributions with $(m,n) = (10,10)$ and $(10,5)$ Murakami concluded that his test is more powerful than the *NEU* test when $m \neq n$. Nevertheless, it should be noted that only the $(10,5)$ case has been considered when $m \neq n$ and that the simulation study is limited to only three distributions.

2.5. *The Cucconi C Test for the Location-Scale Problem*

The Cucconi test statistic is defined as

$$C = \frac{U^2 + V^2 - 2\rho UV}{2(1-\rho^2)},$$

where

$$U = \frac{6\sum_{j=1}^{n} S_j^2 - n(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)(8N+11)/5}},$$

$$V = \frac{6\sum_{j=1}^{n} (N+1-S_j)^2 - n(N+1)(2N+1)}{\sqrt{mn(N+1)(2N+1)(8N+11)/5}}, \quad \rho = \frac{2(N^2-4)}{(2N+1)(8N+11)} - 1$$

and S_j is the rank of Y_j in the pooled sample. Under H_0, $E_0(U) = E_0(V) = 0$ and $VAR_0(U) = VAR_0(V) = 1$. Note that $CORR_0(U,V) = COVAR_0(U,V) = \frac{2(N^2-4)}{(2N+1)(8N+11)} - 1 = \rho$. Under

H_0, (U,V) is centered on (0,0), whereas it is not under H_1, and therefore large values of C speak against H_0. Marozzi (2009) computed a table of critical values that was missing in the literature. It should be noted that without ties it makes no difference whether U and V are computed based on the data of the first or the second sample, whereas in the presence of ties, one obtains very slightly different values for C (Neuhäuser, 2010). In this article, we computed the C statistic as in Marozzi (2009) by always considering the second sample.

2.6. *The Zhang ZH Test for the General Problem*

We also consider a test for the general two-sample problem

$$H_0 : F_1(t) = F_2(t) \text{ for all } t \in (-\infty, \infty) \text{ against } H_0 : F_1(t) \neq F_2(t) \text{ for some } t \in (-\infty, \infty). \tag{2}$$

The Zhang (2006) ZH test statistic is defined as

$$ZH = \frac{1}{N}\left(\sum_{i=1}^{m} \ln\left(\frac{m}{i-0.5}-1\right)\ln\left(\frac{N}{G_i-0.5}-1\right) + \sum_{j=1}^{n} \ln\left(\frac{n}{j-0.5}-1\right)\ln\left(\frac{N}{H_j-0.5}-1\right)\right).$$

It is important to underline that the ZH test is, within the class of tests proposed by Zhang, the analog of the Cramer–Von Mises test and that it is location, scale and shape sensitive. It is as powerful as the traditional Kolmogorov–Smirnov, Cramer–Von Mises, and Anderson–Darling tests for detecting changes in location, and more powerful for detecting changes in scale or shape. It should be noted that the B and $\tilde{B}$ statistics, which use the Euclidean distance between the empirical distribution functions weighted by its variance, are similar to the Anderson–Darling one. Note that small values of ZH speak against H_0.

3. Size and Power Comparison

To derive the theory of locally and globally optimum nonparametric tests, you have to assume to know the distribution underlying the samples, or at least its type (Hajek et al., 1998). Therefore, it is very difficult to derive theoretically optimality properties for nonparametric tests with completely unknown distributions of the samples. In this section, approximations of size and power of the tests are obtained via Monte Carlo simulations with 10000 replicates. The nominal significance level for rejecting H_0 is 0.05.

The null distributions of L, NEU, MUR, PG, C, and ZH statistics can be obtained by computing all possible values of the statistics in all $N!/m!/n!$ possible assignments of the N observations with m observations acting as Xs and n observations acting as Ys which are equally likely under H_0. Because $N!/m!/n!$ increases very rapidly with m and n, it is not computationally friendly to compute the full distribution when m and n are large. For practical purpose, we do not need the exact p-value of the test to draw inference on the null-hypothesis, it is enough to estimate the p-value via Monte Carlo simulations.

Distributions with different properties have been used to simulate the data

(i) bimodal obtained as a mixture of a $N(-1.5,1)$ with probability 0.5 and a $N(1.5,1)$ with 0.5;
(ii) standard normal $N(0,1)$;

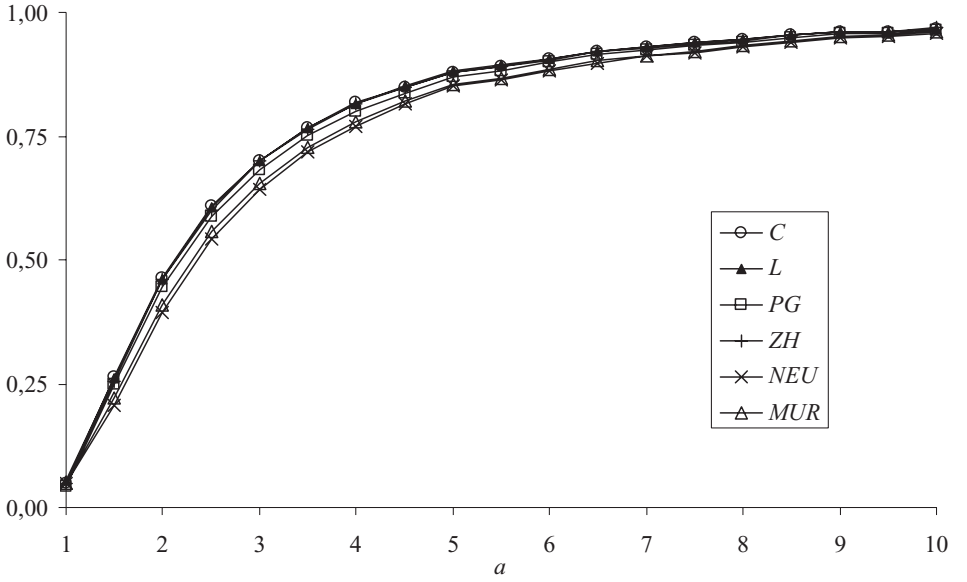

Figure 1. Power estimates with $\alpha = 0.05$ and $(m,n) = (10,5)$ based on two samples uniformly distributed over the intervals (0,1) and (0,a) with $a \geq 1$.

