Behavioral Ecology
doi:10.1093/beheco/ark016
Advance Access publication 17 May 2006

Forum

The unequal variance *t*-test is an underused alternative to Student's *t*-test and the Mann–Whitney *U* test

Graeme D. Ruxton
Division of Environmental and Evolutionary Biology, Institute of Biomedical and Life Sciences, Graham Kerr Building, University of Glasgow, Glasgow G12 8QQ, United Kingdom

Often in the study of behavioral ecology, and more widely in science, we require to statistically test whether the central tendencies (mean or median) of 2 groups are different from each other on the basis of samples of the 2 groups. In surveying recent issues of Behavioral Ecology (Volume 16, issues 1–5), I found that, of the 130 papers, 33 (25%) used at least one statistical comparison of this sort. Three different tests were used to make this comparison: Student's *t*-test (67 occasions; 26 papers), Mann–Whitney *U* test (43 occasions; 21 papers), and the *t*-test for unequal variances (9 occasions; 4 papers). My aim in this forum article is to argue for the greater use of the last of these tests. The numbers just related suggest that this test is not commonly used. In my survey, I was able to identify tests described simply as "*t*-tests" with confidence as either a Student's *t*-test or an unequal variance *t*-test because the calculation of degrees of freedom from the 2 sample sizes is different for the 2 tests (see below). Hence, the neglect of the unequal variance *t*-test illustrated above is a real phenomenon and can be explained in several (nonexclusive ways) ways:

1. Authors are unaware that Student's *t*-test is unreliable when variances differ between underlying populations.
2. Authors are aware of this but consider their samples to have similar variances.
3. Authors believe than the Mann–Whitney *U* test can effectively substitute for Student's *t*-test when variances are unequal.
4. Because the *t* distribution tends to the normal distribution for large sample sizes, authors may consider that their sample sizes are sufficiently large for concerns about unequal variance and nonnormality of the samples to be ignored.

Argument 4 relies on the central limit theorem and would require a combined sample size of at least 30 (Sokal and Rohlf 1987, p. 107); however, in my survey, the majority (47 out of 61) of tests for which sample sizes were provided had a combined sample size below 30. The fallacy of argument 3 has been demonstrated previously on several occasions (e.g., Kasuya 2001; Neuhauser 2002). To explore argument 1 further, imagine that we have 2 sample groups (labeled "1" and "2," with means [μ_1 and μ_2], variances [s_1^2 and s_2^2], and sample sizes [N_1 and N_2]). For the unpaired Student's *t*-test, the *t* statistic is calculated as

$$t = \frac{\mu_1 - \mu_2}{s_p^2\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}}, \tag{1}$$

where the pooled variance s_p^2 is given by

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}. \tag{2}$$

The variances of the 2 samples are pooled in order to achieve the best estimate of the (assumed equal) variances of the 2 populations. Hence, we can see the need for the underlying assumption of equal population variances in this test. The Student's *t*-test performs badly when these variances are actually unequal, both in terms of Type I and Type II errors (Zar 1996). Figure 1 suggests that unequal sample variances are common in behavioral ecology. Although it is true that unequal variances are less problematic if sample sizes are similar, in practice, we often have quite unequal sample sizes (Figure 2). Hence, I suggest that the Student's *t*-test is frequently used in behavioral ecology when one of its important underlying assumptions is violated, and consequently, its performance is unreliable.

The unequal variance *t*-test does not make the assumption of equal variances. Coombs et al. (1996) presented measured Type I errors obtained by simulated sampling from normal distributions for the Student's *t*-test and the unequal variance *t*-test (their result are summarized in Table 1). In the examples in Table 1, we see that the Type I error rate of the unequal variance *t*-test never deviates far from the nominal 5% value, whereas the Type I error rate for the Student's *t*-test was over 3 times the nominal rate when the higher variance was associated with the smaller sample size and less than a quarter the nominal rate when the higher variance was associated with the higher sample size. These results concur qualitatively with other studies of these 2 tests (e.g., Zimmerman and Zumbo 1993). Notice that even when the variances are identical, the unequal variance *t*-test performs just as effectively as the Student's *t*-test in terms of Type I error. The power of the unequal variance *t*-test is similar to that of the Student's *t*-test even when the population variances are equal (e.g., Moser et al. 1989; Moser and Stevens 1992; Coombs et al. 1996). Hence, I suggest that the unequal variance *t*-test performs as well as, or better than, the Student's *t*-test in terms of control of both Type I and Type II error rates whenever the underlying distributions are normal.

