STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–15.
Published online in International Academic Press (www.IAPress.org)

Robust tests for the Behrens-Fisher problem when the underlying distribution is short-tailed symmetric: An application to dopamine and schizophrenia data

Gamze Guven [1,*], Birdal Senoglu [2]

[1] *Department of Statistics, Faculty of Science, Eskisehir Osmangazi University, Turkey*
[2] *Department of Statistics, Faculty of Science, Ankara University, Turkey*

Abstract
In this study, the robust versions of the well-known Welch (*W*) and generalized *p*-value (*GP*) tests, i.e., robust Welch (*RW*) and robust generalized *p*-value (*RGP*) tests, are proposed for testing the equality of two means when the underlying distribution is short-tailed symmetric (*STS*) and variances are unknown and arbitrary. They are based on modified maximum likelihood (*MML*) estimators which have closed forms and are approximately equivalent to the maximum likelihood (*ML*) estimators. Under various scenarios, the proposed and existing tests are compared in terms of Type I error rates and powers via a Monte Carlo simulation study in R software program. Also, robustness of the proposed tests is investigated. Simulation results indicate that proposed tests perform as well as or better than the *W* and *GP* tests, while they satisfactorily control the Type I error rates. Moreover, *RW* and *RGP* tests are generally more robust to departures from the assumed model than their normal-theory counterparts. Finally, a real data set taken from psychology literature is used for illustrative purposes.

Keywords Behrens-Fisher Problem, STS Distribution, MML Estimators, Monte Carlo Simulation, Robustness

AMS 2010 subject classifications 62F03, 62F12, 62F35

DOI: 10.19139/soic-2310-5070-3522

1. Introduction

In the context of statistical applications in various fields, one of the most important problems is to test the equality of two normal population means when the variances are unknown and possibly unequal. This problem is known as the Behrens-Fisher (*BF*) problem. In literature various solutions and tests have been developed for the *BF* problem.

Both within the Bayesian and fiducial frameworks, Fisher [6] suggested a solution for the *BF* problem employing the concept of fiducial distributions. Kim and Cohen [13] reviewed *BF* problem under Fisher's fiducial, Jeffreys' Bayesian and frequentist approaches.

From the frequentist point of view, Welch [29] proposed an approximate solution to the *BF* problem through a test statistic which is approximately distributed as Student's *t*. Several alternative tests were later developed. For example, Saxena and Srivastava [17] proposed a competitor test to the well-known *t* test using the Jackknife estimator of the common population variance. Best and Rayner [4] proposed tests based on Wald, likelihood ratio and score statistics and compared them with the Welch's test. However, they stated that these tests are very similar to the Welch's test in terms of power performance and so they recommended Welch's test. Gupta and Wang [8] modified the Welch's test for a given significance level α to get a test with size α. Singh et al. [19] compared the test

*Correspondence to: Gamze guven (Email: gamzeguven@ogu.edu.tr). Department of Statistics, Faculty of Science, Eskisehir Osmangazi University, Meselik Campus, 26040 Eskisehir, Turkey.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

proposed by Saxena and Srivastava with Welch's test and Cochran-Cox test. Paul et al. [15] made a comprehensive review of the existing tests for the *BF* problem and compared them in terms of size and power. Chen et al. [5] proposed a solution based on the method of variance estimates recovery for the *BF* problem.

Under the generalized *p*-value framework, Tsui and Weerahandi [26] provided a solution to hypothesis testing problems when nuisance parameters are present by developing generalized *p*-value method. Also, they provided an exact solution to the *BF* problem. According to their results *p*-value for the *BF* problem is numerically the same as Jeffreys's Bayesian solution and the *BF* fiducial solution. Witkovský [30] defined a generalized *p*-value test for the *BF* problem.

Based on the studies reviewed so far, *BF* problem has been widely investigated under the normality assumption. Hence, least squares (*LS*) estimators have been used in the above-mentioned test statistics. Different from the earlier studies, this study considers *BF* problem when the underlying distribution is short-tailed symmetric (*STS*) introduced by Tiku and Vaughan [25]. *STS* distribution has received less attention than normal or heavy-tailed distributions in the literature, but it is encountered in many fields such as botany (Iris Setosa data), medicine (coronary heart disease age data and symptom score data), experimental physiology (finger-tapping rate data) and industry (reinforcing bar strength data) and it can also be used for modeling data sets containing inliers, see [27, 2, 3, 9] and references therein for details. Therefore, *STS* distribution is used as an alternative to the normal distribution for data sets that exhibit short-tailed behavior.

It should be remembered that when the normality assumption is not met, the efficiencies of *LS* estimators may decrease and consequently it may cause to decrease in the power of the proposed tests based on them. Therefore, in this study two tests based on modified maximum likelihood (*MML*) estimators are proposed. *MML* methodology is developed by Tiku [21, 22]. These two tests are robust versions of the well-known Welch test and generalized *p*-value test and in the following sections they are called robust Welch (*RW*) test and robust generalized *p*-value (*RGP*) test, respectively. The reason for calling them as "robust" is that *MML* estimators are insensitive to the inliers which are erroneous observations positioned close to the mean, see for example Akkaya and Tiku [1] and Senoglu [18] for details.

To the best of our knowledge, there has been no previous work developing the robust versions of the Welch and generalized p-value tests for the *BF* problem when the underlying distribution is *STS*.

The study is organized as follows. In Section 2, *STS* distribution and *MML* estimators of the *STS* distribution parameters are given. In Section 3, *RW* and *RGP* tests based on *MML* estimators are defined. The Monte Carlo simulation study comparing the *RW* and *RGP* tests with the corresponding tests based on *LS* estimators in terms of simulated Type I error rates and power is presented in Section 4. A real data set is analyzed for illustrative purposes in Section 5. Finally, concluding remarks are given in Section 6.

2. Parameter Estimation for the STS Distribution

In this section, *STS* distribution and *MML* estimators of the parameters of the *STS* distribution are presented.

2.1. STS Distribution

The *STS* distribution was proposed by Tiku and Vaughan [25]. The probability density function (*pdf*) of the *STS* distribution is

$$f(x)=\frac{C}{\sqrt{2\pi}\sigma}\left\{1+\frac{\lambda}{2r}\left(\frac{x-\mu}{\sigma}\right)^2\right\}^r\exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\},\quad -\infty<x<\infty, \tag{1}$$

where r is a constant, $\lambda=r/(r-d)$, $d<r$, and

$$C=\left[\sum_{j=0}^{r}\binom{r}{j}\left(\frac{\lambda}{2r}\right)^j\frac{(2j)!}{2^j j!}\right]^{-1}. \tag{2}$$

The mean and variance of X are given by

$$E(X) = \mu \quad \text{and} \quad V(X) = C \sum_{j=0}^{r} \binom{r}{j} \left(\frac{\lambda}{2r}\right)^{j} \frac{(2(j+1))!}{2^{j+1}(j+1)!}\sigma^2, \tag{3}$$

respectively. It should be noted that λ and therefore the values of the r and d, determine the shapes of the *STS* distribution curves. The *pdf*s of the *STS* distribution for selected certain values of d are plotted in Figures 1 and 2.

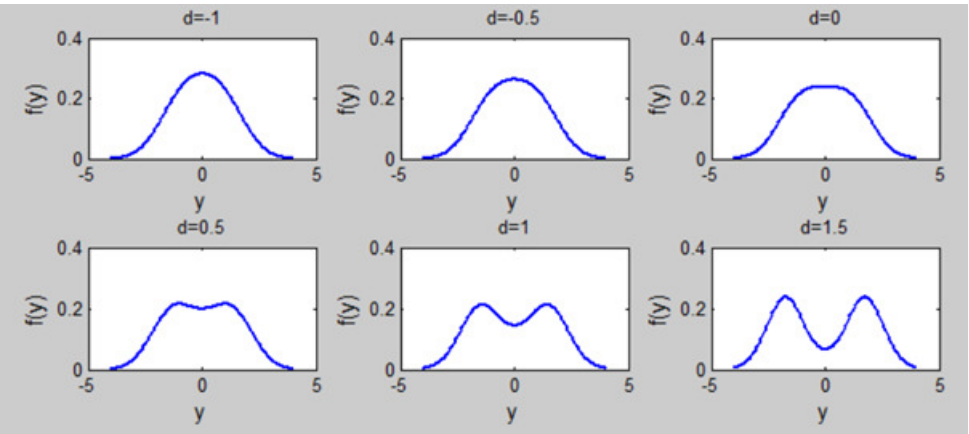

Figure 1. The *pdf* plots of the *STS* distribution for certain values of d when r = 2.