(iii) Laplace double exponential with scale parameter of $1/\sqrt{2}$;
(iv) mixture of a $N(0,1)$ with probability 0.9 and a $N(1,100)$ with 0.1 (mixture10 for short);
(v) mixture of a $N(0,1)$ with probability 0.7 and a $N(1,100)$ with 0.3 (mixture30 for short);
(vi) Student's t with 2 df;
(vii) Student's t with 1 df (Cauchy).

Distribution (i) has light tails, while distributions (iii)–(vii) have moderately heavy to very heavy tails. Distributions (vi) and (vii) have infinite second moment, and distribution (vii) has undefined first moment. As sample size settings, we consider the $m = n$ case with $(m,n) = (10,10)$, (20,20), and (30,30), the $m < n$ case with $(m,n) = (10,30)$, (10,40), and the $m > n$ case with $(m,n) = (30,10)$ and (40,10). Tables 1–4 present size values of the tests (first column of inner cells) and power values for six combinations of (μ,σ) and for a selection of four (m,n) settings. As additional cases we consider

(viii) uniform distribution over the interval (0,a) with $a \geq 1$;
(ix) exponential distribution with parameter $\lambda \geq 1$,

for $(m,n) = (10,10)$ and (10,5). In both cases, the modification of the parameter changes both location and scale allowing the graphical presentation of power results in Figs. 1 to 4.

We underline that p-values of the *PG* test have been always computed exactly, p-values of the *ZH* test have been always approximated by considering a random sample of 1000 permutations (2000 under H_0). For $(m,n) = (10,5)$ and (10,10); p-values of *L*, *NEU*, *MUR*, and *C* have been computed exactly. In the remaining cases, p-values have been approximated by considering a random sample of 1000 permutations (2000 under H_0) for *NEU* and *MUR* and of 1000000 for *L* and *C*.

Table 1
Power estimates with $\alpha = 0.05$ and $(m,n) = (10,10)$

			Normal				
μ	0	0	1	2	1	1	1
σ	1	2	2	2	1	3	5
C	0.049	0.279	0.409	0.763	0.403	0.627	0.870
L	0.049	0.247	0.388	0.750	0.405	0.568	0.816
PG	0.050	0.282	0.413	0.766	0.406	0.631	0.873
ZH	0.053	0.107	0.330	0.782	0.528	0.358	0.528
NEU	0.048	0.212	0.394	0.792	0.462	0.543	0.778
MUR	0.049	0.230	0.389	0.783	0.462	0.567	0.821
			Bimodal				
μ	0	0	2.5	4	1.5	1.5	1.5
σ	1	1.5	1.5	1.5	1	2.5	4
C	0.047	0.195	0.444	0.843	0.245	0.649	0.908
L	0.047	0.173	0.457	0.850	0.258	0.586	0.853
PG	0.048	0.198	0.448	0.845	0.248	0.654	0.911
ZH	0.047	0.082	0.597	0.943	0.409	0.331	0.594
NEU	0.047	0.148	0.534	0.898	0.312	0.553	0.827
MUR	0.047	0.158	0.511	0.891	0.303	0.585	0.866
			Laplace				
μ	0	0	1	2	1	1	1
σ	1	2	2	2	1	3	5
C	0.055	0.179	0.469	0.846	0.561	0.559	0.767
L	0.055	0.171	0.450	0.835	0.553	0.527	0.719
PG	0.055	0.182	0.473	0.848	0.564	0.563	0.770
ZH	0.052	0.082	0.428	0.859	0.637	0.382	0.436
NEU	0.054	0.145	0.479	0.872	0.617	0.517	0.688
MUR	0.054	0.151	0.477	0.872	0.620	0.521	0.709
			Mixture10				
μ	0	0	1	2	1	1	1
σ	1	2	2	2	1	3	5
C	0.050	0.237	0.426	0.897	0.319	0.492	0.861
L	0.050	0.224	0.409	0.889	0.309	0.481	0.865
PG	0.051	0.240	0.430	0.897	0.322	0.496	0.864
ZH	0.052	0.104	0.383	0.922	0.390	0.269	0.562
NEU	0.053	0.195	0.441	0.940	0.360	0.452	0.839
MUR	0.052	0.199	0.440	0.941	0.364	0.451	0.836
			Mixture30				
μ	0	0	3.6	12	1.3	1.3	1.3
σ	1	3	3	3	1	6	18
C	0.051	0.226	0.491	0.911	0.265	0.485	0.897
L	0.052	0.241	0.486	0.904	0.255	0.506	0.890
PG	0.053	0.228	0.493	0.912	0.268	0.490	0.899
ZH	0.054	0.104	0.473	0.923	0.286	0.225	0.549
NEU	0.052	0.209	0.546	0.955	0.298	0.456	0.856
MUR	0.052	0.201	0.544	0.955	0.302	0.440	0.858

(Continued on next page)

Table 1
Power estimates with $\alpha = 0.05$ and $(m,n) = (10,10)$ *(Continued)*

			Student				
μ	0	0	2	4.4	1	1	1
σ	1	2.4	2.4	2.4	1	3.6	9
C	0.049	0.262	0.530	0.902	0.258	0.514	0.890
L	0.049	0.251	0.508	0.894	0.249	0.491	0.864
PG	0.050	0.265	0.533	0.904	0.260	0.518	0.892
ZH	0.052	0.108	0.473	0.912	0.312	0.253	0.562
NEU	0.050	0.215	0.538	0.929	0.291	0.448	0.830
MUR	0.051	0.222	0.536	0.930	0.295	0.458	0.848
			Cauchy				
μ	0	0	3	9	1.5	1.5	1.5
σ	1	3	3	3	1	5	15
C	0.053	0.237	0.501	0.872	0.321	0.479	0.836
L	0.055	0.243	0.493	0.866	0.315	0.487	0.848
PG	0.054	0.241	0.504	0.872	0.323	0.483	0.838
ZH	0.056	0.104	0.436	0.898	0.350	0.238	0.508
NEU	0.055	0.206	0.524	0.935	0.366	0.449	0.814
MUR	0.056	0.206	0.520	0.934	0.369	0.439	0.806

