Psicológica (2004), 25, 103-133.

Inflation of Type I Error Rates by Unequal Variances Associated with Parametric, Nonparametric, and Rank-Transformation Tests

Donald W. Zimmerman*

Carleton University, Canada

It is well known that the two-sample Student t test fails to maintain its significance level when the variances of treatment groups are unequal, and, at the same time, sample sizes are unequal. However, introductory textbooks in psychology and education often maintain that the test is robust to variance heterogeneity when sample sizes are equal. The present study discloses that, for a wide variety of non-normal distributions, especially skewed distributions, the Type I error probabilities of both the t test and the Wilcoxon-Mann-Whitney test are substantially inflated by heterogeneous variances, even when sample sizes are equal. The Type I error rate of the t test performed on ranks replacing the scores (rank-transformed data) is inflated in the same way and always corresponds closely to that of the Wilcoxon-Mann-Whitney test. For many probability densities, the distortion of the significance level is far greater after transformation to ranks and, contrary to known asymptotic properties, the magnitude of the inflation is an increasing function of sample size. Although nonparametric tests of location also can be sensitive to differences in the shape of distributions apart from location, the Wilcoxon-Mann-Whitney test and rank-transformation tests apparently are influenced mainly by skewness that is accompanied by specious differences in the means of ranks.

It is well known that widely-used statistical significance tests, including the two-sample Student t test and the ANOVA F test, are derived under an assumption of homogeneity of variance. Violation of this assumption when sample sizes are unequal substantially alters Type I error probabilities. When a larger variance is associated with a larger sample size, the probability of a Type I error declines below the nominal significance

* Send correspondence to: Donald W. Zimmerman, 1978 134A Street, Surrey, BC V4A 6B6, Canada. Phone: (604) 531-9313, Fax: (604) 535-5354, E-mail: zimmerma@look.ca
The computer programs in this study were written in PowerBasic, version 3.5, PowerBasic, Inc., Carmel, CA. Listings of the programs can be obtained by writing to the author.

level. In contrast, when a larger variance is associated with a smaller sample size, the probability increases, sometimes far above the significance level (see, for example, Alexander and Govern, 1994; Byrk and Raudenbush, 1987; Hsu, 1938; Nelson, 2000; Overall, Atlas, and Gibson, 1995; Scheffe′; 1959; Schneider and Penfield, 1997; Zimmerman and Zumbo, 1993a, 1993b; Wludyka and Nelson, 1999).

At one time, investigators employed preliminary tests of equality of variances before the t test or ANOVA F test (see, for example, Mendes, 2003). If a preliminary test rejected the hypothesis of equality of variances, investigators usually substituted another procedure for t or F tests—often a nonparametric test such as the Wilcoxon-Mann-Whitney test or the Kruskal-Wallis test. More recently it has become apparent that these nonparametric methods, also are strongly affected by variance heterogeneity, although the changes are not as large as for parametric tests (Zimmerman and Zumbo, 1993a, 1993b). The power functions of the Wilcoxon-Mann-Whitney and the Kruskal-Wallis tests are similar in shape to those of the t and F tests, although slightly less in magnitude, and they depend on the ratio of variances, as well as the ratio of sample sizes, in the same way.

It has also been found that it is inefficient to base a decision on preliminary tests of equality of variances (Albers, Boon, and Kallenberg, 2000; Chen and Chen, 1998; Zimmerman, in press), and modern textbooks no longer recommend preliminary tests. Substitution of a separate-variances approximation of the t test (Satterthwaite, 1946; Welch, 1938), usually leads to better results. And in recent years there has been a lot of interest in bootstrap, trimming, and related methods (Keselman, Cribbie, & Zumbo; Keselman, Wilcox, Othman, and Fradette, 2002; Wei-ming, 1999; Westfall and Young, 1993; Wilcox, 2001, 2003).

It is still believed by many researchers that both parametric and nonparametric significance tests are robust to variance heterogeneity when sample sizes are equal. However, investigators have reported changes in the Type I error rates of the Student t test even when sample sizes are equal (e.g., Harwell, 1990; Rogan and Keselman, 1977). These changes, although sometimes substantial, are relatively small compared to the much larger changes that occur when sample sizes differ.

The present study examines a wide variety of non-normal densities. For many skewed distributions, heterogeneity of variance produces significant distortions of the Type I error rates of the t test even when $n_1 = n_2$. Furthermore, the distortion is far more serious after transformation of the initial scores to ranks, and both the Wilcoxon-Mann-Whitney test and the t test on ranks following a rank transformation are adversely affected in the same way.

METHOD

The random number generator used in this study was introduced by Marsaglia, Zaman, and Tsang (1990) and has been described by Pashley (1993, pp. 395-415). Normal variates, N(0,1), were generated by the rejection method of Marsaglia and Bray (1964) and were transformed to have various distribution shapes using inverse distribution functions. Many of the distributions studied were standard continuous probability densities, both skewed and symmetric.

Some occur frequently in applications. The exponential distribution characterizes time and latency measures in psychological research. The Gumbel distribution, or extreme value distribution, characterizes the largest (or smallest) of a number of values. The lognormal distribution applies to random variables that have no values less than zero but a few very large values. The Weibull distribution describes failure rates of equipment. Mixed-normal distributions are often taken as a model of outliers. In some cases, the present study included several distributions from the same family of distributions, having different scale and shape parameters, and consequently different skewness and kurtosis.

Also, the study included various discrete distributions similar to the ones studied by Micceri (1989), which occur frequently in educational and psychological research. The truncated distributions, for example, represent test scores which are concentrated near zero, if a test is too difficult, or near an upper "ceiling" if it is too easy. Bimodal and rectangular distributions also are common in research. Altogether, 25 distribution shapes were included in the study. The algorithms for generating the distributions are described in Table A1 in the appendix.