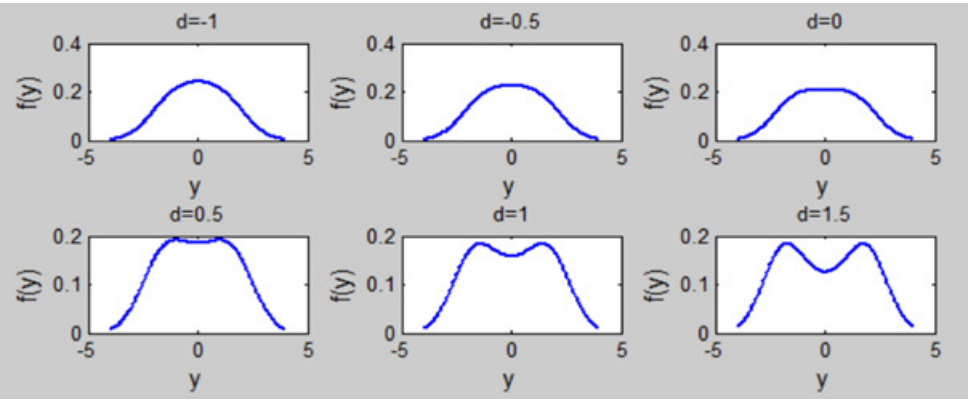

Figure 2. The *pdf* plots of the *STS* distribution for certain values of d when r = 4.

It can be seen from Figures 1 and 2 that the distributions are unimodal when $d \leq 0$, while they are generally multimodal when $d > 0$. In addition, kurtosis values of the *STS* distribution are shown in Table 1.

Table 1. The kurtosis values of the *STS* distribution for certain values of d when $r = 2$ and 4.

$d =$	-1	-0.5	0.0	0.5	1.0	1.5	2.5	3.5
				$r = 2$				
	2.648	2.559	2.437	2.265	2.026	1.711	–	–
				$r = 4$				
	2.541	2.464	2.370	2.255	2.118	1.957	1.591	1.297

In Table 1, the dashed entries are used for $d > r$ since the kurtosis values are defined when $d < r$ and it is clearly seen that the kurtosis values are less than 3.

2.2. MML Estimators

Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ $(i = 1, 2)$ be random samples from the *STS* distributions with parameters μ_i and σ_i. The likelihood (L) and log-likelihood $(\ln L)$ functions are given by

$$L = \left(\frac{C}{\sqrt{2\pi}}\right)^N \prod_{i=1}^{2} \left(\frac{1}{\sigma_i}\right)^{n_i} \prod_{i=1}^{2}\prod_{j=1}^{n_i} \left\{1 + \frac{\lambda}{2r}\left(\frac{x_{ij}-\mu_i}{\sigma_i}\right)^2\right\}^r \exp\left\{-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{x_{ij}-\mu_i}{\sigma_i}\right)^2\right\} \tag{4}$$

and

$$\begin{aligned}\ln L = N[\ln(C) - \ln(\sqrt{2\pi})] - \sum_{i=1}^{2} n_i \ln(\sigma_i) + \sum_{i=1}^{2}\sum_{j=1}^{n_i} r \ln\left\{1 + \frac{\lambda}{2r}\left(\frac{x_{ij}-\mu_i}{\sigma_i}\right)^2\right\} \\ - \frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}\left(\frac{x_{ij}-\mu_i}{\sigma_i}\right)^2,\end{aligned} \tag{5}$$

respectively. Here, $N = n_1 + n_2$.

Derivatives of the Equation (5) with respect to the parameters μ_i and σ_i are given as

$$\frac{\partial \ln L}{\partial \mu_i} = -\frac{\lambda}{\sigma_i}\sum_{j=1}^{n_i} g(z_{ij}) + \frac{1}{\sigma_i}\sum_{j=1}^{n_i} z_{ij} = 0, \tag{6}$$

$$\frac{\partial \ln L}{\partial \sigma_i} = -\frac{n_i}{\sigma_i} - \frac{\lambda}{\sigma_i}\sum_{j=1}^{n_i} z_{ij} g(z_{ij}) + \frac{1}{\sigma_i}\sum_{j=1}^{n_i} z_{ij}^2 = 0, \tag{7}$$

respectively. Here, the standardized observations z_{ij} and the function $g(z_{ij})$ are as follows

$$z_{ij} = \frac{x_{ij}-\mu_i}{\sigma_i} \quad \text{and} \quad g(z_{ij}) = \frac{z_{ij}}{1 + \frac{\lambda}{2r} z_{ij}^2}. \tag{8}$$

From Equations (6) and (7) we get the maximum likelihood (*ML*) estimators of μ_i and σ_i. Obviously, these equations involve the nonlinear function $g(z)$, therefore the solutions can be obtained by using the iterative procedures. However, some convergence problems are encountered when solving these equations. To overcome these problems, *MML* methodology proposed by Tiku [21, 22] is used and the following *MML* estimators of the parameters μ_i and σ_i are obtained

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} \beta_{ij} x_{i(j)}}{m_i} \quad \text{and} \quad \hat{\sigma}_i = \frac{-B_i + \sqrt{B_i^2 + 4n_i C_i}}{2\sqrt{n_i(n_i-1)}}, \tag{9}$$

respectively, where

$$m_i = \sum_{j=1}^{n_i} \beta_{ij}, \ \beta_{ij} = 1 - \lambda\gamma_{ij}, \ \gamma_{ij} = \frac{1 - \frac{\lambda}{2r} t_{i(j)}^2}{\left(1 + \frac{\lambda}{2r} t_{i(j)}^2\right)^2}, \ B_i = \lambda \sum_{j=1}^{n_i} \alpha_{ij} x_{i(j)}, \ \alpha_{ij} = \frac{\frac{\lambda}{r} t_{i(j)}^3}{\left(1 + \frac{\lambda}{2r} t_{i(j)}^2\right)^2}$$

and

$$C_i = \sum_{j=1}^{n_i} \beta_{ij} \left(x_{i(j)} - \hat{\mu}_i\right)^2,$$

see for example Senoglu [18], Balcı and Akkaya [3] and Güven [9] for more detailed information.

It should be noted that α_{ij}, γ_{ij}, and β_{ij} are used as given above when $d < 0$. However, some of the β_{ij} coefficients may become negative when $d > 0$, in which case $\hat{\sigma}_i$ can cease to be real. To overcome this difficulty,

$$\alpha_{ij}^{*} = \frac{\frac{\lambda}{r} t_{i(j)}^{3} + \left(1 - \frac{1}{\lambda}\right) t_{i(j)}}{\left(1 + \frac{\lambda}{2r} t_{i(j)}^{2}\right)^{2}}, \quad \gamma_{ij}^{*} = \frac{\frac{1}{\lambda} - \left(1 - \frac{\lambda}{2r}\right) t_{i(j)}^{2}}{\left(1 + \frac{\lambda}{2r} t_{i(j)}^{2}\right)^{2}} \quad \text{and} \quad \beta_{ij}^{*} = 1 - \lambda \gamma_{ij}^{*}$$

are used in place of α_{ij}, γ_{ij}, and β_{ij}, respectively and then MML estimators are computed from Equation (9), see Tiku and Akkaya [23] for details.

