Removing skewness and kurtosis by transformation when testing for mean equality

Isabel Parra-Frutos Lourdes Molera

ipf@um.es lmolera@um.es

Department of Quantitative Methods for Economy and Business

University of Murcia

Campus de Espinardo, 30100 Murcia

Spain

Corresponding author: Isabel Parra-Frutos (ipf@um.es)

This is an Accepted Manuscript of an article published online by Taylor & Francis in Communications in Statistics - Simulation and Computation on 10 May 2019, available online: http://www.tandfonline.com/10.1080/03610918.2019.1610439

Abstract

A transformation of the Welch statistic to compare means is proposed to correct skewness and kurtosis of parent populations. The results show that this transformation seems to improve the performance of the test in heavy-tailed distributions more than other transformations focused only on skewness. The proposed test outperforms the Welch test in asymmetric heavy-tailed distributions with high heteroscedasticity and it behaves better than the Johnson's transformation trimmed mean Welch test in normal, near-normal and light-tailed distributions. It may also be a better option when some of the distributions are heavy-tailed and some light-tailed.

KEYWORDS: Kurtosis, Skewness, Welch test, Johnson's transformation, Hall's transformation

1. Introduction

The ANOVA *F* test compares means of several groups under the classical assumptions of normality, homoscedasticity and independence. However, it is not robust when the assumptions are violated (Parra-Frutos 2014 and the references therein). To handle unequal variances, an approximate test can be used. The Welch test (Welch 1951) is one of the most valid tests under various conditions of heterogeneity investigated (see Dijkstra and Werter 1981; Wilcox 1989). Unfortunately, this and other heteroscedastic tests do not always control the Type I error rate when distributions are asymmetric heavy-tailed, that is, they cannot handle the problem of non-normality at the same time (Lix, Keselman, and Keselman 1996). The shape of the parent populations may have serious consequences for the performance of a test of equality of means, especially when there is heteroscedasticity (Wilcox 1990, Algina et al. 1994, Cribbie et al. 2007, Cribbie et al. 2012, and Wilcox 2017). We propose a test that can perform well in that situation, that is, in asymmetric heavy-tailed distributions with high heteroscedasticity.

To remove the effect of population skewness on the distribution of a one-sample *t* statistic, some transformations have been proposed, e.g., Johnson's (1978) and Hall's (1992). Both use properties of the data to address a modification of the statistic that is less biased and corrects skewness, leading to a more robust procedure for hypothesis testing. The difference between them is an additional term in Hall`s transformation that makes it monotone and invertible, and therefore suitable for determining confidence intervals (Johnson 1978, Hall 1992).

To remove as well the effect of heavy tails, these transformations have been applied in conjunction with trimmed means in the Behrens-Fisher problem (Luh and Guo 1999, Guo and Luh 2000, and Keselman et al. 2002). The Welch test with trimmed means and Winsorized variances has been shown to provide excellent Type I error control and power even under extreme violations of the normality and variance equality assumptions (Cribbie et al. 2012). This test is outperformed by the Johnson's transformation trimmed mean Welch test (Luh and Guo 1999). However, when using trimmed means the null hypothesis changes to the equality of population trimmed means. Considering the usual mean supposes that the researcher prefers to give equal weight to all the observations. However, if the researcher is interested in giving zero weight to the most extreme values, that is, comparing values that represent the bulk of the data, the trimmed mean must be used. This may be the case of extremely asymmetric heavy-tailed distributions, where the mean loses representativeness in favor of the trimmed mean for applied researchers.

In this paper a new transformation is proposed that not only reduces the effect of skewness but also the effect of kurtosis. Thus, it can be considered the counterpart of the

Johnson’s transformation trimmed Welch test when the usual means are compared. As in Johnson (1978), the new procedure is based on the Cornish-Fisher expansion (Cornish and Fisher 1938), but it includes an additional term that addresses the kurtosis. We focus on the problem of testing equality of the usual means, so the performance of the proposed test is mainly compared to the Welch test, with or without transformations, although the Johnson’s transformation Welch test with robust estimators (i.e. trimmed means and Winsorized variances) is also simulated.

The behavior of the tests is studied for a variety of distributions. The scenario of asymmetric and heavy-tailed distributions has attracted much attention in applied research, and simulation studies have revealed that traditional tests can present serious problems of Type I error control. However, little attention has been paid to light-tailed distributions.

Blanca et al. (2013) show that only 5.5% of the samples of real data from several psychological variables exhibit measures of skewness and kurtosis close to normal distributions. They also find that 65.9% have a skewness measure incompatible with normal distributions. As for the kurtosis, this incompatibility happens in 80.8% of the samples, with 45.7% having negative kurtosis and 35.1% positive. Thus, negative kurtosis seems to be relevant in real data, so simulation studies should include this kind of distribution (Blanca et al. 2013). On the other hand, negative kurtosis is not necessarily accompanied by asymmetry. In this sense, Micceri (1989) found that almost all distributions having low (negative) kurtosis were, at most, moderately asymmetric and, according to Blanca et al. (2013), the figure may be around 95.2%. Indeed, 71.8% of samples with a skewness measure near the normal showed a negative kurtosis measure. Therefore, in addition to heavy-tailed distributions, this simulation study includes light-tailed distributions. Normal and near-normal distributions are also considered.

The simulation results show that the proposed transformation of the Welch statistic corrects problems of non-normality of populations better than other transformations focused only on skewness when there are heavy-tailed distributions. With symmetric heavy-tailed distributions, the new test has a higher Type I error rate than the Welch test but it seems to control. For asymmetric heavy-tailed distributions, the new test outperforms the Welch test if heteroscedasticity is high, with an empirical significance level near to the nominal for large enough samples. However, the Johnson's transformation Welch test with trimmed means provides smaller errors. If distributions are normal, near-normal and light-tailed the performance of the new test is superior to the Johnson's transformation Welch test with robust estimators. When both light-tailed and heavy-tailed distributions are involved, the new test may still be a better procedure.

2. The new transformation of a *t* statistic

Let $t = \sqrt{N}\left(\overline{x} - \mu\right) / s$, the usual statistic to make inferences about the mean μ of a population *X* with finite variance σ^2, where $\overline{x}$ and s^2 are the mean and variance of a sample of *N* independent, identically distributed observations taken from *X*. Johnson (1978) proposes a modification of this *t* statistic that reduces the effect of the skewness in non-normal populations. The modified *t* variable is based on the Cornish-Fisher expansion and is obtained by replacing $\overline{x} - \mu$ in the numerator by

$$T_{Johnson} = \left(\overline{x} - \mu\right) + \lambda + \gamma_1\left(\overline{x} - \mu\right)^2$$

with

$$\lambda = \frac{\mu_3}{6\sigma^2 N},$$

$$\gamma_1 = \frac{\mu_3}{3\sigma^4},$$

where μ_3 is the third central moment of *X*, $\mu_r = E[(X - E(X))^r]$.

This modified variable has an approximate Student's *t* distribution because the adjustment corrects bias and skewness effects on the *t* variable, due to the asymmetry of the population. Thus, the modification results in a more robust procedure.

Hall (1992) adds a new term in Johnson's transformation to obtain a monotone and invertible transformation,

$$T_{Hall} = \left(\overline{x} - \mu\right) + \lambda + \gamma_1 \left(\overline{x} - \mu\right)^2 + \frac{\mu_3^2}{27\sigma^8}\left(\overline{x} - \mu\right)^3.$$

We propose an extended version of Johnson's transformation which not only reduces the effect of the population skewness, but also that of the kurtosis.

The general form of the Cornish-Fisher expansion for a variable *X* is

$$CF(X) = \mu + \sigma Z + \frac{\mu_3}{6\sigma^2}\left(Z^2 - 1\right) + \frac{\sigma K}{24}\left(Z^3 - 3Z\right) + \ldots$$

where *Z* is a standard normal random variable and *K* is the excess of kurtosis of *X* given by $K = \mu_4 / \sigma^4 - 3$. Assuming that all moments of the population exist, the representation of $\overline{x}$ by a Cornish-Fisher expansion to three terms is

$$CF(\overline{x}) = \mu + \frac{\sigma}{\sqrt{N}} Z + \frac{\mu_3}{6N\sigma^2}\left(Z^2 - 1\right) + \frac{\sigma K}{N^{3/2} 24}\left(Z^3 - 3Z\right) + \ldots$$

Following the procedure in Johnson (1978), the proposed new modified *t* variable is given by

$$t_1 = \frac{\sqrt{N}}{s}\left[\left(\overline{x} - \mu\right) + \lambda + \gamma_1(\overline{x} - \mu)^2 + \gamma_2\left(\overline{x} - \mu\right)^3\right].$$

For the derivation of λ, γ_1 and γ_2, the Cornish-Fisher expansion of t_1, $CF(t_1)$, is calculated ignoring the terms of order $O\left(N^{-1}\right)$ except for Z^3, when $O\left(N^{-3/2}\right)$ is used. As in Johnson (1978), the constant λ is selected so that the constant terms in $CF(t_1)$ sum to zero, thereby removing the high-order bias, and γ_1 is chosen so that the coefficient of the Z^2 term in $CF(t_1)$ is zero, hence eliminating the high-order effects of skewness. Similarly, γ_2 is selected so that the coefficient of the Z^3 term in $CF(t_1)$ is zero in order to remove high-order effects of kurtosis. The solutions obtained are

$$\lambda = \frac{\mu_3}{6N\sigma^2}, \qquad \gamma_1 = \frac{\mu_3}{3\sigma^4}, \qquad \gamma_2 = -\gamma_1^2 - \frac{K}{24\sigma^2}.$$

Note that λ and γ_1 are the same as in Johnson (1978) and Hall (1992).

3. Description of tests

According to Luh and Guo (1999), Guo and Luh (2000) and Keselman et al. (2002), the transformations for a Studentized statistic are adapted to test the equality of means. In particular, they are used to transform the Welch statistic.

Let y_{ij}, $i = 1,...,k$ and $j = 1,...,n_i$, denote the *j*-th observation from the *i*-th group, where n_i, $\overline{y}_i$ and s_i^2 are the sample size, mean and variance of the *i*-th group, respectively, and *k* is the number of groups or treatments.

