Journal of Modern Applied Statistical Methods

Volume 2 | Issue 2 Article 5

11-1-2003

A Critical Examination Of The Use Of Preliminary Tests In Two-Sample Tests Of Location

Kimberly T. Perry
Pfizer Inc., Kimberly.t.perry@pfizer.com

DOI: 10.22237/jmasm/1067645100

A Critical Examination Of The Use Of Preliminary Tests In Two-Sample Tests Of Location

Cover Page Footnote
Michael Stoline is acknowledged for his support.

Journal of Modern Applied Statistical Methods
November, 2003, Vol. 2, No. 2, 314-328

1538 – 9472/03/$30.00

A Critical Examination Of The Use Of Preliminary Tests In Two-Sample Tests Of Location

Kimberly T. Perry
Pfizer Inc.
Kalamazoo, Michigan

This paper explores the appropriateness of testing the equality of two means using either a *t* test, the Welch test, or the Wilcoxon-Mann-Whitney test for two independent samples based on the results of using two classes of preliminary tests (i.e., tests for population variance equality and symmetry in underlying distributions).

Key words: *t* test, Welch, Wilcoxon-Mann-Whitney, Levene, preliminary test for variance, triples test, test of symmetry, test selection

Introduction

In practice, the two-sample *t* test is widely used to test the equality of two means. However, it is well known that the assumptions of independence (which will not be discussed in this paper), variance homogeneity and normality must be met for the two-sample *t* test to perform well. Results from Zimmerman and Williams (1989), Gans (1981), Murphy (1976), and Snedecor & Cochran (1967) have demonstrated that the Welch test or the Wilcoxon-Mann-Whitney (WMW) test is more robust in certain cases of variance heterogeneity or non-normality.

Based on the above results for testing the equality of means, we conclude the following:

1. The *t* test is robust when the distributions are symmetric and the variances are equivalent.
2. The Welch test is robust when the distributions are symmetric and the variances are unequal.
3. The Wilcoxon-Mann-Whitney test is robust when the distributions are asymmetric and the variances are equivalent.
4. None of the above three methods are robust when the distributions are asymmetric and the variances are unequal.

Therefore it would be useful to use the results from two classes of preliminary test to determine which of the three tests, the *t* test, the Welch test, or the Wilcoxon-Mann-Whitney test, should be used to test the hypothesis H_o: $\mu_1 = \mu_2$. One class of preliminary tests determines whether the population variances differ, and the other class ascertains if the underlying distributions are symmetric or skewed.

Tests of Variances Used as Preliminary Tests

The goal of the preliminary test for variance heterogeneity is to indicate when to avoid using mean tests that are sensitive to variance heterogeneity.

Many methods for testing variance homogeneity have been developed and compared. Brown and Forsythe (1974), Conover, M.E. Johnson, and M.M. Johnson (1981), Loh (1987), and O'Brien (1979) have conducted simulations to examine the robustness of many popular methods for testing variance homogeneity. The L_{50}, the Levene test using the median, was found to be robust for the non-normal cases and was one of the procedures

Kimberly T. Perry is a Senior Research Advisor, Pfizer Inc., Kalamazoo, Michigan. Her areas of interest are innovated clinical study designs, multiple endpoint analysis, and interim analysis. Email: Kimberly.t.perry@pfizer.com. Michael Stoline is acknowledged for his support.

recommended by Conover et al. (1981) as well as the other authors cited above. Based on the above cited literature, the Levene test using the median might be a robust preliminary test procedure.

Furthermore, Olejnik (1987) conducted a study where the Levene test using the median was compared to the O'Brien procedure (1979) as a preliminary test procedure preceding the means test. His results showed the Levene test and the O'Brien procedure used as preliminary tests of variance homogeneity were only slightly more robust than using the *t* test alone. It is noted that Olejnik (1987) used significance levels of 5% and 10% for testing variance homogeneity in the preliminary test procedure.

It is of interest to examine the performance of the L_{50} test as a preliminary test procedure with a higher significance level. A higher significance level would aid in controlling the Type II error. For this simulation the Levene test at a significance level of 25% was arbitrary selected.

Test of Symmetry Used as Preliminary Tests

Randles, Fligner, Policello, and Wolfe (1980) compared three procedures for testing whether a univariate population is symmetric about some unspecified value compared to a large class of asymmetric distribution alternatives. These are the Triples test, Gupta's skewness test (Gupta, 1967) and Gupta's nonparametric procedure (Gupta, 1967). Their results show that the Triples test is superior to either competitor for testing the hypothesis of symmetry while possessing good power for detecting asymmetric alternative distributions (Randles et al., 1980).

In addition, Cabilio & Masaro (1996) and Perry and Stoline (2002) compared the Triples test to other tests of symmetry and the Triples test continued to perform well both on robustness and power. Based on the above studies, the Triples test is selected as a possible preliminary test of symmetry/skewness prior to the testing of means equality in a test selection procedure. A significance level of 5% for testing of symmetry was arbitrary chosen for this simulation.

Test Selection Procedure

The test selection procedure, hereafter denoted as the TS procedure, will select either a *t* test, the Welch test, or the Wilcoxon-Mann-Whitney test based on the results of the two preliminary tests. One class of preliminary tests determines whether the population variances differ, and the other class ascertains if the underlying distributions are symmetric or skewed. The "recommended" L_{50} test (hereafter denoted Levene test) will be assessed as preliminary test for variance homogeneity, whereas, the Triples test will be assessed as a preliminary test of symmetry/skewness. Based on the results of the two preliminary tests, the TS procedure is constructed in the following way:

1. The *t* test is used to test the equality of means if symmetry is accepted and variance homogeneity is accepted.
2. The Welch test is used to test the equality of means if symmetry is accepted and variance homogeneity is rejected.
3. The Wilcoxon-Mann-Whitney test is used to test the equality of means if symmetry is rejected and variance homogeneity is accepted.
4. The Welch test is used to test the equality of means if symmetry is rejected and variance homogeneity is rejected.

