2016 ■ Vol. 12 ■ no. 1

DOI: 10.20982/tqmp.12.1.p030

Why Welch's test is Type I error robust.

Ben Derrick[a], Deirdre Toher[a] & Paul White[a,✉]

[a]University of the West of England, Bristol, England

Abstract ■ The comparison of two means is one of the most commonly applied statistical procedures in psychology. The independent samples t-test corrected for unequal variances is commonly known as Welch's test, and is widely considered to be a robust alternative to the independent samples t-test. The properties of Welch's test that make it Type I error robust are examined. The degrees of freedom used in Welch's test are a random variable, the distributions of which are examined using simulation. It is shown how the distribution for the degrees of freedom is dependent on the sample sizes and the variances of the samples. The impact of sample variances on the degrees of freedom, the resultant critical value and the test statistic is considered, and hence gives an insight into why Welch's test is Type I error robust under normality.

Keywords ■ Independent samples t-test; Welch's test; Welch's approximation; Behrens-Fisher problem; Equality of means.

✉ Paul.White@uwe.ac.uk

Introduction

One of the most commonly applied hypothesis test procedures in applied research is the comparison of two population means (Wilcox, 1992). For theoretical development purposes, assume two normally distributed populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ are to be compared based upon n_1 and n_2 mutually independent observations. Let $\overline{X}_i$ and S_i^2 denote random variables for sample means and variances respectively ($i = 1, 2$).[1] If the population variances, σ_1^2 and σ_2^2, are assumed to be equal, then an appropriate test statistic is the independent samples t-test, based on (1) and (2).

$$T_1 = \frac{\overline{X}_1 - \overline{X}_2}{\text{StandardError}(\overline{X}_1 - \overline{X}_2)} \tag{1}$$

In the independent samples t-test, the standard error of $(\overline{X}_1 - \overline{X}_2)$, say SE_1, is given by:

$$SE_1 = S_p\sqrt{\frac{2}{\tilde{n}}} \tag{2}$$

where $S_p = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)}}$ and $\tilde{n}$ is the harmonic mean of n_1 and n_2. T_1 is referenced against the t-distribution with degrees of freedom equal to $v_1 = n_1 + n_2 - 2$.

It is known that, when the assumptions of the independent samples t-test are met, the independent samples t-test is an exact test and is the most uniformly powerful test (Sawilowsky & Blair, 1992). The independent samples t-test is an approximate test when population variances are unequal. If sample sizes are unequal and variances are unequal, the probability of rejecting the null hypothesis when it is true deviates from the nominal Type I error rate. This is particularly problematic when the smaller sample size is associated with the larger variance (Zimmerman & Zumbo, 2009; Coombs, Algina, & Oltman, 1996). This gives rise to the dilemma of how to compare means in the presence of unequal variances. This question, applied to two independent random samples from normal populations, is known as the Behrens-Fisher problem. Behrens (1929) and Fisher (1935, 1941) suggested a solution for the problem. It is proposed that the t-test when equal variances cannot be assumed is defined as per (3) and (4).

$$T_2 = \frac{\overline{\text{X}}_1 - \overline{\text{X}}_2}{\text{StandardError}(\overline{\text{X}}_1 - \overline{\text{X}}_2)} \tag{3}$$

In the unequal variances case, the standard error of $(\overline{X}_1 - \overline{X}_2)$, say SE_2 is estimated by:

$$SE_2 = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \tag{4}$$

The formula developed for the degrees of freedom is complex, but it is proposed that an approximation for the degrees of freedom could be given by (5). This is given in most textbooks (e. g., Alfassi, Boger, & Ronen, 2005; Miles & Banyard, 2007).

$$v_2 = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / (n_1 - 1) + \left(\frac{S_2^2}{n_2}\right)^2 / (n_2 - 1)} \tag{5}$$

[1]As standard notation, random variables are shown in upper case, and derived sample values are shown are in lower case.

A numerically equivalent expression for the approximation v_2 is given in (6). This is shown in some textbooks (e. g., Ott & Longnecker, 2001).

$$v_2 = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)\, c^2 + (n_1 - 1)\,(1 - c)^2} \tag{6}$$

where

$$c = \frac{S_1^2 / n_1}{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

The approximation v_2 dates back to a series of papers by Welch (1938, 1947, 1951), independent work by Satterthwaite (1946), works by Fairfield-Smith (1936), and Aspin (1948, 1949). The independent samples t-test corrected for unequal variances is sometimes referred to as the Satterthwaite-Smith-Welch test, the Welch-Aspin-Satterthwaite test, or other interchangeable variations. This may be referred to generically as the unequal variances t-test, or as the separate variances t-test. Usually the unequal variances t-test with the degrees of freedom approximated as above is simply known as Welch's test.