Each replication of the sampling procedure obtained two independent samples of the same size from one of the distributions. The variates were transformed to have mean 0 and standard deviation 1. Next, all scores in one sample were multiplied by a constant, so that the ratio σ_1/σ_2 had a predetermined value. Sample sizes ranged from 20 to 200 under various conditions in the study. The computer code was thoroughly tested, and at least 2 million samples were taken from each distribution to verify that the sample values had mean 0 and standard deviation 1.

For all 25 distributions, the means of 2 million sample values did not deviate from 0 by more than .002, and the standard deviations did not deviate from 1 by more than .006. Also, the means did not deviate from 0 by more than .003 and the standard deviations did not deviate from 2 by more than .008 for samples from N(0,2). The code was further tested to verify that the Type I error probabilities of the t test were close to the nominal

significance level for normal distributions with equal variances and that the Type I error probabilities of the Wilcoxon-Mann-Whitney test were close to the nominal significance level for all 25 distributions when variances were equal. The deviations of these probabilities can be observed in Tables 1 and 2.

On each replication, a Student t test was performed on the scores. Next, the two samples were combined, the $N_1 + N_2 = 2N_1 = 2N_2$ scores were transformed to ranks, and a Student t test was performed on the ranks replacing corresponding scores. Finally, the large-sample normal-approximation form of the Wilcoxon-Mann-Whitney test was performed on the ranks. All significance tests were nondirectional and were evaluated at the .01, .05, and .10 significance levels. Each condition in the study included 50,000 replications of the sampling procedure and subsequent significance tests.

RESULTS OF SIMULATIONS

Table 1 presents data for 25 distributions for $N_1 = N_2 = 25$. The data for the Student t test for the normal distribution is typical of results obtained in many simulation studies in the past. One observes a slight increase in the probability of a Type I error above the nominal significance level as σ_1/σ_2 becomes increasingly more extreme. On the other hand, the probability of a Type I error of the t test performed on ranks replacing the scores increases to a somewhat greater extent. When $\sigma_1/\sigma_2 = 1$, the probability is close to the significance level for both scores and ranks, as would be expected. The Type I error rate of the Wilcoxon-Mann-Whitney test is close to that of the t test on ranks.

The results for the exponential distribution are quite different, and the increase in the Type I error rate is far more extreme. The increase in the case of the t test is considerable for the initial scores, and it is overwhelming for ranks. Again the Wilcoxon-Mann-Whitney test yields a value close to that of the t test on ranks. Moreover, the magnitude of the change apparently is an increasing function of sample size in the case of ranks, although this magnitude is independent of sample size for scores. The probability of a Type I error of the t test on scores is slightly below the nominal significance level when $\sigma_1 = \sigma_2$, and it comes closer to the significance level as sample size becomes larger. However, the probability of the test on ranks is close to the significance level when $\sigma_1 = \sigma_2$, for all sample sizes, as one would expect. This is true for all distributions in the present study.

Table 1. Probability of rejecting H_0 by the t test on scores, the t test on ranks and the Wilcoxon-Mann-Whitney test for various non-normal distributions ($N_1 = N_2 = 25$).

		$\sigma_1/\sigma_2 = 1$			$\sigma_1/\sigma_2 = 2$			$\sigma_1/\sigma_2 = 3$		
distribution	α	t	t on ranks	W	t	t on ranks	W	t	t on ranks	W
normal	.01	.011	.011	.009	.010	.014	.012	.012	.018	.015
	.05	.051	.050	.053	.052	.059	.058	.054	.068	.068
	.10	.100	.099	.102	.103	.113	.110	.105	.127	.126
exponential scale parameter 1	.01	.008	.011	.009	.018	.095	.083	.025	.145	.132
	.05	.046	.050	.049	.063	.228	.227	.070	.303	.302
	.10	.099	.100	.101	.112	.332	.329	.119	.414	.413
lognormal shape parameter 1 scale parameter 1	.01	.004	.011	.010	.024	.201	.184	.045	.315	.292
	.05	.039	.048	.050	.074	.391	.393	.100	.522	.523
	.10	.092	.099	.102	.127	.510	.511	.151	.634	.638
lognormal shape parameter 1 scale parameter .6	.01	.008	.011	.009	.016	.062	.056	.023	.107	.095
	.05	.047	.050	.050	.059	.166	.170	.069	.242	.243
	.10	.098	.100	.098	.110	.255	.264	.120	.345	.346
mixed-normal (.05, 20)	.01	.007	.011	.009	.008	.014	.012	.008	.016	.014
	.05	.045	.050	.050	.044	.056	.057	.048	.067	.068
	.10	.096	.101	.099	.097	.111	.110	.097	.124	.125
mixed-normal (.02, 10)	.01	.004	.010	.010	.005	.014	.012	.006	.018	.014
	.05	.031	.049	.051	.033	.057	.055	.037	.069	.066
	.10	.083	.098	.101	.085	.111	.107	.088	.126	.122
mixed-normal (.01, 20)	.01	.002	.010	.009	.003	.013	.011	.003	.016	.016
	.05	.020	.049	.050	.021	.056	.057	.023	.066	.067
	.10	.062	.099	.099	.064	.110	.111	.066	.124	.124
gamma shape parameter 2	.01	.009	.011	.008	.014	.043	.041	.020	.074	.066
	.05	.050	.052	.050	.057	.131	.136	.065	.184	.185
	.10	.102	.102	.099	.108	.211	.214	.114	.276	.278
gamma shape parameter 5	.01	.010	.011	.009	.012	.024	.020	.015	.038	.033
	.05	.050	.050	.048	.055	.085	.085	.059	.113	.114
	.10	.101	.099	.096	.105	.150	.149	.108	.187	.187