It is noted that robust methods exist for testing H_o: $\mu_1 = \mu_2$ for cases #1-3 above, but no robust method exists for case #4 .

Methodology

This section contains the details describing the two-sample methodology used to test the equality of means and variance homogeneity under selected distributions.

Let $x_{11}, \ldots, x_{1n_1}$ be a random sample with sample size of n_1 from a distribution denoted $f_1(x; \mu_1, \sigma_1)$; and $x_{21}, \ldots, x_{2n_2}$ be a random sample with sample size of n_2 from a distribution denoted $f_2(x; \mu_2, \sigma_2)$. It is assumed that $E(x_{ij}) = \mu_i$ and $Var(x_{ij}) = \sigma_i^2$ for each i= 1, 2 and j = 1,…, n_i. The two samples are assumed to be independent. Let the sample mean and

sample variance for $x_{i1}, \ldots, x_{in_i}$ be denoted as x_i and s_i^2 for $i = 1, 2$, respectively.

Testing the Equality of Means

The *t* test, the Welch test, and the Wilcoxon-Mann-Whitney test procedures of H_o: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$, are now described.

The *t* test is the given as

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{s^2(1/n_1 + 1/n_2)}}, \qquad (1a)$$

$$where\ s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \qquad (1b)$$

is the pooled estimate of σ^2, assuming $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

The Welch test statistic is

$$t_W = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{(s_1^2/n_1) + (s_2^2/n_2)}}, \qquad (2a)$$

which uses Satterthwaite's (1946) approximation for the degrees of freedom:

$$df = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}. \qquad (2b)$$

The Wilcoxon-Mann-Whitney statistic is

$$z = \frac{S - n_1(n_1 + 1)/2 - n_1 n_2/2}{\sqrt{n_1 n_2(n_1 + n_2 + 1)/12}}, \qquad (3)$$

where S is the sum of the ranks assigned to the sample observations from group 1, and *z* is an approximate normal deviate.

The α-level tests of H_o: $\mu_1 = \mu_2$ vs. H_1: $\mu_1 \neq \mu_2$ are $|t| > t_{\alpha/2,\, n_1 + n_2 - 2}$, $|t_w| > t_{\alpha/2, df}$, and $|z| > z_{\alpha/2}$ for the *t* test, the Welch test, and the Wilcoxon-Mann-Whitney test, respectively, where z_α is the upper α-point of the standard unit normal distribution and $t_{\alpha,r}$ is the upper α-point of a *t* distribution with *r* degrees of freedom.

Testing the Equality of Variances

The Levene test of H_o: $\sigma_1^2 = \sigma_2^2$ vs. H_1: $\sigma_1^2 \neq \sigma_2^2$ is now described, assuming the sampling conditions described above hold.

The Levene α-level test is

$$L = \frac{\sum n_i(z_{i.} - z_{..})^2}{\sum\sum(z_{ij} - z_{i.})^2/(n_1 + n_2 - 2)} > F_{\alpha,\, 1,\, n_1 + n_2 - 2}, \qquad (2.5)$$

which is the one-way analysis of variance *F*-test computed on the z_{ij} values, where $z_{ij} = |x_{ij}$ - median of group i|.

Testing of Symmetry

The Triples test, as described in a paper by Randles, Fligner, Policello, and Wolfe (1980), is a test to determine if a distribution is symmetric. The procedure used to obtain the test statistic is outlined in Perry and Stoline (2002) and is not repeated here.

Selected Configurations of Distributions, Sample Sizes and Variance Ratios Used in the Simulation

Type I error rates for testing the homogeneity of means were simulated under a variety of conditions using four probability distributions. Each of these four distributions is classified into one of two groups: (1) symmetric and (2) asymmetric.

The Results section examines the use of the TS procedure using two classes of preliminary tests (i.e., testing for variance homogeneity and testing for symmetry) preceding the test of equality of means, H_o: $\mu_1 = \mu_2$ for the two symmetric distributions: (1) normal and (2) double exponential. In addition, the *Results* section examines the TS procedure for the two asymmetric distributions: (1) lognormal and (2) gamma.

To evaluate the performance of the preliminary test of variance homogeneity, the following standard deviation ratios $R = \sigma_1 / \sigma_2$ are used: 0.25, 0.50, 1.0, 2.0, and 4.0. Clearly the standard deviations are equal when $R = 1$. Sample size configurations (n_1:n_2) used in the simulations are: (10:10), (10:20), (10:40), and

(20:20). This allows for both direct and indirect pairings to be examined.

Direct pairing occurs when either R = 0.25 and 0.50 holds with any of the imbalanced samples (10:20) and (10:40). Direct pairing occurs when the group with the smaller σ is associated with the group with the smaller sample size.

Indirect pairing occurs when either R = 2.0 and 4.0 holds with any of the imbalanced sample sizes (10:20) and (10:40). Indirect pairing occurs when the group with the smaller σ is associated with the group with the larger sample size.