The Welch (1951) *test (W test).* The test statistic is given by

$$T_W = \frac{\sum_{i=1}^{k} w_i T_i^2}{(k-1)\left(1 + \frac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\frac{(1-w_i/W)^2}{n_i-1}\right)},$$

where

$$T_i = \overline{y}_i - \hat{y},$$

$$w_i = \frac{n_i}{s_i^2},$$

$$s_i^2 = \frac{\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^2}{n_i - 1},$$

$$W = \sum_{i=1}^{k} w_i ,$$

$$\hat{y} = \frac{\sum_{i=1}^{k} w_i \overline{y}_i}{W}.$$

The Welch statistic distributes approximately like *F* with $k-1$ and *v* degrees of freedom, where

$$v = \frac{k^2 - 1}{3\sum_{i=1}^{k}\frac{\left(1 - w_i / W\right)^2}{n_i - 1}}.$$

The null hypothesis – population means are equal – is rejected if the computed T_W statistic is greater than $F_{k-1,v;\alpha}$, the $\left(1-\alpha\right)$ percentile of the *F* distribution with $k-1$ and v degrees of freedom.

The Welch test with Johnson's transformation (W.John test). The variable T_i in T_W statistic would be

$$T_{i,Johnson} = \left(\overline{y}_i - \hat{y}\right) + \hat{\lambda}_i + \hat{\gamma}_{1i}\left(\overline{y}_i - \hat{y}\right)^2,$$

where

$$\hat{\lambda}_i = \frac{\hat{\mu}_{3i}}{6 n_i s_i^2},$$

$$\hat{\gamma}_{1i} = \frac{\hat{\mu}_{3i}}{3 s_i^4},$$

$$\hat{\mu}_{3i} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij} - \overline{y}_i\right)^3.$$

The Welch test with Hall's transformation (W.Hall test). The variable T_i would be

$$T_{i,Hall} = \left(\bar{y}_i - \hat{y}\right) + \hat{\lambda}_i + \hat{\gamma}_{1i}\left(\bar{y}_i - \hat{y}\right)^2 + \frac{\hat{\mu}_{3i}^2}{27 s_i^8}\left(\bar{y}_i - \hat{y}\right)^3 .$$

The Welch test with the new transformation (W.new test). The new transformation would lead to

$$T_{i,new} = \left(\bar{y}_i - \hat{y}\right) + \hat{\lambda}_i + \hat{\gamma}_{1i}\left(\bar{y}_i - \hat{y}\right)^2 + \hat{\gamma}_{2i}\left(\bar{y}_i - \hat{y}\right)^3 ,$$

$$\hat{\gamma}_{2i} = -\hat{\gamma}_{1i}^2 - \frac{\hat{K}_i}{24 s_i^2} ,$$

$$\hat{K}_i = \frac{\hat{\mu}_{4i}}{\hat{\mu}_{2i}^2} - 3 ,$$

$$\hat{\mu}_{ri} = \frac{1}{n_i}\sum_{j=1}^{n_i}\left(y_{ij} - \bar{y}_i\right)^r , \qquad r = 2, 4 .$$

R code for the W.new test is provided in the Appendix.

Johnson's transformation Welch test with trimmed means and Winsorized variances (Wt.John test).

The null hypothesis in this test is equal population trimmed means, $\mu_{t1} = \mu_{t2} = \cdots = \mu_{tk}$, where the amount of trimming is given by β and usually set to 20%. Let $y_{i(1)} \le y_{i(2)} \le \cdots \le y_{i(n_i)}$ represent the ordered observations associated with the *i*-th group. Let $g_i = \left[\beta n_i\right]$ the number of observations to be trimmed in each tail of the distribution, where $\left[x\right]$ is the greatest integer equal to or lower than *x*. Thus, $h_i = n_i - 2g_i$ is the effective sample size after trimming. The trimmed sample mean of *i*-th group, $\bar{y}_{ti}$, and $\tilde{s}_{wi}^2$ given by

$$\overline{y}_{ti} = \frac{\sum_{j=g_i+1}^{n_i-g_i} y_{i(j)}}{h_i}, \qquad \tilde{s}_{wi}^2 = \frac{n_i - 1}{h_i - 1} s_{wi}^2,$$

replace the sample mean and the variance, where s_{wi}^2 is the Winsorized variance

$$s_{wi}^2 = \frac{\sum_{j=1}^{n_i}\left(x_{ij} - \overline{y}_{wi}\right)^2}{n_i - 1}, \qquad \overline{y}_{wi} = \frac{\sum_{j=1}^{n_i} x_{ij}}{n_i}, \qquad x_{ij} = \begin{cases} y_{i(g_i+1)} & if \quad y_{ij} \le y_{i(g_i+1)} \\ y_{ij} & if \quad y_{i(g_i+1)} < y_{ij} < y_{i(n_i-g_i)} \\ y_{i(n_i-g_i)} & if \quad y_{ij} \ge y_{i(n_i-g_i)} \end{cases}$$

and $\overline{y}_{wi}$ is the Winsorized mean.

The variable T_i in T_W statistic would be

$$T_{ti,Johnson} = \left(\overline{y}_{ti} - \hat{y}_t\right) + \hat{\lambda}_{ti} + \hat{\gamma}_{1ti}\left(\overline{y}_{ti} - \hat{y}_t\right)^2$$

with

$$\hat{y}_t = \frac{\sum_{i=1}^{k} w_{ti}\overline{y}_{ti}}{W_t}, \qquad w_{ti} = \frac{h_i}{\tilde{s}_{wi}^2}, \qquad W_t = \sum_{i=1}^{k} w_{ti},$$

$$\hat{\lambda}_{ti} = \frac{\tilde{\mu}_{3i}}{6\tilde{s}_{wi}^2 h_i},$$

$$\hat{\gamma}_{1ti} = \frac{\tilde{\mu}_{3i}}{3\tilde{s}_{wi}^4},$$

$$\tilde{\mu}_{3i} = \frac{n_i}{h_i}\hat{\mu}_{w3i},$$

$$\hat{\mu}_{w3i} = \frac{\sum_{j=1}^{n_i}\left(x_{ij} - \overline{y}_{wi}\right)^3}{n_i}.$$

4. Design of the simulation

The Type I error rate of the tests with three and six groups is investigated in a simulation study for different situations: symmetric and asymmetric distributions; light-

and heavy-tailed distributions; homoscedasticity and heteroscedasticity; negative and positive pairing of sample sizes and variances; and small, large, equal and unequal sample sizes. The nominal 5 percent significance level is used throughout. The simulation results shown in the figures are based on 10,000 Monte Carlo replications and those in the tables on 100,000. R software is used.

The samples are taken originally from the distributions given in Table 1. Afterwards, appropriate linear transformations are carried out to generate data sets from the populations with the desired means and variances. When working with trimmed means these transformations imply subtracting the population trimmed mean instead of the population mean.

Equal population means are used to estimate the Type I error rate, and the values 0, 0.5 and 1 in the case of estimating power (they provide a power of 0.654 for the ANOVA *F* test in normal distributions with unit variance and sample sizes of 15). For Johnson's transformation trimmed Welch test all these values correspond to population trimmed means.

The data transformations used do not change skewness and kurtosis of original distributions. Table 1 informs of the skewness $(\alpha_1 = \mu_3 / \mu_2^{3/2})$ and excess of kurtosis (*K*) of each distribution along with the notation of each one. To generate data from the *g*-and-*h* distribution, we use Cribbie et al. (2012) and Hoaglin (1985). All distributions considered are continuous, symmetric or positively skewed, and bell-shaped (except Beta(1/2,1/2), which is U-shaped, and the uniform distribution).

Table 1 Distributions used in the simulation study.

	N(0,1)	t_4	g=0 h= 0.0516	g=0.2 h=0	g=0.81 h=0	g=1 h=0	Beta (1/2,1/2)	Uniform	Beta(2,2)	Beta(2,5)
notation	n	t	gh005	gh020	gh0810	gh10	b0505	u	b22	b25
α_1	0	0	0	0.61	3.8	6.2	0	0	0	0.60
K	0	∞	0.86	0.7	33.3	111	–1.5	–1.2	–0.86	–0.12

In the case of three groups, 17 combinations of distributions are considered in the simulation study. For example, if a combination includes three populations with the same skewness and kurtosis as the Beta(2,2) distribution, it is denoted by b22_b22_b22 (see Table 1). These seventeen combinations cover a wide range of situations: normal distributions (n_n_n); symmetric and near-normal kurtosis distributions (b22_b22_b22 and gh005_gh005_gh005); heavy-tailed distributions (t_t_t, gh020_gh0810_gh0810, gh020_gh0810_gh10 and gh10_ gh0810_gh020); non U-shaped light-tailed distributions (u_u_u, b25_b25_b25 and u_b22_b25); bell- and U-shaped light-tailed distributions (b0505_b0505_b0505, b25_b22_b0505 and b22_b22_b0505); and heavy- and light-tailed distributions (u_gh020_gh0810, b22_gh020_gh10, b22_b25_gh0810 and u_gh0810_gh10).

For each of these combinations, 22 settings are described, corresponding to different samples sizes and variances. Each setting is identified by a number (see Table 2) which is used in the figures below. Specifically, four configurations of sample sizes are studied, from small to large, and from equal to unequal: (15,15,15), (15,20,25), (60,60,60) and (50,60,70). Each of these configurations are considered under homoscedasticity and heteroscedasticity. The following combinations of standard deviations $(\sigma_1, \sigma_2, \sigma_3)$, from mild to extreme heteroscedasticity, are applied: $(1, 1.1, 1.2)$,

$(1, 1.5, 1.75)$ and $(1, 5, 8)$. The reverse of the above has also been included, so positive and negative pairing between variances and sample sizes are considered.

Table 2 Description of the settings used in the simulation study.

Setting	n_1	n_2	n_3	σ_1	σ_2	σ_3
	Homoscedasticity					
1	15	15	15	1	1	1
2	15	20	25	1	1	1
3	60	60	60	1	1	1
4	50	60	70	1	1	1
	Mild heteroscedasticity					
5	15	15	15	1	1.1	1.2
6	15	20	25	1	1.1	1.2
7	60	60	60	1	1.1	1.2
8	50	60	70	1	1.1	1.2
	Moderate heteroscedasticity					
9	15	15	15	1	1.5	1.75
10	15	20	25	1	1.5	1.75
11	60	60	60	1	1.5	1.75
12	50	60	70	1	1.5	1.75
	Extreme heteroscedasticity					
13	15	15	15	1	5	8
14	15	20	25	1	5	8
15	60	60	60	1	5	8
16	50	60	70	1	5	8

Setting	n_1	n_2	n_3	σ_1	σ_2	σ_3
	Mild heteroscedasticity and negative pairing					
17	15	20	25	1.2	1.1	1
18	50	60	70	1.2	1.1	1
	Moderate heteroscedasticity and negative pairing					
19	15	20	25	1.75	1.5	1
20	50	60	70	1.75	1.5	1
	Extreme heteroscedasticity and negative pairing					
21	15	20	25	8	5	1
22	50	60	70	8	5	1

Some other triples of variances and distributions are also considered to refine the conclusions (Tables 3 and 4). In addition, a similar simulation design is used in the case of 6 groups to study how the number of groups affects the performance of the tests (Tables 5 and 6).

