The Two-Sample Trimmed t for Unequal Population Variances

Author(s): Karen K. Yuen

Source: *Biometrika*, Apr., 1974, Vol. 61, No. 1 (Apr., 1974), pp. 165-170
Published by: Oxford University Press on behalf of Biometrika Trust
Stable URL: http://www.jstor.com/stable/2334299

Biometrika (1974), **61**, 1, *p.* 165
With 1 *text-figure*
Printed in Great Britain

The two-sample trimmed *t* for unequal population variances

By KAREN K. YUEN
Department of Biomathematics, University of California, Los Angeles

Summary

The effect of nonnormality on the Welch approximate degrees of freedom t test is demonstrated. A two-sample trimmed t statistic for unequal population variances is proposed and its performance is also evaluated in comparison to the Welch t test under normality and under long-tailed distributions. If the underlying distribution is long-tailed or contaminated with outliers, the trimmed t is strongly recommended.

Some key words: Behrens–Fisher problem; Robust test; t distribution; Trimming; Winsorization.

1. Introduction

Testing the equality of two means from independent samples is a common statistical problem. If the underlying distributions are normally distributed with equal population variances, it is well known that one would use the Student's t test. Unfortunately, this test statistic is sensitive to some nonnormal situations and that leads us to consider other more robust alternatives. Among the alternatives is the two-sample trimmed t, proposed and evaluated by Yuen & Dixon (1973); definitions of trimming will be given later. This statistic can be easily computed and its distribution is satisfactorily approximated by that of a Student's t with the degrees of freedom corresponding to the reduced sample. Results show that the loss of power efficiency in using trimmed t is small under exact normality, while the gain may be appreciable for long-tailed distributions.

In cases where distributions are normal but population variances are unknown, the first exact solution was given by Behrens (1929) and later extended by Fisher (1939) as the fiducial solution. Among many others who have worked on this problem, Welch (1938, 1949) provided an approximate degrees of freedom t solution to his asymptotic series solution. Wang (1971) studied the probabilities of the type I errors of these two Welch solutions for selected cases and concluded that in practice, one can use the Welch approximate degrees of freedom t test without much loss of accuracy. For ease of reference in the rest of this paper, we shall refer to this approximate degrees of freedom t simply as the Welch t test.

In this paper, the lack of robustness of the Welch t test when the underlying distribution is nonnormal is demonstrated. A two-sample trimmed t statistic for unequal population variances is suggested and its performance is evaluated in comparison to the Welch t test under normality and under long-tailed distributions.

2. Performance of the Welch t test for different underlying distributions

The Welch t statistic is given as follows:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{(s_1^2/n_1 + s_2^2/n_2)^{\frac{1}{2}}},$$

where $\bar{x}$ and s^2 are means and variances of the samples. The approximate degrees of freedom f is defined by

$$\frac{1}{f} = \frac{c^2}{n_1 - 1} + \frac{(1-c)^2}{n_2 - 1}, \quad c = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

Probabilities of type I errors of the two-sided test were obtained by Monte Carlo methods for the following underlying distributions: uniform; normal; Student's t; mixtures of normal and normal/uniform, the distribution whose probability density function is the quotient of a normal probability density function by a uniform probability density function; normal/uniform; and Cauchy. Nominal sizes of 0·01, 0·05, and 0·10 were considered. Sample sizes ranged from 10 to 20; 5000 samples were replicated and the results are presented in Table 1.

The Welch t test becomes more conservative as the tail of the underlying distribution gets longer and as the sample sizes get smaller. For very long-tailed distributions like normal/uniform and Cauchy, the probabilities of type I errors are much less than the nominal sizes. On the other hand, the performance of the Welch t test is adequate for short-tailed distributions such as the uniform. Accordingly, it seems appropriate to give most attention to symmetric distributions that have longer tails than the normal.