All the tests maintain the level .05 well and are robust according with both Conover et al. (1981) and Marozzi (2011) indications to consider a test to be robust if its MESL (maximum estimated significance level) does not exceed .1 and .075 respectively. It is interesting to note that no test has a MESL that exceeds .056. In fact, the MESLs for the

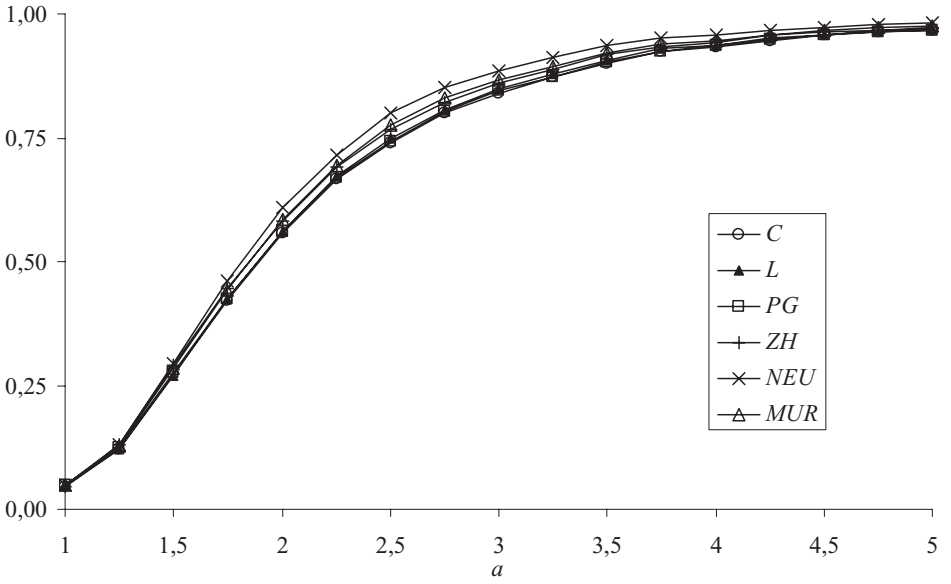

Figure 2. Power estimates with $\alpha = 0.05$ and $(m,n) = (10,10)$ based on two samples uniformly distributed over the intervals (0,1) and $(0,a)$ with $a \geq 1$.

Table 2
Power estimates with $\alpha = 0.05$ and $(m,n) = (10,30)$

			Normal				
μ	0	0	.75	1.5	.75	.75	.75
σ	1	1.5	1.5	1.5	1	2.5	4
C	0.050	0.244	0.477	0.872	0.391	0.791	0.950
L	0.048	0.199	0.436	0.855	0.391	0.697	0.894
PG	0.048	0.238	0.472	0.869	0.384	0.787	0.949
ZH	0.052	0.231	0.521	0.911	0.475	0.798	0.960
NEU	0.051	0.181	0.401	0.842	0.371	0.681	0.896
MUR	0.052	0.200	0.440	0.865	0.423	0.738	0.930
			Bimodal				
μ	0	0	2	3.5	1	1	1
σ	1	1.5	1.5	1.5	1	1.5	3
C	0.051	0.409	0.618	0.923	0.215	0.455	0.954
L	0.052	0.324	0.559	0.911	0.215	0.376	0.888
PG	0.048	0.404	0.614	0.920	0.209	0.448	0.953
ZH	0.051	0.396	0.780	0.982	0.339	0.520	0.973
NEU	0.051	0.304	0.590	0.929	0.211	0.368	0.901
MUR	0.052	0.347	0.630	0.943	0.245	0.416	0.938
			Laplace				
μ	0	0	.75	1.5	.75	.75	.75
σ	1	1.5	1.5	1.5	1	2.5	4
C	0.048	0.169	0.551	0.951	0.547	0.724	0.899
L	0.051	0.143	0.535	0.947	0.551	0.670	0.842
PG	0.047	0.165	0.545	0.950	0.542	0.720	0.895
ZH	0.048	0.152	0.536	0.939	0.561	0.683	0.874
NEU	0.046	0.131	0.488	0.933	0.532	0.621	0.819
MUR	0.046	0.139	0.523	0.941	0.581	0.662	0.858
			Mixture10				
μ	0	0	1	2.5	.75	.75	.75
σ	1	2	2	2	1	2.2	6
C	0.052	0.373	0.544	0.931	0.290	0.528	0.947
L	0.048	0.328	0.510	0.928	0.283	0.482	0.918
PG	0.049	0.367	0.538	0.929	0.286	0.521	0.945
ZH	0.048	0.264	0.455	0.911	0.312	0.407	0.843
NEU	0.051	0.295	0.464	0.918	0.270	0.434	0.906
MUR	0.052	0.310	0.493	0.930	0.321	0.463	0.923
			Mixture30				
μ	0	0	1.8	4	1	1	1
σ	1	2.2	2.2	2.2	1	3	8
C	0.052	0.268	0.507	0.873	0.247	0.470	0.897
L	0.052	0.252	0.503	0.885	0.243	0.455	0.880
PG	0.050	0.263	0.501	0.870	0.242	0.464	0.894
ZH	0.052	0.206	0.451	0.858	0.240	0.367	0.783
NEU	0.050	0.219	0.458	0.888	0.238	0.395	0.851
MUR	0.050	0.219	0.481	0.895	0.280	0.403	0.854