The robustness of a procedure, with respect to the Type I error control, is determined using Bradley's (1978) liberal criterion. Hence, a procedure is deemed robust with respect to the Type I error if the empirical Type I error rate falls within the range

$\alpha \pm \alpha/2$. For $\alpha = 0.05$, the interval is given by $[0.025, 0.075]$, so an empirical rate over 0.075 would indicate a liberal test and one below 0.025 a conservative test.

5. Simulation results

A summary of the simulation results is presented in Figures 1-11 to highlight the differences in the control of the Type I error rate and estimated power of the W.new test in relation to the W, W.John and Wt.John tests. As the the W.Hall behaves similarly to the W.John tests (Type I error rate across the different combinations of distributions is less than 0.0073 and the mean of all differences is about 0.0009, with comparable estimated power), only the W.John test is illustrated. Occasionally, the ANOVA *F* test is also included and its sensitivity to departures from homoscedasticity is shown.

The simulation study reveals that if distributions are normal or symmetric with kurtosis near normal (near-normal distributions), the W.new and W.John tests control the Type I error rate (see Figure 1) and have an estimated power like that of the Welch test (see Figure 2). These three tests improve the ANOVA *F* test in relation to the control of the Type I error rate, with a similar power, when heteroscedasticity is moderate or extreme, especially with negative pairing. Moreover, they behave similarly to the ANOVA *F* test when standard assumptions are satisfied. It is also shown that, even though these tests based on transformations are robust in the near normal distributions, they tend to have an estimated significance level below 0.05 if distributions have negative excess kurtosis, and over 0.05 if positive. On the other hand, while testing equal trimmed means with the Wt.John test leads to control of the Type I error rate, its power is lower in some situations, even though the population trimmed mean and the usual mean are equal in symmetric distributions. Note that the cases of very low power correspond to the settings of extreme heteroscedasticity, where the samples sizes are too small relative to the variances.

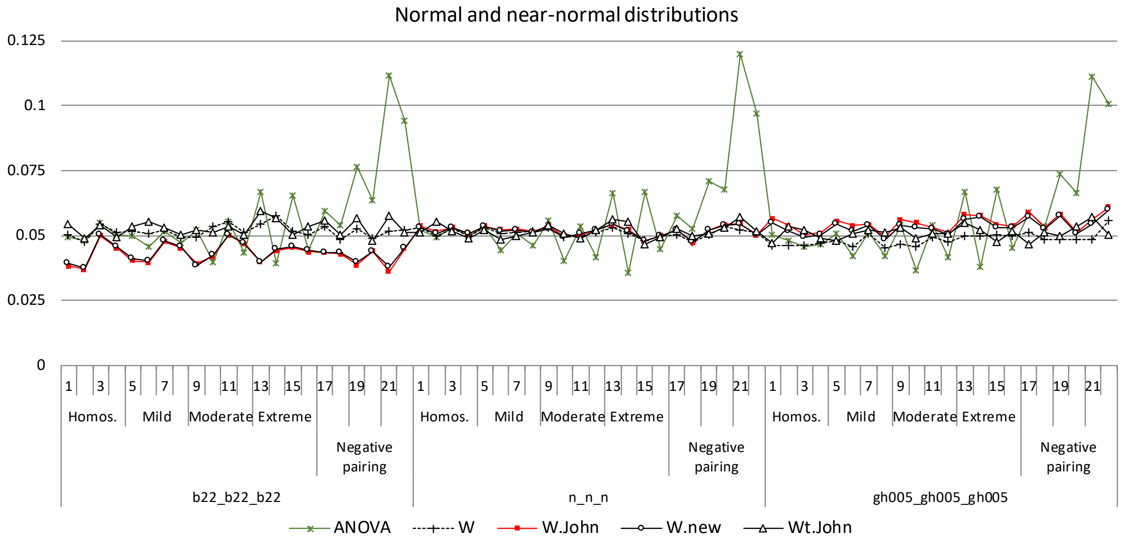

Figure 1 Type I error rate of tests for each of the 22 settings with normal or near-normal distributions.

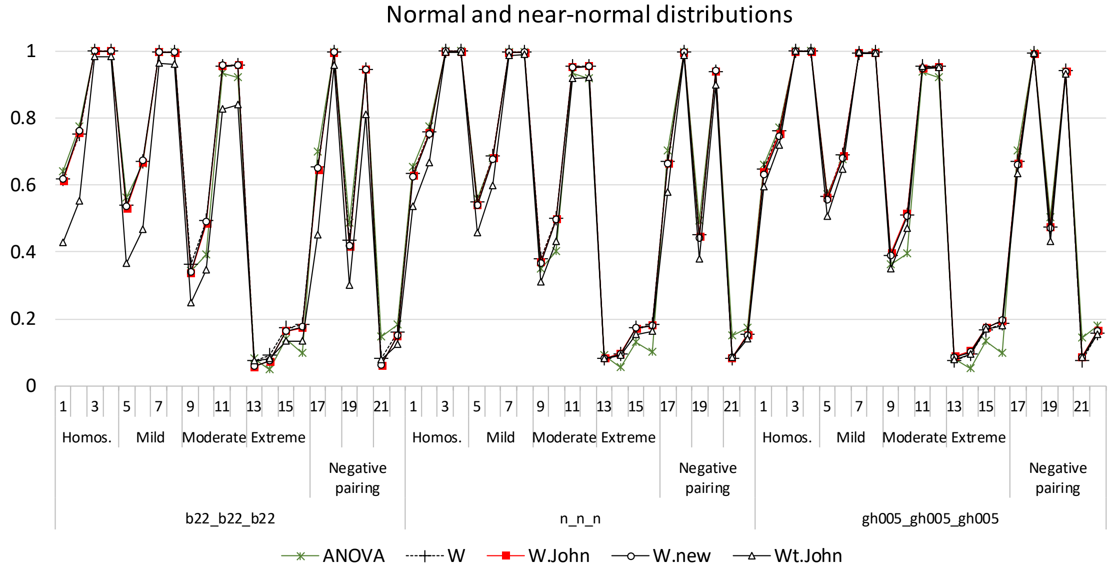

Figure 2 Estimated power of tests for each of the 22 settings with normal or near-normal distributions.

According to the Figure 3 and Figure 4, when distributions are all light-tailed the Type I error rate of the W.John and W.new tests is under 0.05 and falls as the negative excess kurtosis diminishes. All tests control the Type I error rate, but the W.John and

W.new tests seems to behave better than the Wt.John test, with smaller Type I error rates and higher power. Thus, when testing trimmed means the probabilities of errors are higher than when testing the usual means if parent populations are light-tailed. If at least one of the light-tailed distributions is U-shaped, these results are more marked (see Figure 5 and Figure 6).

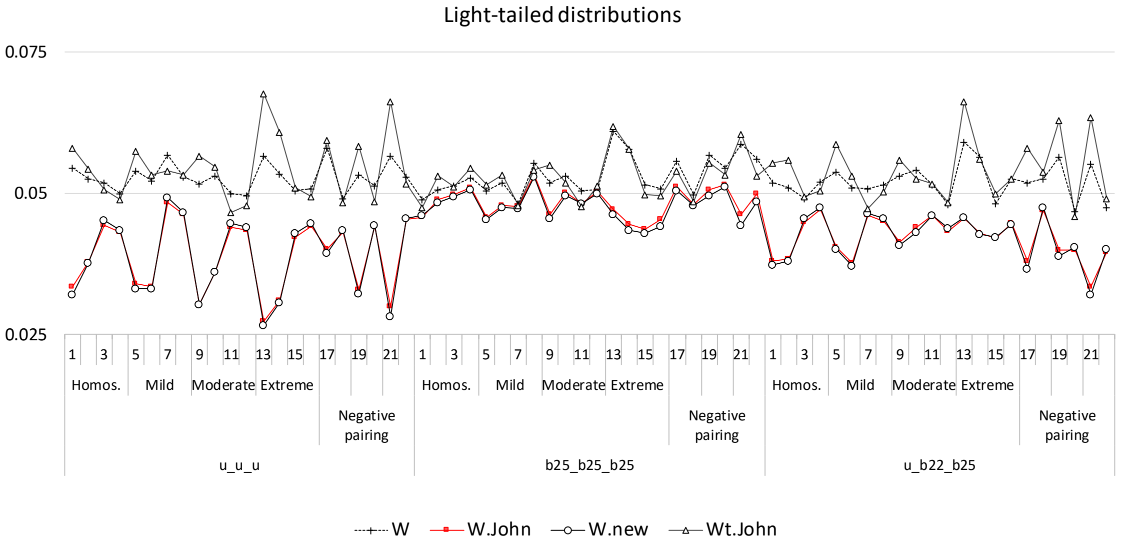

Figure 3 Type I error rate of tests for each of the 22 settings with light-tailed distributions.

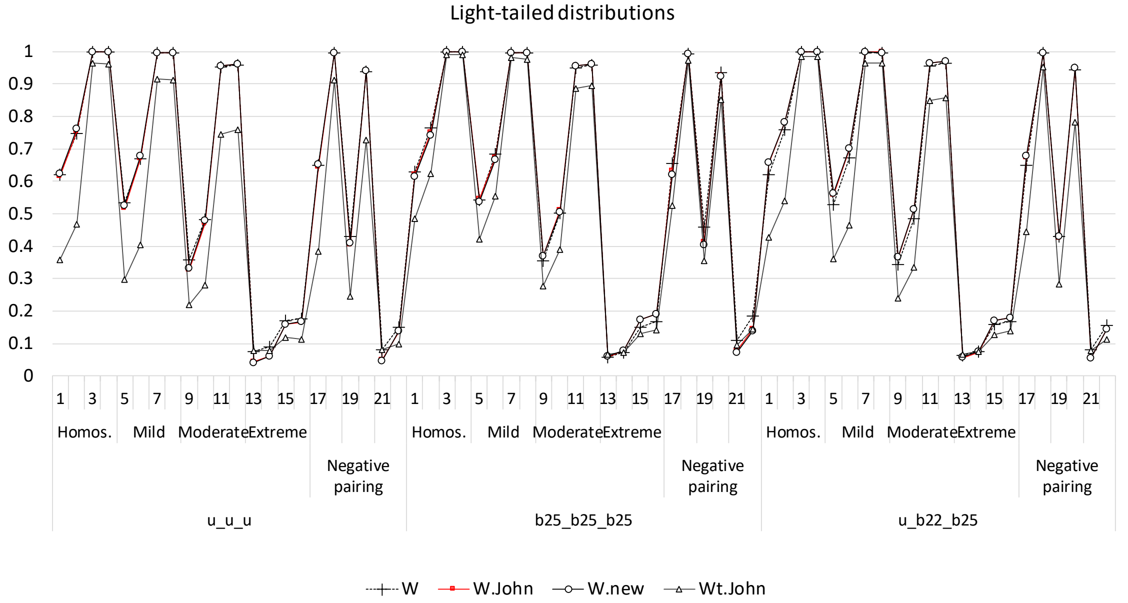

Figure 4 Estimated power of tests for each of the 22 settings with light-tailed distributions.

