Jour. Ind. Soc. Ag. Statistics
Vol. XLI, No. 2 (1989), pp. 206-217

POWER COMPARISONS OF THE STUDENT T-TEST AND TWO APPROXIMATIONS WHEN VARIANCES AND SAMPLE SIZES ARE UNEQUAL

DONALD W. ZIMMERMAN and RICHARD H. WILLIAMS*
Carleton University, Ottawa, Canada

(Receivcd : March, 1988)

SUMMARY

A Monte Carlo study compared the Student t-test for independent groups to the Cochran-Cox and Welch-Satterthwaite 'approximations' of the t-test, using groups having unequal variances and unequal sample sizes. It was found that, when the null hypothesis was true, these two modified versions of the t-test maintained excellent control over the probability of type I errors (the α—level), even when departures from homogeneity of variance were extreme and sample sizes were unequal. It was also found that, when the null hypothesis was false, these modified tests were nearly as powerful as the usual Student t-test. Accordingly, substitution of one of these 'approximations' for the Student t-test under conditions where the latter is known to be inaccurate apparently maintains significance levels at their desired values without appreciable loss of power.

Keywords : Student t-test, Cochran-Cox test, Welch. Satterthwaite test, Type I and Type II Errors, Monte Carlo Method, Power Functions.

Researchers and applied statisticians have been concerned for a long time about the validity of statistical significance tests when assumptions underlying the tests are not satisfied. Standard textbooks frequently discuss at some length the assumptions of normality and homogeneity of

*University of Miami, Florida.

variance which are made in the F-test and the t-test (see, Winer [16]; Hays [7]; Kirk [10].

There is general agreement that the Student t-test, as well as the F-test, are *robust* under violation of the assumption of normality (see classic studies by Box [2]; Scheffè [13]; Boneau [1]. Computer sampling or Monte Carlo studies have shown that the probabilities of type I and type II errors are not appreciably modified for a variety of non-normal populations.

Violation of homogeneity of variance is slightly more complicated, although here also there seems to be general agreement. The Student t-test and the F-test are robust under violation of homogeneity of variance, provided sample sizes are equal. But if sample sizes are unequal, the probability of a type I error (the α—level) is rather severely affected by unequal variances (Hus [9]; Scheffè [13]). See also Boneau [1], Games and Howell [6], and Rogan and Keselman [11] for extensive and detailed discussions of findings in this area of research.

In theoretical statistics there has been another, somewhat independent, line of inquiry that is relevant to these issues. Early in this century, statisticians examined the sampling distribution of the Student t-statistic for independent samples under conditions where population variances are unequal and pooling of sample variances is not feasible. That is, in contrast to the usual test statistic employed in the Student t-test,

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s^2}{N_1} + \frac{s^2}{N_2}}}, \tag{1}$$

where s^2 is a pooled estimate of the population variance, which is distributed as Student's t-test with $N_1 + N_2 - 2$ degrees of freedom, investigators examined the distribution of

$$t' = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}, \tag{2}$$

where s_1^2 and s_2^2 are the (unpooled) sample estimates of the population variance. It was discovered that t' has the so-called Behrens-Fisher distribution (Fisher [4]; Fisher and Yates [5]) and not the Student t-distribution.

Later, Cochran and Cox [3] proposed a modification of the Student t-test for this case of unequal variances. They based their modified test on the t' value given by equation (2) and calculated the critical value of t' by the formula

$$t' \text{ (critical)} = \frac{t_1 \dfrac{s_1^2}{N_1} + t_2 \dfrac{s_2^2}{N_2}}{\dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}}, \qquad (3)$$

where t_1 and t_2 are the usual tabled critical values of the Student t-statistic for $N_1 - 1$ and $N_2 - 1$ degrees of freedom. Accordingly, the critical value of t' varies from one sample to another, depending on the sample variances and relative sample sizes.

Continuing along these lines, Welch [14], [15] and Satterthwaite [12] introduced another version of the t-test based on a different sort of approximation. These investigators employed the same t' statistic calculated according to equation (2), but interpreted this statistic with reference to a modified number of degrees of freedom given by

$$\text{df}' = \frac{[(s_1^2/N_1) + (s_2^2/N_2)]^2}{\dfrac{(s_1^2/N_1)^2}{N_1-1} + \dfrac{(s_2^2/N_2)^2}{N_2-1}}. \qquad (4)$$

Again, the critical value of t' calculated by this method varies from one sample to another.