Table 1 (continued).

distribution		$\sigma_1/\sigma_2 = 1$			$\sigma_1/\sigma_2 = 2$			$\sigma_1/\sigma_2 = 3$		
	α	t	t on ranks	W	t	t on ranks	W	t	t on ranks	W
Laplace	.01	.009	.011	.009	.010	.013	.011	.009	.015	.014
location	.05	.050	.049	.049	.050	.054	.054	.049	.059	.064
parameter 0	.10	.101	.099	.099	.102	.107	.107	.100	.115	.118
scale parameter 1										
logistic	.01	.010	.012	.009	.011	.014	.011	.010	.017	.015
location	.05	.051	.051	.049	.053	.058	.057	.052	.066	.067
parameter 0	.10	.100	.101	.099	.105	.114	.111	.102	.123	.123
scale parameter 1										
logistic	.01	.009	.011	.009	.011	.014	.011	.011	.016	.014
location	.05	.049	.049	.049	.051	.058	.058	.052	.066	.066
parameter 0	.10	.100	.100	.100	.102	.111	.112	.104	.121	.122
scale parameter .5										
logistic	.01	.009	.011	.009	.010	.014	.011	.010	.016	.015
location	.05	.049	.049	.050	.052	.058	.056	.051	.065	.067
parameter 0	.10	.101	.099	.098	.102	.109	.108	.104	.124	.124
scale parameter 2										
Gumbel	.01	.009	.011	.009	.012	.027	.024	.016	.045	.040
location	.05	.049	.049	.048	.055	.092	.091	.059	.126	.129
parameter 0	.10	.099	.099	.099	.107	.162	.161	.109	.206	.207
scale parameter 1										
power function	.01	.010	.011	.009	.014	.050	.045	.016	.068	.058
shape	.05	.049	.048	.050	.054	.140	.140	.059	.169	.165
parameter .5	.10	.099	.099	.099	.104	.221	.221	.109	.255	.252
scale parameter 1										
power function	.01	.010	.011	.009	.014	.041	.036	.017	.061	.053
shape	.05	.050	.051	.051	.056	.123	.121	.059	.157	.156
parameter 3	.10	.100	.099	.102	.106	.199	.197	.108	.242	.243
scale parameter 1										
bimodal	.01	.010	.011	.009	.011	.013	.012	.011	.017	.015
1σ between	.05	.051	.050	.048	.052	.059	.058	.053	.068	.067
modes	.10	.100	.099	.099	.103	.112	.112	.106	.127	.124
bimodal	.01	.010	.011	.009	.011	.015	.012	.012	.019	.016
2σ between	.05	.050	.048	.050	.052	.060	.059	.055	.072	.072
modes	.10	.100	.098	.099	.102	.113	.115	.106	.130	.129

Table 1 (continued).

distribution		$\sigma_1/\sigma_2 = 1$			$\sigma_1/\sigma_2 = 2$			$\sigma_1/\sigma_2 = 3$		
	α	t	t on ranks	W	t	t on ranks	W	t	t on ranks	W
Weibull shape parameter .5 scale parameter 1	.01 .05 .10	.003 .035 .090	.010 .049 .099	.009 .050 .098	.035 .094 .146	.515 .722 .812	.494 .723 .811	.067 .126 .176	.624 .803 .875	.609 .808 .876
Weibull shape parameter 3 scale parameter 1	.01 .05 .10	.011 .049 .099	.011 .050 .097	.010 .049 .099	.011 .051 .100	.014 .059 .114	.013 .059 .114	.012 .054 .106	.019 .072 .131	.016 .070 .130
Weibull shape parameter 2 scale parameter 1	.01 .05 .10	.010 .049 .101	.011 .049 .100	.009 .048 .098	.012 .053 .103	.019 .075 .136	.018 .074 .133	.013 .055 .104	.028 .093 .159	.025 .093 .161
ceiling at $\mu +$ 1.5 σ	.01 .05 .10	.010 .049 .099	.011 .048 .100	.008 .048 .097	.012 .053 .104	.015 .061 .116	.013 .059 .115	.013 .055 .106	.019 .070 .131	.016 .068 .126
ceiling at $\mu +$ .5 σ	.01 .05 .10	.010 .050 .100	.011 .049 .098	.009 .048 .097	.014 .056 .106	.057 .154 .241	.051 .155 .241	.017 .060 .110	.070 .173 .262	.062 .172 .260
triangular range (0, 2)	.01 .05 .10	.012 .051 .101	.011 .051 .100	.010 .051 .100	.011 .052 .102	.014 .059 .113	.011 .057 .112	.013 .054 .103	.018 .068 .125	.016 .070 .126
rectangular range (0,1)	.01 .05 .10	.010 .049 .100	.010 .049 .099	.009 .049 .099	.012 .053 .105	.017 .064 .121	.014 .064 .119	.013 .055 .102	.021 .075 .134	.018 .075 .134

The same pattern of results found for the exponential distribution characterizes many other continuous distributions, as well as several discrete distributions included in the study. For these continuous densities—the power function, three versions of the Weibull, two versions of the lognormal, and the gamma —the inflation of Type I error rates is as pronounced, or more so, as that of the exponential distribution. The same is true for the geometric distribution, the discrete analogue of the exponential distribution. The common feature of all these distributions is skewness. In all these cases, the inflation of the probability of the t test on ranks (and the Wilcoxon-Mann-Whitney test) exceeds that of the t test on scores.

For many other distributions, some increase occurs, but it is not as extreme as the ones listed above. For example, the logistic distribution exhibits changes characteristic of the normal distribution, for both scores and ranks. The same thing is true for several discrete distributions. The common feature of these distributions is symmetry. For every distribution in this table, the Type I error rate of the Wilcoxon-Mann-Whitney test is inflated in the same way when variances are unequal and always remains extremely close to that of the t test performed on ranks.