Originally, an alternative approximation for the degrees of freedom given by Welch, is given in (7):

$$v_3 = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\left(\frac{S_1^2}{n_1}\right)^2 / (n_1 + 1) + \left(\frac{S_2^2}{n_2}\right)^2 / (n_2 + 1)} - 2 \tag{7}$$

The approximation is given in some textbooks (e. g. Frank & Althoen, 1994), rounded down to the nearest integer. However, v_3 is not generally used, and is not numerically equivalent to v_2.

Textbooks frequently recommend the calculation of v_2, rounded down to the nearest integer (e. g. Frank & Althoen, 1994; Ott & Longnecker, 2001). Rounding down tends to produce a conservative test. More generally, some textbooks recommend rounding to the nearest integer (e. g. Alfassi et al., 2005). The rounding requirements appear in textbooks for the purposes of manual calculations. There is a need to use integer degrees of freedom when using statistical tables for critical values. However, the calculation of Welch's test is easy in statistical software such as R and SPSS (Rasch, Kubinger, & Yanagida, 2011). These statistical software would ordinarily conduct the test with non-integer degrees of freedom.

Welch's test better approximates nominal significance levels, and has greater power than the Behrens-Fisher solution (Lee & Gurland, 1975; Best & Rayner, 1987). Fay and Proschan (2010, p. 14) confirm that Welch's solution "is approximately valid for the Behrens-Fisher perspective".

When sample sizes are equal and variances are equal, both the independent samples t-test and Welch's test perform similarly (Zimmerman & Zumbo, 1993; Moser, Stevens, & Watts, 1989). For unequal sample sizes and unequal variances, Welch's test has superior Type I error robustness (Fagerland & Sandvik, 2009). Ruxton (2006) advocates the routine use of Welch's test.

Grimes and Federer (1982, p.10) state that, "In the case of comparing two sample means, the consensus in the literature seems to be the approval of Welch's approximate solution". Thus the most commonly used solution to the Behrens-Fisher problem, is Welch's test with the degrees of freedom calculated by approximation. In a practical environment, Welch's approximation can be used with little loss of accuracy (Wang, 1971; Scheffe, 1970).

It can be seen from (5) that Welch's degrees of freedom, v_2, is a random variable and therefore has its own sampling distribution. Consequently the critical value used in hypothesis testing is also a random variable. In addition, it can be seen from (4) that the sample variances affect both the value of T_2 and the value of v_2.

In this paper; worked examples of the independent samples t-test and Welch's test are provided. The distributions of the degrees of freedom for Welch's test are explored, and the two methods of estimating the standard error of are considered. Simulation is used to identify how the estimated standard error facilitates the Type I error robustness of Welch's test, and provides insight into why the Welch test works in a practical environment.

Worked examples

As part of an investigation into sensitivity when exposed to evidence of "White Privilege", Phillips and Lowery (2015) randomly allocated U.S. participants who self-identified as White/European-American into two groups. The participants completed a survey about equality and their childhood memories ("Experiment 1a"). Prior to completing the survey, Group 1 (n_1 = 54) were given a paragraph to read about "White Privilege", whereas Group 2 (n_2 = 40) were not. Questions on the survey measured participants perceived "life hardship" on a Likert type scale, between 1 = "strongly disagree" and 7 = "strongly agree". The authors performed the independent samples t-test using each participant's mean score.[2] This implies that equality of variance between groups is assumed; this is a seemingly reasonable assumption due to the random assignment of participants. For demonstration purposes, both the independent samples t-test and Welch's test are provided in the present paper. For "Experiment 1a", the published data are as follows; the average participant score for Group 1 is 4.41, (standard deviation of 1.20). The average participant

[2] The published results differ slightly from the calculations given here, due to the use of the published (rounded) sample data in the present paper.

score for Group 2 is 3.82 (standard deviation of 1.20). Thus, $\bar{x}_1 = 4.410$, $s_1^2 = 1.440$, $\bar{x}_2 = 3.820$ and $s_2^2 = 1.440$. Calculations for the independent samples t-test give: $s_p = 1.200$, $se_1 = 0.250$, $t_1 = 2.357$, $v_1 = 92.000$, the p-value using the independent samples t-test is 0.021. Calculations for Welch's test give: $se_2 = 0.250$, $t_2 = 2.357$, $v_2 = 84.186$, the p-value using Welch's test is 0.021. It can be seen that because the two sample variances are equal, $t_1 = t_2$. The degrees of freedom applicable for each test are different, but the impact of this on the critical values of the tests is small. Thus the p-values for both tests are the same to three decimal places. The statistical conclusion made at the 5% significance level, is that the sample mean for Group 1 is significantly greater than the sample mean for Group 2. The authors conclude that perceived "life hardship" is greater when participants are subjected to evidence of "White Privilege".