In addition, approximate values of $t_{i(j)}$ are given by

$$\int_{-\infty}^{t_{i(j)}} \frac{C}{\sqrt{2\pi}} \left(1 + \frac{\lambda}{2r} z^{2}\right)^{r} \exp\left(-\frac{1}{2} z^{2}\right) dz = \frac{j}{n_i + 1}, \quad j = 1, \ldots, n_i.$$

Note that *MML* estimators have closed form and also they satisfy the properties of *ML* estimators asymptotically.

Since the performances of the proposed tests based on *MML* estimators are compared with the corresponding tests based on *LS* estimators in the Monte Carlo simulation study, *LS* estimators of μ_i and σ_i (i.e., $\tilde{\mu}_i$ and $\tilde{\sigma}_i$) are given as follows

$$\tilde{\mu}_i = \bar{x}_i \quad \text{and} \quad \tilde{\sigma}_i = \sqrt{s_i^{2} \Big/ C \sum_{j=0}^{r} \binom{r}{j} \left(\frac{\lambda}{2r}\right)^{j} \frac{(2(j+1))!}{2^{j+1}(j+1)!}}, \tag{10}$$

where $\bar{x}_i$ and s_i^2 are the sample mean and sample variance for the ith sample $(i = 1, 2)$.

Here, it should be noted that the robustness of the *MML* estimators is due to the coefficients β_{ij}. Since β_{ij} coefficients used in calculating the *MML* estimators have inverted umbrella ordering, inliers which are "bad" observations located close to mean receive small weights. Thus, influence of potential inliers is mitigated, and it can be said that the *MML* estimators are more robust to inliers than the *LS* estimators. This feature is illustrated in the application given in Section 5.

3. RW and RGP Tests

Let $X_{i1}, X_{i2}, \ldots, X_{in_i}$ $(i = 1, 2)$ be random samples from the *STS* distributions with parameters μ_i and σ_i. In this section, robust versions of the well-known Welch and generalized p-value tests are obtained for the *BF* problem of testing the hypothesis

$$H_0 : \mu_1 = \mu_2 \tag{11}$$

when σ_1^2 and σ_2^2 are not known and hence possibly unequal.

3.1. RW Test

Welch [29] proposed a test statistic based on *LS* estimators when the underlying distribution is normal. This test statistic is approximately distributed as Student's t. The Welch test is alternatively referred as Welch-Satterthwaite test since the degrees of freedom obtained by Welch can be derived using the Satterthwaite [16] approximation.

Robust version of the Welch test called as *RW* is proposed as follows

$$RW = \frac{(\hat{\mu}_1 - \hat{\mu}_2) - (\mu_1 - \mu_2)}{\sqrt{(\hat{\sigma}_1^{2}/m_1) + (\hat{\sigma}_2^{2}/m_2)}} \tag{12}$$

when the underlying distribution is *STS*. Here, $\hat{\mu}_i$ and $\hat{\sigma}_i$ $(i = 1, 2)$ are given in Equation (9).

The approximate null distribution of RW in (12) is Student's t with approximate degrees of freedom

$$\nu = \frac{\left((\hat{\sigma}_1^{2}/m_1) + (\hat{\sigma}_2^{2}/m_2)\right)^{2}}{(\hat{\sigma}_1^{2}/m_1)^{2}/(n_1 - 1) + (\hat{\sigma}_2^{2}/m_2)^{2}/(n_2 - 1)}, \tag{13}$$

see Lemmas 1 and 2 given below.

For $d \leq 0$ the following asymptotic results hold for the estimators.

Lemma 1. The estimator $\hat{\mu}_i$ is the minimum variance bound (MVB) estimator and is asymptotically normally distributed with mean μ_i and variance σ_i^2/m_i.

Lemma 2. For large n_i, $(n_i-1)\hat{\sigma}_i^2/\sigma_i^2$ is asymptotically distributed as a multiple of a chi-square random variable with n_i-1 degrees of freedom.

See Tiku and Akkaya [23] for proofs of Lemmas

Similar to the earlier studies about normal theory, here the Satterthwaite approximation is used to obtain the degrees of freedom for the proposed RW test. Finally, the p-value for RW test is given by

$$p = Pr(|T_\nu| \geq |RW_{obs}|). \tag{14}$$

Here, T_ν denotes a Student's *t* random variable with ν degrees of freedom, and RW_{obs} is the observed value of the *RW* test statistic computed from data. Also, α denotes the prespecified nominal level. If $p < \alpha$, then H_0 in (11) is rejected.

3.2. RGP Test

The concept of generalized *p*-value introduced byWeerahandi [28] and Tsui and Weerahandi [26] is used in various studies available in the literature. For example, Witkovský [30] defined a generalized *p*-value test for the *BF* problem when the underlying distribution is normal. Krishnamoorthy and Mathew [14] derived exact inference procedures concerning the mean of a single lognormal distribution and for the ratio of the means of two independent lognormal distributions using the idea of generalized *p*-value. Gamage et al. [7] developed a procedure based on the concept of generalized *p*-value for testing the equality of the mean vectors of two multivariate normal populations with unequal covariance matrices. Xu and Wang [31] defined a new generalized test variable and the generalized *p*-value based on this generalized test variable for the *ANOVA* problem.

In this subsection, robust version of the generalized *p*-value test called as *RGP* is obtained in a similar manner with normal theory proposed by Witkovský [30] when the underlying distribution is *STS*. First generalized test variable based on *MML* estimators are defined and then robust generalized *p*-value is obtained.

The generalized test variable based on *MML* estimators can be given as follows

$$GT^2 = \frac{((\hat{\mu}_1-\hat{\mu}_2)-(\mu_1-\mu_2))^2}{(\hat{\sigma}_1^2/m_1)+(\hat{\sigma}_2^2/m_2)}\left(\frac{\sigma_1^2\hat{\sigma}_{1(\text{obs})}^2}{m_1\hat{\sigma}_1^2}+\frac{\sigma_2^2\hat{\sigma}_{2(\text{obs})}^2}{m_2\hat{\sigma}_2^2}\right) \tag{15}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i$ $(i=1,2)$ are given in Equation (9). In addition, $\hat{\sigma}_{1(\text{obs})}^2$ and $\hat{\sigma}_{2(\text{obs})}^2$ are the observed values of $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$, respectively.

Note that the observed value of (15) is as follows

$$GT_{obs}^2 = \left(\hat{\mu}_{1(\text{obs})}-\hat{\mu}_{2(\text{obs})}\right)^2 \tag{16}$$

where $\hat{\mu}_{1(\text{obs})}$ and $\hat{\mu}_{2(\text{obs})}$ are the observed values of $\hat{\mu}_1$ and $\hat{\mu}_2$, respectively.

The null distribution of (15) is given as

$$GT^2 \sim u\left(\frac{(n_1-1)\hat{\sigma}_{1(\text{obs})}^2}{m_1u_1}+\frac{(n_2-1)\hat{\sigma}_{2(\text{obs})}^2}{m_2u_2}\right) \tag{17}$$

where u, u_1 and u_2 are independent chi-square random variables with 1, (n_1-1) and (n_2-1) degrees of freedom, respectively, see Lemmas 1 and 2 given above.

The *p*-value for *RGP* test is calculated using the following Algorithm 1.

Algorithm 1 Computation of the p-value for *RGP* test

Step 1: Compute $\hat{\mu}_1, \hat{\mu}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2$ and GT_{obs}^2 for given data.

Step 2: Generate random numbers from $u \sim \chi_1^2$, $u_1 \sim \chi_{n_1-1}^2$ and $u_2 \sim \chi_{n_2-1}^2$ and compute GT^2 in (17).

Step 3: Let $B = 1$ if $GT^2 > GT_{obs}^2$ else $B = 0$.

Step 4: Repeat the steps 2–3 large number of times M.

Step 6: Compute the p-value as $p = (1/M)\sum_{i=1}^{M} B_i$.

If $p < \alpha$, then H_0 in (11) is rejected.

See also Balcı and Akkaya [3] and Güven [9] for pairwise multiple comparisons and one-way *ANOVA* in the context of *STS* distribution and *MML* methodology, respectively. Moreover, Güven et al. [10] addressed *BF* problem when the underlying distribution is long-tailed symmetric (*LTS*).