Figures 1 through 6 plot more detailed results for selected distributions—the lognormal, rectangular, power function, gamma, logistic, and Weibull—for σ_1/σ_2 ranging from 1 to 3 in increments of .25, for the .05 significance level, and for sample sizes of 20 and 50. In all cases, the pattern of results is similar to that described above. In Figure 1, for the lognormal distribution, the curves for the Wilcoxon-Mann-Whitney test and the rank-transformation test are almost indistinguishable. Since the same is true for all distributions, curves for the Wilcoxon-Mann-Whitney test were omitted in Figures 2 through 6. The size of the change for ranks exceeds that for scores in all cases. The increase for ranks is extreme for the lognormal, power function, gamma, and Weibull distributions, while it is more moderate for the rectangular and logistic distributions. There appears to be a slight or moderate increase for scores in all cases except the power function distribution for N = 50.

Figure 7 plots the probability of rejecting H_0 as a function of sample size for the exponential distribution, when σ_1/σ_2 is fixed at 1.25 or 2.00. Clearly, the extent to which the Type I error rate of the test on ranks exceeds the nominal significance level is an increasing function of sample size, while this is not true for the test on scores.

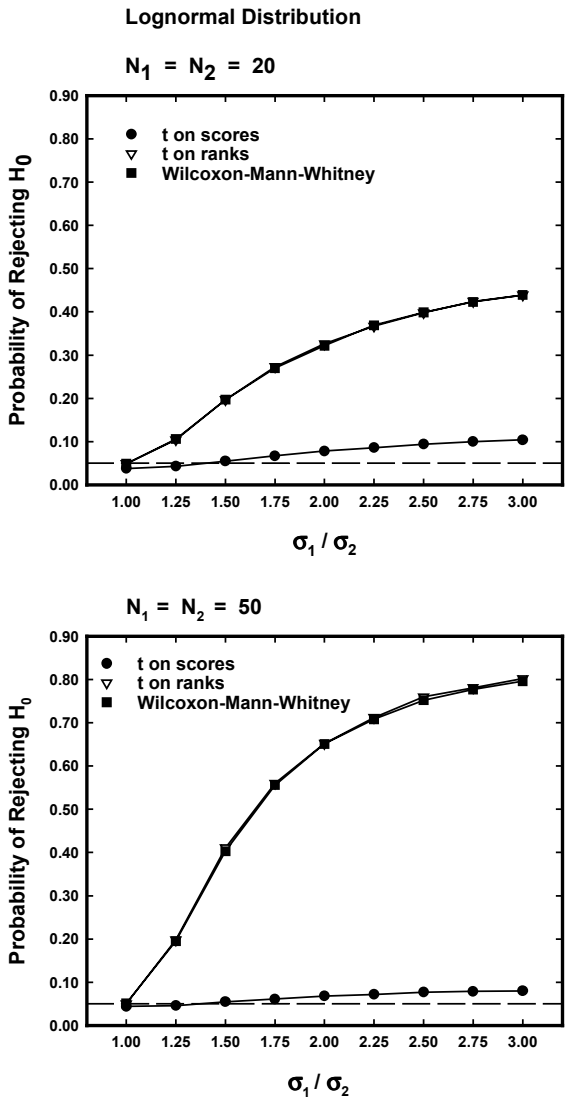

Figure 1. Probability of rejecting H_0 by Student t test on scores, the rank-transformation test (t test on ranks), and the Wilcoxon-Mann-Whitney test, as a function of ratio of standard deviations— lognormal distribution (shape parameter 1, scale parameter 1).

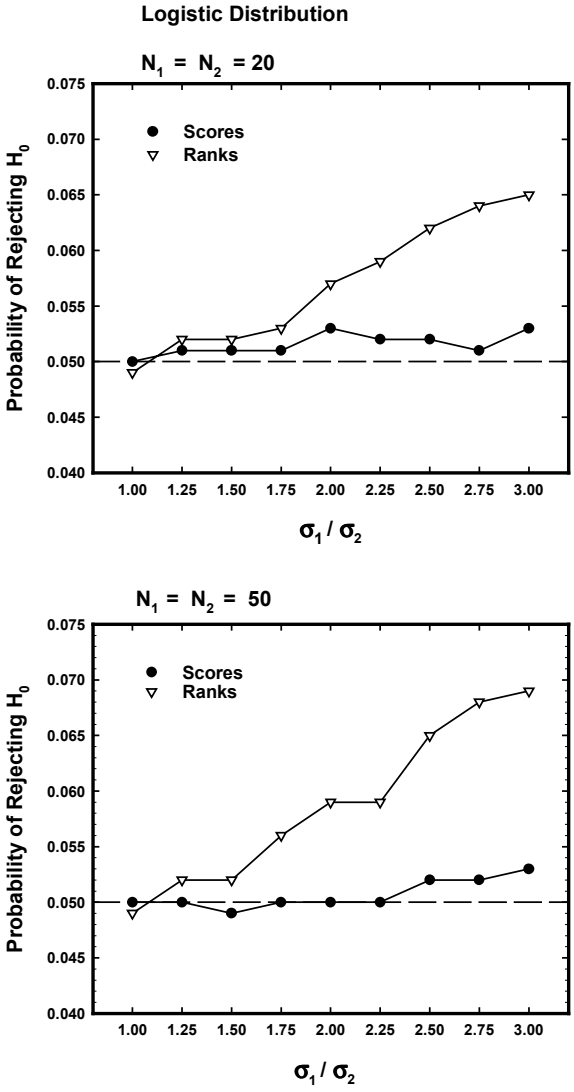

Figure 2. Probability of rejecting H_0 by Student t test on scores and ranks as a function of ratio of standard deviations—logistic distribution (location parameter 0, scale parameter 1).