Phillips and Lowery (2015) replicated this experiment with $n_1 = 49$ and $n_2 = 42$ participants ("Experiment 1b"). The published data shows that the average participant score for Group 1 is 4.53, (standard deviation of 1.52). The average participant score for Group 2 is 3.96, (standard deviation of 1.28). Thus, $\bar{x}_1 = 4.530$, $s_1^2 = 2.310$, $\bar{x}_2 = 3.960$ and $s_2^2 = 1.638$. Calculations for the independent samples t-test give: $s_p = 1.415$, $se_1 = 0.297$, $t_1 = 1.916$, $v_1 = 89.000$, the p-value using the independent samples t-test is 0.059. Calculations for Welch's test give: $se_2 = 0.294$, $t_2 = 1.942$, $v_2 = 88.978$, the p-value using Welch's test is 0.055. In this experiment, the p-values for the two tests are different due to the unequal sample sizes and unequal variances of the two samples. With reference to Experiment 1b, the authors state that participants in Group 1 claim more "life hardship" than participants in Group 2. However, for either test, at the 5% significance level, Experiment 1b alone represents insufficient statistical evidence that there is a difference between Group 1 and Group 2.

Methodology

Simulation is used to investigate Welch's test for Type I error robustness, and the distributional properties of v_2. For both the independent samples t-test and Welch's test, two sided tests are performed with nominal Type I error rate of $\alpha = 0.05$. The aim is to demonstrate deviations from Type I error robustness for the independent samples t-test for unequal variances. The standard error of the independent samples t-test and Welch's test are explored to assess the impact of the standard error on the result of the tests. To achieve these goals, simulations under H_0 for two normally distributed samples are performed as per the layout in Table 1; with n_1 at two levels, n_2 at two levels and σ_2 at two levels. Parameters are selected to cover both "large" and "small" samples and equal and unequal variances. The sample sizes represent extreme scenarios in order to assist in the illustration of the effects.

For each scenario in the simulation design, 10,000 iterations are performed under the condition where H_0 is true.

Results

Welch's degrees of freedom.

The investigation of the distribution of v_2, gives insight into when the degrees of freedom used in Welch's test differ from the degrees of freedom used in the independent samples t-test.

Figure 1 shows the distribution of the degrees of freedom for each of the 8 scenarios simulated (10,000 observations per scenario).

Inspection of Figure 1 shows the greatest discrepancy between v_1 and v_2 to occur when $n_1 \neq n_2$. The simulations demonstrate that $[min\{n_1, n_2\} - 1] \leq v_2 \leq v_1$. This can be proven mathematically using (6). By differentiation, the maximum value of v_2 is found when $s_1^2/s_2^2 = \{(n_1 - 1)n_1\}/\{(n_2 - 1)n_2\}$. The minimum value of v_2 is fixed by the sample with the larger variance. If Sample 1 has the larger variance, then the lower bound is $n_1 - 1$. If Sample 2 has the larger variance, then the lower bound is $n_2 - 1$. Hence, $min\{n_1, n_2\} - 1$ is a very conservative approximation to the degrees of freedom when the smaller sample size is associated with the larger variance. To illustrate these points, see Figure 2 with a fixed variance for Sample 1.

From Figure 2 it can be seen that as s_2^2/s_1^2 tends to zero, the degrees of freedom tends to $n_1 - 1$. As s_2^2/s_1^2 becomes increasingly large, the degrees of freedom asymptotically tends to $n_2 - 1$. The maximum value occurs when $s_1^2/s_2^2 = \{(n_1 - 1)n_1\}/\{(n_2 - 1)n_2\}$. The examples have a total sample size of 30, thus the maximum value of v_2 is 28.

Type I error robustness for the independent samples t-test and Welch's test.

In this section, p-values calculated from performing both the independent samples t-test and Welch's test are considered, as per the simulation design in Table 1. If H_0 is true and if underlying assumptions hold, then the p-values from a valid test procedure are expected to be uniformly distributed (Bland, 2013). Deviations from uniformity give evidence that the test is not Type I error robust. If p-values are consistently less than expected under a uniform distribution, the test gives too many false positives, and is said to be "liberal". If p-values are consistently greater than expected under a uniform distribution, the test is "conservative".