These tests have come to be known as the Cochran-Cox and Welch-Satterthwaite approximations. The term 'approximation' is somewhat misleading in this context. The critical values of the test statistic, t', given by the above formulas are approximations of those given precisely by the Behrens-Fisher distribution. There is considerable evidence, however, that these modified versions are actually more accurate than the usual Student t under some conditions. Welch [14]. [15] regarded this technique as a generalization of the usual Student t-test.

These ideas are not commonly presented in introductory statistics texts in Education and Psychology. See, however, Winer [16], Kirk [10], and Howell [8] for informative discussions. Unfortunately, this entire theoretical development has been carried on more or less independently of the investigations of heterogeneity of variance using Monte Carlo techniques that were mentioned above.

It is notable that the classic Monte Carlo studies like that of Boneau [1] have examined violations of homogeneity of variance with unequal sample sizes by pooling variances and calculating the usual Student t-statistic by equation (1). The widely cited conclusions about unequal variances and unequal sample sizes are based on that method. It is not known how heterogeneous variances together with unequal sample sizes affect the Cochran-Cox and Welch-Satterthwaite versions of the t-test, when the critical value of t', calculated by equations (3) and (4), varies from one sample to another.

The purpose of the present paper is to compare the Student t-test, the Cochran-Cox test, and the Welch-Satterthwaite test with respect to both type I errors and type II errors when variances are unequal and sample sizes are unequal. That is, our intention is to investigate the robustness of these alternatives, or "approximations", under conditions where the usual Student t-test is *not* robust, using the same computer sampling technique to compare all three tests.

Monte Carlo Method for Comparison of t-Test and Approximations

A computer program* obtained two random samples of N_1 and N_2 scores, respectively, based on computer generation of random numbers. These samples were selected from prearranged normally distributed populations having mean 0 and variance 1.

The scores, denoted by x_1 and x_2, were then transformed by adding a constant and multiplying by a constant to produce desired differences in means and variances as needed in different parts of the study. First, a constant c was added to x_1, so that the means of the two groups differed by c. Next, x_2 was multiplied by another constant, k, so that the ratio of the standard deviations of the two groups was k (or the ratio of the variances was k^2). That is, the complete transformation was

$$X_1 = x_1 + c$$

$$X_2 = kx_2$$

This procedure resulted in independent groups of normally distributed scores of size N_1 and N_2, having a difference between means

$$\mu_1 - \mu_2 = c,$$

*A listing of the computer program, written in BASIC, can be obtained by writing to the author.

and a ratio of standard deviations,

$$\sigma_1/\sigma_2 = k.$$

For each pair of samples of N_1 and N_2 scores, the computer performed ths Student t-test for independent groups, the Cochran-Cox modification of the t-test, and the Welch-Satterthwaite modification of the t-test, all on the same scores. Throughout the present study the significance level was .05 and all tests were non-directional. In most cases there were 18 degrees of freedom—that is, sample sizes were $N_1 = 10$, $N_2 = 10$; or $N_1 = 15$, $N_2 = 5$; and so on.

The ratio of sample sizes N_1/N_2, as well as the ratio of standard deviations σ_1/σ_2, was varied systematically. The ratio N_1/N_2 had values of 1/3, 2/3, 1, 3/2, and 3 in different parts of the study, and the ratio σ_1/σ_2 assumed values ranging from 1 to 5 in increments of .5. Furthermore, the difference between means ranged from 0 to 4.5 times the standard error of the difference, given by

$$\text{S.E.} = \sqrt{(\sigma^2/N_1) + (\sigma_2^2/N_2)},$$

in increments of .5 S.E.

Sometimes the larger variance was associated with the larger sample size and sometimes with the smaller sample size. For each combination of parameters in the study, there were 3000 replications of the entire procedure of selecting N_1 and N_2 scores, together with 3000 calculations of each of the three test satistics. From the relative frequency with which a test statistic exceeded the critical value associated with the .05 significance level, the program obtained the probability of type I and type II errors.

The transformations discussed above revealed how these probabilities depend on the degree of variance heterogeneity (the ratio σ_1/σ_2), as well as relative sample sizes (the ratio N_1/N_2). By setting $\mu_1 - \mu_2 = 0$ (the null hypothesis true), the probability of a type I error was found, and by allowing $\mu_1 - \mu_2$ to have non-zero values (the null hypothesis false), the probability of type II errors and power functions were obtained.

In addition to the parameters mentioned above, sample sizes of $N_1 = 12$, $N_2 = 12$; $N_1 = 16$, $N_2 = 8$; and $N_1 = 20$, $N_2 = 4$ were investigated. It was found that the pattern of results was similar independently of the absolute sample sizes.