Generation of Random Realizations

This section contains an outline of how the random realizations are generated for each specified distribution. As before, let $x_{11}, \ldots, x_{1n_1}$ be a random sample of size n_1 from the distribution $f_1(x; \mu_1, \sigma_1)$; and $x_{21}, \ldots, x_{2n_2}$ be a random sample of size n_2 from the distribution $f_2(x; \mu_2, \sigma_2)$, where it is assumed that the two samples are independent.

The random realizations from the standardized distribution f_2 $(x; \mu_2, \sigma_2)$ are generated for each of the selected distributions. For the first sample, f_1 $(x; \mu_1, \sigma_1)$, the random realizations are generated in the same fashion, but shape parameters and scale parameters are adjusted to yield the desired standard deviation ratio $R = \sigma_1/\sigma_2$. Details on each of the four selected distributions are outlined in Perry and Stoline (2002). The IMSL random number generator RNSET, which initializes the seed, is used in all of the simulations.

Testing the Equality of Means Using the TS Procedure

The TS procedure has been described in the Introduction section. Figure 1 is a diagram of how the TS procedure is constructed.

Figure 1. Components of the TS procedure

		H_o: Symmetry	
		Accepted ↓	Rejected ↓
H_o: Variance Homogeneity (H_o: $\sigma_1 = \sigma_2$)	Accepted →	*t* test	WMW test
	Rejected →	Welch test	Welch test

Notes: WMW = Wilcoxon-Mann-Whitney.

Asymmetry is concluded if at least one of the samples is declared skewed. Another alternative would be that skewness is declared significant only if both samples are skewed. It was arbitrary chosen for this simulation to use the former approach with asymmetry being concluded if at least one of the samples is declared skewed.

Results

In this section, the performance of the TS procedure is evaluated. The "TS procedure" denotes the results of the test selection procedure using the 5% Triples test for testing symmetry and the 25% Levene test for testing variance homogeneity.

Symmetric Distributions

For each of the two symmetric distributions (i.e., normal and double exponential) as defined in Perry and Stoline (2002), the simulations are conducted for the four selected sample size combinations (n_1:n_2)= (10:10), (10:20), (10:40), and (20:20). For each of the four sample size combinations, the simulated null rejection rate is generated for the specified ratio $R = \sigma_1/\sigma_2$. These are: (1) R = 0.25, (2) R = 0.50, (3) R = 1 (equal variance), (4) R = 2.0, and (5) R = 4.0.

The results of the simulations for the two symmetric distributions are combined in Table 1. The proportions of rejections are expressed as a percent for the *t* test, the Welch test, the Wilcoxon-Mann-Whitney test, and the TS procedure. These proportions are tabulated for each R grouping combined over <u>all</u> (8) combinations of sample size pairs (4) and distributions (2) for the five categories listed below:

1. $x \leq 2.5$ (extremely conservative)
2. $2.5 < x \leq 4.0$ (conservative)
3. $4.0 < x \leq 6.0$ (robust)
4. $6.0 < x \leq 10.0$ (liberal)
5. $x > 10.0$ (extremely liberal)

The value x represents the percentage of rejections for testing H_o: $\mu_1 = \mu_2$ based on 10,000 simulations for each sample size. Each entry in the following tables denotes the frequency at which $a < x \leq b$ occurs. The outcome of the "test" is defined to be robust if the simulated null rejection rate is > 4.0 and ≤ 6.0.

Equal Variance Cases (R=1)

Table 1 shows, as anticipated, that the t test is robust for the equal variance cases. However, the other procedures are also robust. None of the procedures examined show simulated rejection rates $\leq 4.0\%$ or $> 6\%$.

Unequal Variance Cases

Table 1 shows the *t* test is extremely conservative in 50% of the simulations for the R = 0.25 and 0.50 cases. The WMW test is liberal for the R = 0.50 cases and can be extremely conservative for both the R = 0.25 and the R = 0.50 cases. The Welch test and the TS procedure are robust for both the R = 0.25 and R = 0.50 cases.

For the R = 2.0 cases the *t* test is extremely liberal. The WMW test tends to be liberal and can be extremely liberal. The TS procedure is reasonably robust. The Welsh test is robust.

For the R = 4.0 cases, the *t* test and the WMW test are extremely liberal in 50% of the simulations. The Welsh test and the TS procedure are reasonably robust.

Table 1. Summary Of Symmetric Distributions Using TS Procedure: Frequency (%) Of Simulated Null Rejection Rate (%) With Nominal 5% Level.

R	Test	Extremely Conservative ≤2.5	Conservative 2.5< x ≤4	Robust 4< x ≤6	Liberal 6< x ≤10	Extremely Liberal x > 10
			$\sigma_1 = \sigma_2$			
1.00	t	0.0	0.0	100.0	0.0	0.0
	W	0.0	0.0	100.0	0.0	0.0
	WMW	0.0	0.0	100.0	0.0	0.0
	TS	0.0	0.0	100.0	0.0	0.0
			$\sigma_1 \neq \sigma_2$			
0.50	t	50.0	0.0	50.0	0.0	0.0
	W	0.0	0.0	100.0	0.0	0.0
	WMW	25.0	25.0	0.0	50.0	0.0
	TS	0.0	0.0	100.0	0.0	0.0
0.25	t	50.0	0.0	50.0	0.0	0.0
	W	0.0	0.0	100.0	0.0	0.0
	WMW	25.0	25.0	50.0	0.0	0.0
	TS	0.0	0.0	100.0	0.0	0.0
2.0	t	0.0	0.0	50.0	12.5	37.5
	W	0.0	0.0	100.0	0.0	0.0
	WMW	0.0	0.0	50.0	37.5	12.5
	TS	0.0	0.0	75.0	25.0	0.0
4.0	t	0.0	0.0	37.5	12.5	50.0
	W	0.0	12.5	87.5	0.0	0.0
	WMW	0.0	0.0	0.0	50.0	50.0
	TS	0.0	12.5	87.5	0.0	0.0

Notes: Table is based on the two symmetric distributions (normal and double exponential) and four sample sizes. W = Welch, WMW = Wilcoxon-Mann-Whitney.