Let us now consider convenience of calculation: the unequal variance *t*-test involves calculation of a *t* statistic that is compared with the appropriate value in standard *t* tables. The test statistic for the unequal variance *t*-test (t') is actually slightly simpler than that of the Student's *t*-test:

$$t' = \frac{\mu_1 - \mu_2}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}}. \tag{3}$$

However, the calculation of the degrees of freedom (v) is more involved but not prohibitively so. For the Student's *t*-test, $v = N_1 + N_2 - 2$; for the unequal variance *t*-test, it is given (e.g., Moser and Stevens 1992) by

$$v = \frac{\left(\frac{1}{n_1}+\frac{u}{n_2}\right)^2}{\frac{1}{n_1^2(n_1-1)}+\frac{u^2}{n_2^2(n_2-1)}}, \tag{4}$$

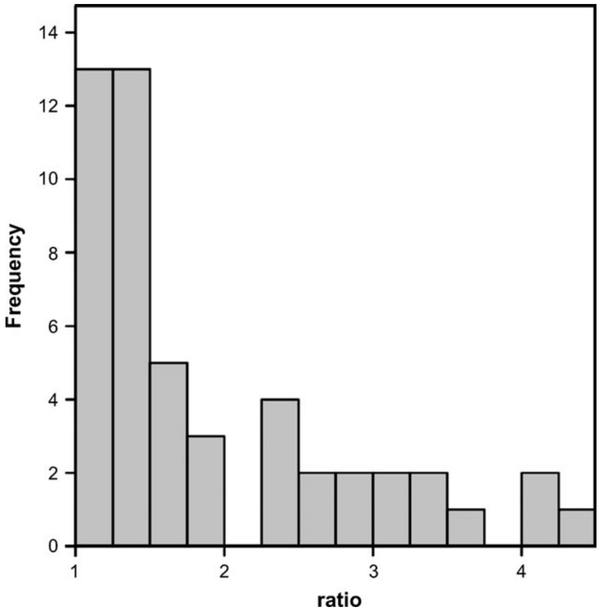

Figure 1
Histogram of the larger divided by the smaller variance for 35 *t*-tests and *U* tests in my sample from Behavioral Ecology for which the variances were provided in the paper. Note for ease of presentation, the following 3 variance ratios were not plotted: 9.0, 9.0, and 21.0.

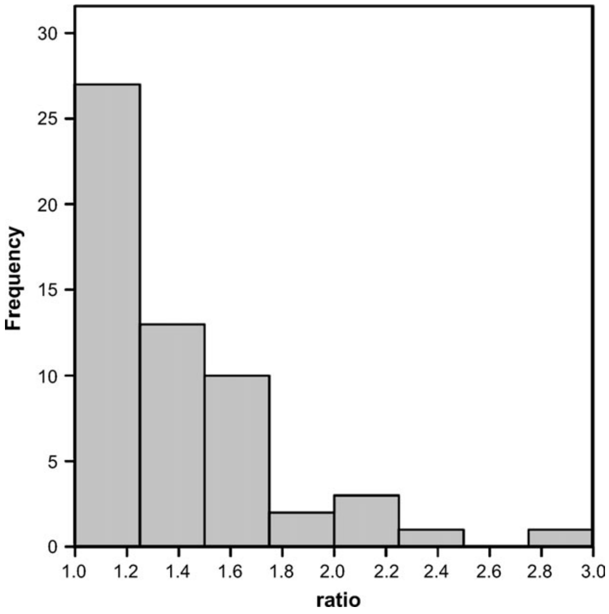

Figure 2
Histogram of the ratio of the highest over the lowest sample size for 61 Student's *t*-tests and Mann–Whitney *U* tests in my survey from Behavioral Ecology for which sample sizes were provided. For ease of presentation, the following four ratios were not plotted: 3.1, 3.3, 4.2, and 5.7.

where

$$u = \frac{s_2^2}{s_1^2}. \quad (5)$$

In general, v calculated from Equation 4 will take a noninteger value; it is conventional to round down to the nearest integer before consulting standard t tables. Hence, the calculation of the unequal variance *t*-test is straightforward. Further, the test is available in several commonly used statistics packages: for example, Excel, Minitab, SPSS, SAS, and SYSTAT. Hence, ease of calculation is not a valid reason for choosing a Student's *t*-test over an unequal variance *t*-test.