Probability of Type I Errors As A Function of Variances and Sample Sizes

Table 1 indicates how the ratio of variances and the ratio of sample sizes jointly determine the probability of type I errors for each of the three significance tests. Each entry in the table is the probability that a test statistic (Student t, Cochran-Cox t', or Welch-Satterthwaite t'') exceeds the critical value associated with the .05 significance level, for a non-directional test of the difference between means, when the true difference is zero. Each entry is based on 3000 pairs of random samples.

Any single column in the table exhibits this probability of a type I error as a function of the ratio σ_1/σ_2, which ranges from 1 to 5 in increments of .5. The columns correspond to the three test statistics (t, t', and t''), as shown in the column headings. The five sections of the table, from left to right, show the functions obtained when N_1 and N_2 have the values indicated. Accordingly, the ratio N_1/N_2 has values, from left to right, of 1/3, 2/3, 1, 3/2, and 3. Inspection of this table discloses that the probability of a type I error for the Student t-test is not modified appreciably by variation in the ratio σ_1/σ_2 when $N_1 = N_2$ (middle section of the table). In other words, the Student t-test is robust under violations of homogeneity of variance (at least over the range $\sigma_1/\sigma_2 = 1$ to $\sigma_1/\sigma_2 = 5$), when sample sizes are equal. However, the probability of a type I errror is changed drastically by variation in the ratio σ_1/σ_2 when N_1 and N_2 are unequal (see two left-hand sections and two right-hand sections). These are essentially the findings previously reported by Hsu [9], Scheffé [13], and others, that are described widely in standard textbooks.

The table shows the systematic influence of the ratio σ_1/σ_2 on the probability of type I errors—or, in other words, departure from the nominal significance level. When the larger variance is associated with the smaller sample size (two left-hand sections), the probability of a type I error increases gradually as the ratio σ_1/σ_2 increases. On the other hand, when the larger variance is associated with the larger sample size (two right-hand sections), the proability of a type I error decreases as the ratio σ_1/σ_2 increases. Furthermore, the larger the discrepancy in sample sizes, the more severe is the effect of unequal variances.

Next, it is apparent from the table that the probability of a type I error for the Welch-Satterthwaite version of the t-test (the columns labelled t'') remains extremely close to the significance level, .05, over the entire range of values of σ_1/σ_2, whatever the value of N_1/N_2. Similarly, the probability of a type I error for the Cochran-Cox version of the t-test is relatively stable over the range of values of σ_1/σ_2, although this prob-

TABLE 1†—PROBABILITY OF TYPE I ERRORS

		$N_1 = 5$ $N_2 = 15$			$N_1 = 8$ $N_3 = 12$			$N_1 = 10$ $N_2 = 10$			$N_1 = 12$ $N_2 = 8$			$N_1 = 15$ $N_2 = 5$		
		t	t'	t''	t	t'	t''	t	t'	t''	t	t'	t''	t	t'	t''
	1.0	.046	.038	.051	.051	.038	.051	.051	.037	.049	.050	.040	.047	.052	.034	.045
	.5	.106	.044	.060	.075	.042	.057	.052	.040	.050	.035	.036	.049	.022	.035	.046
	2.0	.160	.049	.059	.091	.046	.057	.055	.045	.051	.034	.037	.049	.011	.032	.044
	.5	.208	.049	.056	.105	.048	.057	.057	.045	.050	.031	.040	.046	.007	.032	.047
σ_1/σ	3.0	.232	.050	.056	.112	.050	.058	.057	.046	.050	.031	.041	.047	.005	.032	.049
	3.5	.255	.050	.056	.118	.051	.057	.060	.046	.051	.031	.043	.049	.005	.035	.049
	4.0	.271	.051	.056	.122	.052	.055	.062	.047	.052	.030	.044	.048	.005	.038	.050
	4.5	.282	.051	.054	.128	.052	.055	.062	.047	.051	.029	.044	.049	.005	.039	.049
	5.0	.291	.051	.054	.131	.053	.055	.062	.048	.051	.029	.045	.049	.005	.041	.050

†*Note* : t—Student t; t'—Cochran-Cox approximation; t''—Welch-Satterthwaite approximation.

ability appears to be slightly less than .05 when sample sizes are unequal and the larger variance is associated with the larger sample size.

Inspection of Table 1 conveys a strong impression that these modified versions of the t-test, especially the Welch-Satterthwaite version, accomplish their intended goal very well, at least as far as maintaining control over α- levels is concerned. Both approximations apparently control these significance levels much better than the classical Student t–test when variances are heterogeneous and sample sizes are unequal.

