Journal of Applied Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/cjas20

Robustness and power of modified Lepage, Kolmogorov-Smirnov and Crame´r-von Mises two-sample tests

Herbert Büning

Published online: 02 Aug 2010.

Journal of Applied Statistics, Vol. 29, No. 6, 2002, 907-924

Robustness and power of modified Lepage, Kolmogorov-Smirnov and Cramér-von Mises two-sample tests

HERBERT BÜNING, *Institut für Statistik und Ökonometrie der Freien Universität Berlin, Germany*

ABSTRACT *For the two-sample problem with location and/or scale alternatives, as well as different shapes, several statistical tests are presented, such as of Kolmogorov-Smirnov and Cramér-von Mises type for the general alternative, and such as of Lepage type for location and scale alternatives. We compare these tests with the t-test and other location tests, such as the Welch test, and also the Levene test for scale. It turns out that there is, of course, no clear winner among the tests but, for symmetric distributions with the same shape, tests of Lepage type are the best ones whereas, for different shapes, Cramér-von Mises type tests are preferred. For extremely right-skewed distributions, a modification of the Kolmogorov-Smirnov test should be applied.*

1 Introduction

For the well-known two-sample problem, we have a lot of statistical tests, e.g. parametric, non-parametric and robustified ones, which take into account different types of alternatives and underlying distributions. If we assume normality of the data, the *t*-test and the *F*-test are uniformly most powerful for testing the equality of the two means and the two variances, respectively, and it is well known that the *t*-test is robust for non-normal distributions, in contrast to the *F*-test, which is extremely non-robust for non-normal data, see for example Tiku *et al.* (1986) and Büning (1991). The *t*-test, however, is not α-robust in the case of heteroscedasticity of the variances, in contrast to the test of Welch (1937) or the trimmed Welch test introduced by Yuen (1974). Robustified alternatives to the *F*-test are tests of Levene type (see Levene, 1960).

Correspondence: Herbert Büning, Institut für Statistik und Ökonometrie der Freien Universität Berlin, Berlin, Germany.

ISSN 0266-4763 print; 1360-0532 online/02/060907-18
DOI: 10.1080/0266476022013621 2

For location and scale alternatives, the Lepage (1971) test is perhaps the most familiar non-parametric test. It is a combination of a test for location, the Wilcoxon test, and a test for scale—the Ansari-Bradley test. The Lepage test behaves well for symmetric and medium- to long-tailed distributions. We consider modifications of the Lepage test by replacing the Wilcoxon and the Ansari-Bradley statistics by other linear rank statistics for location and scale, respectively. These Lepage-type tests are superior to the classical Lepage test for short-tailed and very-long-tailed distributions (see Büning and Thadewald, 2000).

For general alternatives, with possibly different shapes of the distribution functions of the X- and Y-variables, the Kolmogorov-Smirnov and Cramér-von Mises tests seem to be the most popular. These tests behave well for symmetric and long-tailed distributions. We also consider modifications of both tests by using appropriate weight functions in order to obtain a higher power than the classical counterparts for short-tailed distributions and distributions skewed to the right.

It is the purpose of this paper to present a comprehensive robustness and power study of all these tests for location and scale alternatives with equal and different shapes of the underlying distributions of the X- and Y-variables, assuming short, medium and long tails as well as distributions skewed to the right. In particular, the latter case is very important in statistical practice, as demonstrated by a real data example in Section 2.

2 Model and hypotheses, data example

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random variables with absolutely continuous distributions functions F_X and F_Y, respectively. We wish to test

$$\mathrm{H}_0: F_Y(z) = F_X(z) =: F(z)$$

for all $z \in \mathbb{R}$ versus the general alternative

$$\mathrm{H}_1^G: F_Y(z) \neq F_X(z)$$

for at least one $z \in \mathbb{R}$.

A special case is the location-scale alternative

$$\mathrm{H}_1^{LS}: F_Y(z) = F_X\left(\frac{z-\theta}{\tau}\right)$$

with $\theta \neq 0$, $\tau \neq 1$, $-\infty < \theta < \infty$, $\tau > 0$.

For $\tau = 1$, we have a pure location model, and for $\theta = 0$, we have a pure scale model. In the general case $\mathrm{H}_1^G: F_Y(z) \neq F_X(z)$, the distribution functions F_X and F_Y are different in any way.

Now, we present a real data example which demonstrates that, in practice, we may have the situation that F_X and F_Y are different in both location and scale as well as in shape.

Example 1

Table 1 shows the life expectancy (in years) of industrialized nations and nations with low income reported for 1998 by the *Fischer Almanach* 2001, see also Leinhardt & Wasserman (1979), who present a 1975 *New York Times* publication about the life expectancy and per capita income (in 1974 US dollars) for 105 nations

TABLE 1. Life expectancy (in years) of industrialized nations and nations with low income

Industrialized		Low income		Low income	
Australia	78	Afghanistan	46	Sierra Leone	38
Austria	77	Burma	60	Somalia	47
Belgium	77	Burundi	43	Sudan	55
Canada	79	Cambodia	53	Tanzania	48
Denmark	76	Chad	47	Togo	49
Finland	77	Benin	53	Uganda	40
France	78	Ethiopia	43	Burkina Faso	45
Germany	77	Guinea	47	Sri Lanka	73
Ireland	76	Haiti	54	Yemen	58
Italy	78	India	63	Dem. Rep. Congo	51
Japan	80	Kenya	52		
Netherlands	78	Laos	53		
New Zealand	77	Madagascar	58		
Norway	78	Malawi	39		
Portugal	75	Mali	54		
South Africa	54	Mauritania	54		
Sweden	79	Nepal	58		
Switzerland	79	Niger	49		
Britain	77	Pakistan	64		
United States	77	Rwanda	41		

classified into five categories: industrialized, petroleum exporting, higher income, middle income and low income, from which we choose here only the two categories 'industrialized' and 'low income'. It should be noted that there is a significantly increasing of life expectancy from 1974 to 1998 for nearly all nations listed.

Figure 1 presents the boxplots of the data of Example 1.

We see from the boxplots that the distributions of the data in the two groups of Example 1 seem to be asymmetric and to have different tailweights as well as different variances, the group 'industrialized' is, obviously, very homogenous (with one outlier) in contrast to the group 'low income'.

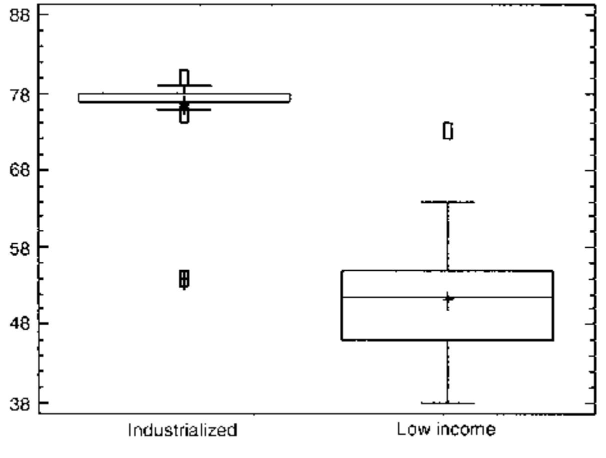

FIG. 1. Boxplots of the data of example 1.