4. Monte Carlo Simulation Study

In this section, performances of the proposed *RW* and *RGP* tests are compared with their normal theory counterparts *W* and *GP* tests in terms of the Type I error rates and powers via Monte Carlo simulation study. Simulations are conducted in R software for the sample sizes $(n_1, n_2) = (5,5)$, $(5,10)$, $(10,10)$, $(10,15)$, $(15,20)$, $(25,25)$ and parameter values $(\sigma_1^2, \sigma_2^2) = (1,1)$, $(1,3)$, $(r,d) = (2,-1)$ and $(r,d) = (2,0)$. As reported in Table 1, kurtosis values of the *STS* distributions are 2.648 and 2.437 when $(r,d) = (2,-1)$ and $(2,0)$, respectively. These two cases are selected in this section to illustrate the performance of the proposed tests under different degrees of short-tailedness.

Table 2. Simulated Type I error rates of the *RW*, *RGP*, *W* and *GP* tests.

$(r,d) = (2,-1)$								
	$(\sigma_1^2, \sigma_2^2) = (1,1)$				$(\sigma_1^2, \sigma_2^2) = (1,3)$			
(n_1, n_2)	*RW*	*RGP*	*W*	*GP*	*RW*	*RGP*	*W*	*GP*
$(5,5)$	0.048	0.027	0.049	0.027	0.046	0.034	0.052	0.031
$(5,10)$	0.045	0.029	0.044	0.029	0.052	0.034	0.051	0.034
$(10,10)$	0.049	0.036	0.049	0.037	0.047	0.039	0.048	0.041
$(10,15)$	0.051	0.045	0.053	0.046	0.050	0.043	0.051	0.042
$(15,20)$	0.050	0.045	0.052	0.046	0.047	0.045	0.050	0.046
$(25,25)$	0.052	0.049	0.052	0.048	0.050	0.047	0.052	0.050
$(r,d) = (2,0)$								
	$(\sigma_1^2, \sigma_2^2) = (1,1)$				$(\sigma_1^2, \sigma_2^2) = (1,3)$			
(n_1, n_2)	*RW*	*RGP*	*W*	*GP*	*RW*	*RGP*	*W*	*GP*
$(5,5)$	0.046	0.028	0.045	0.027	0.046	0.035	0.052	0.032
$(5,10)$	0.055	0.039	0.050	0.034	0.047	0.033	0.047	0.031
$(10,10)$	0.049	0.040	0.047	0.036	0.048	0.041	0.049	0.040
$(10,15)$	0.051	0.043	0.052	0.041	0.051	0.040	0.051	0.041
$(15,20)$	0.052	0.050	0.052	0.045	0.048	0.047	0.049	0.047
$(25,25)$	0.052	0.050	0.050	0.046	0.051	0.047	0.053	0.052

This simulation study is restricted to $d \leq 0$. It should be emphasized that the *MML* estimators are asymptotically equivalent to the ML estimators and their associated minimum variance bounds (*MVB*) are available for $d \leq 0$.

Furthermore, asymptotic results for *MML* estimators and proposed tests are derived for $d \leq 0$. When $d > 0$, estimating equations may have multiple roots, and *MVB* do not generally exist, see [24, 18] for details.

It should be realized that simulated Type I error rates and powers of the *RW* and *W* tests and *p* values of the *RGP* and *GP* tests are calculated based on 3,500 random samples generated from *STS* distribution using inverse transformation method. Also note that 3,500 replications are used when calculating Type I error rates and powers of the *RGP* and *GP* tests for each of the random samples. When calculating the Type I error rates, μ_1 and μ_2 were taken as equal and 0 without loss of generality, see Table 2 for the results.

When the data followed *STS* distribution with $(r, d) = (2, -1)$ and $(r, d) = (2, 0)$, the *RW* and *W* tests control the Type I rate well. However, the Type I error rates of the *RGP* and *GP* tests are significantly smaller than the nominal level $\alpha = 0.05$ when sample sizes are small (equal or unequal). For moderate sample sizes generalized *p*-value based tests (*RGP* and *GP*) are slightly more conservative than Welch type tests (*RW* and *W*). It should be realized that the Type I error rates of the *RGP* and *GP* tests are getting closer to the nominal level $\alpha = 0.05$ as the sample sizes increase, see for example $(n_1, n_2) = (15, 20)$ and $(25, 25)$.

For calculating simulated power values, μ_1 and μ_2 values are determined so that $|\mu_1 - \mu_2|$ is equal to 2δ, see Tables 3a, 3b, 4a and 4b. Obviously, simulated power values reduce to the Type I error rates when δ is equal to zero.

Table 3a. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests when $(r, d) = (2, -1)$ and $(\sigma_1^2, \sigma_2^2) = (1, 1)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.048	0.027	0.049	0.027	0.00	0.045	0.029	0.044	0.029	0.00	0.049	0.036	0.049	0.037
0.32	0.09	0.06	0.09	0.06	0.27	0.12	0.08	0.11	0.08	0.22	0.11	0.09	0.12	0.09
0.64	0.24	0.16	0.24	0.16	0.54	0.26	0.19	0.26	0.19	0.44	0.30	0.26	0.29	0.25
0.96	0.50	0.37	0.50	0.36	0.81	0.50	0.43	0.50	0.42	0.66	0.56	0.51	0.55	0.50
1.28	0.77	0.65	0.76	0.64	1.08	0.74	0.67	0.74	0.66	0.88	0.81	0.77	0.81	0.77
1.60	0.93	0.86	0.93	0.85	1.35	0.89	0.84	0.89	0.84	1.10	0.95	0.93	0.94	0.93
1.92	0.98	0.96	0.98	0.95	1.62	0.99	0.95	0.98	0.95	1.32	0.99	0.99	0.99	0.99
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.051	0.045	0.053	0.046	0.00	0.050	0.045	0.052	0.046	0.00	0.052	0.049	0.052	0.048
0.20	0.10	0.09	0.10	0.09	0.17	0.11	0.10	0.11	0.09	0.13	0.10	0.10	0.10	0.10
0.40	0.29	0.26	0.29	0.25	0.34	0.32	0.29	0.31	0.29	0.26	0.28	0.27	0.28	0.26
0.60	0.56	0.52	0.55	0.51	0.51	0.61	0.58	0.59	0.57	0.39	0.54	0.52	0.54	0.52
0.80	0.81	0.78	0.80	0.77	0.68	0.85	0.84	0.85	0.83	0.52	0.80	0.78	0.79	0.78
1.00	0.95	0.93	0.95	0.93	0.85	0.96	0.96	0.96	0.95	0.65	0.93	0.92	0.92	0.92
1.20	1.00	1.00	1.00	0.99	1.02	0.99	0.99	0.99	0.99	0.78	0.99	0.99	0.98	0.98