The unequal variance *t*-test has no performance benefits over the Student's *t*-test when the underlying population variances are equal. Hence, you might consider that an effective way to conduct your analysis would be to perform an initial test for homogeneity of variance and then perform either a Student's *t*-test when the variances are equal or an unequal variance *t*-test when they are not. The problem with this flexible approach is that the combination of this preliminary test plus whichever of the subsequent tests is ultimately used controls Type I error rates less well than simply always performing an unequal variance *t*-test on every occasion (Gans 1992; Moser and Stevens 1992), this is one reason why it is generally unwise to decide whether to perform one statistical test on the basis of the outcome of another (Zimmerman 2004 and references therein). There are further reasons for not recommending preliminary tests of variances (e.g., Markowski CA and Markowski EP 1990; Quinn and Keough 2002, p. 42). Hence, I suggest avoiding preliminary tests and adopting the unequal variance *t*-test unless an argument based on logical, physical, or biological grounds can be made as to why the variances are very likely to be identical for the 2 populations under investigation.

It is important to remember that although the unequal variance *t*-test is more reliable than the Student's *t*-test in terms of violation of the assumption of homogeneity of variances, it is not necessarily any more reliable than the Student's *t*-test if the assumption of normality of the underlying populations is violated. However, Zimmerman and Zumbo (1993) argue that the unequal variance *t*-test performed on ranked data performs just as well as the Mann–Whitney *U* test (in terms of control of Type I errors) when variances are equal and considerably better than the *U* test when variances are unequal (see Table 2 for an example). This behavior was found when tested with populations coming from 8 different types of nonnormal distribution. Thus, Zimmerman and Zumbo (1993) suggest that the unequal variance *t*-test can effectively replace the Mann–Whitney *U* test if the data are first ranked before the test is applied. There are alternatives to the unequal variance *t*-test that perform even better, in particular, being more robust to nonnormality in the underling populations (e.g., Coombs et al. 1996; Keselman et al. 2004). However, I recommend the unequal variance *t*-test as having the best combination of performance and ease of use.

I have used the name unequal variance *t*-test as this is its most common name in the literature, you may also find in referred to as the Welch test deriving from Welch (1938,

Table 1
Calculated Type I error rate for the *t*-test and unequal variance *t*-test with a nominal α value of 0.05 (adapted from Coombs et al. 1996)

N_1	N_2	s_1	s_2	*t*-test	Unequal
11	11	1	1	0.052	0.051
11	11	4	1	0.064	0.054
11	21	1	1	0.052	0.051
11	21	4	1	0.155	0.051
11	21	1	4	0.012	0.046
25	25	1	1	0.049	0.049
25	25	4	1	0.052	0.048

Table 2

Calculated Type I error rate for the Mann–Whitney *U* test and the unequal variance *t*-test performed on the ranked data from normal distributions with a nominal α value of 0.05 (adapted from Zimmerman and Zumbo 1993)

N_1	N_2	s_1/s_2	*U* test	Unequal
6	18	1	0.052	0.049
6	18	1.5	0.059	0.052
6	18	2	0.085	0.051
6	18	2.5	0.098	0.054
6	18	3	0.108	0.053
6	18	3.5	0.117	0.052
6	18	4	0.104	0.054
18	6	1	0.049	0.052
18	6	1.5	0.038	0.054
18	6	2	0.030	0.056
18	6	2.5	0.028	0.059
18	6	3	0.030	0.064
18	6	3.5	0.025	0.066
18	6	4	0.023	0.063

1947). Welch actually proposed several ways to evaluate the degrees of freedom, and the method I describe in Equations 4 and 5 is sometimes referred to as the Welch Approximate Degrees of Freedom (APDF) test. Note that statistical packages may use other methods for calculating the degrees of freedom. You may also encounter the unequal variances *t*-test called simply the unpooled variances *t*-test or Satterwaite's test or the Welch–Satterthwaite test, after Satterwaite (1946). You may also find it called as the Smith/Welch/Satterwaite test, acknowledging the work in Smith (1936).