There is negligible difference between the p-values when performing the independent samples t-test or Welch's test under equal variances, regardless of sample size. In this case, p-values are approximately uniformly dis-

Table 1 ■ Summary of the simulation design.

Test statistics	T_1, T_2
Degrees of freedom	v_1, v_2
Sample sizes (n_1, n_2)	(5,5), (5,100), (100,5), (100,100)
Standard deviations (σ_1, σ_2)	(1,1), (1,2)
Programming language	R version 3.1.2 (R Development Core Team, 2013)

tributed for both tests (results not shown).

When variances are unequal, Welch's test is not a linear function of the independent samples t-test. Figure 3 is a P-P plot (percentile-percentile plot), for p-values for both the independent samples t-test (T_1) and Welch's test (T_2), with unequal variances. This shows ordered expected p-values from a uniform distribution plotted against ordered observed p-values. Given that for a valid test procedure, observed p-values should be approximately uniformly distributed on (0, 1) then an approximate diagonal would demonstrate Type I error robustness.

Both panels of Figure 3 show that when sample sizes are unequal and variances are unequal, the independent samples t-test is not Type I error robust. When the smaller sample size is associated with the larger variance (left panel, Figure 3), the observed p-values under the independent samples t-test are smaller than expected, and the test is liberal. Conversely, when the larger sample size is associated with the larger variance (right panel, Figure 3), the p-values are larger than expected and the independent samples t-test is conservative, (i.e. the expected Type I error rate is less than the pre-chosen nominal level of significance, α).

The p-values for Welch's test are also given in Figure 3. The simulated p-values for Welch's test, are approximately uniformly distributed. This results in the approximate line of equality observed. Welch's test therefore "corrects" for the fact that the independent samples t-test gives p-values that are not Type I error robust.

To demonstrate the impact of the degrees of freedom, for insight only, the independent samples t-test T_1 but with v_2 degrees of freedom is considered. Likewise, for insight only, Welch's test using statistic T_2 but with v_1 degrees of freedom is considered. These are compared against the standard approaches for the independent samples t-test and Welch's test. Table 2 summarises the Type I error rates observed ($\alpha = .05$, two-sided) for each combination. Bradley's (1978) liberal robustness criteria states that the Type I error rate when the nominal α is .05 should be in the range {0.025, 0.075}.

Table 2 shows that Welch's test (test statistic and degrees of freedom) is Type I error robust across all scenarios simulated. For unequal sample sizes and unequal variances, T_1 used in conjunction with v_1 or v_2, and T_2 used in conjunction with v_1, do not meet liberal robustness criteria. Welch's degrees of freedom therefore represent an important property for controlling Type I error rates. However, clearly the calculation of the test statistic, which takes into account the two separate sample variances, is also important.

Impact of the standard error on the properties of Welch's test.

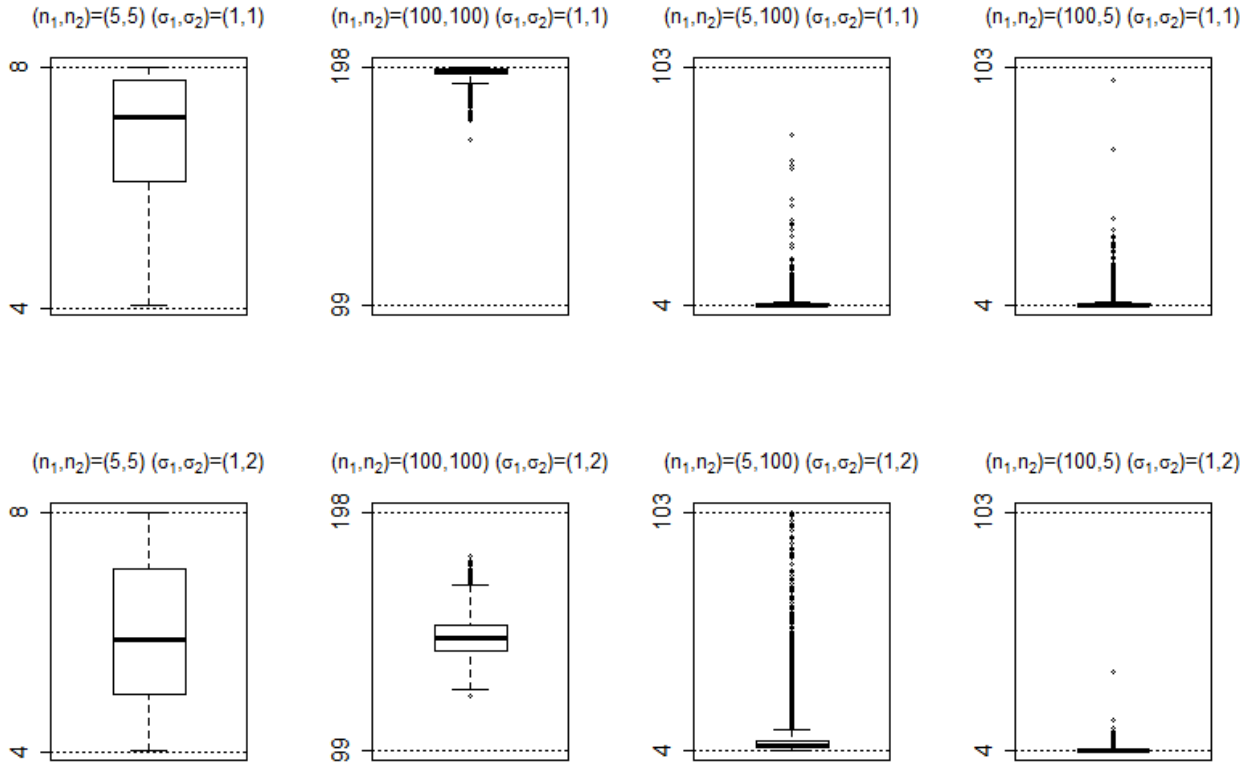
In this section, the impact of the standard error of the test statistics for the independent samples t-test and Welch's test is considered. The corrective properties of Welch's test are, in part, due to the impact of the sample variances on the degrees of freedom, which in turn affects the critical value used in the test. However, Type I error robustness could also be due to the impact of the estimated standard error on the magnitude of the test statistic. Figure 4 and Figure 5 demonstrate how the standard error, SE_1 and SE_2, relate to the critical value and to the absolute values of the test statistic for the independent samples t-test, T_1, and Welch's test, T_2, respectively.