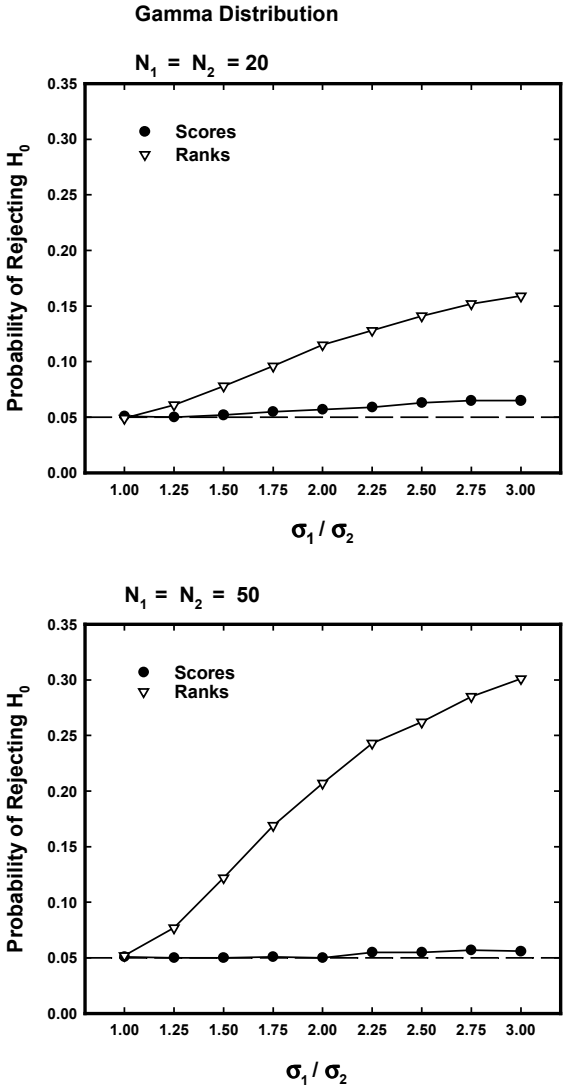

Figure 3. Probability of rejecting H_0 by Student t test on scores and ranks as a function of ratio of standard deviations—gamma distribution (2 convolutions).

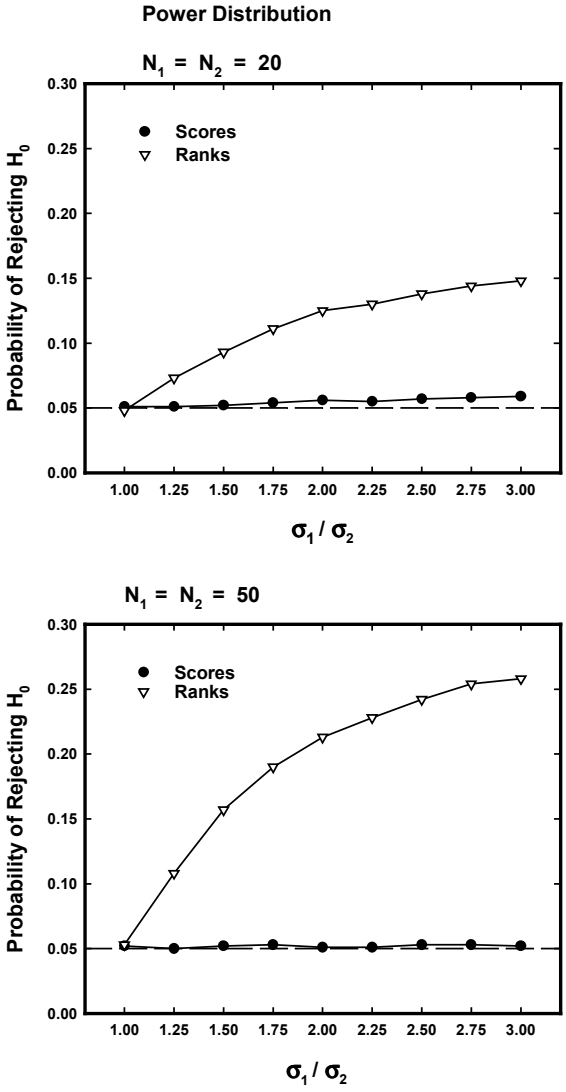

Figure 4. Probability of rejecting H_0 by Student t test on scores and ranks as a function of ratio of standard deviations—power distribution (shape parameter .5, scale parameter 1).