Table 1. *Probability of type I error of the Welch t test*

				Nominal α		
Population 1	Population 2	n_1	n_2	0·01	0·05	0·10
$U[-1, 1]$	$U[-0{\cdot}5, 0{\cdot}5]$	10	10	0·011	0·049	0·097
$U[-1, 1]$	$U[-0{\cdot}5, 0{\cdot}5]$	20	10	0·011	0·052	0·097
$U[-2, 2]$	$U[-0{\cdot}5, 0{\cdot}5]$	20	20	0·011	0·050	0·094
$N(0, 4^2)$	$N(0, 1)$	10	20	0·011	0·050	0·098
$N(0, 2^2)$	$N(0, 1)$	10	10	0·009	0·050	0·097
$4 \times t_8$	t_8	20	10	0·010	0·052	0·100
$2 \times t_5$	t_5	10	10	0·009	0·046	0·098
$0{\cdot}8N(0, 4^2) + 0{\cdot}2N(0, 4^2)/U[0, 1]$	$0{\cdot}8N(0, 1) + 0{\cdot}2N(0, 1)/U[0, 1]$	10	20	0·004	0·031	0·069
$0{\cdot}8N(0, 4^2) + 0{\cdot}2N(0, 4^2)/U[0, 1]$	$0{\cdot}8N(0, 1) + 0{\cdot}2N(0, 1)/U[0, 1]$	20	10	0·004	0·032	0·077
$N(0, 4^2)/U[0, 1]$	$N(0, 1)/U[0, 1]$	20	10	0·001	0·020	0·059
$N(0, 4^2)/U[0, 1]$	$N(0, 1)/U[0, 1]$	10	20	0·002	0·020	0·057
$2 \times$ Cauchy (0, 1)	Cauchy (0, 1)	10	10	0·001	0·016	0·053
Cauchy (0, 1)	Cauchy (0, 1)	20	20	0·002	0·020	0·057
$4 \times$ Cauchy (0, 1)	Cauchy (0, 1)	10	20	0·002	0·019	0·056

U: uniform distribution. N: normal. t_f: Student's t with f degrees of freedom. N/U: normal/uniform.

3. Formulation of a trimmed t statistic

Let $x_1, \ldots, x_n$ be an ordered sample of size n from a population. The g-times trimmed and Winsorized means are defined as

$$\bar{x}_{tg} = \frac{1}{n-2g}(x_{g+1} + x_{g+2} + \ldots + x_{n-g}),$$

$$\bar{x}_{wg} = \frac{1}{n}\{(g+1)x_{g+1} + x_{g+2} + \ldots + x_{n-g-1} + (g+1)x_{n-g}\}.$$

The g-times Winsorized sum of squared deviations is

$$\mathrm{SSD}_{wg} = (g+1)(x_{g+1}-\bar{x}_{wg})^2 + (x_{g+2}-\bar{x}_{wg})^2 + \ldots + (x_{n-g-1}-\bar{x}_{wg})^2 + (g+1)(x_{n-g}-\bar{x}_{wg})^2.$$

We also write $s_{wg}^2 = \mathrm{SSD}_{wg}/(h-1)$, where $h = n-2g$. The trimmed t statistic is given by

$$t = \frac{(\bar{x}_{1tg}-\bar{x}_{2tg})-(\mu_1-\mu_2)}{(s_{1wg}^2/h_1 + s_{2wg}^2/h_2)^{\frac{1}{2}}}.$$

Experience from previous work has led us to consider the possibility that this follows approximately a t distribution with degrees of freedom f calculated from

$$\frac{1}{f} = \frac{c^2}{h_1-1} + \frac{(1-c)^2}{h_2-1}, \quad c = \frac{s_{1wg}^2/h_1}{s_{1wg}^2/h_1 + s_{2wg}^2/h_2}. \tag{1}$$

Since the exact distribution is not easily obtainable analytically, a Monte Carlo technique was used to verify its approximate behaviour.

Normal $(0, \sigma^2)$ samples of size 10 and/or 20 were generated using Chen's technique (1971). The sets of ratio $R = \sigma_1/\sigma_2$ we considered were 0·25, 0·5, 2 and 4. For each pair of samples, the trimmed test statistic was calculated and a check was made to see if the hypothesis of equal means was accepted or rejected at $\alpha = 0{\cdot}01$, 0·05 and 0·10 using the degrees of freedom given in (1). For unequal sample sizes, the amount of trimming was always in fixed proportions. The experiment was repeated 10,000 times and the proportion of observations falling in the critical regions was recorded for different levels of trimming g, where $g = 0, 1, \ldots, [\frac{1}{4}n]$.

Results reveal that the observed sizes are only slightly larger than what was expected. The maximum deviation was 0·007 for all alpha levels we examined. In other words, for the sample sizes, significance levels and R considered, the empirical distribution can be represented reasonably well by that of Student's t with degrees of freedom given in (1). It should be added that for very small sample sizes such as 5, this approximation is not adequate; see Yuen & Dixon (1973).