Both panels of Figure 4 suggest that, when performing the independent samples t-test, the estimated standard error, SE_1, has no apparent relationship with the value of the test statistic, T_1. When the smaller sample size is associated with the larger population variance (left panel, Figure 4), the absolute value of the test statistic has a larger mean and a larger variability. When the larger sample size is associated with the larger population variance (right panel, Figure 4), the absolute value of the test statistic has a smaller mean and a smaller variability. This has the result that more false positives are observed when the smaller sample size is associated with the larger variance.

Both panels of Figure 5 demonstrate the impact of the degrees of freedom on the critical value. In the simulated scenario; the theoretical minimum degrees of freedom is $min(n_1, n_2) = 4$, accordingly the upper bound of the critical value is 2.776; the theoretical maximum degrees of freedom is $v_1 = 98$, accordingly the lower bound of the critical value is 1.984.

It can be seen from both panels of Figure 5 that as Welch's estimate of standard error, SE_2, increases, the absolute value of T_2 decreases. As the estimated standard error becomes large, the impact is far greater on the absolute value of T_2 relative to the critical value. This combination results in fewer false positives being observed as the esti-

Figure 1 ■ Distribution of v_2 for each scenario. The references lines represent the theoretical maximum and minimum values that v_2 can take. The upper reference line is equivalent to v_1.



mated standard error increases.

Discussion

For additional clarity of the above findings, Table 3 summarises theoretical values for each of the combinations in the simulation design. For illustration purposes differences in means are fixed at 1.000, s_1 and s_2 are fixed as σ_1 and σ_2 respectively.

From Table 3, it can be seen that when sample sizes are equal or variances are equal, the test statistics for the independent samples t-test and Welch's test are equivalent. Therefore, the difference in p-values are a direct result of the degrees of freedom used to calculate the critical value.

When variances are not equal, Welch's estimated standard error impacts the critical value, but this effect is smaller than the effect on the value on the test statistic. When the smaller sample size is associated with the larger variance, the effect on the value of the test statistic is exacerbated.

Conclusion

The literature favours Welch's test for a comparison of two means. This paper adds further support to the findings in the literature with respect to the Type I error robustness of Welch's test. The degrees of freedom of Welch's test are a random variable based on the sample size and variance of each sample. The degrees of freedom used in Welch's test are always less than or equal to the degrees of freedom used in the independent samples t-test. The degrees of freedom used in the independent samples t-test and Welch's test are equivalent when $s_1^2/s_2^2 = \{(n_1-1)n_1\}/\{(n_2-1)n_2\}$. The minimum value of Welch's degrees of freedom is $min\{n_1, n_2\} - 1$, this minimum is determined by the sample with the larger variance. Therefore Welch's approximate degrees of freedom are more conservative than the degrees of freedom used in the independent samples t-test, particularly when

Figure 2 ■ Value of v_2 with varying s_2^2, and fixed value $s_1^2 = 1$. Values to the left of $s_2^2 = 1$ have the larger variance associated with Sample 1. Values to the right of $s_2^2 = 1$ have the larger variance associated with Sample 2.