Probability of Type II Errors and Power Functions

In addition to information about the stability of α-levels and type I errors, one would like to have information about type II errors and the power of these versions of the t-test to detect non-zero differences between means. Table 2 exhibits the probability that a test statistic (t, t' or t'') exceeds the critical value associated with the .05 significance level as a function of the effect size, $\mu_1 - \mu_2$, for each significance test. These probabilities are the same as 1 minus the probability of a type II error, or the power of the test. Accordingly, the columns in Table 2 are essentially power functions for each of the three versions of the t-test. Differences between means are expressed in units of the standard error of a difference.

The four sections of the table display power functions for selected combinations of parameters. These include two cases in which the Student t-test turned out to be accurate as far as type I error are concerned (two left-hand sections) and two cases in which the Cochran-Cox and Welch-Satterthwaite tests are far more accurate thn the Student t-test (two right-hand sections).

It is evident from Table 2 that the power of the Cochran-Cox and Welch-Satterthwaite versions compares quite favorably with that of the Student t-test. When $N_1 = N_2 = 10$ and $\sigma_1 = \sigma_2$ (first section), the Student t-test is slightly more powerful than the Cochran-Cox test over the entire range of differences between means. The Welch-Satterthwaite test is very nearly as powerful as the Student t-test, perhaps slightly less over part of the range. These differences are quite small and probably negligible in research practice. It seems clear that there is no substantial loss of power in substituting one of the "approximations" for the Student t-test.

A similar pattern of results holds for the case $N_1 = N_2 = 10$ and $\sigma_1/\sigma_2 = 3$ (second section of Table 2). The power of the Welch-Satterthwaite test is intermediate between that of the Student t-test and that

TABLE2†—POWER FUNCTIONS

	$\sigma_1 = \sigma_2$ $N_1 = 10$ $N_2 = 10$			$\sigma_1/\sigma_2 = 3$ $N_1 = 10$ $N_2 = 10$			$\sigma_1/\sigma_2 = 3$ $N_1 = 5$ $N_2 = 15$			$\sigma_1/\sigma_2 = 3$ $N_1 = 15$ $N_2 = 5$		
$\mu_1 - \mu_2$	t	t'	t''	t	t'	t''	t	t'	t''	t	t'	t''
0.0	.049	.034	.048	.065	.049	.056	.223	.048	.057	.007	.037	.047
0.5	.077	.055	.074	.086	.069	.078	.287	.070	.076	.014	.060	.076
1.0	.165	.132	.159	.165	.141	.149	.434	.125	.136	.036	.118	.151
1.5	.305	.253	.296	.300	.255	.272	.588	.227	.236	.074	.240	.280
2.0	.480	.421	.470	.477	.417	.439	.753	.344	.355	.148	.387	.450
2,5	.662	.602	.653	.660	,598	.621	.881	.486	.493	.267	.576	.638
3.0	.813	.772	.807	.804	.763	.779	.947	.628	.629	.417	.750	.794
3.5	.907	,882	.903	.913	.886	.896	.978	.751	.751	.597	.876	.907
4.0	.969	.955	.966	.964	.950	.955	.994	.847	.846	.756	.949	.964
4.5	.991	.985	.989	.988	.981	.983	.999	.914	.913	.872	.984	.991

†*Note* : t—Student t; t'—Cochran-Cox approximation; t''—Welch-Satterthwaite approximation.

of the Cochran-Cox test over the entire range, although again differences are slight.

Furthermore, for cases where unequal variances and unequal sample sizes are combined (third and fourth sections), the power functions for the Cochran-Cox and Welch-Satterthwaite tests are similar to the cases just described. But the functions for the Student *t*-test now are anomalous, because the α-levels are changed drastically.

Although a cursory glance at Table 2 might lead one to conclude that the usual Student *t*-test is more powerful than the Welch-Satterthwaite and Cochran-Cox tests for $N_1 = 5$ and $N_2 = 15$, this is not ture, because the probability of a type I error for the Student *t*-test is wildly out of control, being .223 rather than .05. An increase in the probability of rejection of a false null hypothesis can be meaningfully identified with an increase in power of a test only if the α- level remains constant. Otherwise expressed, the only situation in which the power of two tests can be compared is when both are valid.

3. Conclusions

Considerable evidence from many investigations over the past several decades indicates that the Student *t*-test is not robust under violation of the assumption of homogeneity of variance when sample sizes are unequal. The results of the Monte Carlo method employed in the present study conform to this well-known finding. More specifically, it was found that the probability of a type I error for the Student *t*-test is an increasing function of the ratio σ_1/σ_2 when sample sizes are unequal and the larger variance is associated with the smaller sample size, and it is a decreasing function of the ratio σ_1/σ_2 when the larger variance is associated with the larger sample size.