(Continued on next page)

Table 2
Power estimates with $\alpha = 0.05$ and $(m,n) = (10,30)$ *(Continued)*

			Student				
μ	0	0	1.1	2.4	.8	.8	.8
σ	1	1.8	1.8	1.8	1	2.2	4.5
C	0.053	0.267	0.489	0.900	0.264	0.506	0.895
L	0.053	0.231	0.460	0.894	0.262	0.461	0.841
PG	0.051	0.262	0.482	0.898	0.258	0.501	0.893
ZH	0.053	0.207	0.446	0.873	0.273	0.425	0.819
NEU	0.052	0.204	0.415	0.871	0.246	0.406	0.821
MUR	0.054	0.216	0.448	0.887	0.289	0.442	0.854
			Cauchy				
μ	0	0	1.5	4	1	1	1
σ	1	2	2	2	1	3	8
C	0.051	0.230	0.454	0.910	0.247	0.520	0.927
L	0.051	0.217	0.450	0.922	0.249	0.496	0.904
PG	0.049	0.226	0.448	0.908	0.241	0.514	0.925
ZH	0.051	0.164	0.368	0.869	0.238	0.374	0.797
NEU	0.052	0.183	0.400	0.912	0.241	0.436	0.880
MUR	0.051	0.183	0.420	0.915	0.283	0.451	0.892

first round of simulations are .055, .055, .055, .056, .055, and .056 respectively for the *C*, *L*, *PG*, *ZH*, *NEU*, and *MUR* tests. The MESLs for the second round of simulations are .052, .055, .052, .052, .052, and .050 respectively.

The first round of simulations show that, in most cases, for a class of distributions considered here, the *C* and *PG* tests are more powerful than the other tests, with the *C*

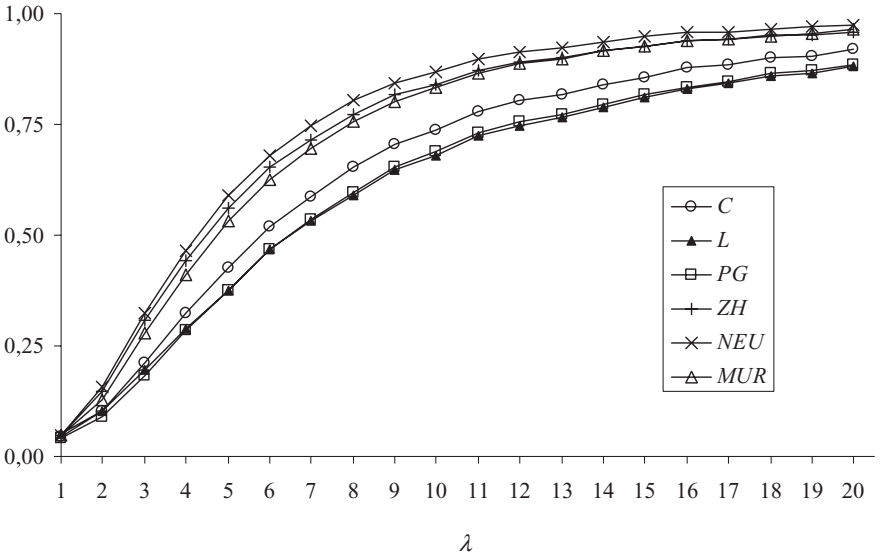

Figure 3. Power estimates with $\alpha = 0.05$ and $(m,n) = (10,5)$ based on two samples exponentially distributed with 1 and $\lambda \geq 1$ parameter.

Table 3
Power estimates with $\alpha = 0.05$ and $(m,n) = (30,10)$.

			Normal				
μ	0	0	1	1.8	.75	.75	.75
σ	1	1.8	1.8	1.8	1	2.5	4
C	0.049	0.234	0.440	0.873	0.376	0.607	0.942
L	0.050	0.238	0.444	0.873	0.374	0.576	0.915
PG	0.048	0.225	0.428	0.865	0.368	0.596	0.939
ZH	0.050	0.024	0.357	0.881	0.463	0.182	0.362
NEU	0.049	0.213	0.575	0.942	0.460	0.634	0.920
MUR	0.049	0.232	0.534	0.925	0.412	0.633	0.939
			Bimodal				
μ	0	0	2	3.5	1	1	1
σ	1	1.5	1.5	1.5	1	1.75	3
C	0.049	0.217	0.434	0.924	0.216	0.384	0.962
L	0.051	0.213	0.502	0.940	0.216	0.395	0.933
PG	0.047	0.209	0.423	0.920	0.211	0.373	0.960
ZH	0.052	0.025	0.527	0.977	0.343	0.106	0.351
NEU	0.052	0.194	0.652	0.982	0.290	0.439	0.937
MUR	0.051	0.216	0.574	0.969	0.247	0.422	0.960
			Laplace				
μ	0	0	1	1.8	.75	.75	.75
σ	1	1.8	1.8	1.8	1	2.5	4
C	0.050	0.138	0.531	0.933	0.551	0.472	0.789
L	0.050	0.149	0.511	0.929	0.554	0.447	0.753
PG	0.047	0.134	0.522	0.930	0.544	0.461	0.780
ZH	0.051	0.023	0.504	0.938	0.566	0.243	0.290
NEU	0.050	0.137	0.661	0.972	0.632	0.543	0.785
MUR	0.051	0.140	0.631	0.965	0.584	0.534	0.800
			Mixture10				
μ	0	0	1.5	3	.75	.75	.75
σ	1	2	2	2	1	3	6
C	0.047	0.225	0.519	0.940	0.294	0.508	0.808
L	0.049	0.248	0.515	0.943	0.289	0.551	0.871
PG	0.046	0.218	0.511	0.939	0.287	0.499	0.803
ZH	0.048	0.036	0.467	0.941	0.317	0.163	0.387
NEU	0.050	0.223	0.639	0.974	0.363	0.570	0.871
MUR	0.050	0.232	0.612	0.967	0.323	0.547	0.845
			Mixture30				
μ	0	0	3	7	1	1	1
σ	1	2.5	2.5	2.5	1	4.5	11
C	0.048	0.188	0.503	0.886	0.249	0.440	0.857
L	0.049	0.246	0.501	0.898	0.249	0.529	0.896
PG	0.046	0.182	0.496	0.881	0.244	0.431	0.851
ZH	0.050	0.036	0.485	0.908	0.239	0.115	0.336
NEU	0.049	0.221	0.657	0.963	0.316	0.522	0.889
MUR	0.049	0.205	0.632	0.953	0.283	0.473	0.866