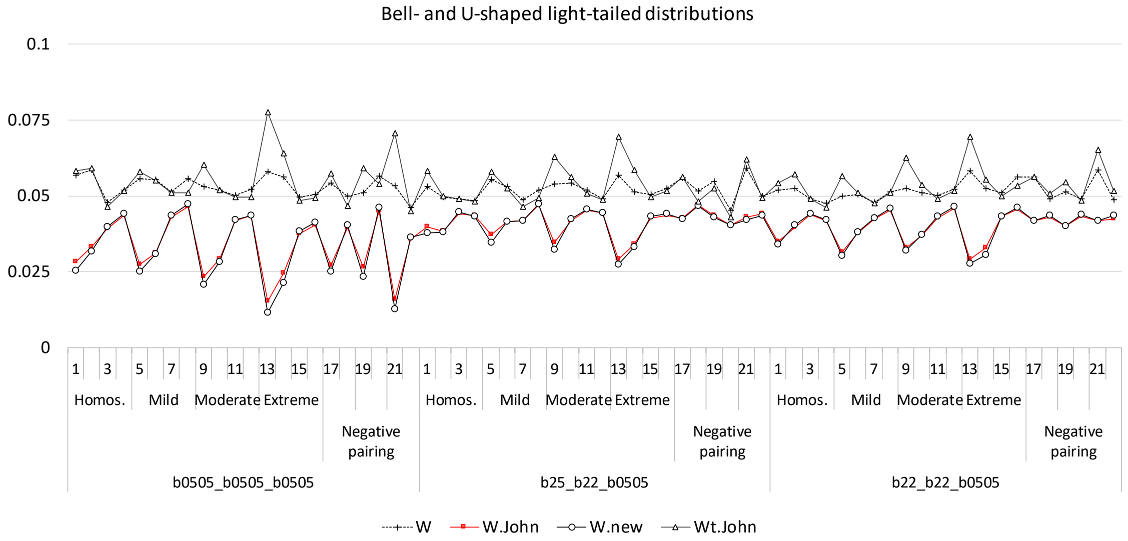

Figure 5 Type I error rate of tests for each of the 22 with bell- and U-shaped light-tailed distributions.

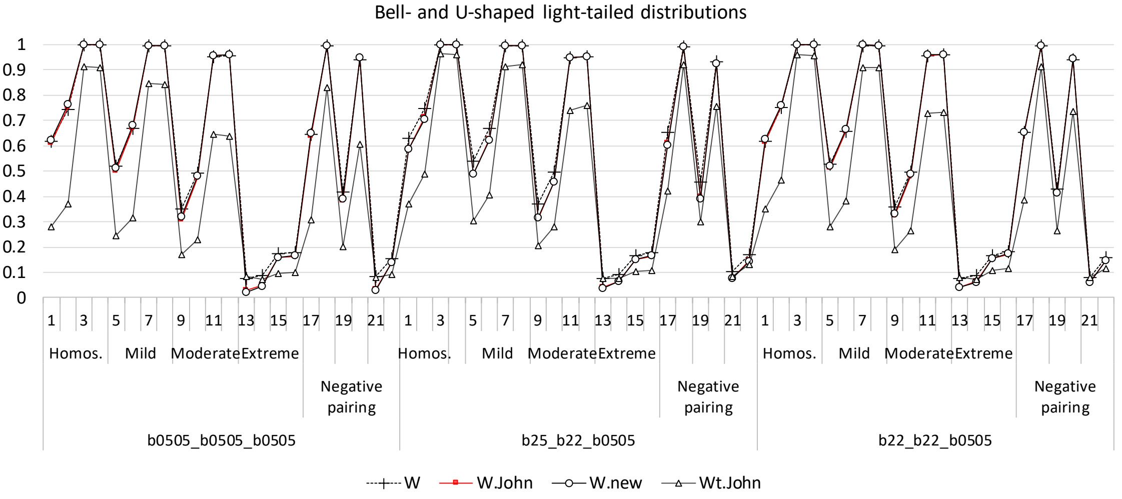

Figure 6 Estimated power of tests for each of the 22 settings with bell- and U-shaped light-tailed distributions.

In light-tailed populations, the estimated significance level of all tests seems to be nearer to the nominal in large samples. However, the Welch and the Wt.John test shows an opposite behavior to that of the W.John and W.new tests as the sample sizes decrease. Specifically, the smaller the sample sizes, the larger the Type I error rate of the Welch test, and especially that of the Wt.John test, and the smaller those of the other two tests. The similar behavior of the W.John and W.new tests in this case suggests that the estimation of the skewness leads to a low probability of rejecting the null hypothesis in light-tailed distributions, especially with small sample sizes.

In the case of heavy-tailed distributions (see Figure 7), all the tests studied appear to be robust if distributions are symmetric. Of the tests for comparing means, the Welch test has the lowest Type I error rate, followed by the W.new test and then by the W.John test. If distributions are asymmetric heavy-tailed, the W.John test shows a liberal behavior in many settings. However, the W.new test seems to control the Type I error rate in many more situations and it improves the Welch test in the case of extreme heteroscedasticity. Here, the W.new test exhibits the smallest Type I error rate of the Welch-based tests for comparing means. Additionally, as large samples reduce the estimated significance level, this may result in the W.new and W.John tests controlling the Type I error rate. In particular, for three gh10 distributions with extreme heteroscedasticity, the rate of the W.new test is about 0.075 for sample sizes (25,30,35) and (30,30,30) and about 0.06 for larger samples like (55,60,65) and (60,60,60) (see Table 3 and Table 4). But if samples sizes are small, it is the negative pairing between variances and skewness that would considerably reduce the estimated significance level of the W, W.John and W.new tests. This could lead the W.new test to control the Type I error rate (see Figure 7).

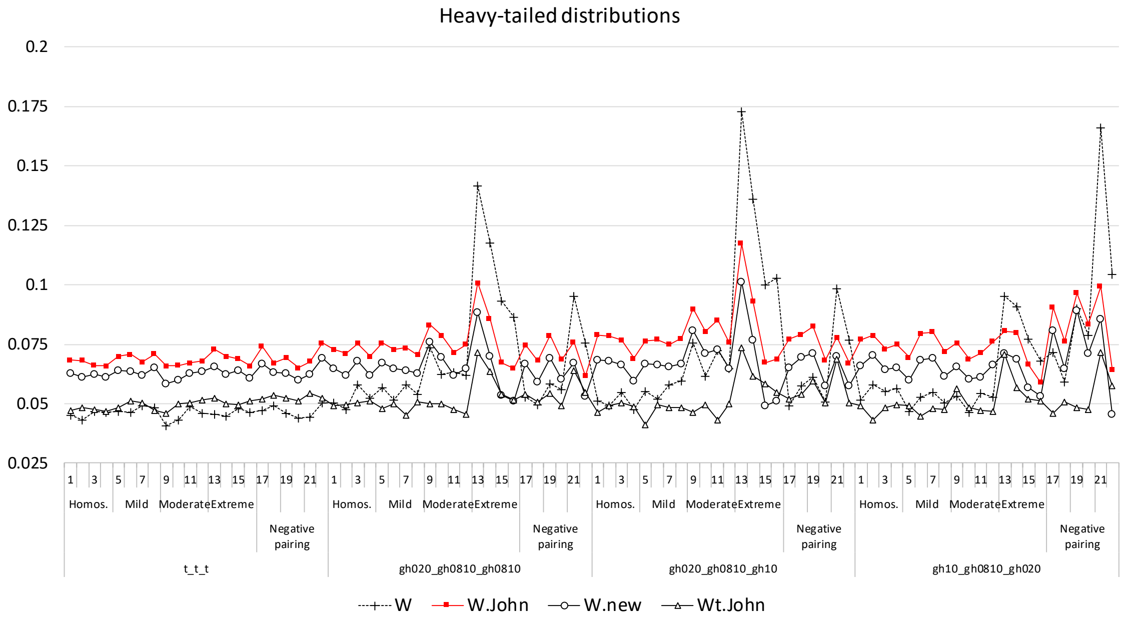

Figure 7 Type I error rate of tests for each of the 22 settings with heavy-tailed distributions.

The power of the transformation-based Welch tests for comparing means is similar in most cases (see Figure 8), although the Type I error rate of the W.new test is always lower. When testing trimmed means, the Wt.John test is a very good option, since it seems to control the Type I error rate quite well and has somewhat more power. In conclusion, with heavy-tailed distributions, the errors when testing equal trimmed means have less probability of happening than when testing equal means.

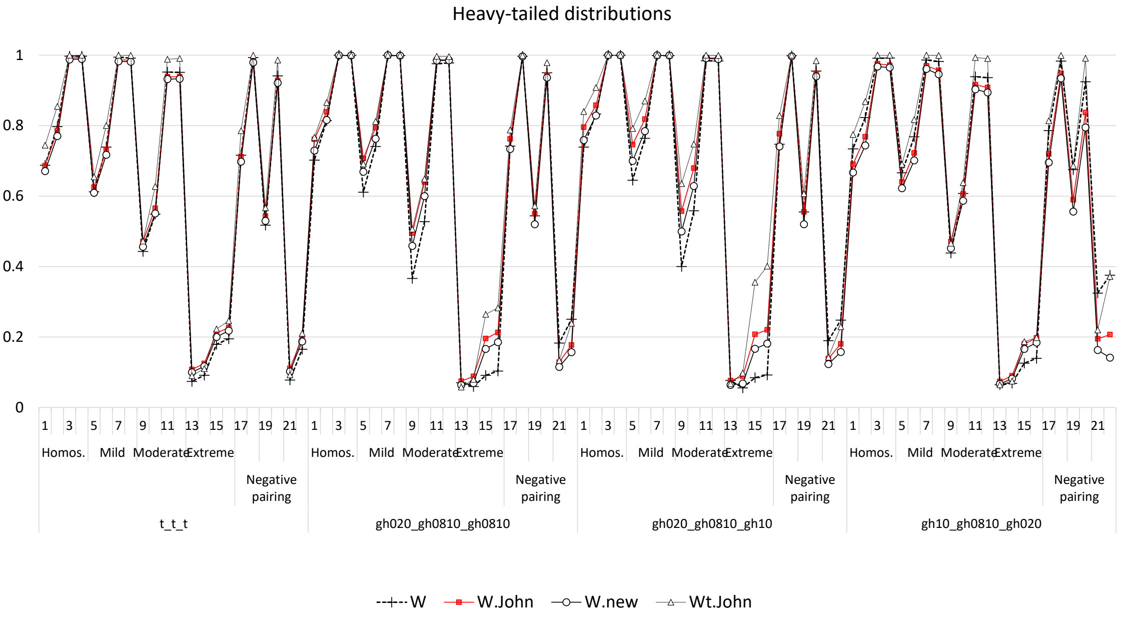

Figure 8 Estimated power of tests for each of the 22 settings with heavy-tailed distributions.