Table 3b. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests when $(r, d) = (2, -1)$ and $(\sigma_1^2, \sigma_2^2) = (1, 3)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.046	0.034	0.052	0.031	0.00	0.052	0.034	0.051	0.034	0.00	0.047	0.039	0.048	0.041
0.49	0.11	0.07	0.11	0.07	0.37	0.11	0.07	0.11	0.07	0.30	0.10	0.08	0.10	0.08
0.98	0.28	0.20	0.28	0.20	0.74	0.29	0.21	0.28	0.22	0.60	0.26	0.24	0.27	0.24
1.47	0.55	0.45	0.55	0.44	1.11	0.59	0.49	0.58	0.49	0.90	0.50	0.46	0.49	0.45
1.96	0.79	0.70	0.78	0.70	1.48	0.85	0.76	0.84	0.76	1.20	0.78	0.75	0.77	0.74
2.45	0.93	0.89	0.93	0.89	1.85	0.96	0.93	0.95	0.92	1.50	0.93	0.92	0.93	0.92
2.94	0.99	0.97	0.98	0.97	2.22	0.99	0.99	0.99	0.99	1.80	0.99	0.98	0.98	0.98
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.050	0.043	0.051	0.042	0.00	0.047	0.045	0.050	0.046	0.00	0.050	0.047	0.052	0.050
0.27	0.09	0.08	0.09	0.08	0.22	0.10	0.09	0.10	0.09	0.17	0.11	0.10	0.10	0.10
0.54	0.30	0.26	0.29	0.26	0.44	0.29	0.27	0.29	0.27	0.34	0.24	0.23	0.24	0.23
0.81	0.59	0.56	0.59	0.55	0.66	0.57	0.55	0.57	0.55	0.51	0.50	0.48	0.48	0.47
1.08	0.83	0.81	0.82	0.80	0.88	0.81	0.79	0.79	0.77	0.68	0.73	0.72	0.71	0.70
1.35	0.96	0.95	0.95	0.94	1.10	0.95	0.94	0.94	0.94	0.85	0.89	0.88	0.88	0.87
1.62	0.99	0.99	0.99	0.99	1.32	0.99	0.99	0.99	0.99	1.02	0.98	0.98	0.97	0.97

It can be seen from Tables 3a and 3b, powers of the proposed *RW* and *RGP* tests are quite close to those of their normal-theory counterparts *W* and *GP* tests, respectively. However, *RW* and *W* tests are more powerful than the corresponding *RGP* and *GP* tests for sample sizes $(n_1, n_2) = (5, 5)$ and $(5, 10)$ whether the variances are equal or unequal, but this difference becomes less pronounced as the sample size increases.

Table 4a. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests when $(r, d) = (2, 0)$ and $(\sigma_1^2, \sigma_2^2) = (1, 1)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.046	0.028	0.045	0.027	0.00	0.055	0.039	0.050	0.034	0.00	0.049	0.040	0.047	0.036
0.38	0.11	0.07	0.11	0.07	0.31	0.11	0.08	0.11	0.08	0.24	0.10	0.09	0.11	0.09
0.76	0.30	0.20	0.29	0.20	0.62	0.29	0.22	0.27	0.21	0.48	0.31	0.27	0.28	0.24
1.14	0.58	0.44	0.56	0.42	0.93	0.55	0.46	0.52	0.44	0.72	0.60	0.54	0.56	0.51
1.52	0.85	0.73	0.83	0.71	1.24	0.81	0.73	0.77	0.70	0.96	0.83	0.79	0.79	0.76
1.90	0.97	0.93	0.96	0.91	1.55	0.94	0.91	0.93	0.89	1.20	0.96	0.94	0.94	0.92
2.28	1.00	0.99	0.99	0.98	1.86	0.99	0.98	0.99	0.98	1.44	0.99	0.99	0.99	0.98
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.051	0.043	0.052	0.041	0.00	0.052	0.050	0.052	0.045	0.00	0.052	0.050	0.050	0.046
0.22	0.11	0.10	0.11	0.09	0.17	0.11	0.10	0.10	0.09	0.13	0.11	0.11	0.10	0.09
0.44	0.30	0.26	0.28	0.25	0.34	0.29	0.26	0.27	0.25	0.26	0.27	0.25	0.25	0.23
0.66	0.60	0.57	0.57	0.53	0.51	0.56	0.54	0.52	0.50	0.39	0.52	0.51	0.49	0.47
0.88	0.85	0.83	0.81	0.78	0.68	0.80	0.78	0.77	0.75	0.52	0.75	0.73	0.70	0.69
1.10	0.96	0.95	0.95	0.93	0.85	0.94	0.93	0.91	0.90	0.65	0.91	0.90	0.88	0.87
1.32	1.00	1.00	0.99	0.99	1.02	0.99	0.99	0.98	0.98	0.78	0.98	0.98	0.96	0.96

Table 4b. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests when $(r, d) = (2, 0)$ and $(\sigma_1^2, \sigma_2^2) = (1, 3)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.046	0.035	0.052	0.032	0.00	0.047	0.033	0.047	0.031	0.00	0.048	0.041	0.049	0.040
0.55	0.11	0.08	0.10	0.07	0.40	0.11	0.09	0.10	0.08	0.35	0.12	0.10	0.11	0.09
1.10	0.28	0.21	0.25	0.19	0.80	0.30	0.23	0.27	0.20	0.70	0.30	0.28	0.28	0.25
1.65	0.56	0.47	0.52	0.45	1.20	0.58	0.50	0.56	0.46	1.05	0.61	0.57	0.59	0.55
2.20	0.81	0.75	0.76	0.72	1.60	0.85	0.79	0.83	0.75	1.40	0.84	0.83	0.82	0.79
2.75	0.95	0.93	0.93	0.91	2.00	0.96	0.94	0.96	0.92	1.75	0.96	0.96	0.95	0.94
3.30	0.99	0.98	0.98	0.98	2.40	0.99	0.99	0.99	0.98	2.10	0.99	0.99	0.99	0.99
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.051	0.040	0.051	0.041	0.00	0.048	0.047	0.049	0.047	0.00	0.051	0.047	0.053	0.052
0.30	0.12	0.10	0.12	0.10	0.25	0.11	0.10	0.11	0.10	0.20	0.10	0.10	0.10	0.09
0.60	0.33	0.31	0.32	0.28	0.50	0.33	0.31	0.31	0.29	0.40	0.30	0.28	0.28	0.27
0.90	0.62	0.58	0.58	0.54	0.75	0.62	0.60	0.57	0.55	0.60	0.56	0.55	0.53	0.52
1.20	0.86	0.84	0.83	0.81	1.00	0.87	0.86	0.84	0.82	0.80	0.81	0.80	0.76	0.75
1.50	0.97	0.96	0.95	0.94	1.25	0.97	0.97	0.96	0.95	1.00	0.94	0.94	0.92	0.92
1.80	0.99	0.99	0.99	0.99	1.50	0.99	0.99	0.99	0.99	1.20	0.99	0.99	0.98	0.98

Tables 4a and 4b show that *RW* and *RGP* tests have slightly higher power than *W* and *GP*, respectively for all sample sizes. Similar to the results in Tables 3a and 3b, the *RW* and *W* tests outperform the corresponding *RGP* and *GP* tests for sample sizes $(n_1, n_2) = (5, 5)$ and $(5, 10)$ while this difference becomes less pronounced for other sample sizes.

Robustness: In most of the real-life problems, it is difficult to feel comfortable that the samples come from the assumed model, therefore robustness of the proposed tests to the misspecification of the assumed model should be examined. In this section, the population model is assumed to be *STS* with $(r, d) = (2, 0)$ and the plausible alternatives to the assumed model, i.e., sample models, are taken as follows

Model I: *STS* distribution with $(r, d) = (2, 1)$.

Model II: Tukey lambda-family defined by the transformation $z = (u^l - (1 - u)^l)/l$ with $l = 1.45$ where u denotes Uniform $(0, 1)$, see Joiner and Rosenblatt [12].

In this robustness study, sample sizes and variance settings are kept the same as described at the beginning of Section 4. Data-generating processes are described as follows: Data sets in two groups are generated from the sample models instead of assumed *STS* model with $(r, d) = (2, 0)$. In other words, the data sets are generated from *STS* distribution with $(r, d) = (2, 1)$ under *Model I* whereas they are generated from the Tukey lambda distribution with $\lambda = 1.45$ under *Model II*. However, all computations in the proposed *RW* and *RGP* tests are still performed under the assumed model *STS* with $(r, d) = (2, 0)$.

See Tables 5a, 5b, 5c and 5d given below for the simulated Type I error rates and powers of the proposed *RW*, and *RGP* tests and their normal theory counterparts *W* and *GP* tests for the sample models given above.