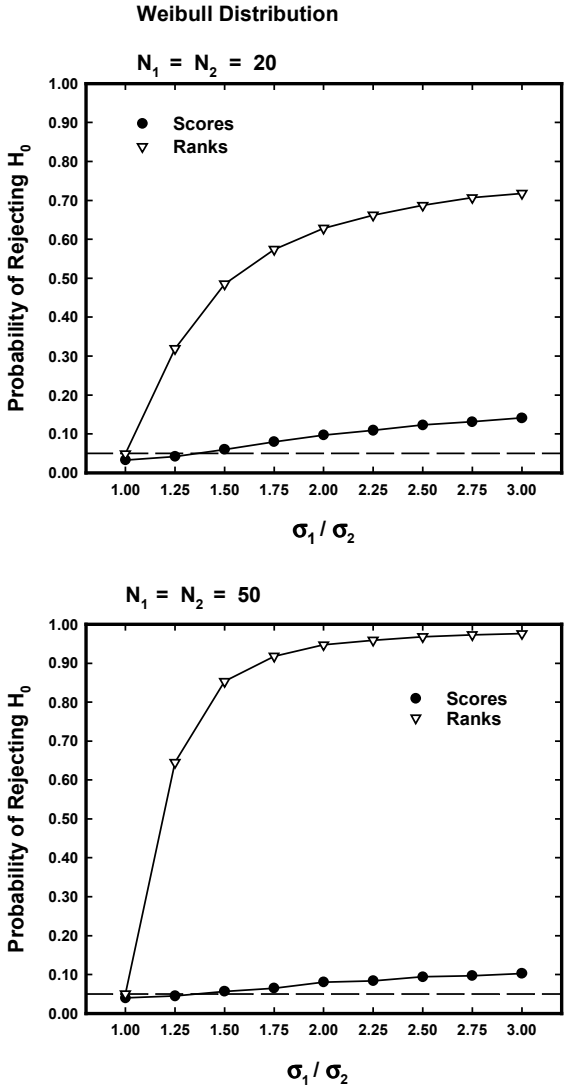

Figure 5. Probability of rejecting H_0 by Student t test on scores and ranks as a function of ratio of standard deviations—Weibull distribution (shape parameter .5, scale parameter 1).

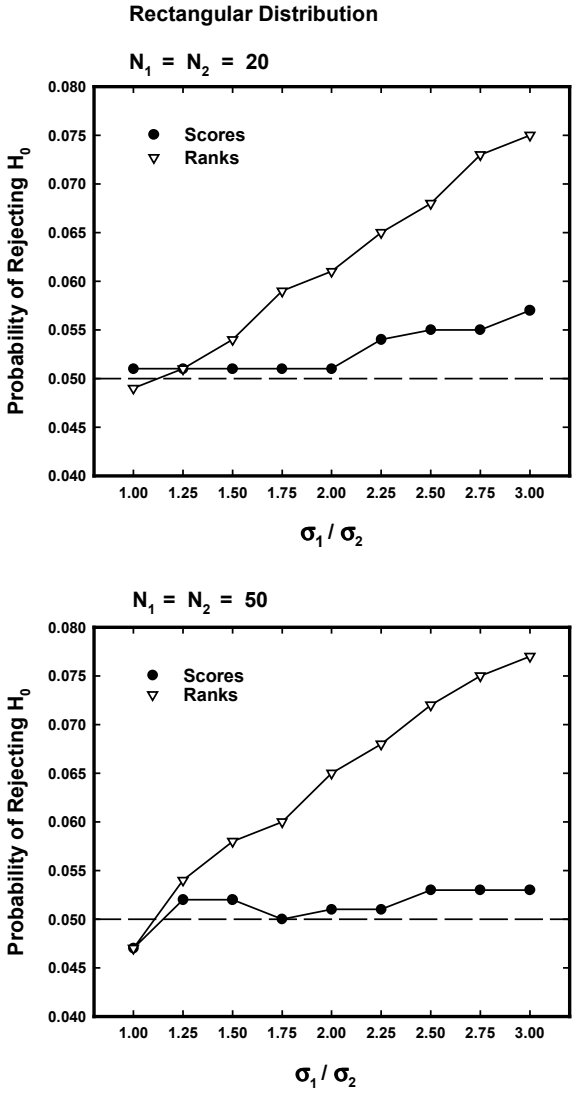

Figure 6. Probability of rejecting H_0 by Student t test on scores and ranks as a function of ratio of standard deviations— rectangular distribution on (0,1).

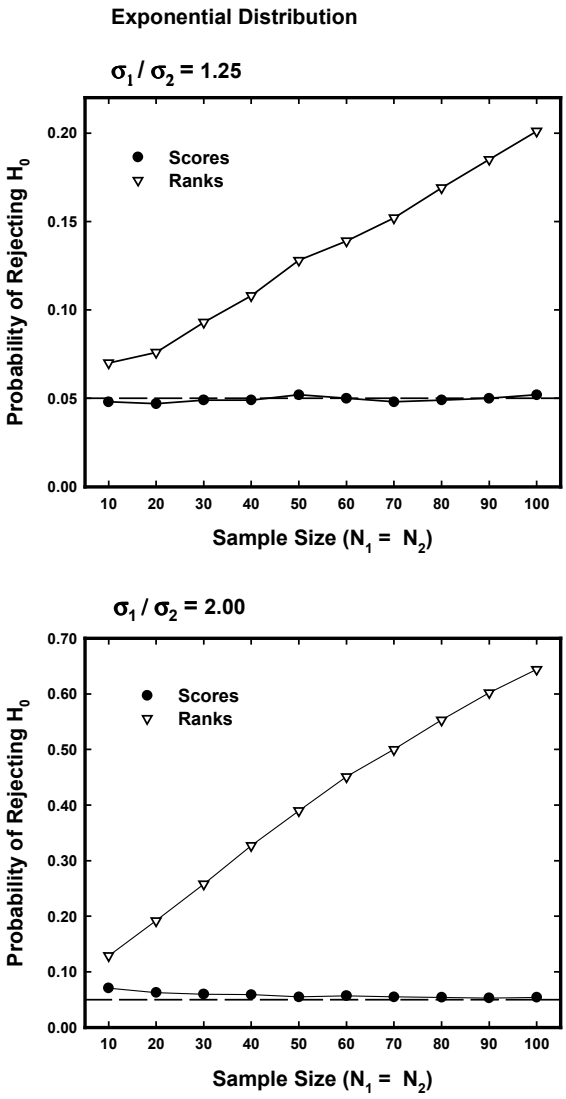

Figure 7. Probability of rejecting H_0 by Student t test on scores and ranks as a function of sample size ($N_1 = N_2$)—exponential distribution (scale parameter 1).