Based on the above simulation results, the Welch test and the TS procedure are reasonably robust for testing the Ho: $\mu_1 = \mu_2$ for the symmetric cases examined.

Results For Asymmetric Distributions

To evaluate the overall performance of the procedures for varying degrees of variance heterogeneity, the results of the simulation for the two asymmetric distributions as defined in Perry and Stoline (2002) are combined in Table 2 using the same format as previously defined for the symmetric distributions.

For the gamma (2,1) distribution the coefficient of skewness ranged from 0.4 when R = 0.25 to approximately 5.7 when R = 4.0. For the lognormal (0, 0.40) distribution, the coefficient of skewness ranged from 0.3 when R = 0.25 to approximately 9.6 when R = 4.0. For each value of R within the gamma and lognorma1 case, a skewness ratio has been calculated. The skewness ratio is the skewness of distribution #1 divided by the skewness of distribution #2 within each gamma and lognormal case. The skewness ratios are displayed in Table 2.

Equal Variance Cases (R=1)

A summary of the simulated null rejection rates for the two asymmetric distributions for the equal variance cases are presented in Table 2. The WMW test and *t* test

are robust for the $R = 1$ cases. The Welsh test is robust for approximately 88% of the $R = 1$ cases. The TS procedure tends to be liberal for approximately 38% of these cases. None of the procedures are extremely liberal, extremely conservative, or conservative.

Unequal Variance Cases

Table 2 shows the Welch test is robust in approximately 75% of the $R = 0.50$ cases. The Welch test can be liberal for some $R = 0.50$ cases. The *t* test is conservative or extremely conservative for approximately 50% of the R = 0.50 cases. Furthermore, the *t* test is liberal in approximately 38% of the simulations for the R = 0.50 cases. The WMW test and the TS procedure are liberal or extremely liberal in at approximately 63% and 50%, respectively, for the R= 0.50 cases.

For the $R = 0.25$ cases, none of the test procedures are robust. The Welch test and the TS procedure tend to be liberal. The *t* test is liberal (50%) as well as extremely conservative (50%). The WMW test is liberal or extremely liberal in approximately 88% of the simulations for the $R = 0.25$ cases.

Table 2 shows all procedures tend to be liberal or extremely liberal for the $R = 2.00$ cases. Furthermore, all procedures are extremely liberal for 100% of the $R = 4$ cases.

In summary for the $R = 1$ cases, the *t* test, the Welsh test, and the WMW test are robust in at least 87% of the simulations. The TS procedure is robust in approximately 63% of the simulations for the R = 1 cases. For the $R = 0.50$ cases, the Welch test is robust for approximately 75% of the simulated cases. For the $R = 0.25$, 2.0 and 4.0 cases, all procedures tend to be liberal. The degree of liberal bias increases as the degree of variance heterogeneity increases.

Frequency (%) Each Means Test Is Used

In addition to the simulated null rejection rates, the TS procedure can report the frequency (%) at which each of the test procedures is used for a given sample size and *R* value. Results for the imbalanced case $n_1 = 10$ and $n_2 = 20$, and the balanced case $n_1 = n_2 = 20$ are summarized for the two symmetric distribution cases combined and the two asymmetric distribution cases combined.

Tables 3 and 4 summarize the frequency (%) at which each of the test procedures is used for the two symmetric distributions cases combined, and the two asymmetric cases combined, respectively. The format for Tables 3 and 4 is as follows. For each *R* value, the frequency at which the *t* test, the Welch-S test, the WMW test, and the Welch-AS test was selected by the TS procedure is reported. In these tables, the *t* test, Welch-S, WMW, and Welch-AS denote the following:

t test: The *t* test was used because the TS procedure concluded $\sigma_1 = \sigma_2$ and symmetry was accepted.

Welch-S: The Welch test was used because the TS procedure concluded $\sigma_1 \neq \sigma_2$ and symmetry was accepted.

WMW: The WMW test was used because the TS procedure concluded $\sigma_1 = \sigma_2$ and symmetry was rejected.

Welch-AS: The Welch test was used because the TS procedure concluded $\sigma_1 \neq \sigma_2$ and symmetry was rejected.

Table 2. Summary Of Asymmetric Distributions Using TS Procedure: Frequency (%) Of Simulated Null Rejection Rate With Nominal 5% Level.