Some new *g*-and-*h* distributions have been simulated to illustrate in more detail the effect of the level of skewness on the Type I error rate (see Figure 9). The Welch test seems to behave similarly across the distribution combinations, in the sense that it controls the Type I error rate except if the heteroscedasticity is high, although the level of heteroscedasticity that causes the lack of control decreases as the skewness increases. However, a general rise in the Type I error rate of the transformation-based Welch tests for comparing means is observed as skewness increases, resulting in an increasing loss of control of Type I error rate even in the situations of homoscedasticity or mild heteroscedasticity. Anyway, it is worth noting that, in the adverse situation of asymmetric heavy-tailed distributions with extreme heteroscedasticity, the W.new test outperforms the Welch test, controlling the Type I error rate for moderate sample sizes if skewness and excess kurtosis are not extremely high.

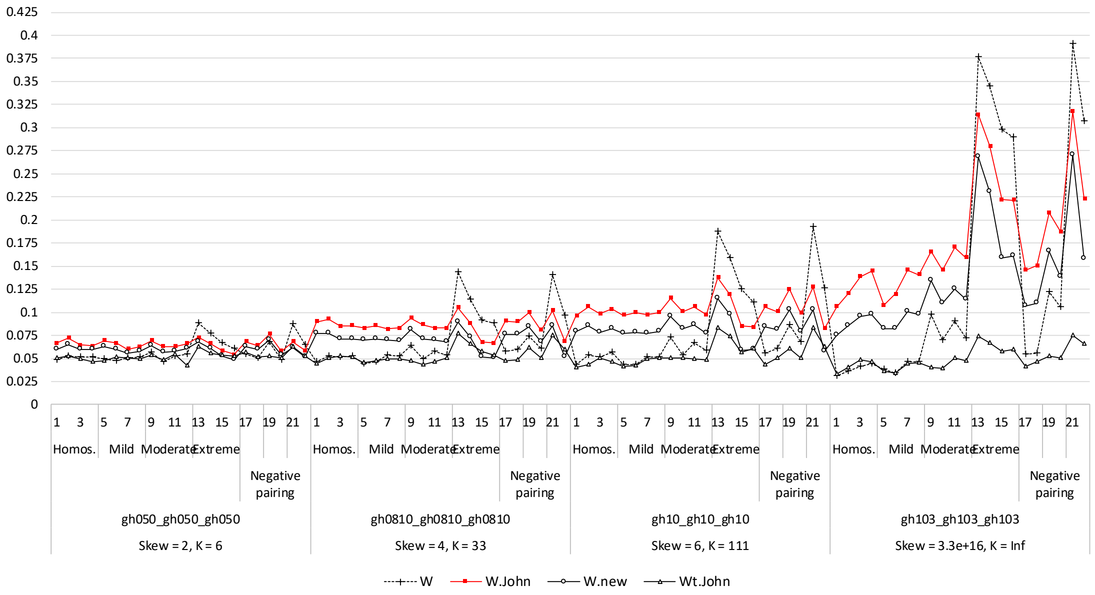

Figure 9 Type I error rate of tests for each of the 22 settings for distributions from low to high skewness. Note. gh050 refers to g=0.5 and h=0 and gh103 to g=1 and h=0.3.

Some new simulation results were also obtained for further analysis of the effect of the heteroscedasticity level. Table 3 shows the Type I error rate of the W.new test for three groups with the same population skewness and excess kurtosis for various levels of skewness and variance ratios in samples of about 30. It also indicates the settings where the W test controls and its maximum Type I error rate. It can be observed that the W.new test controls the Type I error rate for moderate skewness and kurtosis (up to 4 and 33 approx., respectively) regardless of the level of heteroscedasticity. In those cases, a slight increase in the Type I error rate may be observed if there is negative pairing between variances and sample sizes from homoscedasticity to moderate heteroscedasticity. Furthermore, the Type I error rate of the W.new test is decreasing as the variance ratio increases. In contrast, the W test fails to control in heteroscedastic situations, especially with negative pairing.

In larger samples (about 60), a general diminution of the Type I error rate of the W.new test is observed (see Table 4), giving rise to its being controlled in distributions with higher skewness. However, the higher the skewness, the higher the heteroscedasticity that it is necessary to control. For example, when sample sizes are (50,60,70) and (60,60,60), a variance ratio higher than 1:9 is needed in the case of skewness equal to 6.2 and excess kurtosis to 111, and higher than 1:25 in the case of 8 and 257, respectively. On the other hand, the W.new test performs well if there is extreme heteroscedasticity (variance ratio 1:64) up to a skewness coefficient of 13 and excess kurtosis of about 1200. For lower variance ratios, the skewness and excess kurtosis has to be lower to control the Type I error rate.

Table 3 Type I error rate of the W.new test for three groups (with the same population skewness and excess kurtosis) for various levels of heteroscedasticity and skewness, and sample sizes of about 30.

g	0.5	0.81	1	1	1	1	1	1.21	1	1.3	1.4	1.5
h	0	0	0	0.03	0.04	0.05	0.06	0	0.07	0	0	0
Skewness	1.8	3.8	6.2	8.2	9.1	10.1	11.47	11.53	13.1	15.6	22.5	33.5
Exc. kurt.	5.9	33.3	111	257	360	524	795	561	1270	1263	3401	10075
							$(n_1,n_2,n_3) = (25,30,35)$					
$(\sigma_1,\sigma_2,\sigma_3)$							W.new test					
(1,1,1)	.0614*	.0741*	**.0829***	**.0853***	**.0852***	**.0852***	**.0861***	**.0923***	**.0860***	**.0929***	**.0981***	**.0991***
(1,1.21,1.41)	.0610*	.0724*	**.0802***	**.0815***	**.0834***	**.0842***	**.0853***	**.0889***	**.0845***	**.0921***	**.0944***	**.0987***
(1,1.37,1.73)	.0598*	.0735*	**.0830***	**.0822***	**.0848***	**.0853***	**.0866***	**.0945***	**.0873***	**.0980***	**.1027***	**.1114***
(1,1.5,2)	.0601*	.0738*	**.0836***	**.0860***	**.0895***	**.0895***	**.0887***	**.0967***	**.0906***	**.1012***	**.1112**	**.1187**
(1,2,3)	.0599*	.0746*	**.0889**	**.0941**	**.0957**	**.0983**	**.0986**	**.1079**	**.1004**	**.1175**	**.1314**	**.1455**
(1,3,4)	.0567*	.0742	**.0925**	**.0974**	**.1008**	**.1019**	**.1060**	**.1153**	**.1067**	**.1276**	**.1445**	**.1611**
(1,3,5)	.0567*	.0713	**.0873**	**.0972**	**.0990**	**.0999**	**.1010**	**.1096**	**.1063**	**.1244**	**.1381**	**.1546**
(1,4,6)	.0566*	.0690	**.0847**	**.0903**	**.0947**	**.0977**	**.1022**	**.1065**	**.1034**	**.1186**	**.1363**	**.1539**
(1,4,7)	.0563*	.0682	**.0812**	**.0902**	**.0914**	**.0954**	**.0985**	**.1048**	**.1019**	**.1176**	**.1302**	**.1485**
(1,5,8)	.0543*	.0654	**.0786**	**.0854**	**.0871**	**.0902**	**.0934**	**.0969**	**.0958**	**.1080**	**.1238**	**.1423**
(1.41,1.21,1)	.0624*	**.0774***	**.0869***	**.0903***	**.0920***	**.0935***	**.0939***	**.1007***	**.0945***	**.1064***	**.1153***	**.1199**
(1.73,1.37,1)	.0608*	**.0786***	**.0916***	**.0956**	**.0960**	**.0990**	**.0992**	**.1083**	**.1036**	**.1166**	**.1256**	**.1363**
(2,1.5,1)	.0613*	**.0784***	**.0930**	**.0967**	**.1008**	**.1039**	**.1040**	**.1121**	**.1036**	**.1232**	**.1351**	**.1476**
(3,2,1)	.0617*	**.0798**	**.0964**	**.1030**	**.1067**	**.1085**	**.1107**	**.1222**	**.1133**	**.1365**	**.1500**	**.1686**
(4,3,1)	.0583*	**.0756**	**.0939**	**.1037**	**.1056**	**.1085**	**.1118**	**.1241**	**.1154**	**.1386**	**.1588**	**.1778**
(5,3,1)	.0551*	.0735	**.0905**	**.0982**	**.1025**	**.1062**	**.1101**	**.1171**	**.1111**	**.1327**	**.1516**	**.1713**
(6,4,1)	.0553*	.0690	**.0851**	**.0946**	**.0951**	**.1003**	**.1034**	**.1126**	**.1073**	**.1244**	**.1440**	**.1634**
(7,4,1)	.0552*	.0683	**.0827**	**.0915**	**.0960**	**.0987**	**.1018**	**.1062**	**.1035**	**.1221**	**.1394**	**.1597**
(8,5,1)	.0551	.0643	**.0774**	**.0872**	**.0884**	**.0918**	**.0944**	**.1019**	**.0977**	**.1125**	**.1308**	**.1500**
							Welch test					
Max.	**.0759**	**.1196**	**.1582**	**.1725**	**.1774**	**.1813**	**.1865**	**.2141**	**.1886**	**.2365**	**.2674**	**.2988**
							$(n_1,n_2,n_3) = (30,30,30)$					
							W.new test					
(1,1,1)	.0613*	.0737*	**.0814***	**.0843***	**.0837***	**.0853***	**.0851***	**.0881***	**.0859***	**.0934***	**.0961***	**.0974***
(1,1.21,1.41)	.0612*	.0747*	**.0833***	**.0866***	**.0869***	**.0881***	**.0881***	**.0942***	**.0903***	**.0992***	**.1029***	**.1097***
(1,1.37,1.73)	.0603*	**.0751***	**.0855***	**.0894***	**.0901***	**.0928***	**.0924***	**.1011***	**.0933***	**.1054***	**.1125**	**.1216**
(1,1.5,2)	.0596*	**.0759***	**.0870***	**.0924***	**.0929***	**.0948**	**.0970**	**.1052**	**.0988**	**.1109**	**.1224**	**.1333**
(1,2,3)	.0598*	**.0777**	**.0911**	**.0980**	**.1010**	**.1043**	**.1064**	**.1143**	**.1071**	**.1277**	**.1412**	**.1563**
(1,3,4)	.0574*	.0746	**.0926**	**.1011**	**.1042**	**.1057**	**.1087**	**.1195**	**.1113**	**.1325**	**.1486**	**.1682**
(1,3,5)	.0575*	.0730	**.0891**	**.0965**	**.0996**	**.1021**	**.1075**	**.1158**	**.1089**	**.1277**	**.1462**	**.1631**
(1,4,6)	.0549*	.0696	**.0848**	**.0937**	**.0960**	**.0974**	**.1011**	**.1096**	**.1030**	**.1220**	**.1379**	**.1575**
(1,4,7)	.0553*	.0677	**.0823**	**.0900**	**.0913**	**.0948**	**.0981**	**.1050**	**.1038**	**.1175**	**.1343**	**.1530**
(1,5,8)	.0542*	.0645	**.0763**	**.0841**	**.0882**	**.0892**	**.0935**	**.0975**	**.0964**	**.1111**	**.1268**	**.1462**
							Welch test					
Max.	.0732	**.1154**	**.1510**	**.1622**	**.1689**	**.1702**	**.1777**	**.2002**	**.1810**	**.2256**	**.2557**	**.2872**