Table 2 provides a more complete picture of the dependence of the Type I error rate on sample size. All 25 distributions are included in the table, which gives probabilities for sample sizes of 20, 50, and 80 and for ratios of standard deviations of 1.25 and 2.00. The test on ranks exhibits the same dependence on sample size shown in Figure 7 for many other non-normal distributions, both continuous and discrete. These include the lognormal (both versions), gamma (both versions), Gumbel, power function (both versions), Weibull (two of the three versions), and ceiling at $\mu + 1.5\sigma$. In all these cases, the inflation of the Type I error probability is greater when σ_1/σ_2 is larger (2.00 rather than 1.25). Furthermore, it increases more rapidly. All these distributions are skewed. Samples from symmetric distributions, such as the mixed-normal, logistic, bimodal, triangular, rectangular, and Laplace, apparently are not affected in this way.

Further evidence that skewness accounts for inflation of the t statistic is provided by Figures 8 and 9, which are plots of relative frequency distributions of the differences between means. These are based on 50,000 samples, for sample sizes of $N_1 = N_2 = 20$, for a ratio of standard deviations of 3.00, and for a significance level of .05. In the case of the symmetric logistic distribution shown in Figure 8, the differences between means are symmetric about zero, as is expected, and the values of the t statistic also are symmetric about zero. The Type I error rate is .055. However, in the case of the skewed lognormal distribution shown in Figure 9, the distribution of differences between means is negatively skewed, and the values of the t statistic are substantially inflated. In this case the Type I error rate is .106.

Figure 10 shows similar results for the rank transformation applied to samples from a lognormal distribution. The initial scores were transformed to ranks and the t test was performed on the ranks. The sample sizes, ratio of standard deviations, and significance level were the same as in Figure 9. The differences between means of ranks is nearly symmetric about 7 instead of zero, and the t distribution is highly skewed with a mean of about 2 instead of zero. In this case the Type I error rate of the t test on ranks is .444. The distributions of differences between means and of values of the t statistic in Figures 8, 9, and 10 are consistent with the inflated Type I error rates in Tables 1 and 2.

Table 2. Probability of rejecting H_0 by the t test on scores, t test on ranks, and Wilcoxon-Mann-Whitney test as a function of sample size ($N_1 = N_2 = 20$, 50, and 80) for two ratios of standard deviations ($\sigma_1/\sigma_2 =$ 1.25 and 2.00), $\alpha = .05$.

		t			t on ranks			W		
		N_1,N_2			N_1,N_2			N_1,N_2		
distribution	σ_1/σ_2	20	50	80	20	50	80	20	50	80
normal	1.25	.050	.050	.049	.050	.051	.051	.050	.050	.050
	2.00	.054	.050	.053	.058	.059	.059	.057	.057	.058
exponential	1.25	.047	.050	.049	.077	.125	.171	.079	.123	.170
scale parameter 1	2.00	.063	.056	.053	.191	.394	.557	.189	.391	.553
lognormal	1.25	.044	.045	.048	.105	.198	.284	.105	.196	.280
shape parameter 1	2.00	.077	.065	.062	.325	.654	.840	.326	.652	.840
scale parameter 1										
lognormal	1.25	.050	.050	.050	.064	.087	.106	.063	.085	.108
shape parameter 1	2.00	.063	.056	.052	.145	.281	.408	.145	.281	.408
scale parameter .6										
mixed-normal	1.25	.044	.046	.047	.049	.050	.052	.050	.052	.050
(.05, 20)	2.00	.045	.047	.047	.057	.059	.058	.058	.060	.059
mixed-normal	1.25	.033	.038	.041	.052	.052	.050	.051	.050	.051
(.02, 10)	2.00	.033	.037	.039	.056	.058	.057	.056	.057	.059
mixed-normal	1.25	.021	.026	.035	.049	.051	.050	.051	.051	.051
(.01, 20)	2.00	.023	.025	.031	.057	.058	.058	.054	.057	.056
gamma	1.25	.051	.050	.049	.060	.076	.089	.060	.076	.092
shape parameter 2	2.00	.059	.054	.053	.117	.210	.298	.114	.208	.297
gamma	1.25	.051	.050	.051	.053	.060	.064	.052	.058	.062
shape parameter 5	2.00	.055	.052	.050	.078	.112	.146	.079	.113	.147

Table 2 (continued).

		t			t on ranks			W		
		N_1,N_2			N_1,N_2			N_1,N_2		
distribution	σ_1/σ_2	20	50	80	20	50	80	20	50	80
Laplace	1.25	.049	.049	.051	.051	.051	.052	.050	.049	.049
location parameter 0 scale parameter 1	2.00	.050	.051	.050	.056	.057	.054	.055	.055	.058
logistic	1.25	.050	.050	.049	.050	.052	.051	.049	.051	.050
location parameter 0 scale parameter 1	2.00	.052	.049	.051	.057	.058	.059	.055	.057	.058
logistic	1.25	.050	.050	.050	.051	.051	.051	.049	.049	.048
location parameter 0 scale parameter .5	2.00	.050	.051	.050	.055	.058	.057	.057	.057	.058
logistic	1.25	.049	.050	.049	.049	.051	.051	.049	.049	.051
location parameter 0 scale parameter 2	2.00	.051	.050	.048	.057	.059	.056	.056	.058	.056
Gumbel	1.25	.050	.049	.050	.054	.061	.066	.051	.050	.054
location parameter 0 scale parameter 1	2.00	.056	.053	.052	.085	.130	.174	.083	.128	.174
power function	1.25	.053	.051	.050	.076	.111	.137	.075	.107	.138
shape parameter .5 scale parameter 1	2.00	.053	.052	.050	.123	.214	.299	.120	.216	.297
power function	1.25	.051	.051	.051	.061	.078	.095	.060	.076	.092
shape parameter 3 scale parameter 1	2.00	.055	.052	.051	.109	.184	.255	.107	.182	.254
bimodal	1.25	.050	.049	.050	.049	.051	.052	.049	.052	.050
1σ between modes	2.00	.050	.050	.052	.058	.059	.060	.056	.059	.059
bimodal	1.25	.052	.051	.049	.051	.052	.051	.049	.051	.050
2σ between modes	2.00	.052	.053	.051	.058	.061	.060	.058	.059	.060