R	Skewness Ratio Gamma, LN	Test	Extremely Conservative ≤2.5	Conservative 2.5< x ≤4	Robust 4< x ≤6	Liberal 6< x ≤10	Extremely Liberal x > 10
				$\sigma_1 = \sigma_2$			
1.00	1,1	t	0.0	0.0	100.0	0.0	0.0
		W	0.0	0.0	87.5	12.5	0.0
		WMW	0.0	0.0	100.0	0.0	0.0
		TS	0.0	0.0	62.5	37.5	0.0
				$\sigma_1 \neq \sigma_2$			
0.25	0.29,0.23	t	50.0	0.0	0.0	50.0	0.0
		W	0.0	0.0	37.5	62.5	0.0
		WMW	0.0	0.0	12.5	37.5	50.0
		TS	0.0	0.0	37.5	62.5	0.0
0.50	0.50,0.46	t	25.0	25.0	12.5	37.5	0.0
		W	0.0	0.0	75.0	25.0	0.0
		WMW	0.0	12.5	25.0	50.0	12.5
		TS	0.0	0.0	50.0	50.0	0.0
2.0	2.0, 2.39	t	0.0	0.0	0.0	50.0	50.0
		W	0.0	0.0	0.0	75.0	25.0
		WMW	0.0	0.0	0.0	0.0	100.0
		TS	0.0	0.0	0.0	12.5	87.5
4.0	4.04, 7.4	t	0.0	0.0	0.0	0.0	100.0
		W	0.0	0.0	0.0	0.0	100.0
		WMW	0.0	0.0	0.0	0.0	100.0
		TS	0.0	0.0	0.0	0.0	100.0

Notes: Table is based on the two asymmetric distributions [lognormal (0, 0.40) & G(2,1)]and four sample sizes. The skewness ratio is the skewness for distribution #1/distribution #2 for each gamma and lognormal case, respectively, at each R value. W = Welch, WMW = Wilcoxon-Mann-Whitney.

Symmetric Cases

Table 3 contains the frequency (%) at which each of the test procedures is used in the two symmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Includes the Imbalanced and Balanced Cases)

For the R = 1.00 case with equal sample sizes, the t test is known to be robust for the symmetric distributions. Results in Table 3 show that the TS procedure correctly selected the t test for approximately 69% of the simulations. The Welch-S test was incorrectly selected for approximately 22% of the simulations when using the TS procedure. The WMW test was incorrectly selected for only 7% of the simulations when using the TS procedure.

For the R = 1.00 case with unequal sample sizes, Table 3 shows that the TS procedure selected the t test for 70% of the simulations. The TS procedure incorrectly selected the Welch-S test for nearly 23% of the simulations. However, the WMW test was incorrectly selected for less

than 6% of the simulations when using the TS procedure.

Unequal Variances (Includes the Imbalanced and Balanced Cases)

For the $R = 0.50$ and 2.0 cases with equal sample sizes, Table 3 shows the TS procedure correctly selected the Welch-S test for approximately 81% of the simulations. The TS procedure incorrectly selected the *t* test in approximately 10% of the simulations and incorrectly concluded asymmetry in approximately 9% of the simulations.

For the $R = 0.50$ and 2.0 cases with unequal sample sizes, Table 3 shows the TS procedure correctly selected the Welch-S test for about 70%-73% of the simulations. The TS procedure incorrectly selected the *t* test for about 20%-23% of the simulations.

Table 3. Frequency (%) At Which Each Means Test Is Used In The TS Procedure For The Symmetric Distributions.

n_1,n_2	σ_1 , σ_2	*R*	*t* test	Welch-S	WMW	Welch-AS
20,20	$\sigma_1 = \sigma_2$	1.00	68.91	21.96	7.10	2.04
	$\sigma_1 \neq \sigma_2$	0.25	0.09	90.78	<0.01	9.13
		0.50	10.44	80.43	1.30	7.84
		2.00	10.33	80.54	1.28	7.86
		4.00	0.07	90.80	0.02	9.12
10,20	$\sigma_1 = \sigma_2$	1.00	70.30	22.69	5.64	1.38
	$\sigma_1 \neq \sigma_2$	0.25	0.58	92.41	0.05	7.00
		0.50	20.02	72.97	2.06	4.96
		2.00	22.76	70.23	1.92	5.10
		4.00	0.97	92.02	0.15	6.87

For the $R = 0.25$, and 4.0 symmetric cases, the Welch test is known to be robust. Table 3 shows the TS procedure correctly used the Welch-S test for about 90%-92% of the simulations regardless of the sample size configurations. The Welch-AS test was incorrectly used for about 7-9% of the simulations for each of these same cases.

In summary, for the combined symmetric cases, the TS procedure correctly selected the *t* test for approximately 70% of the $R = 1$ cases regardless of the sample size configuration. For the $R = 0.50$ and 2.0 cases, the TS procedure correctly selected the Welch-S test for approximately 81% of the simulations with equal sample sizes and about 70% - 73% of the simulations with unequal sample sizes. For the $R = 0.25$ and 4.0 cases, regardless of sample size configuration, the TS procedure correctly used the Welch-S test for about 90%-92% of the simulations. It is noted for the $R \neq 1$ cases, the TS procedure incorrectly concluded asymmetry for about 7%-9% of the simulations.

Asymmetric Cases

Table 4 contains the frequency (%) at which each of the test procedures is used in the two asymmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Includes the Imbalanced and Balanced Cases)

For the R = 1 case with equal sample sizes, the WMW test is known to be robust for the asymmetric distributions. Results in Table 4 shows the TS procedure correctly selected the WMW test for approximately 42% of the simulations. The TS procedure incorrectly selected the Welch-AS test in approximately 12% of the simulations with homogeneous variances. The t test was incorrectly selected by the TS procedure in approximately 33% of the simulations.

For the R = 1 cases with unequal sample sizes, Table 4 shows the TS procedure correctly selected the WMW test for approximately 31% of the simulations. As also seen for the balanced sample size cases, the TS procedure incorrectly selected the Welch-AS test in approximately 8% of these cases. In addition, the t test was incorrectly selected by the TS procedure in approximately 45% of the simulations.