4. Comparative performance of the trimmed t and Welch t statistics

The probabilities of type I error for the trimmed t, shown in Table 2, were computed under the same long-tailed distributions as presented for the Welch t test in Table 1. Comparative results in these two tables show that when the underlying distribution is long-tailed, the significance probability is reduced; the deviation is greater for the Welch t than for trimmed t. In other words, the trimmed t has a probability of rejecting the null hypothesis when it is actually true, closer to the nominal probability than that of the Welch t.

Comparisons were also made in terms of the power function of a two-sided test for the hypothesis $H_0: \mu_1 = \mu_2$. The underlying distributions studied were normal, and mixtures of normal and normal/uniform, i.e.

$$(1-p)N(\mu_i, \sigma_i^2) + p\frac{N(\mu_i, \sigma_i^2)}{U[0, 1]},$$

where $(1-p)$ is the probability that an observation comes from the Gaussian distribution. Five thousand samples were replicated; the same samples were used for both the trimmed t

and the Welch t tests. For convenience, a nominal size of 5 % was chosen and an example of the results is presented in Fig. 1 for $\delta = 0, 1, 2$ and 3, where

$$\delta = \frac{|\mu_1 - \mu_2|}{(\sigma_1^2/n_1 + \sigma_2^2/n_2)^{\frac{1}{2}}}.$$

Table 2. *Probability of type I error of the trimmed t*

Population 1	Population 2	n_1	n_2	g_1	g_2	Nominal α 0·01	0·05	0·10
$4 \times t_8$	t_8	20	10	2	1	0·008	0·051	0·106
				4	2	0·009	0·052	0·103
$2 \times t_5$	t_5	10	10	1	1	0·009	0·048	0·097
				2	2	0·010	0·047	0·099
$0·8N(0, 4^2)$	$0·8N(0, 1)$	10	20	1	2	0·010	0·048	0·098
$+0·2\dfrac{N(0, 4^2)}{U[0, 1]}$	$+0·2\dfrac{N(0, 1)}{U[0, 1]}$			2	4	0·014	0·057	0·109
$0·8N(0, 4^2)$	$0·8N(0, 1)$	20	10	2	1	0·009	0·047	0·088
$+0·2\dfrac{N(0, 4^2)}{U[0, 1]}$	$+0·2\dfrac{N(0, 1)}{U[0, 1]}$			4	2	0·011	0·050	0·098
$N(0, 4^2)/U[0, 1]$	$N(0, 1)/U[0, 1]$	20	10	2	1	0·003	0·033	0·074
				4	2	0·005	0·038	0·082
$N(0, 4^2)/U[0, 1]$	$N(0, 1)/U[0, 1]$	10	20	1	2	0·007	0·034	0·079
				2	4	0·010	0·045	0·086
$2 \times$ Cauchy (0, 1)	Cauchy (0, 1)	10	10	1	1	0·003	0·024	0·068
				2	2	0·004	0·026	0·069
Cauchy (0, 1)	Cauchy (0, 1)	20	20	1	1	0·002	0·022	0·065
				2	2	0·004	0·030	0·070
				3	3	0·004	0·035	0·078
				4	4	0·005	0·036	0·086
				5	5	0·004	0·038	0·084
$4 \times$ Cauchy (0, 1)	Cauchy (0, 1)	10	20	1	2	0·004	0·026	0·070
				2	4	0·006	0·036	0·078

As expected, the power of trimmed t never exceeds the corresponding power of Student's t under exact normality. For a small amount of trimming, the loss of power is very small. For long-tailed distributions, the trimmed t appears to be generally superior, the amount of power increase depending on the degree of long-tailedness, sample size and level of trimming. In the case of unequal sample size and variance, it is comparatively more powerful if the larger sample size goes with the larger variance than with the smaller variance.