*The Welch test controls the Type I error rate in this setting.

Note. Type I error rate larger than .075 is in bold and larger than .08 it is shaded.

Table 4 Type I error rate of the W.new test for three groups (with the same population skewness and excess kurtosis) for various levels of heteroscedasticity and skewness, and sample sizes of about 60.

g	0.5	0.81	1	1	1	1	1	1.21	1	1.3	1.4	1.5
h	0	0	0	0.03	0.04	0.05	0.06	0	0.07	0	0	0
Skewness	1.8	3.8	6.2	8.2	9.1	10.1	11.47	11.53	13.1	15.6	22.5	33.5
Exc. kurt.	5.9	33.3	111	257	360	524	795	561	1270	1263	3401	10075
$(n_1,n_2,n_3) = (50,60,70)$												
$(\sigma_1,\sigma_2,\sigma_3)$	W.new test											
(1,1,1)	.0582*	.0721*	**.0792***	**.0818***	**.0814***	**.0831***	**.0845***	**.0884***	**.0854***	**.0914***	**.0962***	**.1007***
(1,1.21,1.41)	.0579*	.0701*	**.0769***	**.0797***	**.0803***	**.0825***	**.0830***	**.0871***	**.0825***	**.0917***	**.0957***	**.0982***
(1,1.37,1.73)	.0565*	.0707*	**.0782***	**.0812***	**.0796***	**.0818***	**.0830***	**.0880***	**.0833***	**.0920***	**.0968***	**.1026***
(1,1.5,2)	.0581*	.0684*	**.0782***	**.0804***	**.0802***	**.0810***	**.0842***	**.0885***	**.0853***	**.0929***	**.1005***	**.1061***
(1,2,3)	.0553*	.0671*	**.0783**	**.0823**	**.0827**	**.0855**	**.0855**	**.0931**	**.0867**	**.1002**	**.1113**	**.1214**
(1,3,4)	.0551*	.0640	.0729	**.0803**	**.0817**	**.0837**	**.0851**	**.0925**	**.0890**	**.1002**	**.1126**	**.1273**
(1,3,5)	.0533*	.0613	.0727	**.0766**	**.0772**	**.0815**	**.0833**	**.0883**	**.0854**	**.0970**	**.1067**	**.1219**
(1,4,6)	.0515*	.0578	.0650	.0732	.0745	**.0752**	**.0793**	**.0815**	**.0815**	**.0888**	**.1007**	**.1142**
(1,4,7)	.0515*	.0570	.0656	.0703	.0732	**.0752**	**.0776**	**.0775**	**.0793**	**.0887**	**.0988**	**.1076**
(1,5,8)	.0509*	.0557	.0600	.0657	.0703	.0702	.0710	.0736	**.0762**	**.0794**	**.0890**	**.1014**
(1.41,1.21,1)	.0566*	.0707*	**.0797***	**.0836***	**.0847***	**.0846***	**.0856***	**.0933***	**.0883***	**.0978***	**.1047***	**.1111***
(1.73,1.37,1)	.0573*	.0692*	**.0798***	**.0841***	**.0857***	**.0864***	**.0878***	**.0946**	**.0902***	**.1030**	**.1105**	**.1194**
(2,1.5,1)	.0557*	.0685*	**.0797***	**.0840**	**.0870**	**.0880**	**.0902**	**.0969**	**.0925**	**.1052**	**.1134**	**.1247**
(3,2,1)	.0553*	.0654	**.0779**	**.0841**	**.0855**	**.0877**	**.0921**	**.0997**	**.0930**	**.1071**	**.1188**	**.1349**
(4,3,1)	.0521*	.0614	.0729	**.0796**	**.0825**	**.0819**	**.0874**	**.0916**	**.0910**	**.1028**	**.1155**	**.1332**
(5,3,1)	.0530*	.0587	.0697	**.0761**	**.0778**	**.0787**	**.0826**	**.0873**	**.0860**	**.0973**	**.1116**	**.1253**
(6,4,1)	.0519*	.0565	.0628	.0709	.0739	.0748	**.0759**	**.0789**	**.0793**	**.0878**	**.0999**	**.1148**
(7,4,1)	.0519*	.0549	.0617	.0679	.0702	.0727	**.0750**	.0746	**.0766**	**.0826**	**.0946**	**.1096**
(8,5,1)	.0502*	.0529	.0583	.0637	.0660	.0670	.0700	.0686	.0711	.0741	**.0848**	**.0986**
	Welch test											
Max.	.0654	**.0958**	**.1262**	**.1370**	**.1414**	**.1431**	**.1489**	**.1664**	**.1540**	**.1874**	**.2142**	**.2412**
$(n_1,n_2,n_3) = (60,60,60)$												
	W.new test											
(1,1,1)	.0577*	.0706*	**.0766***	**.0806***	**.0808***	**.0807***	**.0816***	**.0868***	**.0826***	**.0911***	**.0959***	**.0984***
(1,1.21,1.41)	.0577*	.0693*	**.0791***	**.0793***	**.0826***	**.0839***	**.0840***	**.0908***	**.0855***	**.0935***	**.0982***	**.1038***
(1,1.37,1.73)	.0579*	.0698*	**.0776***	**.0827***	**.0828***	**.0841***	**.0859***	**.0918***	**.0860***	**.0957***	**.1025***	**.1088**
(1,1.5,2)	.0570*	.0678*	**.0799***	**.0826***	**.0824***	**.0846***	**.0856***	**.0933**	**.0894***	**.0997**	**.1083**	**.1155**
(1,2,3)	.0549*	.0666*	**.0773**	**.0835**	**.0850**	**.0868**	**.0883**	**.0941**	**.0897**	**.1030**	**.1139**	**.1272**
(1,3,4)	.0521*	.0632	.0728	**.0792**	**.0835**	**.0840**	**.0876**	**.0921**	**.0891**	**.1029**	**.1142**	**.1288**
(1,3,5)	.0536*	.0616	.0718	**.0758**	**.0785**	**.0817**	**.0852**	**.0868**	**.0840**	**.0947**	**.1096**	**.1235**
(1,4,6)	.0508*	.0565	.0664	.0697	.0727	.0739	**.0785**	**.0790**	**.0813**	**.0880**	**.1006**	**.1140**
(1,4,7)	.0516*	.0563	.0632	.0683	.0711	.0727	.0750	**.0775**	**.0776**	**.0865**	**.0964**	**.1094**
(1,5,8)	.0506*	.0528	.0588	.0648	.0668	.0683	.0705	.0699	.0710	**.0773**	**.0878**	**.1004**
	Welch test											
Max.	.0642	**.0916**	**.1185**	**.1297**	**.1346**	**.1353**	**.1414**	**.1587**	**.1452**	**.1771**	**.2020**	**.2297**

*The Welch test controls the Type I error rate in this setting.

Note. Type I error rate larger than .075 is in bold and larger than .08 it is shaded.

If both heavy-tailed and light-tailed distributions are considered, a combination of two effects on the Type I error rate take place. As observed earlier, it looks as if negative excess kurtosis pulls the Type I error rate down and brings closer together the behaviors of the transformation-based Welch tests for comparing means, whereas positive excess kurtosis pulls the Type I error rate up and separates both tests, with the effects being higher as the absolute excess kurtosis increases. In relation to power, light-tailed distributions seem to pull it down for the Wt.John test. So, when both light- and heavy-tailed distributions are present, the resulting Type I error rate and power is a combination of effects and depends on the level of skewness and excess kurtosis of the populations. The general conclusion observed in Figure 10 and Figure 11 seems to be that the W.new test has a smaller Type I error rate than the W.John test, with similar estimated power. The Wt.John test appears to have less power than the W.John and W.new tests in many settings, and no lower Type I error rate appears to be associated in some of those cases. In summary, the superiority of the Wt.John test found when most distributions are heavy-tailed is not clear when these are combined with light-tailed.