Unequal Variances (Imbalanced and Balanced Cases)

For the equal sample size cases, Table 4 shows the TS procedure incorrectly selected the Welsh-S in approximately 50% of the R=0.50 cases and approximately 10% of the R=2.0 cases. Furthermore, the TS procedure incorrectly selected the WMW test in approximately 6% of the R = 0.50 cases and approximately 36% in the R = 2.0 cases. The Welch-AS test was correctly selected for approximately 35% and 47% of the R = 0.50 and 2.0 cases, respectively, when using the TS procedure.

For the R = 0.50 and 2.0 cases with imbalanced sample sizes, results in Table 4 shows the same trends as was seen for the equal sample size cases. The TS procedure incorrectly used the WMW test for approximately 10% of the R = 0.50 and approximately 28% in the R = 2.0 cases; and correctly selected the Welch-AS test for about 25-26% of the R = 0.50 and 2.0 cases.

Results in Table 4 shows for the balanced case that the TS procedure correctly selected the Welch-AS test for approximately 37% of the R = 0.25 cases. The WMW test was incorrectly used for about 2% of the R = 0.25 cases.

Results in Table 4 for the unequal sample size case show that the TS procedure correctly used the Welch-AS test for approximately 35% of the R = 0.25 cases, whereas the WMW test and the Welch-S test were each incorrectly selected for about 20% and 65%, respectively, of the R = 0.25 cases.

The TS procedure incorrectly used the WMW test for approximately 43% of the R = 4.0 cases and the Welch-AS test was correctly used for about 52% of the R = 4.0 equal sample size. For the R= 4.0 unequal sample size cases, the TS procedure incorrectly used the WMW test for approximately 32% of the simulations and the Welsh-AS test was correctly used for approximately 43% of the simulations.

In summary, for the R = 1 cases regardless of the sample size configuration, the TS procedure used the WMW test correctly for about 31%-42% of the simulations. For the R = 0.50 cases, the WMW test was incorrectly selected for about 6%-10% of the simulations when using the TS procedure. The TS procedure generally correctly used the Welch-AS test for about 35%-37% of the 0.25 cases. For the R= 2.0 cases, the TS procedure selected the Welsh-AS test correctly for about 25%-47% of the simulations and the WMW test incorrectly for about 28%-36% of the simulations. The TS procedure selected the Welch-AS test correctly for about 43%-52% of the simulations and the WMW test incorrectly each for about 32%-43% of the simulations for the R= 4.0 cases.

Summary of the TS Procedure Using an Alpha Level of 5% of the Triple's Test

For the cases where variance homogeneity and symmetry each are unknown to the practicing statistician, an overall test using the TS procedure yielded improved results with respect to robustness over using the t test or the Wilcoxon-Mann-Whitney test alone, except for the asymmetry unequal variance cases, where no method maintained the stated Type I error rate. The Welch test is recommended as a robust test for testing H_o: $\mu_1 = \mu_2$ for the symmetric cases examined. The TS procedure is also reasonably robust.

Table 4. Frequency (%) At Which Each Means Test Is Used In The TS Procedure For The Asymmetric Distributions.

n_1,n_2	σ_1 , σ_2	*R*	*t* test	Welch-S	WMW	Welch-AS
20,20	$\sigma_1 = \sigma_2$	1.00	32.98	12.43	42.22	12.38
	$\sigma_1 \neq \sigma_2$	0.25	0.02	63.31	0.02	36.66
		0.50	9.47	49.95	5.90	34.69
		2.00	6.88	9.81	36.21	47.11
		4.00	1.79	2.63	43.26	52.33
10,20	$\sigma_1 = \sigma_2$	1.00	45.02	16.59	30.50	7.90
	$\sigma_1 \neq \sigma_2$	0.25	0.43	64.58	0.20	34.79
		0.50	17.73	46.61	9.79	25.88
		2.00	21.73	24.83	28.32	25.13
		4.00	9.74	15.74	31.55	42.98

The performance of the TS procedure was also evaluated by the frequency at which the TS procedure selected the most appropriate test of means. For the symmetric equal variance cases, the TS procedure correctly selected the *t* test for approximately 70% of the simulated. For the symmetric cases with unequal variances (*R* = 0.25, 0.50, 2.0, and 4.0), the frequency at which the Welch test was correctly selected was about 70%-92% for the TS procedure. Asymmetry was incorrectly concluded for about 7%-9% of the simulated symmetric cases when using the TS procedure.

The TS procedure correctly concluded asymmetry for about 35%-96% of the simulated cases for the families of asymmetric distributions examined. For the asymmetric equal variance cases, the TS procedure correctly selected the Wilcoxon-Mann-Whitney test for about 31%-42% of the simulations. For the asymmetric cases with unequal variances, the TS procedure correctly concluded asymmetry and variance heterogeneity for about 25%-52% of the simulations.

Results showed that the TS procedure concluded symmetry too often (for 45%-62% of the asymmetric cases with equal variances).

Since the TS procedure examined in this simulation study concluded symmetry too often, it would be of interest to examine the performance of an TS procedure using the Triples test for testing of symmetry at a higher significance level such as $\alpha = 0.25$.

Further Investigation of the TS Procedure Using an Alpha Level of 25% for the Test of Symmetry

As the results above showed that the TS procedure was concluding symmetry too often, the simulations were repeated using the TS procedure with the alpha level set at 25% for the Triples test. To compare the TS procedure using the Triples test at alpha level 25% versus 5%, only the results of the frequency (%) at which each means test is used are displayed.