(Continued on next page)

Table 3
Power estimates with $\alpha = 0.05$ and $(m,n) = (30,10)$. *(Continued)*

				Student			
μ	0	0	1.7	3.2	.8	.8	.8
σ	1	2.2	2.2	2.2	1	3	6
C	0.053	0.241	0.525	0.894	0.276	0.493	0.894
L	0.054	0.274	0.509	0.884	0.273	0.515	0.909
PG	0.051	0.236	0.516	0.890	0.270	0.483	0.889
ZH	0.054	0.034	0.450	0.893	0.285	0.127	0.363
NEU	0.054	0.247	0.639	0.952	0.337	0.528	0.904
MUR	0.054	0.247	0.616	0.942	0.299	0.511	0.900
				Cauchy			
μ	0	0	2.5	5.5	1	1	1
σ	1	2.5	2.5	2.5	1	4	10
C	0.051	0.201	0.482	0.871	0.242	0.442	0.838
L	0.049	0.246	0.462	0.872	0.243	0.506	0.893
PG	0.049	0.194	0.473	0.867	0.236	0.434	0.834
ZH	0.046	0.036	0.434	0.890	0.234	0.124	0.378
NEU	0.051	0.225	0.613	0.952	0.311	0.507	0.890
MUR	0.049	0.212	0.591	0.941	0.277	0.475	0.864

test to be preferred because it is computationally simpler. The *NEU* and *MUR* tests behave quite similarly but are slightly less powerful than the *C* and *PG* tests. The power of the *ZH* test is often poor (with some exceptions) and this test seems to be the worst one. However, it should be noted that this result is not surprising since this is the only test designed

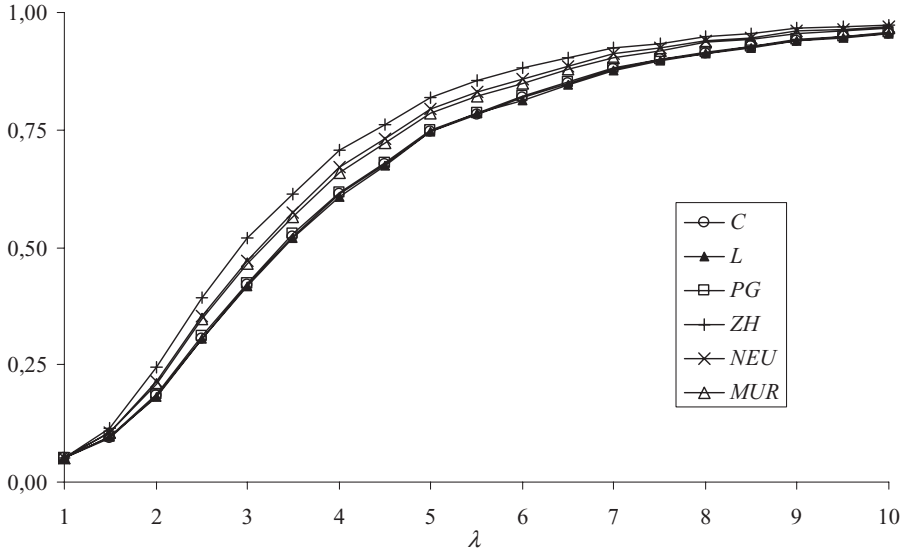

Figure 4. Power estimates with $\alpha = 0.05$ and $(m,n) = (10,10)$ based on two samples exponentially distributed with 1 and $\lambda \geq 1$ parameter.

Table 4
Power estimates with $\alpha = 0.05$ and $(m,n) = (30,30)$

			Normal				
μ	0	0	.5	1	.5	.5	.5
σ	1	1.3	1.3	1.3	1	1.75	2.5
C	0.049	0.168	0.401	0.871	0.364	0.707	0.968
L	0.051	0.142	0.384	0.865	0.367	0.641	0.928
PG	0.050	0.170	0.403	0.872	0.366	0.709	0.968
ZH	0.053	0.115	0.423	0.909	0.446	0.629	0.940
NEU	0.051	0.136	0.412	0.889	0.403	0.652	0.940
MUR	0.049	0.149	0.414	0.887	0.403	0.680	0.959
			Bimodal				
μ	0	0	1.1	2.2	.75	.75	.75
σ	1	1.3	1.3	1.3	1	1.4	2
C	0.051	0.301	0.552	0.933	0.243	0.543	0.971
L	0.052	0.246	0.523	0.933	0.251	0.473	0.920
PG	0.052	0.304	0.554	0.934	0.245	0.545	0.971
ZH	0.050	0.208	0.677	0.991	0.412	0.525	0.960
NEU	0.050	0.240	0.601	0.973	0.306	0.513	0.942
MUR	0.050	0.272	0.603	0.970	0.298	0.536	0.966
			Laplace				
μ	0	0	.5	1	.5	.5	.5
σ	1	1.3	1.3	1.3	1	1.75	2.5
C	0.050	0.114	0.490	0.945	0.511	0.631	0.884
L	0.050	0.109	0.484	0.943	0.513	0.599	0.843
PG	0.051	0.115	0.492	0.945	0.512	0.632	0.885
ZH	0.051	0.079	0.477	0.942	0.527	0.548	0.786
NEU	0.050	0.099	0.514	0.957	0.557	0.607	0.842
MUR	0.049	0.102	0.516	0.957	0.558	0.620	0.860
			Mixture10				
μ	0	0	.75	1.5	.5	.5	.5
σ	1	1.5	1.5	1.5	1	1.8	3
C	0.047	0.243	0.522	0.927	0.264	0.531	0.909
L	0.046	0.232	0.512	0.929	0.264	0.517	0.907
PG	0.047	0.245	0.524	0.928	0.266	0.534	0.909
ZH	0.049	0.139	0.468	0.927	0.290	0.372	0.718
NEU	0.047	0.219	0.539	0.951	0.295	0.510	0.901
MUR	0.047	0.223	0.540	0.951	0.297	0.514	0.900
			Mixture30				
μ	0	0	1.2	2.8	.7	.7	.7
σ	1	1.8	1.8	1.8	1	2.3	4.5
C	0.050	0.249	0.495	0.905	0.242	0.481	0.880
L	0.051	0.267	0.511	0.908	0.240	0.511	0.910
PG	0.050	0.251	0.498	0.905	0.243	0.483	0.881
ZH	0.050	0.134	0.424	0.935	0.231	0.311	0.698
NEU	0.051	0.244	0.539	0.966	0.274	0.489	0.892
MUR	0.050	0.229	0.529	0.966	0.277	0.467	0.869