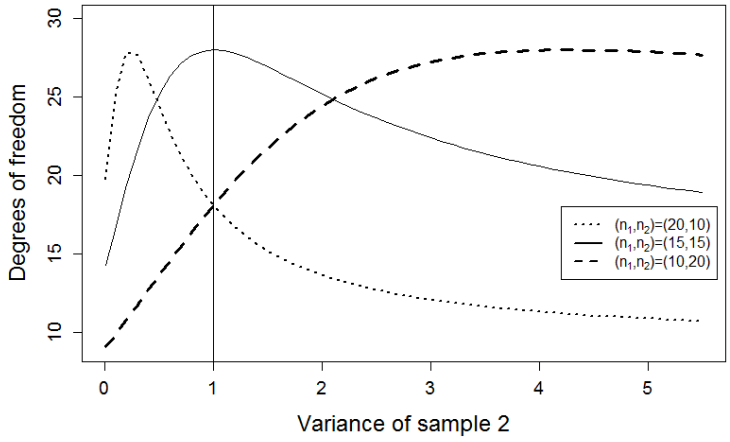


Table 2 ■ Type I error rates for each combination of test statistic with degrees of freedom. Type I error robust combinations are highlighted in bold.

(n_1, n_2)	(σ_1, σ_2)	T_1 with v_1	T_1 with v_2	T_2 with v_1	T_2 with v_2
5,5	1,1	**0.050**	**0.045**	**0.050**	**0.045**
	1,2	**0.056**	**0.047**	**0.056**	**0.047**
5,100	1,1	**0.053**	0.012	0.110	**0.056**
	1,2	0.001	0.000	0.093	**0.060**
100,5	1,1	**0.050**	0.011	0.108	**0.055**
	1,2	0.295	0.153	0.118	**0.052**
100,100	1,1	**0.049**	**0.049**	**0.049**	**0.049**
	1,2	**0.050**	**0.049**	**0.050**	**0.049**

the smaller sample size is associated with the larger variance. When performing Welch's test, the estimated standard error impacts the magnitude of the test statistic. Under the null hypothesis, it is the estimated standard error when performing Welch's test, which is the most influential factor on the result of the test. For Welch's test, the probability of making a Type I error decreases as the standard error increases. This paper gives insight in to why Welch's test is Type I error robust for normally distributed data, in scenarios when the independent samples t-test is not. Additionally, it is shown that in situations when the independent samples t-test is Type I error robust, Welch's test is also. In a practical environment for the comparisons of two means from assumed normal populations, a general rule to preserve Type I error robustness is, if in doubt use Welch's test.

References

Alfassi, Z. B., Boger, Z., & Ronen, Y. (2005). *Statistical treatment of analytical data.* CRC Press.

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika. 2*, 88–96. doi:10.2307/2332631

Aspin, A. A. (1949). Tables for use in comparisons whose accuracy involves two variances, separately estimated. *Biometrika. 36*, 290–296. doi:10.2307/2332668

Behrens, W. U. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. *Landwirtschaftliche Jahrbucher. 68*, 807–837.

Best, D. J. & Rayner, J. C. W. (1987). Welch's approximate solution for the behrens-fisher problem. *Technometrics. 29*(2), 205–210. doi:10.2307/1269775

Figure 3 ■ P-values for the independent samples t-test, T_1, and Welch's test, T_2. The left panel shows the smaller sample size associated with the larger variance. The right panel shows the larger sample size associated with the larger variance.

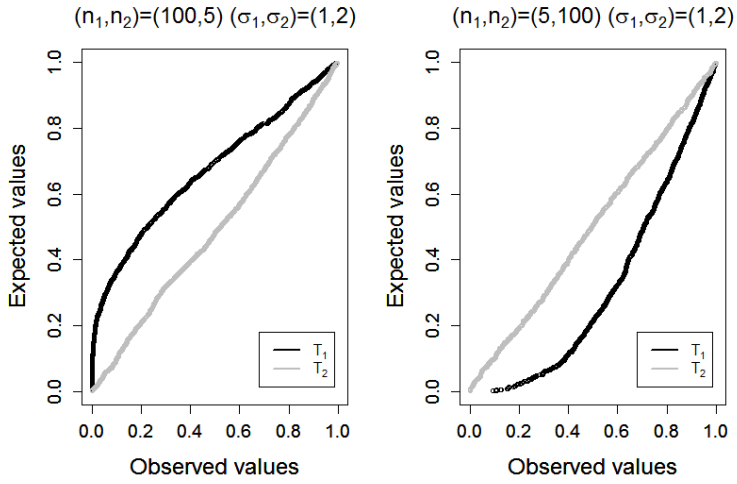


Table 3 ■ Components of the tests for each scenario in the simulation design.