3 Tests

3.1 Tests for general alternatives

For the general alternative H_1^G from Section 2 we consider the following non-parametric tests.

(1) *Kolmogorov-Smirnov test (KS)*
Let $X_{(1)}, \ldots, X_{(m)}$ and $Y_{(1)}, \ldots, Y_{(n)}$ be the order statistics of $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ and let $\hat{F}_X$ and $\hat{F}_Y$ be the usual empirical distribution functions for the X- and Y-samples, respectively. Furthermore, let $Z_{(1)}, \ldots, Z_{(N)}$ be the order statistics of the combined sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, $N = m + n$. Then the Kolmogorov-Smirnov statistic KS is defined by

$$KS = \max_{1 \leqslant i \leqslant N} |\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)})|$$

The corresponding test rejects H_0 if $KS \geqslant k_{1-\alpha}(m, n)$. Critical values $k_{1-\alpha}$ are reported by Büning & Trenkler (1994), where the asymptotic null distribution of KS is also given.

(2) *Cramér-von Mises test (CM)*
The Cramér-von Mises statistic CM is defined as follows:

$$CM = \frac{mn}{N^2} \sum_{i=1}^{N} (\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)}))^2$$

The corresponding test rejects H_0 if $CM \geqslant c_{1-\alpha}\ (m, n)$. Exact and asymptotically critical values $c_{1-\alpha}$ can be found in Burr (1963).

(3) *Modified versions of KS and CM*
We can modify the test statistics KS and CM by introducing appropriate weights to the differences $|\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)})|$ and $(\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)}))^2$, respectively, in the following way.

Let

$$t(x) = \lambda \hat{F}_X(x) + (1 - \lambda) \hat{F}_Y(x)$$

with $\lambda = m/N$ and let $M = mn/N$. Then we define

$$KS1 = \max_{1 \leqslant i \leqslant N} \frac{\sqrt{M}\,|\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)})|}{\sqrt{t(Z_{(i)})\,(1 - t(Z_{(i)}))}}$$

and

$$CM1 = \frac{mn}{N^2} \sum_{t=1}^{N} \frac{(\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)}))^2}{t(Z_{(i)})\,(1 - t(Z_{(i)}))} \quad \text{for } t(Z_{(i)}) \neq 0 \text{ or } 1$$

If $\hat{F}_X(x) = \hat{F}_Y(x) = 0$ or 1, i.e. $t(x) = 0$ or 1, we arbitrarily define the denominator of $KS1$ to be equal to $\sqrt{N - 1}/N$ and that of $CM1$ equal to $(N - 1)/N^2$. Obviously, the denominators $\sqrt{t(1 - t)}$ and $t(1 - t)$ place a high weight on the upper and lower part of the underlying distribution. The test $KS1$ has been studied by Canner (1975), Wilcox (1989, 1997) and Büning & Chakraborti (1999).

Secondly, we consider the test statistics

$$KS2 = \max_{1 \leq i \leq N} \frac{\sqrt{M}\,|\hat{F}_X(Z_{(i)}) - \hat{F}_Y(Z_{(i)})|}{\sqrt{t(Z_{(i)})\,(2 - t(Z_{(i)}))}}$$

and

$$CM2 = \frac{mn}{N^2} \sum_{i=1}^{N} \frac{(\hat{F}_X(Z_{(i)} - \hat{F}_Y(Z_{(i)}))^2}{t(Z_{(i)})\,(2 - t(Z_{(i)}))} \quad \text{for } t(Z_{(i)}) \neq 0$$

If $\hat{F}_X(x) = \hat{F}_Y(x) = 0$ i.e. $t(x) = 0$, we define the denominators of $KS2$ and $CM2$ as in the case of $KS1$ and $CM1$.

Now, the denominators $\sqrt{t(2 - t)}$ and $t(2 - t)$ emphasize the lower part of the underlying distribution.

Further weight functions can be chosen as $(t - 0.5)^{-2}$ for the symmetric case and $(1 - t^2)^{-1}$ for the asymmetric case. Here, we restrict our attention to the first two weight functions.

Simulated critical values of the test statistics $KS1$, $KS2$, $CM1$ and $CM2$ for selected samples sizes m, n with $10 \leq m, n \leq 50$ can be found in Büning (2001a).

3.2 Tests for location and scale alternatives

For the location and scale alternative from H_1^{LS} Section 2 we consider the following Lepage-type tests, which are based on two linear rank tests, one for location and one for scale.

The general form of the two-sample linear rank statistic L_N is given by $(N = m + n)$:

$$L_N = \sum_{k=1}^{N} a(N, k)\, V_k \text{ with scores } a(N, k) \in \mathbb{R} \text{ and}$$

$$V_k = \begin{cases} 1, \text{ if } Z_{(k)} \text{ belongs to the } X\text{-sample} \\ 0, \text{ otherwise} \end{cases}$$

where again, $Z_{(k)}$ is the kth order statistic of the combined two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$.

Under H_0, we get

$$E(L_N) = \frac{m}{N} \sum_{k=1}^{N} a(N, k)$$

and

$$\text{Var}(L_N) = \frac{mn}{N(N-1)} \sum_{k=1}^{N} (a(N, k) - \bar{a}_N)^2$$

with

$$\bar{a}_N = \frac{1}{N} \sum_{k=1}^{N} a(N, k),$$

see for example Büning & Trenkler (1994, p.128).

Now, we consider one test statistic of location

$$L_{N,\ell} = \sum_{k=1}^{N} a_\ell(N,k)\, V_k \text{ with scores } a_\ell(N,k) \in \mathbb{R}.$$

and one test statistic for scale

$$L_{N,s} = \sum_{k=1}^{N} a_s(N,k)\, V_k \text{ with scores } a_s(N,k) \in \mathbb{R}^+.$$

Then the Lepage-type statistic is defined by

$$LP = \left(\frac{L_{N,\ell} - E(L_{N,\ell})}{(\mathrm{Var}(L_{N,\ell}))^{1/2}}\right)^2 + \left(\frac{L_{N,s} - E(L_{N,s})}{(\mathrm{Var}(L_{N,s}))^{1/2}}\right)^2$$

Randles & Wolfe (1979, p. 259) give sufficient conditions under which the test statistics $L_{N,\ell}$ and $L_{N,s}$ are uncorrelated under H_0:

(1) $a_\ell(N,k) + a_\ell(N, N-k+1) = \text{constant}$ and
(2) $a_s(N,k) - a_s(N, N-k+1) = 0$, $k = 1, \ldots, N$.