If sample sizes are equal, there is apparently little dependence on the ratio σ_1/σ_2, Otherwise expressed, the probability of a type I error is close to the nominal significance level, α, when *either* variances are equal *or* sample sizes are equal. Only the simultaneous failure of both of these conditions modifies the significance level to any considerable degree.

Next, the present study disclosed that the so-called Cochran-Cox and Welch-Satterthwaite "approximations" of the *t*-test maintained excellent control over the probability of type I error, even under extreme violations of homogeneity of variance and even when sample sizes were unequal. In the case of the Welch-Satterthwaite test, especially, this probability remained quite close to the nominal significance level, .05, when the ratio σ_1/σ_2 was as large as 5 and the ratio N_1/N_2 was as large as 3.

Finally, it was found that the power to reject a false null hypothesis of the Cochran-Cox and Welch-Satterthwaite "approximations" was nearly as great as the Student t-test. This was true under those conditions where the Student t-test was accurate with respect to type I errors, as well as conditions where the Student t-test was decidedly inaccurate. In some cases there were slight differences in power in favor of the Student t-test. However, it seems reasonable to conclude from the results in Table 1 and Table 2 taken together that the slight differences in power of the respective tests is not too important, considering the vastly greater superiority of the Welch-Satterthwaite and Cochran-Cox tests in maintaining control of the significance level when variances and sample sizes are unequal. In any event, there is certainly no substantial loss of power in substituting one of these so-called "approximations" for the Student t-test.

Because of concern about unequal variances and sample sizes, many textbooks advise researchers to maintain equal N's whenever feasible. For example, Hays (1981) remarked : ". . . when the variances are quite unequal the use of different sample sizes can have serious effects on the conclusions. The moral should be plain : given the usual freedom about sample size in experimental work, *when in doubt use samples of the same size.*" Many other authors have made similar recommendations.

However, there are research areas in Psychology and Education where variances of different groups of subjects are unequal, but sample size is fixed and not under the investigator's control. Also, even in experimental studies initiated with equal sample sizes in the cells, attrition can lead to unequal sample sizes. The present findings suggest that, in these circumstances, the Welch-Satterthwaite or Cochran-Cox approximations of the t-test can be performed without substantial modification of the probability of either type I or type II errors for whatever sample sizes happen to be available for analysis.

REFERENCES

[1] Boneau, C. A. (1960) : The effects of violations of assumptions underlying the t test. *Psychological Bulletin* 57 : 49-64.

[2] Box, G. E. P. (1953) : Non-normality and tests on variance. *Biometrika* 40 : 318-335.

[3] Cochran, W. G., and Cox, G. M. (1957) : *Experimental Designs* (2nd ed.). Wiley, New York.

[4] Fisher, R. A. (1935). *The Design of Experiments.* Oliver Boyd, Edinburgh.

[5] Fisher, R. A., and Yates, F. (1953) : *Statistical Tables for Biological, Agricultural, and Medical Research* (4th ed.). Oliver & Boyd, Edinburgh.

[6] Games, P. A. and Howell, J. F. (1976) : Pairwise multiple comparison procedures with unequal n's and/or variances : A Monte Carlo study. *Journal of Educational Statistics* **1** : 113-125.

[7] Hays, W. L. (1981) : *Statistics* (3rd ed.). Holt, Rinehart, & Winston, New York.

[8] Howell, D. C. (1987) : *Statistical Methods for Psychology* (2nd ed.). Duxbury Press, Boston.

[9] Hsu, P. L. (1938) : Contributions to the theory of "student's" t-test as applied to the problem of two samples. *Statistical Research Memoirs* 2 : 1-24.

[10] Kirk, R. E. (1982) : *Experimental Design* (2nd ed.). Brooks-Cole, Monterey, Calif.

[11] Rogan, J. C. and Keselman, H. J. (1977) : Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? : An investigation via a coefficient of variation. *American Educational Research Journal* **14** : 493-498.

[12] Satterthwaite, F. E. (1946) : An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2** : 110-114.

[13] Scheffè, H. A. (1959) : *The Analysis of Variance.* Wiley, New York.

[14] Welch, B. L. (1938) : The significance of the difference between two means when the population variances are unequal. *Biometrika* 29 : 350-362.

[15] Welch, B. L. (1947) : The generalization of Student's problem when several different population variances are involved. *Biometrika* **34** : 29-35.

[16] Winer, B. J. (1971) : *Statistical Principles in Experimental Design* (2nd ed.). McGraw-Hill, New York.