Statistical Papers
https://doi.org/10.1007/s00362-024-01531-7

SHORT COMMUNICATION

# Welch's *t* test is more sensitive to real world violations of distributional assumptions than student's *t* test but logistic regression is more robust than either

David Curtis[1]

Received: 3 August 2023 / Revised: 8 January 2024

## Abstract

It has previously been pointed out that Student's *t* test, which assumes that samples are drawn from populations with equal standard deviations, can have an inflated Type I error rate if this assumption is violated. Hence it has been recommended that Welch's *t* test should be preferred. In the context of carrying out gene-wise weighted burden tests for detecting association of rare variants with psoriasis we observe that Welch's test performs unsatisfactorily. We show that if the assumption of normality is violated and observations follow a Poisson distribution, then with unequal sample sizes Welch's *t* test has an inflated Type I error rate, is systematically biased and is prone to produce extremely low *p* values. We argue that such data can arise in a variety of real world situations and believe that researchers should be aware of this issue. Student's *t* test performs much better in this scenario but a likelihood ratio test based on logistic regression models performs better still and we suggest that this might generally be a preferable method to test for a difference in distributions between two samples.

This research has been conducted using the UK Biobank Resource.

**Keywords** Welch's *t* test · Student's *t* test · Likelihood ratio test · Logistic regression · Psoriasis

✉ David Curtis
d.curtis@ucl.ac.uk

[1] UCL Genetics Institute, UCL, Darwin Building, Gower Street, WC1E 6BT London, England

Published online: 04 March 2024

## 1 Introduction

Rasch and colleagues recommended that Welch's *t* test should be used in preference to Student's *t* test or Wilcoxon's *U* test because, when tested over a range of simulated samples with varying skewness and kurtosis, Student's *t* test had an inflated Type 1 error rate when samples were drawn from populations which had unequal standard deviations whereas Wilcoxon's *U* test could often have lower power and sometimes inflated type 1 error rate (Rasch et al. 2011). Likewise, Delacre and colleagues argued that Welch's *t* test should be used in preference to Student's *t* test because if the assumption of equal standard deviations is violated then, if sample sizes are also unequal, Student's *t* test can have an inflated Type I error rate (Delacre et al. 2017, 2022). They argue that even when the assumption of equal standard deviations is justified then Student's *t* test is only slightly more powerful and that the differences in Type II error rate are small, so that overall Welch's *t* test is to be generally preferred because it is more robust with respect to Type I error. They provide an example where the ratio of standard deviations (SDR) between the populations is 2 and with sample sizes of 40 and 60, and the smaller sample being drawn from the population with the larger standard deviation, then for a normally distributed variable the Type I error rate at $\alpha=0.05$ is 0.083 for Student's *t* test. They also provide results for a number of combinations of different distributions, including double exponential, chi-squared and normal skewed, some of which show similar effects. Both Student's *t* test and Welch's *t* test make the assumption that the variable is normally distributed, although they are recognised to be often fairly robust to departures from this assumption and it may sometimes not be formally tested. If samples are drawn from two different populations then Welch's *t* test is expected to have a Type I error rate at $\alpha=0.05$ of 0.05 if both populations have the same mean, whereas for Student's *t* test this is only true if the two populations have the same mean and the same standard deviation.

Here, I draw attention to scenarios relevant to the real world where Welch's *t* test is anti-conservative, systematically biased and prone to produce extremely small p values even when samples are drawn from the same population. This can occur when the variable is not normally distributed.

The phenomenon came to attention when carrying out a weighted burden test of rare genetic variants within genes, to test for association with psoriasis in the UK Biobank sample (Szustakowski et al. 2021). UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51,119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11,527/001).