Under conditions (1) and (2) LP has, under H_0, asymptotically a χ^2-distribution with 2 degrees of freedom.

Let us consider some examples of Lepage-type tests based on specially chosen two-sample scores where, in the parentheses, the tailweight of the (symmetric) distributions is indicated for which the location and scale tests have high power. Obviously, all the test statistics for location and scale satisfy the sufficient conditions (1) and (2). Thus, the statistics are uncorrelated and the resulting Lepage-type statistic is, under H_0, asymptotically χ^2-distributed with 2 degrees of freedom.

Power comparisons of such Lepage-type tests are also presented by Büning & Thadewald (2000).

(1) *Wilcoxon/Ansari-Bradley test (medium tails)*
The classical test of Lepage (1971) is a combination of the Wilcoxon test (W) for location and the Ansari-Bradley test (AB) for scale alternatives.

The scores of the Wilcoxon statistic are given by $a_W(N,k) = k$ and the scores of the Ansari-Bradley statistic by $a_{AB}(N,k) = (N+1)/2 - |k - (N+1)/2|$.

The resulting location and scale test is abbreviated by $LP1$.

(2) *Gastwirth/Gastwirth test (short tails)*
Gastwirth (1965) proposed a test for location based on scores:

$$a_{G\ell}(N,k) = \begin{cases} k - \dfrac{N+1}{4} & \text{if} \quad k \leqslant \dfrac{(N+1)}{4} \\[2ex] 0 & \text{if} \quad \dfrac{(N+1)}{4} < k < \dfrac{3(N+1)}{4} \\[2ex] k - \dfrac{3(N+1)}{4} & \text{if} \quad k \geqslant \dfrac{3(N+1)}{4} \end{cases}$$

and a test for scale based on scores:

$$a_{G,s}(N,k) = \begin{cases} \dfrac{N+1}{4} - k & \text{if} \quad k \leqslant \dfrac{(N+1)}{4} \\ 0 & \text{if} \quad \dfrac{(N+1)}{4} < k < \dfrac{3(N+1)}{4} \\ k - \dfrac{3(N+1)}{4} & \text{if} \quad k \geqslant \dfrac{3(N+1)}{4} \end{cases}$$

The resulting Lepage-type test is abbreviated by $LP2$.

(3) *Van der Waerden/Klotz test (medium tails)*
The scores of the location test of Van der Waerden are given by: $a_{\mathrm{vdW}} = \Phi^{-1}(k/(N+1))$ and the scores of the scale test of Klotz by $a_K = (\Phi^{-1}(k/(N+1)))^2$.
The resulting Lepage-type test is abbreviated by $LP3$.

(4) *LT/Mood test (long tails)*
For long-tailed distributions and location alternatives we apply a test, called the LT-test, which has much higher power than the 'classical' Median test, see Büning (1997). The scores are chosen analogously to Huber's Ψ-function referring to M-estimates, see Huber (1964), and they are given by

$$a_{LT}(N,k) = \begin{cases} -\left(\left[\dfrac{N}{4}\right]+1\right) & \text{if} \quad k < \left[\dfrac{N}{4}\right]+1 \\ k - \dfrac{N+1}{2} & \text{if} \quad \left[\dfrac{N}{4}\right]+1 \leqslant k \leqslant \left[\dfrac{3(N+1)}{4}\right] \\ \left[\dfrac{N}{4}\right]+1 & \text{if} \quad k > \left[\dfrac{3(N+1)}{4}\right] \end{cases}$$

$[x]$ denotes the greatest integer less than or equal to x.
The scores of the scale test of Mood are defined by

$$a_M(N,k) = \left(k - \frac{N+1}{2}\right)^2$$

The resulting Lepage-type test is abbreviated by $LP4$.
Goria (1982) proposed similar modified Lepage tests and presented some results on the asymptotic relative efficiencies of the tests.

3.3 *Tests for location*

For the pure location model, we consider five tests, the t-test, the trimmed t-test, the Welch test, the trimmed Welch test and a rank version of the Welch test following a proposal of Conover & Iman (1981).

(1) *t-test*
The t-statistic is given by

$$t = \frac{\bar{X} - \bar{Y}}{\left(\frac{(m-1)S_X^2 + (n-1)S_Y^2}{m+n-2}\left(\frac{1}{m} + \frac{1}{n}\right)\right)^{1/2}}$$

with

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i,\; S_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2 \textit{ and analogously for } \bar{Y} \textit{ and } S_Y^2$$

Assuming normality of the data, i.e. $X_i \sim N(\mu_X, \sigma_X^2), i = 1, \ldots, m, Y_j \sim N(\mu_Y, \sigma_Y^2)$, $j = 1, \ldots, n$, and equal variances, $\sigma_X^2 = \sigma_Y^2 = \sigma^2$, then the t-statistic has, under $H_0: \mu_X = \mu_Y$, a t-distribution with $m + n - 2$ degrees of freedom. Under normality, the t-test is the uniformly most powerful unbiased test for one- and two-sided alternatives.

(2) *Trimmed t-test*
Let

$$\bar{X}_{g_1} = \frac{1}{m - 2g_1}\sum_{i=g_1+1}^{m-g_1} X_{(i)}$$

and

$$\bar{Y}_{g_2} = \frac{1}{n - 2g_2}\sum_{j=g_2+1}^{n-g_2} Y_{(j)}$$

be the trimmed means of the X- and Y-variables, respectively, where g_1 and g_2 are the numbers of trimmed variables and let

$$\bar{X}_{w_1} = \frac{1}{m}[g_1 X_{(g_1+1)} + g_1 X_{(m-g_1)} + \sum_{i=g_1+1}^{m-g_1} X_{(i)} \text{ and}$$

$$\bar{Y}_{w_2} = \frac{1}{n}[g_2 Y_{(g_2+1)} + g_2 Y_{(n-g_2)} + \sum_{j=g_2+1}^{n-g_2} Y_{(j)}$$

be the winsorized means of the two variables. Furthermore, let

$$SSD_{w_1}(X) = g_1(X_{(g_1+1)} - \bar{X}_{w_1})^2 + g_1(X_{(m-g_1)} - \bar{X}_{w_1})^2 + \sum_{i=g_1+1}^{m-g_1}(X_{(i)} - \bar{X}_{w_1})^2 \text{ and}$$

$$SSD_{w_2}(Y) = g_2(X_{(g_2+1)} - \bar{Y}_{w_2})^2 + g_2(Y_{(n-g_2)} - \bar{Y}_{w_2})^2 + \sum_{j=g_2+1}^{n-g_2}(Y_{(j)} - \bar{Y}_{w_2})^2$$

be the winsorized sum of squared deviations of the X- and Y-variables. Then the trimmed t-statistic of Yuen (1974) is given by

$$t_g = \frac{\bar{X}_{g_1} - \bar{Y}_{g_2}}{((SSD_{w_1}(X) + SSD_{w_2}(Y))/(h_1 + h_2 - 2)(1/h_1 + 1/h_2))^{1/2}}$$

where $h_1 = m - 2g_1$ and $h_2 = n - 2g_2$. The usual trimmed fractions g_1/m and g_2/n are 10%, 15% or 20%.