5. Discussion

This paper extends the discussion of the two-sample trimmed t statistic to include cases where the population variances are unknown. This new statistic is reasonably robust with respect to long-tailedness and heteroscedasticity. The ordinary Student's t table can be used to find the critical values. If the underlying distribution is intrinsically long-tailed or contaminated with outliers, a situation which we feel occurs frequently, the trimmed t is strongly recommended. The premium, loss of power, one pays in using this test is not high when there is exact normality. If the investigator has some idea from previous experiments how his

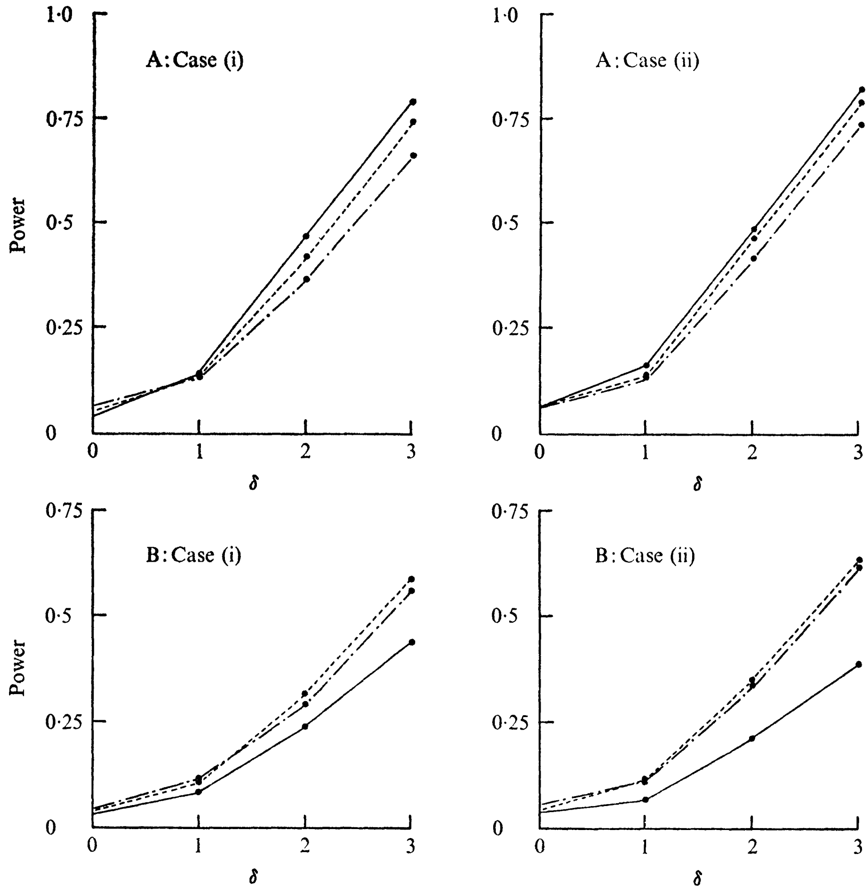

Fig. 1. Power of the Welch t and trimmed t.

A: Under normality. B: Under $0{\cdot}8N(0, \sigma^2) + 0{\cdot}2N(0, \sigma^2)/U[0, 1]$

Case (i): $n_1 = n_2 = 10$; $\sigma_1 = 2$, $\sigma_2 = 1$. Case (ii): $n_1 = 20$, $n_2 = 10$; $\sigma_1 = 4$, $\sigma_2 = 1$.

—— Welch t test; ---- trimmed t, $g = 1$; —·—·— trimmed t, $g = 2$.

measurements are distributed, he may be able to choose an optimum level of trimming, g. Otherwise, g should be chosen depending on the long-tailedness of the underlying distribution.

The author wishes to express her sincere thanks to Dr M. B. Brown for his valuable comments on the original manuscript. The computer work was done at the Health Sciences Computing Facility, UCLA, sponsored by a National Institutes of Health Special Research Resources Grant.

References

Behrens, W. V. (1929). Ein Beitrag zur Fehlerberechnung bei wenigen Beobachtungen. *Landwirtsch, Jahrbucher*. **68**, 807–37.

Chen, E. H. (1971). A random normal number generator for **32**-bit-word computers. *J. Am. Statist. Assoc.* **66**, 400–3.

Fisher, R. A. (1939). The comparison of samples with possibly unequal variances. *Ann. Eugen.* **9**, 174–80.

WANG, Y. Y. (1971). Probabilities of the type I errors of the Welch tests for the Behrens-Fisher problem. *J. Am. Statist. Assoc.* **66**, 605–8.
WELCH, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350–62.
WELCH, B. L. (1949). Appendix to A. A. Aspin's tables. *Biometrika* **36**, 293–6.
YUEN, K. K. & DIXON, W. J. (1973). The approximate behaviour and performance of the two-sample trimmed *t*. *Biometrika* **60**, 369–74.

[*Received July* 1973. *Revised August* 1973]