A Re-examination of the Robustness of the two Independent Means T-test to Population Non-Normality

R. Clifford Blair and Steven W. Phillippy

University of South Florida

The consequences of violating the t-test's assumptions of population normality and homogeniety of variance have been widely investigated and reported in the literature (Pearson, 1931; Bartlett, 1935; Welch, 1937; Daniels, 1938; Quensel, 1947; Gayen, 1950a, 1950b; David and Johnson, 1951; Norton, 1952; Horsnell, 1953; Box 1954a, 1954b; Box and Andersen, 1955; Scheffe, 1959; Srivastava, 1959; Boneau, 1960, 1962; Baker, Hardyck and Petrinovich, 1966; Games and Lucas, 1966; Neave and Granger, 1968). Although most studies of this type have focused primarily on type I errors, some have examined effects upon power as well. It has generally been claimed that the results of these studies demonstrate that the t-test is quite robust (Box, 1953) to deviations of populations from normality. This is said to be especially true when samples are drawn from populations having the same non-normal shapes, and in the case of equal (or approximately equal) sample sizes, for the same to be true when variances are heterogeneous.

Educational and psychological researchers have come to regard these conclusions concerning the robustness of the t-test as tenets of faith. This is especially true of the normality assumption as can be seen from this statement by the authors of one of the most popular statistical tests used by educational and psychological research workers. "Violation of the assumption of normality in the t-test of $H_0: \mu_1 - \mu_2 = 0$ has been shown to have only trivial [italics added] effects on the level of significance and power of the test and hence should be no cause for concern." (Glass and Stanley, 1970, p. 297) Recently, however, this position has been challenged by Bradley (1977) who asserts that "...the strength of the evidence for robustness [of the t-test] appears to derive partly from selectivity in investigating only the more familiar population shapes..." (p. 150). Bradley (1977) further implies that many populations of interest in the social sciences deviate from normality to a far greater degree than do the familiar functions that have been investigated in the past. If this be the case, then the conclusion reached by Glass and Stanley (and many others) is at least premature.

Large sample population estimates in the form of data collected in research contexts suggests the "mixed normal" distribution as a viable population model for many social science phenomena (Allport, 1934; Bradley, 1968, 1976, 1977). This distribution appears to be fairly common in certain research areas and arises where, for perfectly valid reasons, some discrete causal variable is left uncontrolled (Bradley, 1977). Because of its rather bizarre shape (and based on evidence presented by Bradley, 1968, 1976), this distribution should present a more stringent test of the t-test's robustness than have other commonly investigated distributions.

The purpose of this study is to reinvestigate the robustness of the two independent means t-test to departures from normality. By using mixed normal distributions as the sampled populations, it is hoped that evidence will be generated that will either (a) support Glass and Stanley's contention that non-normal population shape alone can have only trivial effects on the t statistic, or (b) support Bradley's position that previous studies have been too selective in their choice of populations and therefore have not provided a stringent enough examination of the t-test's robustness. At the same time, we will examine some of the conclusions reached by Boneau (1960) who studied this subject and whose research is widely cited as evidence of the t-test's robustness.

Methodology

The general method of investigation was computer simulation which was carried out as follows. Let A_i be a number selected at random from a (pseudo) normal p.d.f. with mean zero and standard deviation one. Let B_i be a number selected at random from a (pseudo) uniform p.d.f. with end points at 0 and 1 inclusive. C represents a number between 0 and 1 while m and n are a pair of specified integers. X_i, the random variable of interest, is defined as follows: if $B_i < C$ then $X_i = A_i$. If $B_i \geq C$ then $X_i = nA_i + m$. This means that

$p(X_i \sim N\ [0,1]) = C$ and $p(X_i \sim N\ [m,n^2]) = 1\text{-}C$. Hence the term "mixed normal" distribution. In this study, m and n took the values 22 and 100 respectively while C was taken to be .95.

The sample sizes (n_1, n_2) investigated were 3,9; 6,6; 9,27; 18,18; 15,45; 30,30; 27,81 and 54,54. For each set of sample sizes (n_1, n_2) 5,000 independent pairs of samples were drawn from the population, and t values were computed for each sample pair. Type I error rates were assessed by computing the proportion of t values that fell outside the appropriate critical value.

Power functions were determined by the methods outlined above except that constants were added to the scores of the designated "treatment" group. This constant represents the value $\mu_1 - \mu_2$. In order to make comparisons between the power functions generated in this study and those calculated under normal theory, the values of $\mu_1 - \mu_2$ were chosen by substituting values for the ES term in the equation $\mu 1 \quad \mu 2 = \sigma$ (ES) where σ is the common standard deviation of the two sampled populations and ES was the effect size (see Cohen, 1977, for a discussion of this term). Values for ES were chosen from tables provided by Cohen(1977).