Under normality of the data, the distribution of t_g can, under H_0, be approximated by a t-distribution with $\upsilon = h_1 + h_2 - 2$ degrees of freedom.

(3) *Welch test*
The statistic V of Welch (1937) is defined by

$$V = \frac{\bar{X} - \bar{Y}}{\left(\dfrac{S_X^2}{m} + \dfrac{S_Y^2}{n}\right)^{1/2}}$$

In contrast to the t-statistic, the variances of both samples are estimated separately in the denominator of V. Assuming normality of the data, the Welch statistic has, under $H_0: \mu_X = \mu_Y$, approximately a t-distribution with

$$\upsilon = \frac{(S_X^2/m + S_Y^2/n)^2}{S_X^4/(m^2(m-1)) + S_Y^4/(n^2(n-1))}$$

degrees of freedom.

For asymptotic results of the V-test with respect to level α and power β, see for example Staudte & Sheather (1990, p. 180f).

(4) *Trimmed Welch test*
The trimmed Welch statistic V_g, proposed by Yuen (1974), is defined by

$$V_g = \frac{\bar{X}_{g1} - \bar{Y}_{g2}}{(SSD_{w_1}(X)/h_1 + SSD_{w_2}(Y)/h_2)^{1/2}}$$

Under normality of the data V_g has, under H_0, approximately a t-distribution with

$$\upsilon = \frac{(SSD_{w_1}(X)/h_1 + SSD_{w_2}(Y)/h_2)^2}{SSD_{w_1}^2(X)/(h_1^2(h_1-1)) + SSD_{w_2}^2(Y)/(h_2^2(h_2-1))}$$

degrees of freedom.

The V_g-test is a robustified version of the Welch test. V_g protects against outliers and unequal variances.

(5) *Welch rank test*
Following a proposal of Conover & Iman (1981) we can modify the Welch statistic V by replacing the original observations by its ranks in the combined ordered sample. The resulting test statistic is called the Welch rank test and it is denoted by VR. Because an extreme outlier is replaced by rank 1 or rank N, the test based on VR might be more efficient than the V-test for long-tailed distributions.

3.4 *Tests for scale*

For scale alternatives, the parametric counterpart to the location t-test is the F-test, the statistic of which is defined by $F = S_x^2/S_Y^2$. The F-test is, under normality, the uniformly most powerful test for one-sided alternatives $\sigma_X^2 < \sigma_Y^2$ or $\sigma_X^2 > \sigma_Y^2$, but it is extremely α – non-robust with respect to departures from normality, see for example Tiku *et al.* (1986, p. 146) and Büning (1991, p. 148). A robustified version of the F-statistic is the statistic of Levene (1960), which is nothing else than the square of the t-statistic applied on transformed samples of the Xs and Ys. To go into more detail: let $X_i' = |X_i - \bar{X}|$ and $Y_j' = |Y_j - \bar{Y}|$, $i = 1, \ldots, m$, $j = 1, \ldots, n$

be transformations of the random variables $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$, respectively. Furthermore, let $\bar{X}'$ and $\bar{Y}'$ be the arithmetic means of the X_i' – and Y_j' – variables. The Levene statistic is then defined by

$$LV1 = \frac{(\bar{X}' - \bar{Y}')^2}{S'^2(1/m + 1/n)}$$

with the pooled variance

$$S'^2 = \left(\sum_{i=1}^{m} (X_i' - \bar{X}')^2 + \sum_{j=1}^{n} (Y_j' - \bar{Y}')^2 \right) / (m + n - 2)$$

The statistic $LV1$ has, under H_0, approximately a F-distribution with $(1, m + n - 2)$ degrees of freedom, see Levene (1960). Miller (1968) has shown that $LV1$ is asymptotically distribution-free if the underlying distribution function is symmetric or at least has its median equal to its mean.

If we replace the arithmetic means in the definition of the transformed variables X_i' and Y_j' by the sample medians, M_X and M_Y, respectively, we obtain a robustified version of the Levene test studied in detail by Brown & Forsythe (1974). We abbreviate the robustified version of $LV1$ by $LV2$. The test $LV2$ is generally asymptotically distribution-free, provided that the first two moments of the underlying distribution exist, see Miller (1968).

4 Robustness and power study

4.1 *Tailweight, skewness and Levy distance*

For our robustness and power study we select seven distributions, the uniform (Uni), the normal (Norm), the double exponential (Dexp), the Cauchy (Cau), the exponential (Exp) and two contaminated normal distributions, $CN1 = 0.9N(0,1) + 0.1N(0,3^2)$ (symmetric) and $CN2 = 0.5N(1,2^2) + 0.5N(-1,1)$ (skewed to the right). These seven distributions cover a wide range of short-tailed up to very long-tailed distributions, as well as asymmetric ones. In Table 2 values of two measures S and T for skewness and tailweight, respectively, are presented for the distributions above. The measures are defined by

$$S = \frac{x_{0.975} - x_{0.5}}{x_{0.5} - x_{0.025}}$$

TABLE 2. Values of S and T for some distributions

Distribution	S	T
Uniform	1	1.267
Normal	1	1.704
Logistic	1	1.883
CN1	1	1.991
Double exp.	1	2.161
Cauchy	1	5.263
CN2	1.769	1.693
Exponential	4.486	1.883

TABLE 3. Levy distances

Distribution	Uniform	*CN*1	Logistic	Dexp	Cauchy
Levy distance	0.0897	0.0135	0.0163	0.0766	0.0932

and

$$T = \frac{x_{0.975} - x_{0.025}}{x_{0.875} - x_{0.125}}$$

where x_p is the p-quantile of distribution function F. Obviously, $S < 1$, if F is skewed to the left; $S = 1$, if F is symmetric; and $S > 1$, if F is skewed to the right. $T \geqslant 1$, the longer the tails the greater T. The measures S and T are location and scale invariant.

Looking at Table 2, we see that the Cauchy distribution has very long tails and the exponential is extremely right-skewed.

Deviations from the ideal model, e.g. from the normal distribution, may be described by the Prohorov-, Kolmogorov- or Levy-distance of two distributions F and G. Here, we restrict our attention to the Levy distance d_L, which is defined by

$$d_L(F,G) = \inf\{\delta \mid G(x-\delta) - \delta \leqslant F(x) \leqslant G(x+\delta) + \delta, \forall x \in \mathbb{R}\}$$

Obviously, d_L is a metric. $\sqrt{2}d_L(F,G)$ is the maximum distance between the graphs of F and G, measured along a 45° direction, see for example. Büning (1991, p. 28). Notice that the Levy distance is not location and scale invariant. Values of the distance d_L between the standard normal distribution and some other symmetric distributions are given in Table 3. Because d_L is not location and scale invariant, the corresponding densities f are all scaled in such a way that $f(0) = 1/\sqrt{2\pi}$ as for the standard normal.