Tables 5 and 6 summarize the frequency (%) at which each of the test procedures is used for the two symmetric distributions cases combined, and the two asymmetric cases combined, respectively. The format for Tables 5 and 6 is the same as described above in section "Frequency (%) at Which Each Mean Test is Used."

Frequency (%) Each Means Test is Used For Symmetric Cases

Table 5 contains the frequency (%) at which each of the test procedures is used in the two symmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Imbalanced and Balanced Cases)

For the R = 1.00 case with equal sample sizes, the t test is known to be robust for the symmetric distributions. Results in Table 5 show that the TS procedure correctly selected the t test for approximately 46% of the simulations. The Welch-S test was incorrectly selected for approximately 14% of the simulations when using the TS procedure. The WMW test was incorrectly selected for only 32% of the simulations when using the TS procedure.

For the R = 1.00 case with unequal sample sizes, Table 5 shows that the TS procedure selected the t test for approximately 48% of the simulations. The TS procedure incorrectly selected the Welch-S test for approximately 15% of the simulations. However, the WMW test was incorrectly selected for about 30% of the simulations when using the TS procedure.

Unequal Variances (Imbalanced and Balanced Cases)

For the R = 0.50 and 2.0 cases with equal sample sizes, Table 5 shows the TS procedure correctly selected the Welch-S test for approximately 58% of the simulations. The TS procedure incorrectly selected the t test in approximately 2% of the simulations and incorrectly concluded asymmetry in approximately 40% of the simulations.

For the R = 0.50 and 2.0 cases with unequal sample sizes, Table 5 shows the TS procedure correctly selected the Welch-S test for about 54%-57% of the simulations. The TS procedure incorrectly selected the t test for about 6%-9% of the simulations.

For the R = 0.25, and 4.0 symmetric cases, the Welch test is known to be robust. Table 5 shows the TS procedure correctly used the Welch-S test for about 60%-63% of the simulations regardless of the sample size configurations. The Welch-AS test was incorrectly used for about 37%-40% of the simulations for each of these same cases.

Table 5. Frequency (%) At Which Each Means Test Is Used In The TS Procedure For The Symmetric Distributions.

n_1,n_2	σ_1 , σ_2	R	t test	Welch-S	WMW	Welch-AS
20,20	$\sigma_1 = \sigma_2$	1.00	46.38	13.77	32.42	7.44
	$\sigma_1 \neq \sigma_2$	0.25	0.00	60.15	0.01	39.85
		0.50	2.38	57.77	1.78	38.08
		2.00	2.46	57.69	1.67	38.19
		4.00	0.00	60.15	0.00	39.81
10,20	$\sigma_1 = \sigma_2$	1.00	47.89	15.14	29.97	7.01
	$\sigma_1 \neq \sigma_2$	0.25	0.02	63.00	0.01	36.98
		0.50	6.24	56.78	4.31	32.68
		2.00	9.25	53.77	6.22	30.76
		4.00	0.11	62.92	0.07	36.91

In summary, for the combined symmetric cases, the TS procedure correctly selected the *t* test for approximately 47% of the $R = 1$ cases regardless of the sample size configuration. For the $R = 0.50$ and 2.0 cases, the TS procedure correctly selected the Welch-S test for approximately 58% of the simulations with equal sample sizes and about 54%-57% of the simulations with unequal sample sizes. For the $R = 0.25$ and 4.0 cases, regardless of sample size configuration, the TS procedure correctly used the Welch-S test for about 60%-63% of the simulations. It is noted for the $R \neq 1$ cases, the TS procedure incorrectly concluded asymmetry for about 37%-40% of the simulations.

Frequency (%) Each Means Test is Used For Asymmetric Cases

Table 6 contains the frequency (%) at which each of the test procedures is used in the two asymmetric distributions combined for the balanced and imbalanced cases, respectively.

Equal Variances (Imbalanced and Balanced Cases)

For the $R = 1$ case with equal sample sizes, the WMW test is known to be robust for the asymmetric distributions. Results in Table 6 show the TS procedure correctly selected the WMW test for approximately 67% of the simulations. The TS procedure incorrectly selected the Welch-AS test in approximately 22% of the simulations with homogeneous variances. The *t* test was incorrectly selected by the TS procedure in approximately 8% of the simulations.

For the $R = 1$ cases with unequal sample sizes, Table 6 shows the TS procedure correctly selected the WMW test for approximately 60% of the simulations. As also seen for the balanced sample size cases, the TS procedure incorrectly selected the Welch-AS test in approximately 19% of these cases. In addition, the *t* test was incorrectly selected by the TS procedure in approximately 15% of the simulations.

Unequal Variances (Imbalanced and Balanced Cases)

For the equal sample size cases, Table 6 shows the TS procedure incorrectly selected the Welsh-S in approximately 25% of the R=0.25 cases. Furthermore, the TS procedure incorrectly selected the WMW test in approximately 12% of the $R = 0.50$ cases and approximately 43% in the $R = 2.0$ cases. The Welch-AS test was correctly selected for approximately 67% and 55% of the $R = 0.50$ and 2.0 cases, respectively, when using the TS procedure.