Results

Table 1 gives the one-tailed type I error rates obtained from the procedures outlined above. Column headings in this

Table 1

One-Tailed Type I Error Rates for a Mixed Normal Distribution for which $p(X_i \sim N[0,1]) = .95$ and $p(X_i \sim N[22,100])=.05$

		Nominal α Level			
n_1, n_2	Tail	.05	.025	.01	.005
3,9	upper	.085	.023	.009	.005
	lower	.028	.016	.005	.003
6,6	either	.028	.014	.005	.002
9,27	upper	.084	.039	.020	.008
	lower	.009	.003	.001	.000
18,18	either	.028	.008	.002	.001
15,45	upper	.080	.044	.022	.011
	lower	.002	.001	.000	.000
30,30	either	.040	.011	.002	.000
27,81	upper	.064	.035	.015	.008
	lower	.009	.000	.000	.000
54,54	either	.050	.021	.004	.001

table show (1) sample sizes employed, (2) whether the test was conducted in the upper or lower tail of the distribution and (3) nominal significance levels.

There are several interesting points to be made concerning the results in this table: (1) The deviations of type I error rates from nominal significance levels are generally greater than those found in studies that investigated more familiar distributions (for example, compare Table 1 with Boneau's (1960) Table 1, remembering that we are dealing only with the nonnormality of a single shape, and not with heterogeneity of either shape or variance). It is difficult to decide whether or not to characterize the more pronounced of these deviations as "large" since to some extent that determination must rest in the eye of the beholder, but it seems safe to say that most researchers would not characterize them as "trivial." (2) Two-tailed type I error rates for nominal significance levels that are twice those shown in Table 1 can be obtained by summing the upper and lower tail values given for a particular sample size combination under a particular nominal significance level. For example, the two-tailed type 1 error rate for samples of sizes 3 and 9 at $\alpha = .05$ is obtained as $.023 + .016 = .039$. When this is done throughout Table 1, it is noted that unequal sample sizes often yield better two-tailed results than one-tailed results, while the opposite is often true of equal sample sizes. (3) Since samples were drawn from a single population, Boneau's (1960) conclusion that the t-test will yield probability statements that "are accurate to a high degree," provided that samples are drawn from populations having approximately the same shape, is highly questionable. (4) Since samples were drawn from a single population, Boneau's (1960) statement that "If the sample sizes are unequal, one is in no difficulty provided the variances are compensatingly equal" (p. 62) seems blatently incorrect. (5) Many researchers would find questionable Boneau's conclusion that samples of sizes thirty are sufficiently large to undo even the most extreme violations of the underlying assumptions.

Table 2 shows selected power function results obtained from the same distribution examined in Table 1. In this table, power is shown in the body of the table as the percentage of t statistics falling in appropriate critical regions. Power is appraised under the condition $\mu_1 > \mu_2$ and $\mu_1 < \mu_2$, and these results are compared with power calculated for the same effect size under normal theory. (Normal theory calculations are taken from Cohen (1977, pp. 36-37).) Several points concerning these results should be noted: (1) For local alternatives, the t-test may show considerably more power under this distribution than it does under the normal p.d.f. This result does not conflict with statistical theory as some might believe since even though the t-test is the uniformly most powerful (UMP) unbiased test under

Table 2

Selected One-Tailed Power Functions for a Mixed Normal Distribution for which $p(X_i \sim N[0,1]) = .95$ and $p(X_i \sim N[22,100]) = .05$, $\alpha = .025$

n_1, n_2	Source	Alternative Hypothesis		Effect Size = $\frac{\mu_1 \quad \mu_2}{\sigma}$										
				.10	.20	.30	.40	.50	.60	.70	.80	1.00	1.20	1.40
8,18	normal	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	.06	.09	.14	.21	.31	.41	.53	.64	.83	.94	.98
	mixed normal	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	.08	.25	.36	.44	.51	.56	.62	.70	.80	.89	.94
30,30	normal	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	.07	.12	.21	.33	.47	.63	.76	.86	.97	*	*
	mixed normal	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	.10	.24	.34	.45	.57	.67	.76	.84	.93	.97	.99
27,81	normal	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	.07	.14	.26	.42	.60	.75	.87	.94	.99	*	
	mixed normal	$\mu_1 > \mu_2$		.09	.19	.32	.47	.62	.76	.86	.93	.98	.99	*
	mixed normal	$\mu_1 < \mu_2$		.05	.22	.36	.52	.67	.78	.86	.93	.98	.99	
54,54	normal	$\mu_1 > \mu_2$	$\mu_1 < \mu_2$	.08	.18	.34	.53	.73	.87	.95	.98	*		
	mixed normal			.11	.24	.41	.59	.75	.86	.93	.97	.99	*	

* Indicates that values exceed .995.

normal theory, this does not indicate that it cannot show even greater power under some other distribution. (2) At higher points along the power function, the t-test may show slightly less power under this distribution than would be obtained under normal theory. (3) In the unequal sample size case, the t-test may show greater power when the one-tailed test is in one direction than it does for a test in the other direction.