(Continued on next page)

Table 4
Power estimates with $\alpha = 0.05$ and $(m,n) = (30,30)$ *(Continued)*

				Student			
μ	0	0	.8	1.6	.6	.6	.6
σ	1	1.6	1.6	1.6	1	1.8	2.9
C	0.049	0.272	0.529	0.897	0.289	0.510	0.899
L	0.050	0.257	0.522	0.895	0.287	0.496	0.884
PG	0.050	0.274	0.531	0.898	0.291	0.513	0.900
ZH	0.050	0.154	0.449	0.883	0.299	0.373	0.746
NEU	0.049	0.237	0.529	0.915	0.321	0.486	0.873
MUR	0.049	0.247	0.530	0.915	0.327	0.491	0.880
				Cauchy			
μ	0	0	1.2	2.5	.8	.8	.8
σ	1	1.8	1.8	1.8	1	2.2	4
C	0.051	0.238	0.519	0.899	0.304	0.506	0.873
L	0.051	0.253	0.531	0.903	0.307	0.523	0.888
PG	0.051	0.240	0.521	0.900	0.306	0.508	0.875
ZH	0.048	0.128	0.427	0.893	0.294	0.331	0.662
NEU	0.050	0.230	0.555	0.940	0.348	0.507	0.875
MUR	0.050	0.216	0.547	0.939	0.350	0.492	0.859

for the general two-sample problem (2) whereas the other tests are designed just for the location-scale problem (1). By analyzing separately the results with respect to the $m = n$, $m < n$, and $m > n$ cases as well as the light/normal tailed and heavy tailed distributions cases, we conclude that for $m = n$ the *C* and *PG* tests are preferable both for light/normal and heavy tailed distributions; for $m < n$ the *ZH* test is preferable for light/normal tailed distributions whereas the *C* and *PG* tests for heavy tailed distributions; for $m > n$, the *NEU* test is preferable both for light/normal tailed and heavy tailed distributions. The second round of simulations show that the tests behave very similarly for the uniformly distributed samples. The behavior is more different for the exponential case especially with $(m,n) = (10,5)$ when the *NEU*, *MUR*, and (surprisingly) *ZH* tests are more powerful than the other ones.

Marozzi (2008) concluded that the Lepage test was slightly more powerful than the Cucconi test. The reason was that in the simulations by Marozzi (2008) to compute the *p*-value of the Cucconi test, asymptotic critical values have been used. Since the asymptotic critical values are greater than the exact ones (Marozzi, 2009), the Cucconi test with asymptotic critical values was conservative and less powerful than the Cucconi test with exact critical values. The simulations in the present article and in Marozzi (2009) show that the Cucconi test has a size close to α and is generally more powerful than the Lepage test.

So far, we considered only continuous distributions, but in practice ties may occur. Note that data modifications and rounding can lead to ties. Therefore, it is of interest to study how the tests behave in the presence of ties and to this aim we performed some additional simulations. The usual way of dealing with ties is to assign average ranks (Neuhäuser and Senske, 2004). Data with ties have been generated by simulating from the binomial distribution with parameters 10 and 0.5 and $m = n = 6$. The binomial distribution has been relocated and rescaled so that the mean is 0 and the standard deviation is 1. 10000 Monte

Table 5
Power estimates with $\alpha = 0.05$ and $(m,n) = (6,6)$ in the presence of ties

μ	0	2	0	2	2	3
σ	1	1	3	3	4	3
C	0.045	0.825	0.393	0.528	0.546	0.633
L	0.040	0.812	0.223	0.426	0.463	0.573
PG	0.044	0.819	0.342	0.503	0.528	0.613
ZH	0.066	0.931	0.128	0.427	0.335	0.572
NEU	0.042	0.885	0.209	0.479	0.461	0.626
MUR	0.042	0.886	0.197	0.466	0.437	0.610

Carlo simulations were performed. To accurately compute the *p*-values in the presence of ties, for each simulated data set, we generated the whole conditional permutation distribution of all the test statistics. Therefore, exact *p*-values have been computed. Table 5 shows that the Cucconi test is generally preferable to the other tests in the presence of ties.