Weighted burden analysis involves assigning a score to each genetic variant based on its rarity and predicted functional effect and then for each research subject summing up the scores of the variants they carry within a particular gene to produce an overall score for the gene-wise variant burden (Curtis 2016). Association testing can be carried out to see if this score differs between cases and controls. When the method was first developed association testing was done using a *t* test but in order to account for ancestral diversity in the UK Biobank sample the method was adapted to use logistic regression instead, incorporating sex and population principal components as covariates (Curtis 2021). This approach was applied by using R to analyse the weighted burden score for each gene and each subject in 2,944 cases with psoriasis and 197,683 controls. As was not unexpected, given the fairly low heritability of psoriasis, the results of the full logistic regression analysis incorporating principal components and sex conformed closely to those expected under the null hypothesis that no gene demonstrated an increased variant burden associated with disease. However, as a sanity check two additional tests were performed for each gene, one being a likelihood ratio test based on a logistic regression model with no covariates and the other being a *t* test. The likelihood ratio test demonstrated moderate inflation, presumably due to population stratification. However for a number of genes the *t* test produced extremely low p values, with 393 out of 20,637 genes producing a p value less than $10^{-10}$ and with one gene producing a p value of $10^{-102}$. Further investigation revealed that this was a consequence of the default *t* test implemented in R being Welch's *t* test rather than Student's *t* test and when the same data was analysed instead using Student's *t* test then results conformed much more closely to those obtained using the likelihood ratio test from the logistic regression analysis. Results of the gene-wise weighted burden analyses are provided in Supplementary Table 1.

In order to better understand this phenomenon, simulated data was generated to compare the performance of Welch's *t* test, Student's *t* test and logistic regression.

## 2 Method

R code was written to generate simulated case and control samples with the ability to specify for each sample the sample size, population mean and population variance. Although the weighted burden score for each gene is a quantitative measure, in fact many subjects may not carry any variants at all whereas others may carry a small number of variants with very high scores so that in reality the distribution of scores is somewhat similar to a Poisson distribution. Therefore, the option was provided to simulate data using either a Poisson or normal distribution. It also seemed possible that the observed problems with the *t* test might be related to the disparity in sample sizes so simulations were performed with equal and unequal sample sizes. All simulations were carried out under the null hypothesis, with population means and standard deviations being equal for cases and controls. Each simulated dataset was analysed using Welch's *t* test, Student's *t* test and a likelihood ratio test based on a logistic regression model.

For this last test, the log likelihood was obtained for two logistic regression models, one being a model in which case-control status was predicted from the score and the other being a null model, and then twice the difference in log likelihoods was treated as a chi-squared statistic with one degree of freedom. The p value obtained for each of the three tests was then converted to a signed log p (SLP), defined as the logarithm base 10 of the two-tailed p value, given a positive sign if the case mean score was higher or a negative one if the control mean was higher. For each set of simulations a quantile-quantile (QQ) plot was generated of the observed versus expected distribution of SLPs.

We may note in passing that there are conceptual differences between the *t* tests and the test based on logistic regression. For the *t* tests, one assumes that there are two populations and one seeks to test whether the mean of a quantitative variable differs between them. In a logistic regression framework, one tests whether the value of a quantitative variable influences the probability of being assigned to one of two possible outcomes. However from the point of view of a test for association these underlying notions of causality have no bearing on the performance of the tests.

## 3 Results

Exploratory analyses revealed that all three tests performed well if either the two sample sizes were equal or if the normal distribution was used. However with unequal sample sizes and the Poisson distribution Welch's *t* test had an inflated Type I error rate, was systematically biased towards regarding lower means in the smaller sample as statistically significant and could occasionally produce extremely small p values. Summary results illustrating these phenomena are shown in Table 1, based on sets of 10,000 simulations. Figure 1 shows the full QQ plots obtained from 100,000 simulations of 100 cases and 900 controls using a Poisson distribution with a mean score of 0.05.

In this example, at $\alpha = 0.05$ Welch's *t* test has a Type I error rate of 0.082 compared with 0.046 for Student's *t* test and 0.056 for the likelihood ratio test. Welch's *t* test is systematically biased, with a mean SLP of -0.23 instead of the desired value of 0. Student's *t* test and the likelihood ratio test produce mean SLPs of 0.017 and –0.045 respectively. Most worryingly, Welch's *t* test is prone to produce extreme low p values. It produces an SLP less than –4 (equivalent to $p = 10^{-4}$) for 878 out of 100,000 replicates compared with the 5 which should be expected by chance. By contrast, Student's *t* test yields 42 SLPs greater than 4 and the likelihood ratio test is slightly conservative, with only 3 SLPs having an absolute value greater than 4. Welch's *t* test produces an SLP of less than –8 (equivalent to $p = 10^{-8}$) for 635 out of 100,000 replicates.