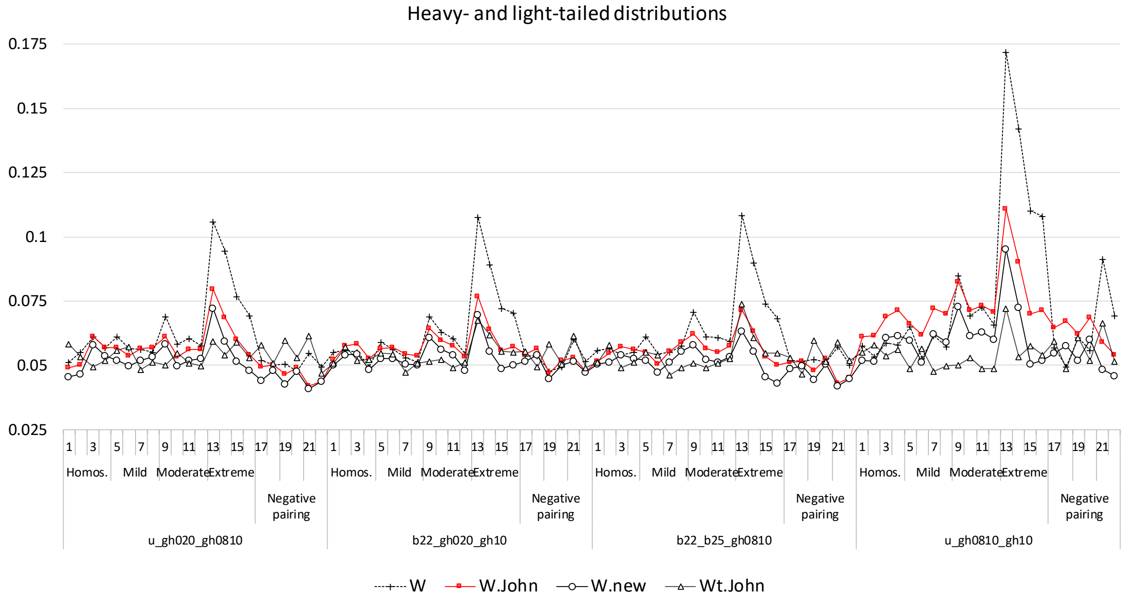

Figure 10 The Type I error rate of tests for each of the 22 settings with heavy- and light-tailed distributions.

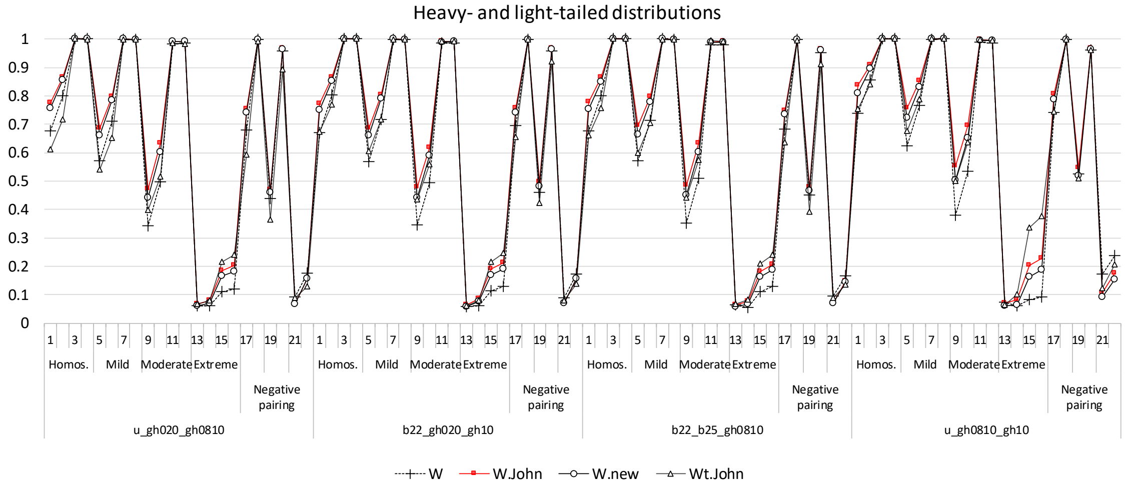

Figure 11 Estimated power of tests for each of the 22 settings with heavy- and light-tailed distributions.

In the case of six groups the W.new test still controls the Type I error rate for the symmetric distributions considered, whereas the W.John test does not control it in the case of six t_4-Student distributions. If distributions are skewed, the estimated significance level of the transformation-based Welch tests for comparing means seems to increase with the number of groups, leading to a loss of control in many settings, especially with homoscedasticity and low levels of heteroscedasticity. However, the W.new test is again superior to the Welch test if heteroscedasticity is high.

To illustrate the behavior of the W.new test with six skewed distributions, some simulation results have been included in Table 5 (sample sizes of about 30) and Table 6 (sample sizes of about 60) for a wide range of heteroscedasticity levels and several combinations of distributions with the same skewness and kurtosis. It is observed a general increase of the Type I error rate compared to the case of three groups. If sample sizes are of about 30, the W.new test controls the Type I error rate for all heteroscedasticity patterns considered for low skewness, unlike the Welch test. As in the case of three groups, the Welch test does not control the Type I error rate if distributions are asymmetric heavy-tailed and there is high heteroscedasticity, but the W.new test can control for large sample sizes depending on skewness. In summary, for skewed distributions and six groups it is observed that the W.new test appears to control the Type I error rate in settings where the Welch test does not control it, with the presence of heteroscedasticity, but it needs higher heteroscedasticity as the skewness increases. For sample sizes of about 60 (see Table 6) the W.new test seems to control the Type I error rate for skewness of 3.8 (and excess kurtosis of 33) and variance ratio higher than 1:9. For skewness of 6.2 (and excess kurtosis of 111) the variance ratio for which it controls is somewhat greater than 1:64.

Table 5 Type I error rate of the W.new test for six groups (with the same population skewness and excess kurtosis) for various levels of heteroscedasticity and skewness, and sample sizes of about 30.

g	0.5	0.81	0.5	0.81
h	0	0	0	0
Skewness	1.8	3.8	1.8	3.8
Excess kurtosis	5.9	33.3	5.9	33.3
$(n_1,n_2,n_3,n_4,n_5,n_6)$	(25,27,29,31,33,35)		(30,30,30,30,30,30)	
$(\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5,\sigma_6)$	W.new test			
(1, 1, 1, 1, 1, 1)	.0657*	**.0923***	.0678*	**.0924***
(1, 1.08, 1.17, 1.25, 1.33, 1.41)	.0667*	**.0919***	.0674*	**.0918***
(1, 1.15, 1.29, 1.44, 1.59, 1.73)	.0658*	**.0909***	.0655*	**.0918**
(1, 1.2, 1.4, 1.6, 1.8, 2)	.0661*	**.0909***	.0665*	**.0904**
(1, 1.4, 1.8, 2.2, 2.6, 3)	.0640*	**.0898**	.0665*	**.0917**
(1, 1.6, 2.2, 2.8, 3.4, 4)	.0656*	**.0896**	.0632*	**.0898**
(1, 1.8, 2.6, 3.4, 4.2, 5)	.0646*	**.0896**	.0659	**.0887**
(1, 2, 3, 4, 5, 6)	.0628*	**.0879**	.0635	**.0892**
(1, 2.2, 3.4, 4.6, 5.8, 7)	.0631	**.0857**	.0630	**.0859**
(1, 2.4, 3.8, 5.2, 6.6, 8)	.0619	**.0866**	.0640	**.0852**
(1.41, 1.33, 1.25, 1.17, 1.08, 1)	.0660*	**.0929**		
(1.73, 1.59, 1.44, 1.29, 1.15, 1)	.0681*	**.0923**		
(2, 1.8, 1.6, 1.4, 1.2, 1)	.0675*	**.0929**		
(3, 2.6, 2.2, 1.8, 1.4, 1)	.0661*	**.0924**		
(4, 3.4, 2.8, 2.2, 1.6, 1)	.0649*	**.0906**		
(5, 4.2, 3.4, 2.6, 1.8, 1)	.0657	**.0916**		
(6, 5, 4, 3, 2, 1)	.0659	**.0896**		
(7, 5.8, 4.6, 3.4, 2.2, 1)	.0636	**.0877**		
(8, 6.6, 5.2, 3.8, 2.4, 1)	.0632	**.0869**		
	Welch test			
Max.	**.0865**	**.1494**	**.0849**	**.1396**

*The Welch test controls the Type I error rate in this setting.

Note. Type I error rate larger than .075 is in bold and larger than .08 it is shaded.

Table 6 Type I error rate of the W.new test for six groups (with the same population skewness and excess kurtosis) for various levels of heteroscedasticity and skewness, and sample sizes of about 60.

g	0.5	0.81	1	1	0.5	0.81	1
h	0	0	0	0.03	0	0	0
Skewness	1.8	3.8	6.2	8.2	1.8	3.8	6.2
Excess kurtosis	5.9	33.3	111	257	5.9	33.3	111
$(n_1,n_2,n_3,n_4,n_5,n_6)$	(55,57,59,61,63,65)				(60,60,60,60,60,60)		
$(\sigma_1,\sigma_2,\sigma_3,\sigma_4,\sigma_5,\sigma_6)$	W.new test						
(1, 1, 1, 1, 1, 1)	.0605*	**.0801***	**.0965***	**.1039***	.0613*	**.0780***	**.0966***
(1, 1.08, 1.17, 1.25, 1.33, 1.41)	.0610*	**.0802***	**.0963***	**.1038***	.0614*	**.0781***	**.0960***
(1, 1.15, 1.29, 1.44, 1.59, 1.73)	.0603*	**.0789***	**.0957**	**.1028**	.0594*	**.0789***	**.0949**
(1, 1.2, 1.4, 1.6, 1.8, 2)	.0604*	**.0795***	**.0941**	**.1011**	.0603*	**.0778***	**.0944**
(1, 1.4, 1.8, 2.2, 2.6, 3)	.0606*	**.0750**	**.0932**	**.1014**	.0592*	**.0763**	**.0916**
(1, 1.6, 2.2, 2.8, 3.4, 4)	.0592*	.0738	**.0896**	**.0992**	.0576*	.0740	**.0904**
(1, 1.8, 2.6, 3.4, 4.2, 5)	.0569*	.0716	**.0869**	**.0961**	.0568*	.0708	**.0865**
(1, 2, 3, 4, 5, 6)	.0559*	.0703	**.0848**	**.0934**	.0569*	.0696	**.0842**
(1, 2.2, 3.4, 4.6, 5.8, 7)	.0555*	.0669	**.0822**	**.0933**	.0551*	.0685	**.0824**
(1, 2.4, 3.8, 5.2, 6.6, 8)	.0563*	.0660	**.0806**	**.0901**	.0548*	.0652	**.0803**
(1.41, 1.33, 1.25, 1.17, 1.08, 1)	.0608*	**.0812***	**.0967**	**.1050**			
(1.73, 1.59, 1.44, 1.29, 1.15, 1)	.0595*	**.0797***	**.0964**	**.1031**			
(2, 1.8, 1.6, 1.4, 1.2, 1)	.0603*	**.0803***	**.0946**	**.1026**			
(3, 2.6, 2.2, 1.8, 1.4, 1)	.0582*	**.0757**	**.0934**	**.1004**			
(4, 3.4, 2.8, 2.2, 1.6, 1)	.0559*	.0721	**.0898**	**.0974**			
(5, 4.2, 3.4, 2.6, 1.8, 1)	.0561*	.0717	**.0872**	**.0977**			
(6, 5, 4, 3, 2, 1)	.0551*	.0702	**.0849**	**.0949**			
(7, 5.8, 4.6, 3.4, 2.2, 1)	.0535*	.0666	**.0828**	**.0910**			
(8, 6.6, 5.2, 3.8, 2.4, 1)	.0535*	.0654	**.0800**	**.0885**			
	Welch test						
Max.	.0707	**.1127**	**.1532**	**.1673**	.0712	**.1088**	**.1480**

*The Welch test controls the Type I error rate in this setting.