This comparison study, as several other ones, see for example, Büning (2002), Büning and Thadewald (2000), and Podgor and Gastwirth (1994), does not find a clear winner for all the situations considered. Some tests have good power, but no single test dominates. Since different tests are more powerful against different alternatives this is not surprising (Podgor and Gastwirth, 1994). According to our results, when a difference in location as well as one in variability is possible, both the Cucconi test and the Podgor–Gastwirth test are recommended because in most cases, for a class of distributions considered here, they are more powerful than the other tests. We prefer the Cucconi test because it is easier to compute. The Neuhauser and Murakami tests perform quite well, but they are slightly less powerful and above all much more difficult to compute because no formulae are available for the moments of some statistics necessary for applying them (therefore, to apply these tests, one has to generate the permutation distribution of the statistics by simulation or complete enumeration, and compute the necessary moments, Marozzi, 2009). Since our results are based on simulations, generalizations require caution.

4. Examples

For the purpose of illustration of the procedures, we use two classic data sets. Zhang (2006) reanalyzed patch clamp data from *L*-type Ca^{++} channels analyzed by Baumgartner et al. (1988). They were interested in whether the native proteins occurring in ventricular myocytes behave the same as recombinant channels. The empirical open probabilities of the channels shall be compared. See Table 6. Zhang (2006) estimated the *p*-value by using

Table 6
Values of the empirical open probabilities for cardiac and recombinant channel

		Cardiac		
.0166	.0247	.0295	.0588	.0642
		Recombinant		
.0178	.0182	.0202	.0393	.0906

Table 7
Platelet counts (per cubic millimeter) of two groups of newborn babies

Case subjects									
120	124	215	90	67	95	190	180	135	399
Control subjects									
12	20	112	32	60	40				

100,000 permutations of the pooled sample. This is very surprising, because $m = n = 5$ and then it suffices to consider $10!/5!/5! = 252$ permutations of the pooled sample to compute the p-values exactly. The exact p-values of the L, NEU, MUR, PG, C, and ZH tests are .89683, .75397, .85714, .92857, .94444, and .87302 respectively. The results speak in favor of the null hypothesis of no difference between the populations underlying the cardiac and the recombinant samples, and are consistent with those of Zhang (2006) and Baumgartner et al. (1988).

As the second example, we analyze the subset of the data obtained by Karpatkin et al. (1981) in their study of the effect of maternal steroid therapy on platelet counts of newborn babies, that was analyzed by Hollander and Wolfe (1999). Table 7 reports platelet counts (per cubic millimeter) of two groups of newborn babies that were delivered vaginally. Mothers of the babies were diagnosed with autoimmune thrombocytopenia purpura (ATP), those of the first group were treated with the corticosteroid prednisone and those of the second group were not. The aim of the treatment is to raise platelet counts of babies to safe levels during the delivery, in order to lower the possibility of intracranial hemorrhage for the baby. A woman with ATP produces antibodies to her own platelets and because of transplacental passage of antiplatelet antibodies during pregnancy, their babies are often born with low platelet counts, with a big concern that a vaginal delivery could cause intracranial hemorrhage.

It is expected that the prednisone by crossing the placenta enters the baby circulation and prevents splenic removal of those baby's platelets which are coated by the mother's antibodies. The primary aim of the study is whether or not predelivery maternal prednisone therapy increases newborn baby platelet counts. Even if the main statistical problem is detecting a possible difference in locations for the treated and not treated distribution populations, there is also some worries that the steroid therapy could rather largely increase the variability in newborn baby platelet counts. Therefore, we regard changes in location and in scale as treatment effects by applying the tests considered before. By considering all the $16!/10!/6! = 8008$ possible permutations of the pooled sample, we obtain the following exact p-values of .00337, .00125, .00150 .00250, .00250, and .00250 for the L, NEU, MUR, PG, C, and ZH tests respectively. Each test gives strong evidence that there are differences in locations or scales (or both) between treated and not treated distribution populations. Note that the significance of the ZH test may be due to shape difference as well as location and/or scale differences. The practitioner should take into account this finding in interpreting the increase in platelet counts resulting from corticosteroid treatment.

In this section, two classical data sets have been considered for the purpose of mere illustration of the procedures described in section 2. The practitioner, when facing a location-scale problem has many procedures at his/her disposal but will not use all these procedures together. The Cucconi and the Podgor–Gastwirth tests in most cases, for a class of distributions considered in section 3, are more powerful than the other tests. Since the Cucconi test

is easier to be computed, this is the test we suggest to use (with special care in the presence of ties).

5. Conclusion

The two-sample location-scale problem arises very often in many contexts like climate dynamics, bioinformatics, medicine, and finance when a test for jointly detecting location and scale changes is more appropriate than a test for the Beherens–Fisher problem. For example, in many biomedical situations, the treatment can change location and scale simultaneously and then the difference in variability indicates that the treatment produces an effect (as in the second actual example discussed in section 4). Since in practice, the normal assumption is often not fulfilled or the observations are too few to rely on the central limit theorem we suggested to address this problem within the nonparametric framework. Other arguments in favor of this framework are the possible presence of outliers, heavy tails and skewness and the fact that the data may be not normally distributed even after some preprocessing (as may happen for example with microarray data). In such situations, a nonparametric test is generally more robust and powerful than a parametric test. The most familiar nonparametric test for the location-scale problem is due to Lepage (1971). In this article, we considered also several other nonparametric tests due to Cucconi (1968), Podgor and Gastwirth (1994), Neuhäuser (2000), Zhang (2006), and Murakami (2007). So far, all these tests have not been compared among themselves. Moreover, the power has not been studied in detail for the Neuhäuser test and the Murakami test. Therefore, these tests have been compared for the jointly detection of location and scale changes by means of a very detailed simulation study.

The comparison study, as several other ones, see for example, Büning (2002), Büning and Thadewald (2000), and Podgor and Gastwirth (1994), did not find a clear winner for all the situations considered. Some tests have good power, but no single test dominates. Since different tests are more powerful against different alternatives this is not surprising (Podgor and Gastwirth 1994). According to our results, when a difference in location as well as one in variability is possible, both the Cucconi test and the Podgor–Gastwirth test are recommended because in most cases, for a class of distributions considered here, they are more powerful than the other tests. We prefer the Cucconi test because it is easier to compute. The Neuhauser test and the Murakami test perform quite well, but they are slightly less powerful and above all much more difficult to compute.

