Comparative Power of Student *T* Test and Mann-Whitney *U* Test for Unequal Sample Sizes and Variances

DONALD W. ZIMMERMAN
Carleton University

ABSTRACT

A computer program generated power functions of the Student *t* test and Mann-Whitney *U* test under violation of the parametric assumption of homogeneity of variance for equal and unequal sample sizes. In addition to depression and elevation of nominal significance levels of the *t* test observed by Hsu and by Scheffé, the entire power functions of both the *t* test and the *U* test were depressed or elevated. When the smaller sample was associated with a smaller variance, the *U* test was more powerful in detecting differences over the entire range of possible differences between population means. When sample sizes were equal, or when the smaller sample had the larger variance, the *t* test was more powerful over this entire range. These results show that replacement of the *t* test by a nonparametric alternative under violation of homogeneity of variance does not necessarily maximize correct decisions.

THERE IS CONSIDERABLE EVIDENCE that the *t* test and *F* test are robust under violation of parametric assumptions. These statistical tests are scarcely affected by non-normality of population distributions (Bartlett, 1935; Boneau, 1960; Hsu & Feldt, 1969; Pearson, 1931; see also Glass, Peckham, & Sanders, 1972; Govindarajulu & Leslie, 1971; Huber, 1972). Furthermore, it is generally agreed that the tests are robust under violation of the assumption of homogeneity of variance, provided sample sizes are equal (Box, 1953; Hsu, 1938; Rogan & Keselman, 1977; Scheffé, 1959; see also Glass, Peckham, & Sanders, 1972).

If sample sizes differ, then inequality of variances can have a pronounced effect on significance levels and on the probability of Type I errors. For example, nominal significance levels of the *t* test are lowered if the smaller sample is less variable than the larger and are raised if the smaller sample is more variable than the larger (Hsu, 1938; Scheffé, 1959). Based on these findings, many modern statistics texts deemphasize normality but advise researchers to be cautious about homogeneity of variance when sample sizes are unequal.

The present paper examines this concept of robust statistics from another point of view. When assumptions of parametric statistical tests such as *t* and *F* are not satisfied, researchers frequently turn to nonparametric alternatives to analyze their data (Gibbons, 1985; Siegel, 1956). It has been assumed implicitly that this procedure increases the likelihood of correct decisions. Unfortunately, investigations of robust statistics have not examined the relative merits of parametric and nonparametric methods under violation of parametric assumptions.

In the present study, a computer program obtained repeated random samples from known populations, some of which violated the homogeneity assumption and performed both Student *t* tests for independent groups and Mann-Whitney *U* tests on the sample values. First, the results reported by Hsu (1938) and by Scheffé (1959), indicating modification of the significance levels of the *t* test with unequal sample sizes and variances, were replicated. The probabilities of

Type I errors found in the present Monte Carlo study were close to those found by these investigators.

Next, both Type I and Type II errors of the *t* test and *U* test, applied to the same repeated random samples, were examined. The entire power functions of the two statistical tests were plotted from computer-generated probabilities and compared. (See Winer, 1971, and Hoel, 1971, for discussions of power functions of statistical tests.) This approach, we shall see, throws a somewhat different light on the findings mentioned above and on the concept of robust statistical tests.

Method

Computer Sampling Procedure

A computer program[1] repeatedly selected pairs of random samples from pairs of normally distributed populations and performed two-tailed Student *t* tests and Mann-Whitney *U* tests on the samples. The program calculated the probability that each test statistic exceeded the critical value associated with the .05 significance level.

In some instances the populations were equally variable ($\sigma_1 = \sigma_2$), and in other instances the population standard deviations differed by a factor of 5 ($\sigma_1 = 5\sigma_2$). In part of the study, sample sizes were equal ($N_1 = N_2 = 10$); in another part of the study, the smaller sample was less variable than the larger ($N_1 = 16$, $N_2 = 4$); and, in the third part of the study, the smaller sample was more variable than the larger ($N_1 = 4$, $N_2 = 16$).

Accordingly, each relation between variances (equality or inequality) was combined with each of the three relations between sample sizes. For each of these six combinations of parameters, the difference between population means varied in increments of one-half the standard error of the difference, from 0 to 4.5 times the standard error, to obtain power functions. Finally, for each combination of parameters and for each of the 10 degrees of difference between means, there were 2,000 replications of the sampling procedure.

Results and Discussion

Estimated Probabilities of Type I Errors

Table 1 shows the probability that each test statistic (*t* and *U*) exceeds the critical value associated with the .05 significance level when random samples are repeatedly drawn from populations with equal means. Each entry in the table in the columns labeled 1, 2, and 3 is based on 2,000 pairs of random samples. These values give an indication of the stability of the probability estimates obtained from repeated samples. The entry in the column labeled *M* is the mean of the three estimates in columns 1, 2, and 3. For these sample sizes the probability estimates apparently are quite stable for 2,000 replications.

