
We recommend you cite the published version. The publisher’s URL is: http://www.tqmp.org/RegularArticles/vol12-1/p030/index.html

Refereed: Yes

(no note)

Disclaimer

UWE has obtained warranties from all depositors as to their title in the material deposited and as to their right to deposit such material.

UWE makes no representation or warranties of commercial utility, title, or fitness for a particular purpose or any other warranty, express or implied in respect of any material deposited.

UWE makes no representation that the use of the materials will not infringe any patent, copyright, trademark or other property or proprietary rights.

UWE accepts no liability for any infringement of intellectual property rights in any material deposited but will remove such material from public view pending investigation in the event of an allegation of any such infringement.

PLEASE SCROLL DOWN FOR TEXT.
Why Welch’s test is Type I error robust.

Ben Derricka, Deirdre Tohera & Paul Whitea✉

aUniversity 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(µ1, σ2 1) and N(µ2, σ2 2) to be compared based upon n1 and n2 mutually independent observations. Let \( \bar{X}_i \) and \( S_i^2 \) denote random variables for sample means and variances respectively (i = 1, 2).\(^1\) If the population variances, \( \sigma_1^2 \) and \( \sigma_2^2 \), are assumed to be equal, then an appropriate test statistic is the independent samples t-test, based on (1) and (2).

\[
T_1 = \frac{\bar{X}_1 - \bar{X}_2}{\text{StandardError}(\bar{X}_1 - \bar{X}_2)} \quad (1)
\]

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

\[
SE_1 = S_p \sqrt{\frac{2}{\bar{n}}} \quad (2)
\]

where \( S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 - 1)(n_2 - 1)}} \) and \( \bar{n} \) is the harmonic mean of \( n_1 \) and \( n_2 \). \( T_1 \) is referenced against the t-distribution with degrees of freedom equal to \( \nu_1 = n_1 + n_2 - 2 \).

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, & Olzman, 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).

\[
T_2 = \frac{\bar{X}_1 - \bar{X}_2}{\text{StandardError}(\bar{X}_1 - \bar{X}_2)} \quad (3)
\]

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

\[
SE_2 = \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \quad (4)
\]

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).

\[
\nu_2 = \frac{\left( \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\left( \frac{S_1^2}{n_1} \right)^2 / (n_1 - 1) + \left( \frac{S_2^2}{n_2} \right)^2 / (n_2 - 1)} \quad (5)
\]

\(^1\)As standard notation, random variables are shown in upper case, and derived sample values are shown in lower case.
A numerically equivalent expression for the approximation $\nu_2$ is given in (6). This is shown in some textbooks (e.g., Ott & Longnecker, 2001).

$$\nu_2 = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1) c^2 + (n_1 - 1)(1 - c)^2}$$

where

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

The approximation $\nu_2$ 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 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, $\nu_2$, 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 $\nu_2$.

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. 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
score for Group 2 is 3.82 (standard deviation of 1.20). Thus, $x_1 = 4.410$, $s^2_1 = 1.440$, $x_2 = 3.820$ and $s^2_2 = 1.440$. Calculations for the independent samples t-test give: $s_p = 1.200$, $se_1 = 0.250$, $t_1 = 2.357$, $n_1 = 92.000$, 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 3.90, (standard deviation of 1.28). Thus, $x_1 = 4.530$, $s^2_1 = 2.310$, $x_2 = 3.900$ and $s^2_2 = 1.638$. Calculations for the independent samples t-test give: $s_p = 1.415$, $se_1 = 0.297$, $t_1 = 1.916$, $n_1 = 89.000$, 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 $\sigma_2$ 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 $s^2_2/s^2_1 = [(n_1 - 1)n_1]/[(n_2 - 1)n_2]$. 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 $s^2_2/s^2_1$ tends to zero, the degrees of freedom tends to $n_1 - 1$. As $s^2_2/s^2_1$ becomes increasingly large, the degrees of freedom asymptotically tends to $n_2 - 1$. The maximum value occurs when $s^2_2/s^2_1 = [(n_1 - 1)n_1]/[(n_2 - 1)n_2]$. 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.

<table>
<thead>
<tr>
<th>Test statistics</th>
<th>( T_1, T_2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degrees of freedom</td>
<td>( \nu_1, \nu_2 )</td>
</tr>
<tr>
<td>Sample sizes ((n_1, n_2))</td>
<td>((5,5), (5,100), (100,5), (100,100))</td>
</tr>
<tr>
<td>Standard deviations ((\sigma_1, \sigma_2))</td>
<td>((1,1), (1,2))</td>
</tr>
<tr>
<td>Programming language</td>
<td>R version 3.1.2 (R Development Core Team, 2013)</td>
</tr>
</tbody>
</table>

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 \( n_1 = 98 \), 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 $v_1$.

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 $\sigma_1$ and $\sigma_2$ 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.

<table>
<thead>
<tr>
<th>$(n_1, n_2)$</th>
<th>$(\sigma_1, \sigma_2)$</th>
<th>$T_1$ with $v_1$</th>
<th>$T_2$ with $v_1$</th>
<th>$T_1$ with $v_2$</th>
<th>$T_2$ with $v_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,5</td>
<td>1,1</td>
<td>0.050</td>
<td>0.050</td>
<td>0.045</td>
<td>0.045</td>
</tr>
<tr>
<td></td>
<td>1,2</td>
<td>0.056</td>
<td>0.056</td>
<td>0.047</td>
<td>0.047</td>
</tr>
<tr>
<td>5,100</td>
<td>1,1</td>
<td>0.053</td>
<td>0.012</td>
<td>0.110</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>1,2</td>
<td>0.001</td>
<td>0.000</td>
<td>0.093</td>
<td>0.060</td>
</tr>
<tr>
<td>100,5</td>
<td>1,1</td>
<td>0.050</td>
<td>0.011</td>
<td>0.108</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>1,2</td>
<td>0.295</td>
<td>0.153</td>
<td>0.118</td>
<td>0.052</td>
</tr>
<tr>
<td>100,100</td>
<td>1,1</td>
<td>0.049</td>
<td>0.049</td>
<td>0.049</td>
<td>0.049</td>
</tr>
<tr>
<td></td>
<td>1,2</td>
<td>0.050</td>
<td>0.049</td>
<td>0.050</td>
<td>0.049</td>
</tr>
</tbody>
</table>

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 into 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
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.

<table>
<thead>
<tr>
<th>((n_1, n_2))</th>
<th>((s_1, s_2))</th>
<th>Independent samples t-test</th>
<th>Welch's test</th>
</tr>
</thead>
<tbody>
<tr>
<td>((n_1, n_2))</td>
<td>((s_1, s_2))</td>
<td>test statistic</td>
<td>critical value</td>
</tr>
<tr>
<td>5,5</td>
<td>1,1</td>
<td>1.581</td>
<td>2.306</td>
</tr>
<tr>
<td>5,100</td>
<td>1,1</td>
<td>2.182</td>
<td>1.983</td>
</tr>
<tr>
<td>100,5</td>
<td>1,1</td>
<td>2.182</td>
<td>1.983</td>
</tr>
<tr>
<td>100,100</td>
<td>1,1</td>
<td>2.065</td>
<td>1.983</td>
</tr>
<tr>
<td>100,100</td>
<td>1,2</td>
<td>4.472</td>
<td>1.972</td>
</tr>
</tbody>
</table>


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.
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.


Citation


Copyright © 2016 Derrick, Toher, & White. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

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