As helpfully pointed out by an anonymous referee, the problematic performance of Welch's *t* test with the Poisson distribution is ameliorated if the sample size is increased. For example, with 1,000 cases and 9,000 controls the Type I error rate at $\alpha = 0.05$ falls from 0.082 to 0.055 and the mean SLP increases from –0.23 to -0.049.

Examination of the QQ plots shows that although there are clearly problems with Welch's *t* test it is also the case that Student's *t* test does not perform as it should and appears inferior to the likelihood ratio test. In particular, with Student's *t* test there is deflation of the negative SLPs and inflation of the positive SLPs, meaning that no negative SLP is less than –2.3 whereas one positive SLP exceeds 6.

**Table 1** Summary results for sets of 10,000 simulations showing Type I error rates and mean SLP for the three statistical tests. For all simulations the mean and variance of the generating distribution were set to 0.05. For a test to perform well, the Type I error rate at α=0.05 should be 0.05 and the mean SLP across simulations should be 0. All three tests show adequate performance in all scenarios except that when case and control sample sizes are unequal and a Poisson distribution is used then Welch's t test has a Type I error rate of 0.078 and a mean SLP of -0.2149

|  |  |  | Type I error rate at α=0.05 |  |  | Mean SLP |  |  |
|---|---|---|---|---|---|---|---|---|
| Number of controls | Number of cases | Generating distribution | Welch's *t* test | Student's *t* test | Likelihood ratio test | Welch's *t* test | Student's *t* test | Likelihood ratio test |
| 500 | 500 | Normal | 0.051 | 0.051 | 0.051 | -0.0085 | -0.0085 | -0.0085 |
| 500 | 500 | Poisson | 0.050 | 0.050 | 0.052 | 0.0074 | 0.0074 | 0.0075 |
| 900 | 100 | Normal | 0.051 | 0.050 | 0.050 | 0.0012 | 0.0006 | 0.0006 |
| 900 | 100 | Poisson | 0.078 | 0.047 | 0.054 | -0.2149 | 0.0327 | -0.0288 |

## 4 Discussion

To gain insight into the differential performance of Welch's *t* test and Student's *t* test when sample sizes differ it is helpful to look at the formulae they use to estimate the standard error of the mean difference between samples. Each test depends on taking the difference between the observed means and dividing the difference by this standard error to produce the *t* statistic. Using the notation N1 and N2 to denote sample sizes and s1 and s2 to denote sample standard deviations, to obtain the standard error for the difference in means Student's test uses this formula:

$$((1/N1+1/N2) * ((N1-1) * s1^2 + (N2-1) * s2^2) / (N1+N2-2))^{1/2}$$

This essentially uses a weighted average of the sample variances, which begins with multiplying each variance by its sample size.

By contrast Welch's test uses this to obtain the standard error of the mean difference:

$$(s1^{2} / N1+s2^{2} / N2)^{1/2}$$

In Student's formula each variance is multiplied by the sample size whereas in Welch's formula each variance is divided by its sample size. This explains how the bias arises in Welch's test - if the larger variance occurs in the larger sample then after division by the sample size it will make only a relatively small contribution to the standard error of the mean difference, tending to result in a larger *t* statistic. Using a Poisson distribution to generate the sample scores leads to the sample with the higher mean score tending to have a higher variance and so simulations in which the larger sample has the higher mean score will be more likely to produce a statistically significant *t* statistic, accounting for the observed systematic bias. For Student's test the situation is somewhat reversed, explaining the slight deflation of negative SLPs and inflation of positive SLPs, but the magnitude of effect is small in comparison and no extreme *p* values are generated.

It is clear that Welch's *t* test is unsuitable for the simulated datasets used for Fig. 1. It could be argued that it is well known that that both *t* tests depend on the assumption of normally distributed data and hence it should be obvious that use of either would be inappropriate. To this one could counter that there is a general view that *t* tests are in fact quite robust to departures from normality, that both tests work acceptably even on Poisson distributed data when sample sizes are similar and that the Student's *t* test does in fact perform reasonably well in all situations.