Table 6. Frequency (%) For Means Test In The TS Procedure For The Asymmetric Distributions.

n_1,n_2	σ_1 , σ_2	*R*	*t* test	Welch-S	WMW	Welch-AS
20,20	$\sigma_1 = \sigma_2$	1.00	8.05	3.48	66.95	21.53
	$\sigma_1 \neq \sigma_2$	0.25	0.01	24.85	0.03	75.12
		0.50	3.43	17.52	11.85	67.17
		2.00	1.16	1.49	42.72	54.65
		4.00	0.16	0.22	45.08	54.55
10,20	$\sigma_1 = \sigma_2$	1.00	14.78	6.34	60.04	18.85
	$\sigma_1 \neq \sigma_2$	0.25	0.22	27.49	0.40	71.90
		0.50	7.15	18.80	20.83	53.23
		2.00	5.36	6.27	44.27	44.11
		4.00	1.88	3.05	39.18	55.90

For the R = 0.50 and 2.0 cases with imbalanced sample sizes, results in Table 6 shows the same trends as was seen for the equal sample size cases. The TS procedure incorrectly used the WMW test for approximately 21% of the R = 0.50 and approximately 44% in the R = 2.0 cases; and correctly selected the Welch-AS test for about 44%-53% of the R = 0.50 and 2.0 cases.

Results in Table 6 shows for the balanced case that the TS procedure correctly selected the Welch-AS test for approximately 75% of the R = 0.25 cases. The Welch-S test was incorrectly used for about 25% of the R = 0.25 cases.

Results in Table 6 for the unequal sample size case show that the TS procedure correctly used the Welch-AS test for approximately 72% of the R = 0.25 cases, whereas the Welch-S test was incorrectly selected for about 27% of the R = 0.25 cases.

The TS procedure incorrectly used the WMW test for approximately 45% of the R = 4.0 cases and the Welch-AS test was correctly used for about 55% of the R = 4.0 equal sample size. For the R= 4.0 unequal sample size cases, the TS procedure incorrectly used the WMW test for approximately 39% of the simulations and the Welsh-AS test was correctly used for approximately 56% of the simulations.

In summary, for the R = 1 cases regardless of the sample size configuration, the TS procedure used the WMW test correctly for about 60%-67% of the simulations. For the R = 0.50 cases, the WMW test was incorrectly selected for about 12%-21% of the simulations when using the TS procedure. The TS procedure generally correctly used the Welch-AS test for about 72%-75% of the 0.25 cases. For the R= 2.0 cases, the TS procedure selected the Welsh-AS test correctly for about 44%-55% of the simulations and the WMW test incorrectly for about 43%-44% of the simulations. The TS procedure selected the Welch-AS test correctly for about 55%-56% of the simulations and the WMW test incorrectly each for about 39%-45% of the simulations for the R= 4.0 cases.

Conclusion

For the TS procedure using the Triples test with an alpha level of 5%, results showed that the TS procedure concluded symmetry too often (for 45%-62% of the asymmetric cases with equal variances).

For the TS procedure using the Triples test at an alpha level of 25%, results showed that the TS procedure concluded asymmetry for the symmetric distributions in 37%-40% of the R≠ 1 cases.

Recommendations for alternative approaches in the future, would be to examine the performance of an TS procedure which concludes asymmetry at an alpha level between 5% and 25% (i.e., 15%) or concludes asymmetry only if both samples were judged to be nonsymmetric at α = 0.25. In addition, there was a trend, especially in the asymmetric distributions, of concluding variance homogeneity too often for the R ≠ 1 cases. Therefore, it would be recommended to increase alpha level for testing of variance homogeneity to a higher alpha level beyond α = 0.25.

References

Brown, M. B. & Forsythe, A. B. (1974, June). Robust tests for the equality of variances. *Journal of the American Statistical Association*, *69* (346), 364-367.

Brown, M. B. & Forsythe, A.B. (1974b). The small behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129-132.

Cabilio, P. & Masaro, J. (1996). A simple test of symmetry about an unknown median. *The Canadian Journal of Statistics*, *24*(3), 349-361.

Conover, W. J., Johnson, M. E., & Johnson, M.M. (1981). A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data. *Technometrics*, *23* (4), 351-361.

Gans, D. J. (1981). Use of a preliminary test in comparing two sample means. *Communication in Statistics B, Simulation & Computation*, *B10* (2), 163-174.

Gupta, M. K. (1967). An asymptotically nonparametric test of symmetry. *Annals of Mathematical Statistics*, *38*, 849-866.

IMSL. (1989, January). *Math/Library User's Manual (Version 1.1)*. Houston, Texas: Author.

IMSL. (1989, December). *Math/Library User's Manual (Version 1.1)*. Houston, Texas: Author.

Loh, W. (1987, May 21). Some modifications of Levene's test of variance homogeneity. *Journal of Statistical Computation & Simulation*, *28*, 213-226.

Murphy, B. P. (1976). Comparison of some two sample means tests by simulation. *Communication in Statistics B, Simulation and Computation*, *B5* (1), 23-32.

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, *74*, 877-880.

Olejnik, S. (1987). Conditional ANOVA for mean differences when population variances are unknown. *Journal of Experimental Education*, *55*, 141-148.

Perry, K. T. & Stoline, M. R. (2002). A comparison of the D'Agostino S_U test to the Triples test for testing of symmetry versus asymmetry as a preliminary test to testing the equality of means. *Journal of Modern Applied Statistical Methods*, *1* (2), 316-325.

Randles, R. H., Fligner, M. A., Policello II, G.E., & Wolfe, D.A. (1980, March). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, *75* (369), 168-172.

Snedecor, G. W. & Cochran, W. G. (1967). *Statistical methods*. Ames, Iowa: The Iowa State University Press.

Zimmerman, D. W. & Williams, R. H. (1989). Power comparisons of the student t-test and two approximations when variances and sample sizes are unequal. *Journal of Indian Society Agricultural Statistics*, *41* (2), 206-217.