Of course, tailweight T and Levy distance d_L are completely different concepts, e.g. the uniform and the Cauchy distribution have nearly the same distance to the normal but very different tailweight.

4.2 α-robustness (robustness of validity)

Consider the model and hypotheses from Section 2. For a nominal level α the critical region C_α of a test statistic T_n for testing $H_0: F_Y = F_X = F$ may be uniquely determined by $P_F(T_n \in C_\alpha \mid H_0) = \alpha$. We now assume, under H_0, a distribution function $G_Y = G_X = G$ for the data and determine the actual level $\alpha^\star$ of the test, i.e. $a^\star = P_G(T_n \in C_\alpha \mid H_0)$.

Definition 1

$r_\alpha(\alpha, \alpha^\star) := |\alpha - \alpha^\star|/\alpha$ is called the α-robustness measure of the test statistic T_n. Of course, $r_\alpha(\alpha, \alpha^\star)$ also depends on the sample size n. We say that the test based on T_n is α-robust w.r.t. G if $r_\alpha(\alpha, \alpha^\star)$ is 'not too large', that is it may be smaller than 0.1 (strong criterion) or smaller than 0.5 (liberal criterion), see Bradley (1978).

For our power study in Section 4.3 we, at first, investigate the α-robustness of all the tests from Section 3 in order to guarantee a meaningful power comparison of these tests. The non-parametric tests of Kolmogorov-Smirnov- and Cramér-von Mises type are, of course, distribution-free and those of Lepage type are asymptotically distribution-free.

TABLE 4. Critical values $c_{1-\alpha}$ of KS- and CM-type tests, $m = 20$, $n = 30$, $\alpha = 5\%$

Tests	KS	CM	$KS1$	$KS2$	$CM1$	$CM2$
$c_{1-\alpha}$	0.380	0.461	2.886	1.786	2.463	0.791

TABLE 5. Actual levels $\alpha^\star$ of tests with nominal level $\alpha = 5\%$, $m = 20$, $n = 30$

	Distributions						
Tests	Uni	Norm	$CN1$	Dexp	Cau	$CN2$	Exp
KS	4.7	4.9	4.9	4.9	4.8	4.8	5.1
CM	4.8	5.3	4.9	5.0	5.1	5.2	4.8
$KS1$	5.1	5.4	5.2	5.4	5.0	5.3	5.2
$KS2$	4.8	5.2	4.9	5.2	5.1	5.4	5.2
$CM1$	4.7	4.9	4.8	4.9	4.6	5.2	5.1
$CM2$	4.6	5.1	4.6	5.0	4.5	5.3	5.0
$LP1$	5.0	4.7	4.7	5.0	4.7	4.7	5.0
$LP2$	4.6	4.7	4.6	4.5	4.5	4.8	4.7
$LP3$	4.1	4.4	4.4	4.2	4.4	4.7	4.5
$LP4$	4.7	5.0	4.9	4.9	4.7	4.7	5.0
t	4.8	5.2	4.9	4.8	2.3	4.9	5.1
t_g	4.8	5.1	4.8	4.7	4.3	5.0	4.9
V	4.9	5.1	4.9	4.7	2.2	5.0	4.8
V_g	5.0	5.3	5.0	4.6	3.8	5.0	4.8
VR	4.9	4.9	5.2	5.0	5.0	5.1	4.9
F	0.6	5.2	16.0	13.1	33.4	6.5	6.9
$LV1$	5.3	5.3	5.2	6.0	16.8	8.2	4.0
$LV2$	3.7	4.3	4.2	4.5	1.9	4.9	5.1

The calculation of the actual levels $\alpha^\star$ is carried out via Monte Carlo Simulation (10 000 replications). Critical values of the tests of Kolmogorov and Cramér-von Mises type used in the study are listed in Table 4 for sample sizes $m = 20$, $n = 30$ as in our data Example 1 from Section 2. The values for $KS1$, $KS2$, $CM1$ and $CM2$ are obtained by Monte Carlo simulation (100 000 replications) and those of KS and CM can be found in Büning & Trenkler (1994) and Burr (1963), respectively.

In Table 5 values of $\alpha^\star$ of the 18 tests from Section 3 are presented for seven distributions described in Section 4.1. The nominal level is 5%.

From table 5 we can state the following.

- The approximation of the distribution of the Lepage type tests by the Chi-square distribution works well for the seven distributions, except for the test $LP3$, which is moderately conservative.
- For the Cauchy distribution, all the parametric tests—with the exception of VR—are not α-robust, all the tests for location and the test $LV2$ for scale are very conservative, and the tests F and $LV1$ are extremely anticonservative.
- For the other distributions we obtain that t, t_g, V and V_g maintain the level α quite well, the α-robustness measure $r_\alpha(\alpha, \alpha^\star)$ always takes values smaller than 0.1, i.e. it fulfils the strong criterion of Bradley (1978). The test $LV2$ is moderately α-robust, in all cases it is conservative. The approximation of the distribution of $LV2$ by the F-distribution does not work as well for this unbalanced case ($m = 20$, $n = 30$) as in the balanced case with larger sample

sizes $m = n = 40$, see Tiku *et al.* (1986, p. 146) and Büning (1991, p. 158). For the α-robustness measure we get $r_a(\alpha, \alpha^\star) \leqslant 0.26$, i.e. the liberal criterion of Bradley is still fulfilled. The tests F and $LV1$ are extremely α-non-robust for deviations of the normal distribution. The α-robustness measure $r_\alpha(\alpha, \alpha^\star)$ of the F-test takes the value 2.20 for $CN1$ and the value 2.38 for the exponential distribution.

- Considering the results of α-robustness of the tests, we present in our power tables in Section 4.3, as a test for scale, only the $LV2$-test, and as a test for location, only the VR-test. The power study has shown that, besides its α-robustness, the test VR has mostly higher power than its location counterparts.

4.3 *Power comparison*

Now, let us study via Monte Carlo simulation (10 000 replications) the power of 12 tests selected from Section 3. Again, we consider the seven distributions given in Section 4.1.

For our power comparison we assume for all the distributions considered (except for the exponential with $E(X) = 1/\lambda = 1$) that, under $H_0 : F_Y(z) = F_X(z)$, we have $E(X_i) = E(Y_j) = 0$ and $\mathrm{Var}(X_i) = \mathrm{Var}(Y_j) = 1, i = 1, \ldots, m, j = 1, \ldots, n$.

Under $H_1^{LS} : F_Y(z) = F_X((z - \theta)/\tau)$, we set $Y_j^\star = \theta + \tau Y_j, j = 1, \ldots, n$, with varying location parameter θ and scale parameter τ, namely, $(\theta, \tau) = (0, 1)$, $(0.6, 1.6)$, $(0.8, 1.2)$ and $(0.4, 2.0)$. The last two vectors emphasize differences in location and in scale of the X- and Y-variables, respectively. The nominal level α of all tests is 5%.