References

Baumgartner, W., Weiss, P., Schindler, H. (1998). A nonparametric test for the general two-sample problem. *Biometrics* 54:1129–1135.

Brownie, C., Boos, D. D., Hughes-Oliver, J. (1990). Modifying the t and ANOVA F tests when treatment is expected to increase variability relative to controls. *Biometrics* 46:259–266.

Büning, H. (2002). Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramèr-Von Mises two-sample tests. *Journal of Applied Statistics* 29(6):907–924.

Büning, H., Thadewald, T. (2000). An adaptive two-sample location-scale test of lepage type for symmetric distributions. *Journal of Statistical Computation and Simulation* 65:287–310.

Conover, W. J., Johnson, M. E., Johnson, M. M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics* 23:351–361.

Cucconi, O. (1968). Un nuovo test non parametrico per il confronto tra due gruppi campionari. *Giornale degli Economisti* XXVII:225–248.

Fletcher, D., MacKenzie, D., Villouta, E. (2005). Modelling skewed data with many zeros: A simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics* 12:45–54.

Freidlin, B., Zheng G., Li, Z., Gastwirth, J. L. (2002). Trend tests for case control studies of genetic markers: Power, sample size and robustness. *Human Heredity* 53:146–152.

Giles, P. J., Kipling, D. (2003). Normality of oligonucleotide microarray data and implications for parametric statistical analyses. *Bioinformatics* 19:2254–2262.

Hall, P., Yao, Q. (2003). Inference in ARCH and GARCH models with heavy-tailed errors. *Econometrica* 71:285–317.

Hajek, J., Sidak, Z., Sen, P. K. (1998). *Theory of Rank Tests.* 2nd ed. New York: Academic Press.

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

Karpatkin, M., Porges, R. F., Karpatkin, S. (1981). Platelet counts in infants of women with autoimmune thrombocytopenia: effects of steroid administration to the mother. *New England Journal of Medicine* 305:936–939.

Kwoon, M., Jhun, J. G., Ha, K. J. (2007). Decadal change in east Asian summer monsoon circulation in the mid-1990s. *Geophysical Research Letters* 34:L21706.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley: Hoboken, NJ.

Lepage, Y. (1971). A combination of Wilcoxon's and Ansari-Bradley's statistics. *Biometrika* 58:213–217.

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

Lunde, A., Timmermann, A. (2004). Duration dependence in stock prices: an analysis of bull and bear markets. *Journal of Business and Economic Statistics* 22:253–273.

Magusin, A. (2003). Complementary techniques of clustering and composite pattern analysis to Saccharomyces cerevisiae gene expression. *Applied Bioinformatics* 2(3 Supplement):S37–S46.

Marozzi, M. (2004a). A bi-aspect nonparametric test for the two-sample location problem. *Computational Statistics and Data Analysis* 44:639–648.

Marozzi, M. (2004b). A bi-aspect nonparametric test for the multi-sample location problem. *Computational Statistics and Data Analysis* 46:81–92.

Marozzi, M. (2007). Multivariate tri-aspect non-parametric testing. *Journal of Nonparametric Statistics* 19:269–282.

Marozzi, M. (2008). The Lepage location-scale test revisited. *Far East Journal of Theoretical Statistics* 24:137–156.

Marozzi, M. (2009). Some notes on the location-scale Cucconi test. *Journal of Nonparametric Statistics* 21:629–647.

Marozzi, M. (2011). Levene type tests for the ratio of two scales. *Journal of Statistical Computation and Simulation* 81(7):815–826.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin* 105(1):56–166.

Muccioli, C., Belford, R., Podgor, M., Sampaio, P., de Smet, M., Nussenblatt, R. (1996). The diagnosis of intraocular inflammation and cytomegalovirus retinitis in HIV-infected patients by laser flare photometry. *Ocular Immunology & Inflammation* 4(2):75–81.

Murakami, H. (2007). Lepage type statistic based on the modified Baumgartner statistic. *Computational Statistics and Data Analysis* 51:5061–5067.

Neuhäuser, M. (2000). An exact two-sample test based on the Baumgartner-Weiss-Schindler statistic and a modification of Lepage's test. *Communication in Statistics – Theory and Methods* 29:67–78.

Neuhäuser, M. (2010). *Computer-Intensive und Nichtparametrische Statistische Tests.* München: Oldenbourg.

Neuhäuser, M., Senske, R. (2004). The Baumgartner-Weiss-Schindler test for the detection of differentially expressed genes in replicated microarray experiments. *Bionformatics* 20:3553–3564.

O'Brien, P. C. (1988). Comparing two samples: Extensions of the t, rank sum, and log-rank tests. *Journal of the American Statistical Association* 83:52–61.

Podgor, M. J., Gastwirth, J. L. (1994). On non-parametric and generalized tests for the two-sample problem with location and scale change alternatives. *Statistics in Medicine* 13:747–758.

Rice, K. L., Rubins, J. B., Lebahn, F., Parenti, C. M., Duane, P. G., Kuskowski, M., Joseph, A. M., Niewoehner, D. E. (2000). Withdrawal of chronic systemic corticosteroids in patients with COPD. *American Journal of Respiratory and Critical Care Medicine* 162:174–178.

Yang, T., Chen, X., Zhang, Z. C. (2009). Spatio-temporal changes of hydrological processes and underlying driving forces in Guizhou region, Southwest China. *Stochastic Environmental Research and Risk Assessment* 23:1071–1087.

Zhang, J. (2006). Powerful two-sample tests based on the likelihood ratio. *Technometrics* 48:95–103.

Zhang, Q, Xu, C. Y., Yang, T. (2009). Variability of water resource in the yellow river basin of past 50 years, china. *Water Resource Management* 23:1157–1170.