The importance of considering whether or not to pool variances extends beyond the simple case of comparing 2 groups. Julious (2005) argues against the standard practice of using the pooled variance across all groups when performing a comparison between 2 groups from several used in an analysis of variance. Indeed, Julious (2005) argues that using a pooled variance across more than 2 groups can be even more serious than the issues covered in this paper. No matter the number of groups, the decision as to whether to pool or not also needs careful consideration in the construction of randomization tests as well as the analytic tests considered here.

IN CONCLUSION: A STEP-BY-STEP SUMMARY

If you want to compare the central tendency of 2 populations based on samples of unrelated data, then the unequal variance *t*-test should always be used in preference to the Student's *t*-test or Mann–Whitney *U* test. To use this test, first examine the distributions of the 2 samples graphically. If there is evidence of nonnormality in either or both distributions, then rank the data. Take the ranked or unranked data and perform an unequal variance *t*-test. Draw your conclusions on the basis of this test. Note that some packages (e.g., SPSS) perform a Student's *t*-test and unequal variances *t*-test simultaneously and provide output for both. The experimenter ought to have decided which test they consider most appropriate beforehand and thus look at the output for that test alone, ignoring the other.

In presenting the outcome of the unequal variance *t*-test, provide a suitable reference for the adoption of the test and its exact formulation (e.g., Moser et al. 1989 or this paper) as well as providing the mean, variance, and number of samples in each group, the calculated t' value, the calculated degrees of freedom (v), and finally the *P* value.

Thanks to Steven Julious and 2 anonymous referees for helpful comments on a previous version.

Address correspondence to G.D. Ruxton. E-mail: g.ruxton@bio.gla.ac.uk.

Received 10 November 2005; revised 18 April 2006; accepted 19 April 2006.

REFERENCES

Coombs WT, 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. Rev Educ Res 66:137–79.

Gans DJ. 1991. Preliminary tests on variances. Am Stat 45:258.

Julious SA. 2005. Why do we use pooled variance analysis of variance. Pharm Stat 4:3–5.

Kasuya E. 2001. Mann-Whitney U-test when variances are unequal. Anim Behav 61:1247–9.

Keselman HJ, Othman AR, Wilcox RR, Fradette K. 2004. The new and improved two-sample *t* test. Psychol Sci 15:47–51.

Markowski CA, Markowski EP. 1990. Conditions for the effectiveness of a preliminary test of variance. Am Stat 44:322–6.

Moser BK, Stevens GR. 1992. Homogeneity of variance in the two-sample means test. Am Stat 46:19–21.

Moser BK, Stevens GR, Watts CL. 1989. The two-sample *t*-test versus Satterwaite's approximate *F* test. Commun Stat Theory Methodol 18:3963–75.

Neuhauser M. 2002. Two-sample tests when variances are unequal. Anim Behav 63:823–5.

Quinn GP, Keough MJ. 2002. Experimental design and data analysis for biologists. Cambridge: Cambridge University Press.

Satterwaite FE. 1946. An approximate distribution of estimates of variance components. Biometrics Bull 2:110–4.

Smith H. 1936. The problem of comparing the results of two experiments with unequal errors. J Counc Sci Ind Res 9:211–2.

Sokal RR, Rohlf FJ. 1987. Introduction to biostatistics. 2nd ed. New York: Freeman.

Welch BL. 1938. The significance of the difference between two means when the population variances are unequal. Biometrika 29:350–62.

Welch BL. 1947. The generalisation of students problem when several different population variances are involved. Biometrika 34: 23–35.

Zar JH. 1996. Biostatistical analysis. 3rd ed. London: Prentice Hall International.

Zimmerman DW, Zumbo BN. 1993. Rank transformations and the power of the Student *t*-test and Welch t'-test for non-normal populations. Can J Exp Psychol 47:523–39.

Zimmerman DW. 2004. A note on preliminary tests of equality of variances. Br J Math Stat Psychol 57:173–81.