Table 6 presents power values of the tests for location and scale alternatives w.r.t. the same type of distributions of the X- and Y-variables. Table 7 is concerned with different distribution functions (shapes) of X and Y where the first line below 'Distributions' marks the distribution function of the X-variable and the second line that of the Y-variable.

A look at Table 6 shows that we have to distinguish between the three parameter vectors $(\theta, \tau) = (0.6, 1.6)$, $(0.8, 1.2)$ and $(0.4, 2.0)$.

- For $(\theta, \tau) = (0.6, 1.6)$, the Lepage type tests and the Levene test $LV2$ are the best ones for the uniform and the normal distribution. $LP2$ has the highest power for the uniform distribution and $LP4$ has the highest power for the normal and all symmetric distributions with longer tails than the normal, except for the Cauchy where KS or CM are preferred. Here, $LV2$ is very poor. For $CN2$ (slightly right-skewed) the test with the highest power is $LV2$, and for the exponential, the test $KS2$ seems to be the best, although the differences in power of all tests are small for the exponential with the exception of $LV2$.
- For $(\theta, \tau) = (0.8, 1.2)$, with an emphasis on location, it is not surprising that the scale test $LV2$ has very low power for all distributions, whereas the location test VR behaves very well—for the normal it is the best one. Again, $LP2$ is remarkably superior to the other tests for the uniform distribution. For all symmetric distributions with longer tails than the normal and for both distributions skewed to the right, the KS- and CM-type tests are preferred. The classical tests KS and CM are the best for long-tailed distributions.
- For $(\theta, \tau) = (0.4, 2.0)$, with an emphasis on scale, the Levene test $LV2$ is, on the whole, a test with high power in comparison to the other tests, as would be expected, except for the exponential, and of course, for the Cauchy where it is very poor again. (Here, the actual level $\alpha^\star$ is 1.9%, so that a power

TABLE 6. Power of some tests for location-scale alternatives (in %), $m = 20$, $n = 30$, $\alpha = 5\%$

		Distributions						
Tests	θ, τ	Uni	Norm	*CN*1	Dexp	Cau	*CN*2	Exp
KS	0.6, 1.6	36.2	36.1	47.0	51.8	43.9	24.1	93.1
	0.8, 1.2	47.8	57.9	75.8	82.7	73.4	60.7	97.6
	0.4, 2.0	37.2	27.0	31.0	29.4	25.2	15.5	86.1
CM	0.6, 1.6	35.0	38.3	49.3	53.3	42.4	26.8	95.4
	0.8, 1.2	58.4	65.4	81.7	84.2	71.4	67.7	97.2
	0.4, 2.0	36.4	26.1	30.1	29.3	22.4	15.0	92.5
*KS*1	0.6, 1.6	54.8	37.8	46.0	46.4	37.1	23.2	99.3
	0.8, 1.2	66.3	56.8	73.4	77.7	65.9	62.4	99.9
	0.4, 2.0	64.7	34.7	35.9	29.2	22.5	20.8	95.7
*KS*2	0.6, 1.6	12.4	16.5	25.1	31.6	26.8	13.7	99.8
	0.8, 1.2	56.1	48.0	65.5	71.7	59.7	65.7	100
	0.4, 2.0	27.5	12.6	13.5	14.1	11.8	13.0	97.7
*CM*1	0.6, 1.6	40.5	39.3	49.3	51.2	38.3	27.0	97.5
	0.8, 1.2	66.0	66.0	80.9	82.0	57.0	69.1	98.7
	0.4, 2.0	52.1	31.9	33.2	30.2	21.0	19.1	95.0
*CM*2	0.6, 1.6	23.1	27.9	38.5	43.5	34.0	20.8	98.5
	0.8, 1.2	57.6	60.9	77.8	80.4	66.5	70.8	99.5
	0.4, 2.0	34.1	19.3	20.5	20.8	15.7	14.6	95.8
*LP*1	0.6, 1.6	57.4	47.6	51.7	49.6	34.4	27.6	94.7
	0.8, 1.2	53.7	56.5	72.5	73.4	52.9	59.6	97.8
	0.4, 2.0	81.4	62.6	57.3	48.0	32.7	47.4	89.5
*LP*2	0.6, 1.6	75.2	41.4	38.1	28.9	14.2	24.1	98.7
	0.8, 1.2	75.6	45.8	53.8	40.7	17.9	52.1	99.7
	0.4, 2.0	94.9	66.8	52.4	38.7	18.8	54.1	94.8
*LP*3	0.6, 1.6	69.1	48.1	47.6	46.4	26.2	29.4	97.3
	0.8, 1.2	54.7	55.0	69.9	67.5	42.1	58.2	99.0
	0.4, 2.0	94.1	68.9	53.8	46.7	22.2	55.9	93.4
*LP*4	0.6, 1.6	63.6	50.4	54.0	55.4	41.2	30.9	89.7
	0.8, 1.2	36.3	52.2	71.3	77.1	64.5	53.1	95.9
	0.4, 2.0	90.9	70.7	62.2	54.6	34.4	57.0	80.8
VR	0.6, 1.6	28.1	35.5	46.7	48.6	35.7	26.8	94.8
	0.8, 1.2	64.7	68.4	81.7	82.4	62.1	68.9	94.6
	0.4, 2.0	12.7	15.4	20.2	21.8	16.4	7.9	93.6
*LV*2	0.6, 1.6	65.6	47.1	30.9	28.0	3.4	39.2	22.8
	0.8, 1.2	13.6	10.3	7.7	7.7	1.9	9.2	7.1
	0.4, 2.0	93.8	80.4	58.7	55.3	6.2	69.3	43.8

comparison intrinsically becomes meaningless.) The tests *LP*2, *LP*3 and *LP*4 are slightly superior to *LV*2 for the uniform and normal distributions, respectively. The *KS*- and *CM*-type tests do not work as well now.

For the case of different shapes we select seven 'realistic' combinations of the distribution functions F_X and F_Y, taking into consideration different tailweight and strength of skewness for F_X and F_Y. Table 7 presents the power results.