		Independent samples t-test			Welch's test		
(n_1, n_2)	(s_1, s_2)	test statistic	critical value	p-value	test statistic	critical value	p-value
5,5	1,1	1.581	2.306	0.153	1.581	2.306	0.153
	1,2	1.000	2.306	0.347	1.000	2.571	0.363
5,100	1,1	2.182	1.983	0.031	2.182	2.776	0.095
	1,2	1.107	1.983	0.271	2.041	2.571	0.097
100,5	1,1	2.182	1.983	0.031	2.182	2.776	0.095
	1,2	2.065	1.983	0.041	1.111	2.776	0.329
100,100	1,1	7.071	1.972	<0.001	7.071	1.972	< 0.001
	1,2	4.472	1.972	< 0.001	4.472	1.976	< 0.001

Bland, M. (2013). Do baseline p-values follow a uniform distribution in randomised trials? *PloS one*, *8*(10), e76010. doi:10.1371/journal.pone.0076010

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*. *31*(2), 144–152. doi:10.1111/j.2044-8317.1978.tb00581.x

Coombs, W. T., Algina, J., & Oltman, D. (1996). Univariate and multivariate omnibus hypothesis tests selected to control type i error rates when population variances are not necessarily equal. *Review of Educational Research*. *66*(2), 137–79. doi:10 . 3102 / 00346543066002137

Fagerland, M. W. & Sandvik, L. (2009). Performance of five two-sample location tests for skewed distributions with unequal variances. *Contemporary Clinical Trials*. *30*(5), 490–496. doi:10.1016/j.cct.2009.06.007

Fairfield-Smith, H. (1936). The problem of comparing the results of two experiments with unequal errors. *Journal of the Council for Scientific and Industrial Research*, *9*, 211–212.

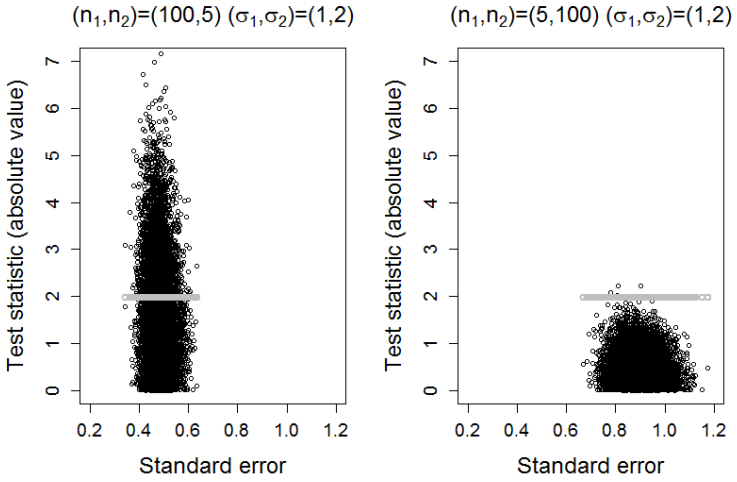
Fay, M. P. & Proschan, M. A. (2010). Wilcoxon-mann-whitney or t-test? *On assumptions for hypothesis tests and multiple interpretations of decision rules. Statistics Surveys*. *4*(1), 1. doi:10.1214/09-SS051

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics*. 391–398. doi:10.1111/j.1469-1809.1935.tb02120.x

Fisher, R. A. (1941). The asymptotic approach to behrens' integral, with further tables for the d test of significance. *Annals of Eugenics*. *11*, 141–172. doi:10.1111/j.1469-1809.1941.tb02281.x

Frank, H. & Althoen, S. C. (1994). *Statistics: concepts and applications*. Cambridge University Press.

Figure 4 ■ Simulated values of the standard error, SE_1, against the absolute value of the test statistic, T_1, for the independent samples t-test. The critical value, a constant at 1.984, has been superimposed. The left panel shows the smaller sample size associated with the larger variance. The right panel shows the larger sample size associated with the larger variance.



Grimes, B. A. & Federer, W. T. (1982). *Comparison of means from populations with unequal variances*. Biometrics Unit Technical Reports: Number BU-762-M.