Conclusions and Discussion

Because the population considered in this study probably models reality quite well in some research contexts, the results tend to support Bradley's contention that we have erred in examining only the more familiar population shapes. Authors and researchers who, like Glass and Stanley (1970), have dismissed population shape as a point of concern in relation to the t-test have probably done so prematurely. Likewise, some of Boneau's (1960) recommendations which were based on his investigations of more moderately nonnormal populations seem questionable or in some cases incorrect.

The reader might object at this point that with the exception of some moderate type I error rate inflations and some small power losses, the t-test behaved quite well in this study. After all, the larger type I error rate deviations were in the conservative direction and in spite of this there were substantial gains in power. But the purpose of this study was not to determine whether the t-test should be used under this distribution (if it had been the case, the t-test would have been compared with competitor statistics such as the locally most powerful test, if one exists, or a nonparametric test), but rather to determine whether the type I error rates and power of the t-test can be affected in a nontrivial manner by population shape alone. Since this study seems to have resolved this issue in the affirmative, the question now becomes whether the effects on the t-test under other distributions will be so favorable. In other words, the t-test is vulnerable to population shapes and there are no guarantees that this vulnerability will not manifest itself in much more unpleasant ways when other "real" populations are studied.

Finally, the results of this study have not shown the t-test to be anything other than the remarkably robust test that it is generally believed to be. It has shown, however, that in our enthusiasm for this statistic we have overstated the case for robustness, and in doing so have blinded researchers to some very real problems that may exist in many research contexts.

References

Allport, F. H. The J-Curve hypothesis of conforming behavior. *Journal of Social Psychology*, 1934, 5, 141-183.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. Weak measurements vs. strong statistics: an empirical critique of S. S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, 1966,26, 291-309.

Bartlett, M. S. The effect of non-normality on the t-distribution. *Proceedings of the Cambridge Philosophical Society*, 1935, 31, 223-231.

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

Boneau, C. A. A comparison of the power of the U and t tests. *Psychological Review*, 1962, 69, 246-256.

Box, G. E. P. Non-normality and tests on variances. *Biometrika*, 1953, 40, 318-335.

Box, G. E. O. Some theorems on quadratic forms applied in the study of analysis of variance problems. I, Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 1954a, 25, 290-302.

Box, G. E. P. Some theorems on quadratic forms applied in the study of analysis of variance problems. II, Effects of inequality of variance and of correlation between errors in the two-way classification. *Annals of Mathematical Statistics*, 1954b, 25, 484-498.

Box, G. E. P. & Andersen, S. L. Permutation theory in the derivation of robust criteria and the study of departures from assumption. Journal of the Royal Statistical Society, (Series B) 1955, 17, 1-34.

Bradley, J. V. Studies in research methodology. [a monograph], Dissertation Abstracts (B), 1968, 28, 4815-4816.

Bradley, J. V. Probability; Decision; Statistics. Englewood Cliffs, N. J.: Prentice-Hall, 1976.

Bradley J. V. A common situation conducive to bizarre distribution shapes. The American Statistician, 1977, 31, 147-150.

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, Inc., 1977.

Daniels, H. E. The effect of departures from ideal conditions other than nonnormality on the t and z tests of significance. Proceedings of the Cambridge Philosophical Society, 1938, 34, 321-328.

David, F. N. & Johnson, N. L. The effect of non-normality on the power function of the F-test in the analysis of variance. Biometrika, 1951, 38, 43-57.

Games, P. A. & Lucas, P. A. Power of the analysis of variance of independent groups on non-normal and normally transformed data. Educational and Psychological Measurement, 1966, 26, 311-327.

Gayen, A. K. The distribution of the variance ratio in random samples of any size drawn from non-normal universes. Biometrika, 1950a, 37, 236-255.

Gayen, A. K. Significance of difference between the means of two non-normal samples. Biometrika, 1950b, 37, 399-408.

Glass,G. V. & Stanley, J. C. Statistical methods in education and psychology. Englewood Cliffs, N. J. : Prentice-Hall, 1970.

Horsnell, G. The effect of unequal group variances on the F-test for the homogeneity of group means. Biometrika, 1953, 40, 128-136.

Neave, H. R. & Granger, C. W. J. A monte carlo study comparing various two-sample tests for differences in mean. Technometrics, 1968, 10, 509-522.

Norton, D. W. An empirical investigation of the effects of non-normality and heterogeneity upon the F-test of analysis of variance. Unpublished doctoral dissertation, State University of Iowa, 1952.

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

Quensel, C. E. The validity of the z-criterion when the variates are taken from different normal populations. Skand. Aktuarietids, 1947, 30, 44-55.

Scheffe, H. The analysis of variance. New York: John Wiley & Sons, Inc., 1959.

Srivastava, A. B. L. Effect of non-normality on the power of the analysis of variance test. Biometrika, 1959, 46, 114-122.

Welch, B. L. The significance of the difference between two means when the population variances are unequal. Biometrika, 1937, 29, 350-362.