Table 5a. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests for the plausible alternatives to *STS* distribution with $(r, d) = (2, 0)$ under Model I and $(\sigma_1^2, \sigma_2^2) = (1, 1)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.036	0.023	0.044	0.025	0.00	0.046	0.032	0.051	0.035	0.00	0.035	0.028	0.052	0.038
0.40	0.07	0.05	0.09	0.06	0.34	0.09	0.07	0.10	0.07	0.27	0.08	0.07	0.10	0.08
0.80	0.20	0.14	0.22	0.14	0.68	0.22	0.16	0.22	0.17	0.54	0.26	0.23	0.26	0.21
1.20	0.44	0.34	0.45	0.33	1.02	0.48	0.38	0.44	0.36	0.81	0.59	0.52	0.52	0.47
1.60	0.73	0.62	0.72	0.58	1.36	0.79	0.69	0.73	0.63	1.08	0.85	0.81	0.78	0.73
2.00	0.92	0.87	0.91	0.82	1.70	0.94	0.90	0.90	0.84	1.35	0.97	0.96	0.94	0.91
2.40	0.99	0.98	0.98	0.95	2.04	0.99	0.98	0.98	0.96	1.62	0.99	0.99	0.98	0.98
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.031	0.025	0.048	0.039	0.00	0.032	0.028	0.052	0.048	0.00	0.028	0.027	0.049	0.046
0.22	0.07	0.06	0.09	0.08	0.19	0.07	0.07	0.09	0.08	0.16	0.08	0.07	0.10	0.09
0.44	0.22	0.19	0.22	0.20	0.38	0.25	0.23	0.25	0.22	0.32	0.27	0.25	0.25	0.24
0.66	0.46	0.42	0.42	0.38	0.57	0.53	0.50	0.47	0.44	0.48	0.57	0.56	0.50	0.49
0.88	0.75	0.71	0.68	0.63	0.76	0.81	0.79	0.72	0.70	0.64	0.83	0.81	0.74	0.73
1.10	0.93	0.91	0.87	0.84	0.95	0.96	0.95	0.91	0.89	0.80	0.96	0.96	0.90	0.89
1.32	0.99	0.98	0.96	0.95	1.14	0.99	0.99	0.97	0.96	0.96	0.99	0.99	0.98	0.98

Table 5b. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests for the plausible alternatives to *STS* distribution with $(r, d) = (2, 0)$ under Model I and $(\sigma_1^2, \sigma_2^2) = (1, 3)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.048	0.035	0.051	0.036	0.00	0.038	0.027	0.048	0.034	0.00	0.033	0.026	0.046	0.035
0.57	0.08	0.06	0.09	0.06	0.40	0.08	0.05	0.09	0.06	0.35	0.08	0.07	0.10	0.08
1.14	0.18	0.15	0.21	0.15	0.80	0.20	0.15	0.21	0.16	0.70	0.21	0.19	0.23	0.20
1.71	0.38	0.35	0.43	0.33	1.20	0.45	0.35	0.42	0.33	1.05	0.46	0.46	0.44	0.40
2.28	0.66	0.65	0.71	0.58	1.60	0.75	0.64	0.70	0.59	1.40	0.75	0.73	0.69	0.65
2.85	0.87	0.87	0.89	0.82	2.00	0.91	0.86	0.86	0.80	1.75	0.92	0.90	0.86	0.84
3.42	0.96	0.97	0.98	0.95	2.40	0.99	0.97	0.97	0.94	2.10	0.99	0.98	0.97	0.96
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.036	0.031	0.052	0.045	0.00	0.031	0.026	0.050	0.043	0.00	0.030	0.028	0.045	0.045
0.31	0.08	0.07	0.10	0.08	0.25	0.08	0.07	0.10	0.08	0.22	0.07	0.07	0.10	0.09
0.62	0.23	0.21	0.23	0.20	0.50	0.22	0.21	0.22	0.21	0.44	0.24	0.23	0.24	0.23
0.93	0.53	0.50	0.47	0.44	0.75	0.50	0.48	0.45	0.42	0.66	0.53	0.51	0.46	0.45
1.24	0.80	0.77	0.72	0.69	1.00	0.78	0.75	0.69	0.68	0.88	0.83	0.82	0.74	0.73
1.55	0.95	0.94	0.90	0.88	1.25	0.94	0.93	0.88	0.87	1.10	0.95	0.95	0.89	0.89
1.86	0.99	0.99	0.98	0.97	1.50	0.99	0.99	0.97	0.96	1.32	0.99	0.99	0.97	0.97

Table 5c. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests for the plausible alternatives to *STS* distribution with $(r, d) = (2, 0)$ under Model II and $(\sigma_1^2, \sigma_2^2) = (1, 1)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.039	0.031	0.052	0.034	0.00	0.047	0.031	0.054	0.037	0.00	0.029	0.021	0.049	0.037
0.10	0.07	0.05	0.09	0.05	0.07	0.08	0.06	0.09	0.06	0.06	0.07	0.05	0.09	0.07
0.20	0.20	0.14	0.23	0.14	0.14	0.17	0.12	0.18	0.14	0.12	0.20	0.16	0.22	0.19
0.30	0.45	0.34	0.47	0.33	0.21	0.33	0.26	0.34	0.26	0.18	0.47	0.41	0.44	0.39
0.40	0.78	0.65	0.75	0.59	0.28	0.59	0.48	0.55	0.46	0.24	0.78	0.72	0.69	0.64
0.50	0.96	0.92	0.93	0.84	0.35	0.86	0.76	0.78	0.69	0.30	0.95	0.93	0.87	0.83
0.60	0.99	0.99	0.99	0.97	0.42	0.98	0.94	0.92	0.86	0.36	0.99	0.99	0.97	0.95
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.026	0.022	0.045	0.037	0.00	0.027	0.023	0.053	0.046	0.00	0.022	0.020	0.051	0.044
0.05	0.06	0.05	0.08	0.07	0.04	0.05	0.04	0.08	0.07	0.03	0.05	0.04	0.08	0.07
0.10	0.16	0.14	0.19	0.16	0.08	0.16	0.14	0.19	0.17	0.06	0.13	0.12	0.16	0.15
0.15	0.40	0.36	0.39	0.35	0.12	0.40	0.38	0.38	0.35	0.09	0.36	0.34	0.33	0.32
0.20	0.70	0.66	0.62	0.57	0.16	0.69	0.66	0.59	0.56	0.12	0.62	0.60	0.53	0.51
0.25	0.91	0.88	0.80	0.77	0.20	0.90	0.88	0.78	0.76	0.15	0.83	0.82	0.73	0.71
0.30	0.98	0.98	0.92	0.91	0.24	0.98	0.98	0.92	0.91	0.18	0.96	0.95	0.87	0.86

Table 5d. Simulated power values of the *RW*, *RGP*, *W* and *GP* tests for the plausible alternatives to *STS* distribution with $(r, d) = (2, 0)$ under Model II and $(\sigma_1^2, \sigma_2^2) = (1, 3)$.

δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*	δ	*RW*	*RGP*	*W*	*GP*
$(n_1, n_2) = (5, 5)$					$(n_1, n_2) = (5, 10)$					$(n_1, n_2) = (10, 10)$				
0.00	0.043	0.035	0.055	0.036	0.00	0.030	0.023	0.051	0.032	0.00	0.027	0.022	0.050	0.037
0.13	0.07	0.05	0.08	0.06	0.11	0.08	0.06	0.09	0.06	0.08	0.06	0.05	0.09	0.08
0.26	0.15	0.13	0.19	0.14	0.22	0.25	0.19	0.26	0.19	0.16	0.16	0.15	0.20	0.18
0.39	0.32	0.28	0.37	0.28	0.33	0.57	0.44	0.52	0.42	0.24	0.39	0.36	0.39	0.35
0.52	0.57	0.55	0.63	0.51	0.44	0.88	0.79	0.80	0.70	0.32	0.68	0.65	0.62	0.58
0.65	0.83	0.83	0.85	0.75	0.55	0.99	0.97	0.95	0.91	0.40	0.91	0.89	0.81	0.78
0.78	0.99	0.99	0.98	0.93	0.66	0.99	0.99	0.99	0.98	0.48	0.99	0.99	0.95	0.93
$(n_1, n_2) = (10, 15)$					$(n_1, n_2) = (15, 20)$					$(n_1, n_2) = (25, 25)$				
0.00	0.028	0.024	0.051	0.043	0.00	0.025	0.023	0.049	0.042	0.00	0.023	0.021	0.046	0.044
0.07	0.06	0.05	0.09	0.07	0.06	0.06	0.05	0.09	0.08	0.05	0.06	0.05	0.09	0.08
0.14	0.18	0.16	0.20	0.18	0.12	0.20	0.18	0.21	0.20	0.10	0.19	0.18	0.21	0.20
0.21	0.44	0.42	0.41	0.37	0.18	0.46	0.44	0.42	0.39	0.15	0.46	0.45	0.42	0.41
0.28	0.75	0.72	0.65	0.61	0.24	0.79	0.77	0.67	0.65	0.20	0.78	0.77	0.67	0.65
0.35	0.95	0.93	0.84	0.82	0.30	0.96	0.95	0.87	0.85	0.25	0.94	0.93	0.84	0.83
0.42	0.99	0.99	0.95	0.94	0.36	0.99	0.99	0.96	0.96	0.30	0.99	0.99	0.95	0.95