We see that similar results are obtained as for the case of equal shapes, but now

TABLE 7. Power of some tests for different shapes (in %), $m = 20$, $n = 30$, $\alpha = 5\%$

		Distributions						
Distrib.	θ, τ	Norm *CN*1	Norm Dexp	Norm Cau	*CN*1 Dexp	Dexp Cau	Norm *CN*2	Exp *CN*2
KS	0.6, 1.6	35.1	37.9	36.0	48.7	50.4	17.8	69.6
	0.8, 1.2	65.8	69.1	64.4	78.6	78.7	45.3	18.5
	0.4, 2.0	21.8	18.9	21.4	28.2	31.2	14.2	91.1
CM	0.6, 1.6	40.5	42.5	40.6	52.6	52.7	20.6	56.9
	0.8, 1.2	73.1	74.2	68.7	83.1	79.9	55.7	15.8
	0.4, 2.0	22.4	21.1	24.1	29.2	31.5	12.6	82.6
*KS*1	0.6, 1.6	34.3	34.8	34.1	45.8	46.6	22.8	88.5
	0.8, 1.2	65.0	65.5	59.0	74.7	73.4	52.9	36.8
	0.4, 2.0	24.2	20.1	25.3	30.5	32.4	22.6	98.0
*KS*2	0.6, 1.6	22.8	26.9	21.8	29.3	29.0	9.8	93.8
	0.8, 1.2	64.2	64.2	52.6	68.2	63.8	55.3	50.7
	0.4, 2.0	8.8	9.7	11.9	13.0	15.7	8.6	99.1
*CM*1	0.6, 1.6	40.3	41.4	41.4	51.7	52.5	22.4	65.8
	0.8, 1.2	73.3	73.3	67.7	81.6	77.9	59.6	19.8
	0.4, 2.0	23.7	22.1	27.1	31.7	35.0	16.3	89.0
*CM*2	0.6, 1.6	33.3	35.9	32.7	42.9	42.7	14.3	79.1
	0.8, 1.2	72.9	72.6	63.6	79.0	73.8	58.8	29.1
	0.4, 2.0	14.7	15.3	18.5	21.2	23.5	10.3	94.8
*LP*1	0.6, 1.6	36.3	33.5	38.4	49.1	56.8	25.3	85.6
	0.8, 1.2	62.8	62.6	55.3	72.3	68.2	47.7	35.8
	0.4, 2.0	38.0	28.6	40.8	47.7	59.5	43.9	97.2
*LP*2	0.6, 1.6	27.4	25.0	35.0	33.9	41.2	30.2	95.6
	0.8, 1.2	51.0	43.2	33.4	45.8	35.4	54.9	55.1
	0.4, 2.0	36.9	31.2	49.5	45.9	55.9	52.0	99.7
*LP*3	0.6, 1.6	35.4	35.2	48.4	48.4	60.9	23.5	82.2
	0.8, 1.2	61.1	59.0	57.4	68.1	66.8	50.0	33.8
	0.4, 2.0	41.0	36.8	55.9	50.0	64.2	46.9	96.4
*LP*4	0.6, 1.6	36.5	37.5	49.3	54.7	67.2	22.9	86.0
	0.8, 1.2	59.1	61.3	62.6	74.4	78.3	36.4	37.8
	0.4, 2.0	43.5	37.1	53.4	55.8	68.0	50.5	97.5
VR	0.6, 1.6	41.2	41.7	36.3	48.9	43.0	22.1	36.3
	0.8, 1.2	74.7	74.4	65.7	82.1	73.9	60.2	13.2
	0.4, 2.0	18.2	18.2	18.1	22.0	19.1	7.1	52.5
*LV*2	0.6, 1.6	16.8	16.2	13.5	28.7	19.0	36.3	51.2
	0.8, 1.2	3.4	3.1	4.8	7.1	7.8	7.2	21.1
	0.4, 2.0	45.4	41.7	23.9	55.8	29.5	68.3	76.0

the gain in power of the *KS*- and *CM*-type tests compared with the Lepage-type tests is remarkable for $(\theta, \tau) = (0.6, 1.6)$ and $(0.8, 1.2)$, a result that is not surprising. For parameter vector $(0.4, 2.0)$, however, the power of the *KS*- and *CM*-type tests is much less than its competitors of the Lepage type, especially in comparison to *LP*3 and *LP*4, which seem to be the best ones for combinations of symmetric long-tailed distributions. The location test *VR* has the highest power among all tests for $(\theta, \tau) = (0.6, 1.6)$ and $(0.8, 1.2)$ if F_X and F_Y have medium tails. *LV*2 is the best one for $(\theta, \tau) = (0.4, 2.0)$ in all cases where the Cauchy is not included. For the combination of two asymmetric distributions, *LP*2 and *KS*2 are the clear winners among all the parametric and non-parametric tests.

Now, let us interpret the power results of the tests for the symmetric distributions from Table 6 in connection with the Levy distance $d_L(F, G)$ of two distributions F and G, see Table 3.

On the one hand, the d_L(Norm, Uni) and d_L(Norm, Cau) take nearly the same values but the power behaviour of all the tests for the normal distribution in comparison with the uniform distribution is, obviously, not the same as that in comparison with the Cauchy distribution. Fixing a special alternative (θ, τ) for the normal, there is a remarkable gain in power of some tests for the uniform distribution with the same vector (θ, τ) but a great loss of power for the Cauchy distribution and vice versa.

On the other hand, d_L(Norm, *CN*1) and d_L(Norm, Dexp) take very different values but, for a fixed vector (θ, τ), the power behaviour of all the tests under *CN*1 and the double exponential is nearly the same. Thus, the Levy distance is not an appropriate indicator for power behaviour in contrast to the tailweight of the distribution.

Further power studies on two-sample tests in the case of location and scale alternatives with the same and with different shapes of distribution functions F_x and F_Y can be found in Fligner & Policello (1981), Goria (1982), Podgor & Gastwirth (1994), Fueda & Ôhori (1995), Arrenberg (1996), Magel & Wibowo (1997), Büning & Chakraborti (1999) and Büning & Thadewald (2000).

4.4 Results for the data example

Now, let us study the data example in Section 2. What is an appropriate test for these data? In order to make a decision for one of the tests from Tables 6 and 7 we calculate the measures $\hat{S}$ and $\hat{T}$ separately for the *X*- and *Y*-variables, where $\hat{S}$ and $\hat{T}$ are estimates of the measures of skewness S and tailweight T from Section 4.1, obtained by replacing the p-quantile x_p by its estimate

$$\hat{x}_p = \begin{cases} x_{(1)} & \text{if} \quad p < 0.5/m \\ (1-\lambda)x_{(i)} + \lambda x_{(i+1)} & \text{if} \quad 0.5/m \leqslant p \leqslant 1 - 0.5/m \\ x_{(m)} & \text{if} \quad p > 1 - 0.5/m \end{cases}$$

with $i = [mp + 0.5]$ and $\lambda = mp + 0.5 - i$.

$\hat{y}_p$ is defined analogously. We get for Example 1

$$x\text{-observations:}\ \ \hat{S}_X = 0.130,\ \hat{T}_X = 8.667$$

$$y\text{-observations:}\ \ \hat{S}_Y = 1.453,\ \hat{T}_Y = 1.806$$

Looking at Table 2, we see that the distribution of the *X*-variable seems to be *extremely left*-skewed with *very heavy* tails, whereas the distribution of the *Y*-variable is *slightly right skewed* with *medium* tails. That means that both distributions F_X and F_Y are obviously of different type. Therefore, we recommend the use of the *CM*-test or perhaps the *CM*1-test. We apply the *CM*-test and obtain $CM = 3.73$. Table 4 shows that the (asymptotically) critical value of *CM* is equal to 0.461. Thus, H_0 has to be rejected at level $\alpha = 0.05$, a decision we might have expected.