It is not hard to imagine real world circumstances, outside of genetics, where a researcher might be dealing with similar results. The unequal sample sizes and Poisson distribution could occur in cohort studies when an outcome is relatively uncommon and exposures are also somewhat rare, for example testing whether adults currently taking antidepressants had witnessed more car traffic accidents or whether the number of school exclusions during childhood was associated with increased risk of incarceration by age 30. It might be that in an epidemiological study in which a number of variables were analysed then some could approximate a Poisson distribution without the researcher being aware. In any event, the problems with Welch's test

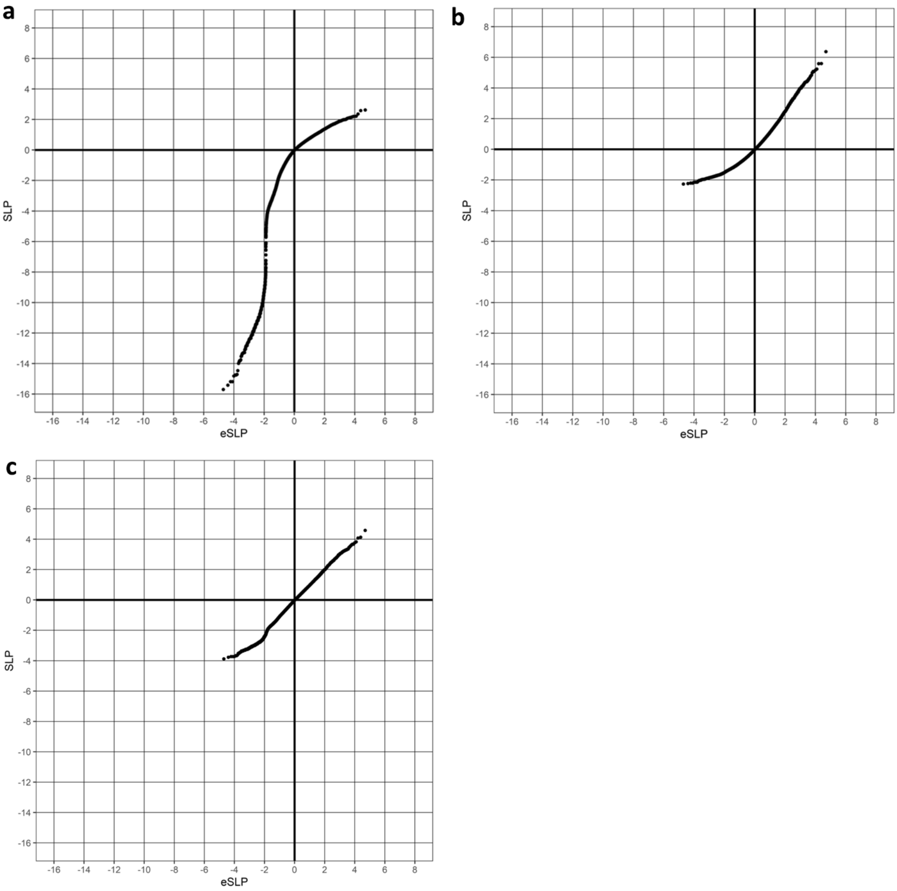

**Fig. 1** QQ plots showing the signed log p values (SLPs) plotted against the values expected under the null hypothesis (eSLP) obtained from 100,000 simulations with 900 controls, 100 cases and a Poisson distribution with mean = 0.05. Each point represents the result for one simulation. For a test to perform well, the points should tend to lie symmetrically on the x = y line in the segment between (-5,-5) and (5,5). **a** Results for Welch's *t* test. **b** Results for Student's *t* test. **c** Results for likelihood ratio test using logistic regression models

in this situation seem far more severe than those previously reported for Student's test, implying that researchers should at least bear this in mind as a potential issue. It may be of concern that when a *t* test is performed in R then the default implementation is in fact Welch's test rather than Student's test and the results presented here suggest that perhaps Student's test should be routinely preferred, or at least that researchers should be encouraged to specifically choose one test or the other.