It can be seen from Tables 5a–5d that that *W* test controls Type I error rate very well and *RGP* test is conservative for all the cases considered. Type I error rates of the *RW* test, except for $(n_1, n_2, \sigma_1^2, \sigma_2^2) = (5, 10, 1, 1)$ and $(n_1, n_2, \sigma_1^2, \sigma_2^2) = (5, 5, 1, 3)$, and *GP* test, except for large sample sizes, are smaller than the nominal level $\alpha = 0.05$.

The *RW* and *RGP* tests outperform the *W* and *GP* tests in terms of power except for the case of $(n_1, n_2, \sigma_1^2, \sigma_2^2) = (5, 5, 1, 3)$. Furthermore, *RW* test is more powerful than the *RGP* test for all cases. Consequently, the proposed tests are robust against departures from the assumed model.

5. Application

In this section, the application of proposed tests for *BF* problem is illustrated for dopamine and schizophrenia data obtained from the psychology literature, see Sternberg et al. [20] and Hand et al. [11]. For this purpose, a total of $N = 25$ hospitalized schizophrenic patients were divided into two groups by the hospital staff: psychotic $(n_1 = 10)$ and non-psychotic $(n_2 = 15)$. Samples of cerebrospinal fluid were taken from each patient and assayed for the dopamine b-hydroxylase (DBH) activity. Then, it is desired to investigate whether there is a difference between patient groups in terms of DBH activity. In other words, it is evaluated the null hypothesis

$$H_0 : \mu_1 = \mu_2$$

against the alternative

$$H_1 : \mu_1 \neq \mu_2.$$

The units of the data are nmol/(ml)(h)/(mg) of protein, see Table 6 given below.

Table 6. Dopamine and schizophrenia data $(N = 25)$

Psychotic $(n_1 = 10)$									
0.0150	0.0204	0.0208	0.0222	0.0226	0.0245	0.0270	0.0275	0.0306	0.0320
Nonpsychotic $(n_2 = 15)$									
0.0104	0.0105	0.0112	0.0116	0.0130	0.0145	0.0154	0.0156	0.0170	0.0180
0.0200	0.0200	0.0210	0.0230	0.0252					

Firstly, quantile-quantile (Q-Q) plots are provided for each group (psychotic and nonpsychotic) by plotting the sample quantiles versus theoretical quantiles when $(r, d) = (2, 0)$, see Figure 3.

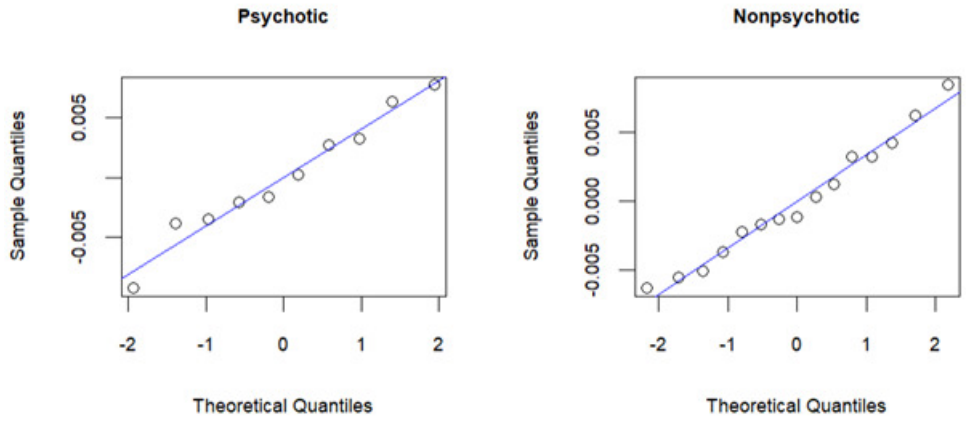

Figure 3. STS Q-Q plots for psychotic and nonpsychotic groups when $(r, d) = (2, 0)$.

It can be seen from Figure 3 that Q-Q plot based on *STS* distribution with $(r, d) = (2, 0)$ closely approximates a straight line for each group, therefore *STS* with $(r, d) = (2, 0)$ fits psychotic and nonpsychotic patient groups.

As it is known that Q-Q plot is a graphical method, therefore, we also use the Kolmogorov-Smirnov (*KS*) test which is well-known goodness of fit test to test the null hypothesis that a sample of data set follows a specified distribution. Calculated values of the *KS* test statistics and the corresponding p-values for each group are given in Table 7.

Table 7. Calculated values of the KS test and the corresponding p-values.

	Test Statistics	p-value
psychotic	0.1155	0.9380
nonpsychotic	0.1235	0.8240

It can be seen from Table 7 that the p-values of the *KS* test are greater than the nominal level $\alpha = 0.05$ for both groups. When these results are evaluated together with the Q-Q plots, distribution of each data set in groups fits well to *STS* distribution with $(r, d) = (2, 0)$, i.e., the data do not contradict the *STS* distribution assumption.

Secondly, the *MML* estimates of the parameters are calculated, see Table 8. Note that formulas required to calculate the value of $\hat{\mu}_i$ and $\hat{\sigma}_i$ $(i = 1, 2)$ are given in Equation (9).

Table 8. The *MML* estimates of the parameters μ_1, μ_2, σ_1 and σ_2.

MML Estimates			
$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1$	$\hat{\sigma}_2$
0.02434	0.01671	0.00357	0.00318

Values of β_{ij}, m_i, B_i and C_i $(i = 1, 2; n_1 = 10; n_2 = 15)$ which are necessary to obtain the $\hat{\mu}_i$ and $\hat{\sigma}_i$ $(i = 1, 2)$ are found as follows:

$$\begin{aligned}
\beta_{1j}: &\ 0.98379 \quad 0.76908 \quad 0.49943 \quad 0.21909 \quad 0.02731 \quad 0.02731 \quad 0.21909 \quad 0.49943 \quad 0.76908 \quad 0.98379,\\
\beta_{2j}: &\ 1.03797 \quad 0.91062 \quad 0.75363 \quad 0.57050 \quad 0.37343 \quad 0.18776 \quad 0.05093 \quad 0.00000 \quad 0.05093 \quad 0.18776\\
&\ 0.37343 \quad 0.57050 \quad 0.75363 \quad 0.91062 \quad 1.03797,
\end{aligned}$$

$$m_1 = 4.99740; \quad m_2 = 7.76968,$$

$$B_1 = 0.02515; \quad B_2 = 0.03693$$

$$C_1 = 0.00020; \quad C_2 = 0.00026.$$

It should be realized that β_{ij} values decrease until the middle value and then increase in a symmetric fashion. Therefore, the middle observations receive small weights and the nature of β_{ij} coefficients make *MML* estimators robust to inliers in a sample.