It should be noted that if we delete the extreme outlier 54 (South Africa) from the x-sample we obtain $\hat{S}_X = 1.500$, $\hat{T}_X = 1.667$, i.e. the distribution of the X-variable is now classified as slightly right-skewed with medium tails. We see, that the measures of skewness and tails, $\hat{S}$ and $\hat{T}$, respectively, are strongly influenced by such extreme outliers.

5 Conclusions

In this paper, we studied several parametric and non-parametric tests for location and scale alternatives, including different shapes of the distribution functions F_X and F_Y of the X- and Y-variables, respectively. For our simulation study we consider a broad range of distributions with different strengths of skewness and different tailweights. Clearly, there is no clear winner for all distributions considered. If we can assume symmetric distribution functions F_X and F_Y of the same type but with differences in location *and* scale we would recommend the Lepage-type test $LP2$ for distributions with short tails and the Lepage-type test $LP4$ for those with medium up to long tails. In the case of different shapes we would prefer the CM- or $CM1$-test. If both distributions are extremely right-skewed, in general we recommend the $KS2$-test.

A more convincing way for selecting an appropriate test for a given data set is to apply an adaptive test based on Hogg's concept, see Hogg (1974)), i.e. first, to classify the unknown distribution function with respect to the measures of skewness and tailweight for the combined two samples, $\hat{S}$ and $\hat{T}$, and second, to use an appropriate test for this classified type of distribution. It can be shown that this two-staged adaptive test is, under H_0, distribution-free over the class of all continuous distribution functions, see Randles & Wolfe (1979, p. 388) and Büning (1991, p. 219). Büning & Thadewald (2000) proposed such an adaptive test for location-scale alternatives and Büning (2001b) suggested an adaptive test for general alternatives.

REFERENCES

ARRENBERG, J. (1996) A note on the nonparametric test based on the -version of the Cramér-von Mises statistic, *Statistical Papers*, 37, pp. 95-104.

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

BROWN, M. B. & FORSYTHE, A. B. (1974) The small sample behavior of some statistics which test the equality of several means, *Technometrics*, 16, pp. 129-132.

BÜNING, H. (1991) *Robuste und adaptive Tests* (Berlin, De Gruyter).

BÜNING, H. (1997) Robust analysis of variance, *Journal of Applied Statistics*, 24, pp. 319-332.

BÜNING, H. (2001a) Kolmogorov-Smirnov and Cramér-von Mises type two-sample tests with various weight functions, *Communications in Statistics—Simulation and Computation*, 30, pp. 847-865.

BÜNING, H. (2001b) An adaptive distribution-free test for the general two-sample problem, *Diskussionsbeiträge des Fachbereichs Wirtschaftswissenschaft der Freien Universität Berlin*, Volkswirtschaftliche Reihe, Nr.2001/07, to be published in *Computational Statistics*.

BÜNING, H. & TRENKLER, G. (1994) *Nichtparametrische statistische Methoden* (Berlin, De Gruyter).

BÜNING, H. & CHAKRABORTI, S. (1999) Power comparison of several two-sample tests for general alternatives, *Allgemeines Statistisches Archiv*, 83, pp. 190-210.

BÜNING, H. & THADEWALD, Th. (2000) An adaptive two-sample location-scale test of Lepage-type for symmetric distributions, *Journal of Statistical Computation and Simulation*, 65, pp. 287-310.

BURR, E. J. (1963) Distribution of the two-sample Cramér-von Mises criterion for small equal samples, *Annals of Mathematical Statistics*, 35, pp. 95-101.

CANNER, P. L. (1975) A simulation study of one- and two-sample Kolmogorov-Smirnov statistics with a particular weight function, *Journal of the American Statistical Association*, 70, pp. 209-211.

CONOVER, W. J. & IMAN, R. L. (1981) Rank transformation as a bridge between parametric and nonparametric statistics, *The American Statistician,* 35, pp. 124-133.

FLIGNER, M. A. & POLICELLO II, G. E. (1981) Robust rank procedures for the Behrens-Fisher problem, *Journal of the American Statistical Association,* 76, pp. 162-168.

FUEDA, K. & ÔHORI, K. (1995) Versatile two-sample rank tests based on Wilcoxon test, *Bulletin of Informatics and Cybernetics*, 27, pp. 159-164.

GASTWIRTH, J. L. (1965) Percentile modifications of two sample rank tests, *Journal of the American Statistical Association,* 60, pp. 1127-1141.

GORIA, M. N. (1982) A survey of two-sample location-scale problem, asymptotic relative efficiency of some rank tests, *Statistica Neerlandica*, 36, pp. 3-13.

HOGG, R. V. (1974) Adaptive robust procedures. A partial review and some suggestions for future applications and theory, *Journal of the American Statistical Association,* 69, pp. 909-927.

HUBER, P. J. (1964) Robust estimation of a location parameter, *Annals of Mathematical Statistics*, 35, pp. 73-101.

LEINHARDT, S. & WASSERMAN, S. S. (1979) Teaching regression: an exploratory approach, *The American Statistician*, 33, pp. 196-203.

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

LEVENE, H. (1960) Robust tests for equality of variances, *In:* I. Olkin (Ed.) *Contributions to Probability and Statistics* (Palo Alto, Stanford University Press), pp. 278-292.

MAGEL, R. C. & WIBOWO, S. H. (1997) Comparing the powers of the Wald-Wolfowitz and Kolmogorov-Smirnov tests, *Biometrical Journal*, 39, pp. 665-675.

MILLER, R. G. (1968) Jackknifing variances, *Annals of Mathematical Statistics*, 39, pp. 567-582.

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, pp. 747-758.

RANDLES, R. H. & WOLFE, D. A. (1979) *Introduction to the Theory of Nonparametric Statistics* (New York, Wiley).

STAUDTE, R. G. & SHEATHER, S. J. (1990) *Robust Estimation and Testing* (New York, Wiley).

TIKU, M. L., TAN, W. Y. & BALAKRISHNAN, N. (1986) *Robust Inference* (New York, Marcel Dekker).

WELCH, B. L. (1937) The significance of the difference between two means when the population variances are unequal, *Biometrika*, 29, pp. 350-362.

WILCOX, R. R. (1989) Percentage points of a weighted Kolmogorov-Smirnov statistic, *Communications in Statistics- Simulation and Computation*, 18, pp. 237-244.

WILCOX, R. R. (1997) *Introduction to Robust Estimation and Hypothesis Testing* (London, Academic Press).

YUEN, K. K. (1974) The two-sample trimmed t for unequal population variances, *Biometrika*, 61, pp. 165-170.