Note. Type I error rate larger than .075 is in bold and larger than .08 it is shaded.

6. Concluding remarks

A new transformation of the Welch statistic is proposed to correct the effects of skewness and kurtosis of the parent populations on the Type I error rate. It can be considered as an extension of Johnson’s transformation. This paper focuses primarily on the comparison of the W.new test with the W.John, W.Hall and Welch tests, since they all have the same null hypothesis of equal means, and also with the Wt.John test, which tests the equality of a more robust central tendency measure (the trimmed mean) and has been shown to perform well in heavy-tailed distributions.

According to the simulation results, the W.new test competes very well with conventional procedures when standard assumptions are met. If distributions are heavy-tailed, it has a lower Type I error rate than the W.John and the W.Hall tests, with no loss of power. On the other hand, the W.new test performs better than the Wt.John test when distributions are normal, near-normal or light-tailed, since it has a similar or lower Type I error rate with higher power.

Although in symmetric distributions the population trimmed mean matches the usual mean, testing equal trimmed means leads to a lower power, except when populations are heavy-tailed. Using trimmed means in symmetric populations seems to cause a higher power loss when the kurtosis is lower. So, trimming means is not recommendable in symmetric distributions, unless populations are heavy-tailed.

In the case of asymmetric distributions, the behavior of the W.new test depends on the level of the skewness and heteroscedasticity. The Welch test controls the Type I error rate except for asymmetric heavy-tailed distributions with extreme heteroscedasticity, in which case the W.new test performs better, and even controls in certain scenarios.

The Wnew.test shows a special ability to control the Type I error rate if there is high skewness accompanied by high heteroscedasticity in large enough samples. The Type I error rate of the W.new test diminishes as sample sizes increase, but, given a skewness level, the test requires smaller samples to control in higher heteroscedasticity scenarios. On the other hand, given the samples sizes, it needs higher heteroscedasticity the higher the skewness. Furthermore, a negative pairing between variances and skewness may reduce the estimated significance level, especially in small samples.

The Type I error rate of the W.new test seems to rise with the number of groups so that, even for moderate skewness and kurtosis, a larger variance ratio for the same sample sizes is required to control the Type I error rate. Otherwise, the samples sizes needed to control the Type I error rate are getting larger.

For asymmetric heavy-tailed populations, the Wt.John test has an estimated significance level lower than the W.new test and nearer to the nominal level. However, it seems to be inflated if heteroscedasticity is extreme. In general, it also has slightly better power than the W.new test. This superiority of the Wt.John test can be lost if some distributions have negative excess kurtosis.

In summary, the Welch test based on the new transformation might be an alternative to take into consideration when testing for mean equality, because it performs well in a wide range of settings. Of note is its remarkable ability to control the Type I error rate in the adverse situation of extreme heteroscedasticity with asymmetric heavy-tailed distributions, given an appropriate sample size. Only when the skewness and kurtosis are very high does the W.new test provide Type I error rates quite far from the nominal level for extreme heteroscedasticity and with reasonable sample sizes. Nevertheless, in this case, applied researchers can consider comparing trimmed means more appropriate when the bulk of the distributions is studied.

7. References

Algina, James, T. C. Oshima, and Wen-Ying Lin. 1994. ‘Type I Error Rates for Welch’s Test and James’s Second-Order Test under Nonnormality and Inequality of Variance When There Are Two Groups’. *Journal of Educational and Behavioral Statistics* 19 (3): 275–91. doi:10.2307/1165297.

Blanca, María J., Jaume Arnau, Dolores López-Montiel, Roser Bono, and Rebecca Bendayan. 2013. ‘Skewness and Kurtosis in Real Data Samples’. *Methodology:*

European Journal of Research Methods for the Behavioral and Social Sciences 9 (2): 78–84. doi:10.1027/1614-2241/a000057.

Bradley, James V. 1978. 'Robustness?' *British Journal of Mathematical and Statistical Psychology* 31 (2): 144–52. doi:10.1111/j.2044-8317.1978.tb00581.x.

Cornish, E. A., and R. A. Fisher. 1938. 'Moments and Cumulants in the Specification of Distributions'. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute* 5 (4): 307–20. doi:10.2307/1400905.

Cribbie, Robert A., Lisa Fiksenbaum, Harvey J. Keselman, and Rand R. Wilcox. 2012. 'Effect of Non-Normality on Test Statistics for One-Way Independent Groups Designs'. *The British Journal of Mathematical and Statistical Psychology* 65 (1): 56–73. doi:10.1111/j.2044-8317.2011.02014.x.

Cribbie, Robert A., Rand R. Wilcox, Carmen Bewell, and Harvey J. Keselman. 2007. 'Tests for Treatment Group Equality When Data Are Nonnormal and Heteroscedastic'. *Journal of Modern Applied Statistical Methods* 6 (1): 117–32. doi:10.22237/jmasm/1177992660.

Dijkstra, Jan B., and Paul S. P. J. Werter. 1981. 'Testing the Equality of Several Means When the Population Variances Are Unequal'. *Communications in Statistics - Simulation and Computation* 10 (6): 557–69. doi:10.1080/03610918108812235.

Guo, Jiin-Huarng, and Wei-Ming Luh. 2000. 'An Invertible Transformation Two-Sample Trimmed t-Statistic under Heterogeneity and Nonnormality'. *Statistics & Probability Letters* 49 (1): 1–7. doi:10.1016/S0167-7152(00)00022-5.

Hall, Peter. 1992. 'On the Removal of Skewness by Transformation'. *Journal of the Royal Statistical Society. Series B (Methodological)* 54 (1): 221–28.

Hoaglin, David C. 1985. 'Summarizing Shape Numerically: The g-and-h Distributions'. IN (libro).

Johnson, Norman J. 1978. ‘Modified t Tests and Confidence Intervals for Asymmetrical Populations’. *Journal of the American Statistical Association* 73 (363): 536–44. doi:10.2307/2286597.

Keselman, Harvey J., Rand R. Wilcox, Abdul Othman, and Katherine Fradette. 2002. ‘Trimming, Transforming Statistics, And Bootstrapping: Circumventing the Biasing Effects Of Heterescedasticity And Nonnormality’. *Journal of Modern Applied Statistical Methods* 1 (2). doi:10.22237/jmasm/1036109820.

Lix, Lisa M., Joanne C. Keselman, and Harvey J. Keselman. 1996. ‘Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test’. *Review of Educational Research* 66 (4): 579–619. doi:10.3102/00346543066004579.

Luh, Wei-Ming, and Jiin-Huarng Guo. 1999. ‘A Powerful Transformation Trimmed Mean Method for One-Way Fixed Effects ANOVA Model under Non-Normality and Inequality of Variances’. *British Journal of Mathematical and Statistical Psychology* 52 (2): 303–20. doi:10.1348/000711099159125.

Micceri, Theodore. 1989. ‘The Unicorn, the Normal Curve, and Other Improbable Creatures’. *Psychological Bulletin* 105 (1): 156–66. doi:10.1037/0033-2909.105.1.156.

Parra-Frutos, Isabel. 2014. ‘Controlling the Type I Error Rate by Using the Nonparametric Bootstrap When Comparing Means’. *British Journal of Mathematical & Statistical Psychology* 67 (1): 117–32. doi:10.1111/bmsp.12011.

Welch, B. L. 1951. ‘On the Comparison of Several Mean Values: An Alternative Approach’. *Biometrika* 38 (3/4): 330–36. doi:10.2307/2332579.

Wilcox, Rand R. 1989. ‘Adjusting for Unequal Variances When Comparing Means in One-Way and Two-Way Fixed Effects ANOVA Models’. *Journal of Educational and Behavioral Statistics* 14 (3): 269–78. doi:10.3102/10769986014003269.

———. 1990. ‘Comparing the Means of Two Independent Groups’. *Biometrical Journal* 32 (7): 771–80. doi:10.1002/bimj.4710320702.

———. 2017. *Introduction to Robust Estimation and Hypothesis Testing*. 4th ed. Amsterdam: Elsevier Inc.

APPENDIX

```
Wnew.test <- function (formula, data, alpha = .05) {
  dp = as.character(formula)
  y = data[, dp[[2L]]]
  group = data[, dp[[3L]]]
  x.levels <- levels(factor(group))
  y.n <- y.means <- y.vars <- m3 <- m4 <- w <- NULL
  for (i in x.levels) {
    samplei <- y[group == i]
    y.n[i] <- length(samplei)
    y.vars[i] <- var(samplei)
    w[i] <- y.n[i]/y.vars[i]
    y.means[i] <- mean(samplei)
    m3[i] <- sum((samplei - y.means[i])^3)/y.n[i]
    m4[i] <- sum((samplei - y.means[i])^4)/y.n[i]
  }
  U = sum(w)
  w_y = sum(w * y.means)/U
  J = length(x.levels)
  kurt <- m4/((y.n-1)*y.vars/y.n)^2 - 3
  # Transformation
  lambda <- m3/(6 * y.n * y.vars)
  gamma1 <- m3/(3 * y.vars^2)
  gamma2 <- -gamma1^2 - kurt/(24 * y.vars)
  T <- (y.means - w_y) + lambda + gamma1 * (y.means - w_y)^2 + gamma2 * (y.means -
w_y)^3

  # Statistic
  A <- sum(w * T^2)/(J - 1)
  B <- 2 * (J - 2)/(J^2 - 1) * sum((1 - w/U)^2/(y.n - 1))
  Ftest <- A/(B + 1)
  df1 <- J - 1
  df2 <- (J^2 - 1) / (3* sum((1 - w/U)^2/(y.n - 1)))
  p.value <- pf(Ftest, df1, df2, lower.tail = FALSE)

  result <- list()
  result$alpha <- alpha
  result$statistic <- Ftest
  result$df <- c(df1, df2)
  result$criticalvalue <- qf(alpha, df1, df2, lower.tail = FALSE)
  result$p.value <- p.value
  result
}
```