Finally, the value of the *RW* test statistics, its degrees of freedom and the p-values for the *RW* and *RGP* tests are calculated based on the *MML* estimates given above. They are given as follows

$$RW = 3.89036; \quad \nu = 17.62319; \quad p\text{-value for the } RW \text{ test: } 0.00111; \quad p\text{-value for the } RGP \text{ test: } 0.00314.$$

Since the p-values of the *RW* and *RGP* tests are considerably smaller than the nominal level $\alpha = 0.05$, null hypothesis $H_0 : \mu_1 = \mu_2$ is rejected, i.e., there is a difference between psychotic and nonpsychotic patients in terms of *DBH* activity.

6. Conclusion

In this study, the proposed *RW* and *RGP* tests are compared with their normal-theory counterparts, *W* and *GP* tests, in terms of Type I error rates and powers when the underlying distribution is *STS* for the *BF* problem. Our results indicate that the *RGP* and *GP* tests are conservative while *RW* and *W* tests maintain the appropriate Type I error control for small sample sizes when $(r, d) = (2, -1)$ and $(2, 0)$. Proposed *RW* and *RGP* tests exhibit very

similar performance to their normal-theory counterparts *W* and *GP*, respectively, in terms of power under *STS* distribution with $(r, d) = (2, -1)$. However, proposed tests *RW* and *RGP* are slightly more powerful than their counterparts *W* and *GP*, respectively, when the underlying distribution is *STS* with $(r, d) = (2, 0)$, which has lower kurtosis than *STS* distribution with $(r, d) = (2, -1)$. According to the results of the robustness study, the *RW* and *RGP* tests are generally more robust than their normal theory counterparts as they maintain high power in the presence of deviations from the assumed model. Overall, *RW* and *RGP* may be regarded as useful alternatives to the corresponding normal-theory tests when the underlying distribution is *STS* and the variances are unknown and unequal.

References

1. A. D. Akkaya, and M. L. Tiku, *Robust estimation and hypothesis testing under short-tailedness and inliers*, Test, vol. 14, pp. 129–150, 2005.
2. A. D. Akkaya, and M. L. Tiku, *Autoregressive models with short-tailed symmetric distributions*, Statistics, vol. 42, no. 3, pp. 207–221, 2008.
3. S. Balci, and A. D. Akkaya, *Robust pairwise multiple comparisons under short-tailed symmetric distributions*, Journal of Applied Statistics, vol. 42, no. 11, pp. 2293–2306, 2015.
4. D. J. Best, and J. C. W. Rayner, *Welch's approximate solution for the Behrens–Fisher problem*, Technometrics, vol. 29, no. 2, pp. 205–210, 1987.
5. C. Chen, Y. Li, K. Liang, and J. Du, *A test for the Behrens–Fisher problem based on the method of variance estimates recovery*, Communications in Statistics–Theory and Methods, vol. 52, no. 18, pp. 6444–6455, 2023.
6. R. A. Fisher, *The fiducial argument in statistical inference*, Annals of Eugenics, vol. 6, no. 4, pp. 391–398, 1935.
7. J. Gamage, T. Mathew, and S. Weerahandi, *Generalized p-values and generalized confidence regions for the multivariate Behrens–Fisher problem and MANOVA*, Journal of Multivariate Analysis, vol. 88, no. 1, pp. 177–189, 2004.
8. A. K. Gupta, and Y. Wang, *Some tests with specified size for the Behrens–Fisher problem*, Communications in Statistics–Theory and Methods, vol. 28, no. 3–4, pp. 511–517, 1999.
9. G. Güven, *Testing the equality of treatment means in one-way ANOVA: Short-tailed symmetric error terms with heterogeneous variances*, Hacettepe Journal of Mathematics and Statistics, vol. 51, no. 6, pp. 1736–1751, 2022.
10. G. Güven, S. Acitas, H. Şamkar, and B. Senoglu, *RobustBF: An R package for robust solution to the Behrens–Fisher problem*, R Journal, vol. 13, no. 2, pp. 713–733, 2021.
11. D. J. Hand, F. Daly, A. D. Lunn, K. J. McConway, and E. Ostrowski, *A handbook of small data sets*, Chapman & Hall, London, 1st ed., 1994.
12. B. L. Joiner, and J. R. Rosenblatt, *Some properties of the range in samples from Tukey's symmetric lambda distributions*, Journal of the American Statistical Association, vol. 66, no. 334, pp. 394–399, 1971.
13. S. H. Kim, and A. S. Cohen, *On the Behrens–Fisher problem: A review*, Journal of Educational and Behavioral Statistics, vol. 23, no. 4, pp. 356–377, 1998.
14. K. Krishnamoorthy, and T. Mathew, *Inferences on the means of lognormal distributions using generalized p-values and generalized confidence intervals*, Journal of Statistical Planning and Inference, vol. 115, no. 1, pp. 103–121, 2003.
15. S. Paul, Y. G. Wang, and I. Ullah, *A review of the Behrens–Fisher problem and some of its analogs: Does the same size fit all?*, Revstat Statistical Journal, vol. 17, no. 4, pp. 563–597, 2019.
16. F. E. Satterthwaite, *An approximate distribution of estimates of variance components*, Biometrics Bulletin, vol. 2, no. 6, pp. 110–114, 1946.
17. K. K. Saxena, and O. P. Srivastava, *A new approximation to the critical point of t-distribution*, Statisztikai Szemle, vol. 64, pp. 1239–1244, 1986.
18. B. Senoglu, *Estimating parameters in one-way analysis of covariance model with short-tailed symmetric error distributions*, Journal of Computational and Applied Mathematics, vol. 201, no. 1, pp. 275–283, 2007.
19. P. Singh, K. K. Saxena, and O. P. Srivastava, *Power comparisons of solutions to the Behrens–Fisher problem*, American Journal of Mathematical and Management Sciences, vol. 22, no. 3–4, pp. 233–250, 2002.
20. D. E. Sternberg, D. P. VanKammen, P. Lerner, and W. E. Bunney, *Schizophrenia: Dopamine β-hydroxylase activity and treatment response*, Science, vol. 216, no. 4553, pp. 1423–1425, 1982.
21. M. L. Tiku, *Estimating the mean and standard deviation from a censored normal sample*, Biometrika, vol. 54, no. 1–2, pp. 155–165, 1967.
22. M. L. Tiku, *Estimating the parameters of log-normal distribution from censored samples*, Journal of the American Statistical Association, vol. 63, no. 321, pp. 134–140, 1968.
23. M. L. Tiku, and A. D. Akkaya, *Robust estimation and hypothesis testing*, New Age International (P) Limited, New Delhi, 2004.
24. M. L. Tiku, M. Q. Islam, and A. S. Selçuk, *Nonnormal regression. II. Symmetric distributions*, Communications in Statistics–Theory and Methods, vol. 30, no. 6, pp. 1021–1045, 2001.
25. M. L. Tiku, and D. C. Vaughan, *A family of short-tailed symmetric distributions*, Technical Report, McMaster University, Canada, 1999.
26. K. W. Tsui, and S. Weerahandi, *Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters*, Journal of the American Statistical Association, vol. 84, no. 406, pp. 602–607, 1989.
27. D. C. Vaughan, *The generalized secant hyperbolic distribution and its properties*, Communications in Statistics–Theory and Methods, vol. 31, no. 2, pp. 219–238, 2002.

28. S. Weerahandi, *Testing regression equality with unequal variances*, Econometrica: Journal of the Econometric Society, vol.55, no. 5, pp. 1211–1215, 1987.
29. B. L. Welch, *The significance of the difference between two means when the population variances are unequal*, Biometrika, vol. 29, no. 3–4, pp. 350–362, 1938.
30. V. Witkovský, *On the Behrens–Fisher distribution and its generalization to the pairwise comparisons*, Discussiones Mathematicae Probability and Statistics, vol. 22, no. 1–2, pp. 73–104, 2002.
31. L. W. Xu, and S. G. Wang, *A new generalized p-value for ANOVA under heteroscedasticity*, Statistics & Probability Letters, vol. 78, no. 8, pp. 963–969, 2008.