It is evident that inequality of variances has only a slight effect on the probability of a Type I error (expected to be .05) when sample sizes are equal but a large effect when sample sizes are unequal. The computer-generated probabilities for *t* are lower than the nominal significance level, .05, when the smaller sample is less variable and higher than .05 when the smaller sample is more variable. These results are consistent with the findings reported by Hsu (1938) and Scheffé (1959). Also, it is apparent that the probabilities for *U* are also modified in the same way as those for *t*, to a lesser degree.

Type II Errors and Power Functions

Figure 1 plots the probability that each test statistic exceeds the critical value associated with the .05 significance level as a function of the difference between means, expressed in units of one-half the standard error of a difference. That is, the figure displays power functions of the two statistical tests. The minimum value of each function, for zero difference, corresponds to the probabilities given in Table 1.

These functions show how the probability of rejecting the null hypothesis when it is false, or 1 minus the probability of a Type II error, depends on the difference between means. The two power functions in each section of the figure compare the power of the *t* test and *U* test over the entire range of differences between means.

When variances are equal and sample sizes are equal (upper left-hand section of Figure 1), the power function of *t* slightly dominates the one of *U*, as expected. In this case, when parametric assumptions are satisfied, the assymptotic relative efficiency of the *U* test is known to be .955. When variances are unequal and sample sizes are equal (lower left-hand section of Figure 1), the difference in favor of the *t* test actually is slightly larger. If variances are equal and sample sizes are unequal (upper middle and upper right-hand sections of Figure 1), *t* still slightly dominates *U*.

When sample sizes are unequal and the smaller sample has the smaller variance (lower middle section of Figure 1), the *U* test is more powerful than the *t* test over the entire range of differences between means. However, when sample sizes are unequal and the smaller sample has the larger variance (lower right-hand section of Figure 1), the *t* test is more powerful than the *U* test over the entire range of differences between means.

Robust Parametric Tests and Nonparametric Alternatives

From the results displayed in Figure 1 and in Table 1 it is evident that the depression of nominal significance levels of *t* noted by previous investigators is associated with an overall depression of the entire power function of the *t* test. Furthermore, there is a similar although less marked depression of the power function of the *U*

TABLE 1—Probability of Type I Errors, *t* Test, and *U* Test.

	U test				*t* test			
	1	2	3	*M*	1	2	3	*M*
Equal variances								
$N_1 = N_2 = 10$	.045	.038	.040	.041	.049	.046	.045	.047
$N_1 = 16, N_2 = 4$	.049	.049	.049	.049	.049	.048	.045	.048
$N_1 = 4, N_2 = 16$	.053	.045	.046	.048	.056	.042	.049	.049
Unequal variances								
$\sigma_1 > \sigma_2$								
$N_1 = N_2 = 10$	.074	.078	.075	.075	.067	.066	.061	.065
$N_1 = 16, N_2 = 4$	.005	.008	.005	.006	0	.001	.001	.001
$N_1 = 4, N_2 = 16$	.142	.135	.127	.134	.355	.351	.327	.344

NOTE: Entries are computer-generated probabilities that test statistics (*t* and *U*) exceed critical values associated with .05 significance level.