Lee, A. F. S. & Gurland, J. (1975). Size and power of tests for equality of means of two normal populations with unequal variances. *Journal of the American Statistical Association. 70*(352), 933–941. doi:10.1080/01621459.1975.10480326

Miles, J. & Banyard, P. (2007). *Understanding and using statistics in psychology: a practical introduction*. Sage.

Moser, B. K., Stevens, G. R., & Watts, C. L. (1989). The two-sample t test versus satterthwaite's approximate f test. *Communications in Statistics-Theory and Methods. 18*(11), 3963–3975. doi:10.1080/03610928908830135

Ott, R. L. & Longnecker, M. (2001). *An introduction to statistical methods and data analysis*. Pacific Grove, CA: Duxbury.

Phillips, L. T. & Lowery, B. S. (2015). The hard-knock life? whites claim hardships in response to racial inequity. *Journal of Experimental Social Psychology, 61*, 12–18. doi:10.1016/j.jesp.2015.06.008

R Development Core Team. (2013). *R: a language and environment for statistical computing*. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org

Rasch, D., Kubinger, K., & Yanagida, T. (2011). *Statistics in psychology using r and spss*. John Wiley and Sons.

Ruxton, G. (2006). The unequal variance t-test is an underused alternative to student's t-test and the mann-whitney u test. *Behavioral Ecology. 17*(4), 688–690. doi:10.1093/beheco/ark016

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin. 2*, 110–114. doi:10.2307/3002019

Sawilowsky, S. S. & Blair, R. C. (1992). A more realistic look at the robustness and type ii error properties of the t-test to departures from population normality. *American Psychological Association. 111*(2), 352–360. doi:10.1037/0033-2909.111.2.352

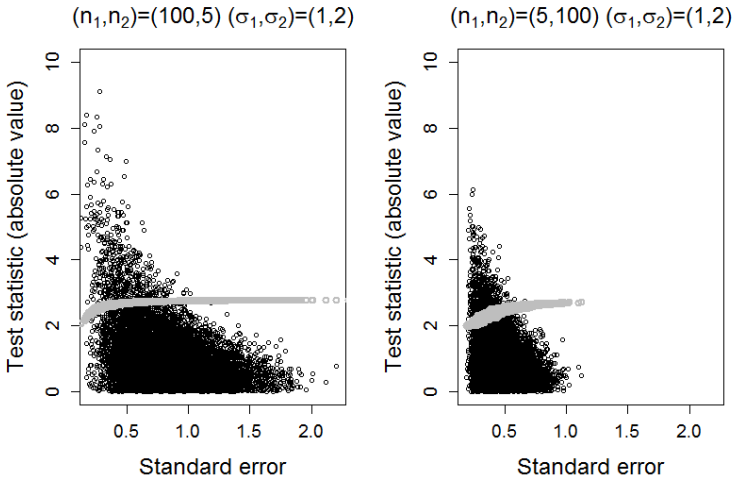
Scheffe, H. (1970). Practical solutions of the behrens-fisher problem. *Journal of the American Statistical Association. 65*, 1501–1508. doi:10.1080/01621459.1970.10481179

Wang, Y. Y. (1971). Probabilities or the type I errors of the welch tests for the behrens-fisher problem. *Journal of the American Statistical Association. 66*, 605–608. doi:10.1080/01621459.1971.10482315

Welch, B. L. (1938). The significance or the difference between two means when the population variances are unequal. *Biometrika. 29*, 350–362. doi:10.2307/2332010

Welch, B. L. (1947). The generalization of 'student's' problem when several different population variances are involved. *Biometrika. 34*, 28–35. doi:10.2307/2332510

Figure 5 ■ Properties of Welch's test. The critical values have been superimposed. The left panel shows the smaller sample size associated with the larger variance. The right panel shows the larger sample size associated with the larger variance.



Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika. 38*, 330–336. doi:10.2307/2332579

Wilcox, R. R. (1992). Why can methods for comparing means have relatively low power, and what can you do to correct the problem? *Current Directions in Psychological Science. 1*(3), 101–105.

Zimmerman, D. W. & Zumbo, B. D. (1993). Rank transformations and the power of the student t-test and welch t-test for non-normal populations. *Canadian Journal of Experimental Psychology. 47*(3), 523–39. doi:10.1037/h0078850

Zimmerman, D. W. & Zumbo, B. D. (2009). Hazards in choosing between pooled and separate-variances t-tests. *Psicológica: Revista de metodología y psicología experimental, 30*(2), 371–390.

Citation

Derrick, B., Toher, D., & White, P. (2016) Why Welch's test is Type I error robust. . *The Quantitative Methods for Psychology, 12*(*1*), 30-38.

Received: 28/08/2015 ~ Accepted: 12/10/2015