Of course, the QQ plots in Fig. 1 beg the question as to whether either *t* test should be used to compare the means between two groups, given the superior performance of the likelihood ratio test. The likelihood ratio test based on logistic regression has been shown to have acceptable Type I error rates for testing association with a single genetic variant in a variety of situations (Ma et al. 2013). The present examples demonstrate that this also applies in the situation of a weighted burden test with rare variants and unequal sample sizes. Although having the correct Type I error rate is crucial for any statistical test, a secondary consideration is the power of the test and in theory this likelihood ratio test will be the most powerful only if the key assumption of the logistic regression model is met, which is that the predictor variable is linearly correlated with the logarithm of the odds. This means that it is possible that there could be situations in which one or other *t* test could have both the correct Type I error rate and higher power than the likelihood ratio test. In practice, logistic regression is widely used for testing association in genetic studies because it allows inclusion of relevant covariates. A *t* test can be calculated by hand and is relatively easy to understand and teach but nowadays software such as R, in which the likelihood ratio test can be easily implemented, is readily available. Although the concepts underlying logistic regression may be more complex and may not reflect real world causal relationships, based on the findings presented here an argument could be made that it should be the default method for testing whether the distribution of a quantitative variable differs between two populations. Further work could be undertaken to systematically investigate the advantages and disadvantages of each approach.

**Supplementary Information** The online version contains supplementary material available at https://doi.org/10.1007/s00362-024-01531-7.

**Acknowledgements** This research has been conducted using the UK Biobank Resource under Application Number 51119. The author wishes to acknowledge the staff supporting the High Performance Computing Cluster, Computer Science Department, University College London. The author wishes to thank the participants who volunteered for the UK Biobank project.

**Author contributions** David Curtis conceived the study, carried out the analyses and wrote the manuscript.

**Funding** No external funding was received.

**Code and data availability** The genetic and phenotype data is available on application to UK Biobank. Scripts and software used to carry out the genetic analyses are available at: https://github.com/davenomiddlenamecurtis. Code in R to carry out the simulations and produce the QQ plots is at https://github.com/davenomiddlenamecurtis/TestTTest.

## Declarations

**Ethical approval** UK Biobank had obtained ethics approval from the North West Multi-centre Research Ethics Committee which covers the UK (approval number: 11/NW/0382) and had obtained informed consent from all participants. The UK Biobank approved an application for use of the data (ID 51119) and ethics approval for the analyses was obtained from the UCL Research Ethics Committee (11527/001).

**Competing interests** I declare no competing interests.

## References

Curtis D (2016) Pathway analysis of whole exome sequence data provides further support for the involvement of histone modification in the aetiology of schizophrenia. Psychiatr Genet 26:223–227. https://doi.org/10.1097/YPG.0000000000000132

Curtis D (2021) Analysis of 200 000 exome-sequenced UK Biobank subjects illustrates the contribution of rare genetic variants to hyperlipidaemia. J Med Genet. https://doi.org/10.1136/jmedgenet-2021-107752. jmedgenet-2021-107752

Delacre M, Lakens D, Leys C (2017) Why psychologists should by default Use Welch's t-test instead of Student's t-test. Int Rev Social Psychol 30(1):92–101. https://doi.org/10.5334/IRSP.82

Delacre M, Lakens D, Leys C (2022) Correction: why psychologists should by default Use Welch's t-test instead of Student's t-test. Int Rev Social Psychol 35(1). https://doi.org/10.5334/IRSP.661/

Ma C, Blackwell T, Boehnke M, Scott LJ (2013) Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. Genet Epidemiol 37(6):539–550. https://doi.org/10.1002/GEPI.21742

Rasch D, Kubinger KD, Moder K (2011) The two-sample t test: pre-testing its assumptions does not pay off. Stat Pap 52(1):219–231. https://doi.org/10.1007/S00362-009-0224-X/METRICS

Szustakowski JD, Balasubramanian S, Kvikstad E, Khalid S, Bronson PG, Sasson A, Wong E, Liu D, Wade Davis J, Haefliger C, Katrina Loomis A, Mikkilineni R, Noh HJ, Wadhawan S, Bai X, Hawes A, Krasheninina O, Ulloa R, Lopez AE, Team U-ER (2021) Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nat Genet 53(7):942–948. https://doi.org/10.1038/s41588-021-00885-0