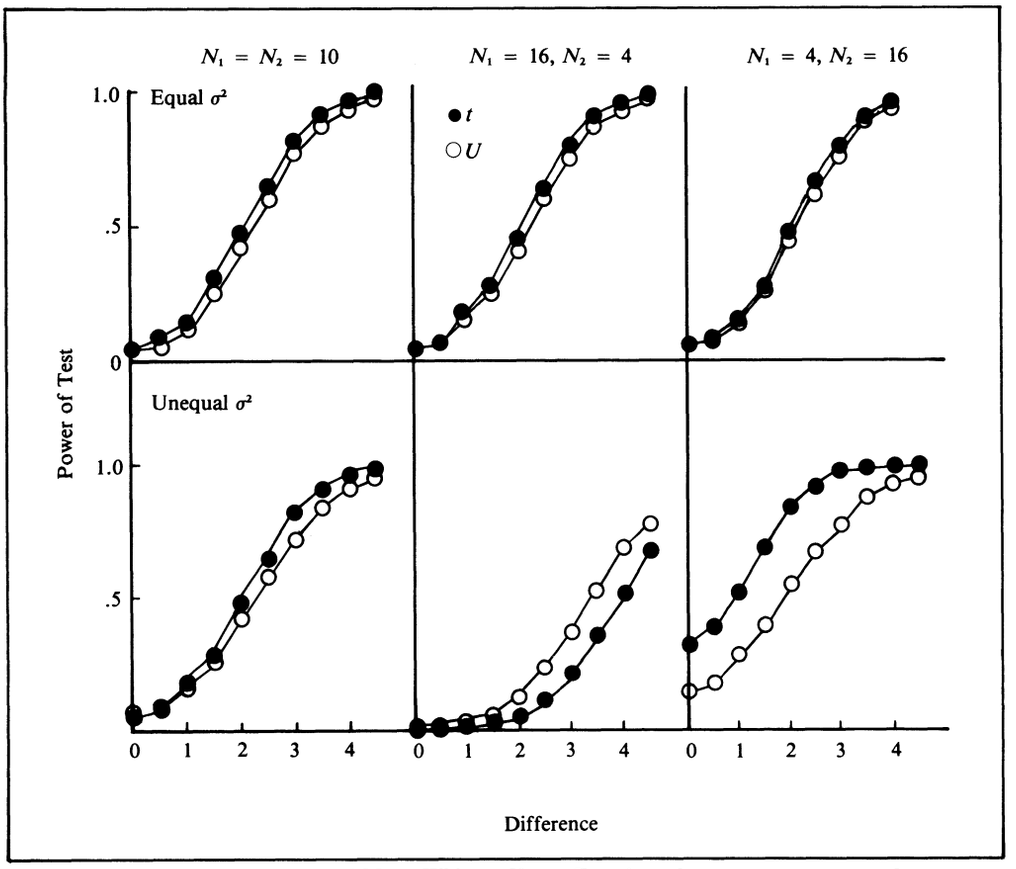

Figure 1—Power of Student *t* test and Mann-Whitney *U* test. Functions show computer-generated probabilities that test statistics (*t* and *U*) exceed critical values associated with .05 significance level. Differences between means are expressed in units of one-half a standard error of the difference.

test. As a consequence, the *U* test turns out to be more powerful than the *t* test when sample sizes are unequal and the smaller sample has the smaller variance.

Similarly, it is evident that the elevation of the nominal significance levels of *t* noted previously is associated with elevation of the entire power function of the *t* test. A similar effect of lesser magnitude again occurs for *U*, so that superiority of the *t* test becomes even more pronounced when sample sizes are unequal and the smaller sample has the larger variance.

From the standpoint of a researcher trying to decide whether to use the *t* test or a nonparametric alternative, the implications of the present study are somewhat anomalous. Investigators have generally assumed that

when a parametric significance test such as t or F is inappropriate due to violation of distributional assumptions, correct decisions are maximized by employing nonparametric methods. The present study shows that there are circumstances where the t test leads to incorrect decisions because of violation of assumptions but where employing the Mann-Whitney U test in its place leads to even poorer results.

First, when sample sizes are equal, it makes no sense to substitute the U test for the t test if homogeneity of variance is violated. In that situation, t retains its superiority and in fact the differences in power between t and U become somewhat larger. When sample sizes are unequal, the α levels for both t and U are modified if homogeneity of variance is violated, although the change for t is greater than the change for U. Also, the power of both tests is modified.

If avoidance of Type II errors is a major consideration, the preferred course of action depends on relations between sizes of variances and sizes of samples. If the smaller sample has the smaller variance, clearly the U test is the best choice (see Figure 1). But in all other cases examined in the present study, the t test was more powerful than the U test over the entire range of differences between means.

As a practical matter, perhaps the best rule to follow is to design experiments so that sample sizes are equal unless there are compelling reasons for doing otherwise. As long as sample sizes are equal, violation of assumptions of normality and homogeneity of variance does not alter the superiority of the t test. At least present evidence strongly suggests that this is true for distributions that are likely to be encountered in most research in psychology and the social sciences.

NOTE

1. A listing of the computer program, written in BASIC, can be obtained by writing to the author at the following address: Dr. Donald W. Zimmerman, Psychology Department, Carleton University, Ottawa, Ontario, Canada K1S 5B6.

REFERENCES

Bartlett, M. S. (1935). The effect of non-normality on the t test. *Proceedings of the Cambridge Philosophical Society, 31*, 223.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t test. *Psychological Bulletin, 57*, 49–64.

Box, G. E. P. (1953). Nonnormality and tests on variance. *Biometrika, 40*, 318–335.

Gibbons, J. D. (1985). *Nonparametric methods for quantitative analysis* (2nd ed.). Columbus, OH: American Sciences Press.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *Review of Educational Research, 42*, 237–288.

Govindarajulu, Z., & Leslie, R. T. (1971). *Annotated bibliography on robustness studies of statistical procedures.* U.S. Department of Health, Education, and Welfare, Ser. 2, No. 51.

Hoel, P. G. (1971). *Introduction to mathematical statistics* (4th ed.). New York: Wiley.

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.

Hsu, T. C., & Feldt, L. S. (1969). The effect of limitations on the number of criterion score values on the significance of the F test. *American Educational Research Journal, 6*, 515–527.

Huber, P. J. (1972). Robust statistics: A review. *Annals of Mathematical Statistics, 43*, 1041–1067.

Pearson, E. S. (1931). The analysis of variance in cases of non-normal variation. *Biometrika, 23*, 114–133.

Rogan, J. C., & 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.

Scheffé, H. (1959). *The analysis of variance.* New York: Wiley.

Siegel, S. (1956). *Nonparametric statistics.* New York: McGraw-Hill.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.