Table 2 (continued).

		t			t on ranks			W		
		N_1,N_2			N_1,N_2			N_1,N_2		
distribution	σ_1/σ_2	20	50	80	20	50	80	20	50	80
Weibull shape parameter .5 scale parameter 1	1.25	.043	.045	.045	.321	.651	.836	.319	.647	.835
	2.00	.097	.076	.069	.630	.946	.994	.628	.945	.994
Weibull shape parameter 3 scale parameter 1	1.25	.050	.052	.050	.049	.052	.053	.050	.052	.051
	2.00	.054	.052	.049	.060	.062	.061	.060	.060	.061
Weibull shape parameter 2 scale parameter 1	1.25	.049	.049	.049	.051	.055	.059	.052	.054	.057
	2.00	.053	.051	.052	.070	.091	.112	.069	.091	.108
ceiling at $\mu + 1.5\sigma$	1.25	.051	.051	.050	.050	.053	.053	.050	.052	.053
	2.00	.052	.052	.050	.059	.062	.063	.059	.062	.062
ceiling at $\mu + .5\ \sigma$	1.25	.050	.050	.052	.100	.179	.257	.104	.180	.256
	2.00	.056	.052	.050	.135	.243	.348	.135	.245	.344
triangular range (0, 2)	1.25	.052	.049	.051	.051	.051	.051	.050	.050	.051
	2.00	.054	.050	.052	.060	.060	.060	.060	.057	.056
rectangular range (0,1)	1.25	.052	.050	.050	.052	.052	.052	.051	.052	.052
	2.00	.054	.051	.052	.063	.064	.066	.063	.064	.065

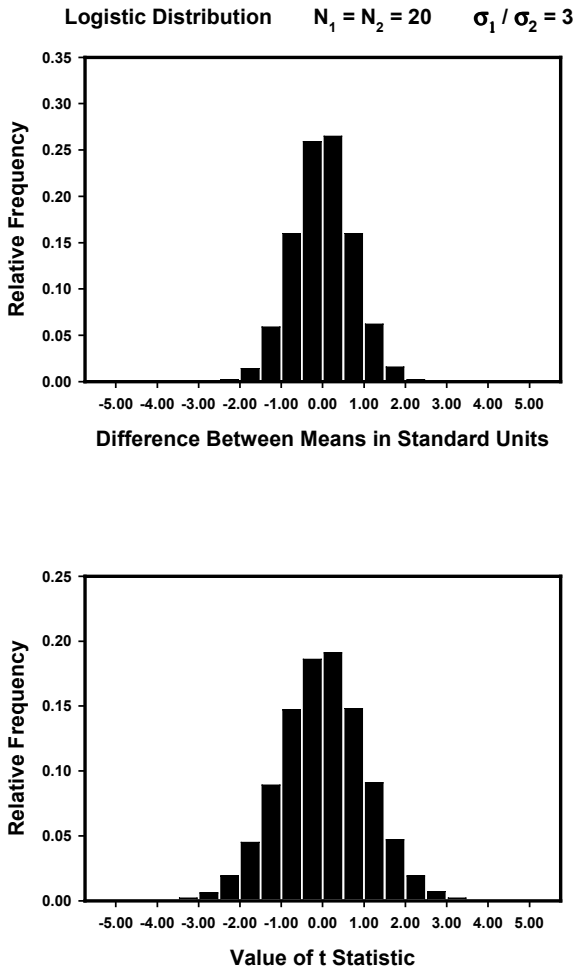

Figure 8. Relative frequency distributions of differences between means and values of t statistic— logistic distribution

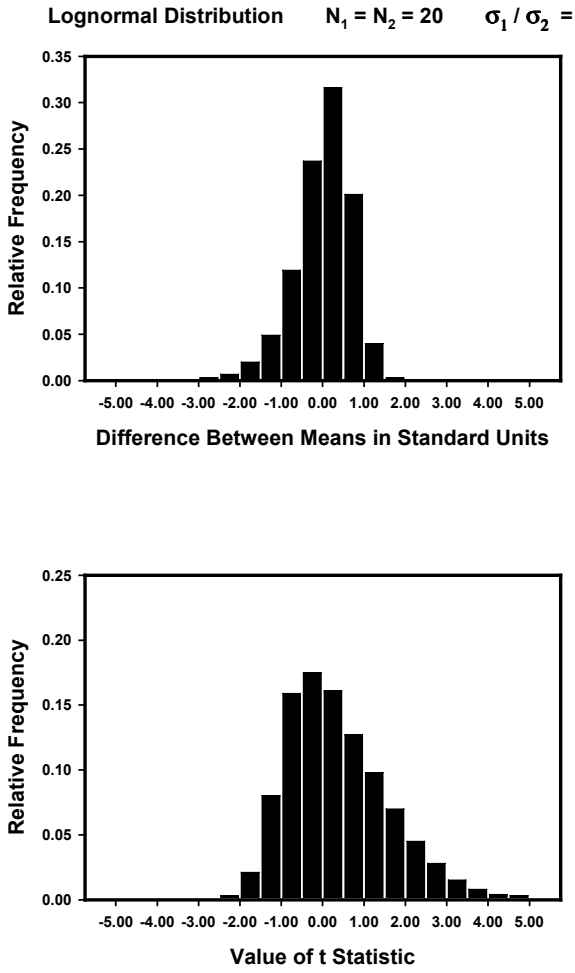

Figure 9. Relative frequency distributions of differences between means and values of t statistic— lognormal distribution

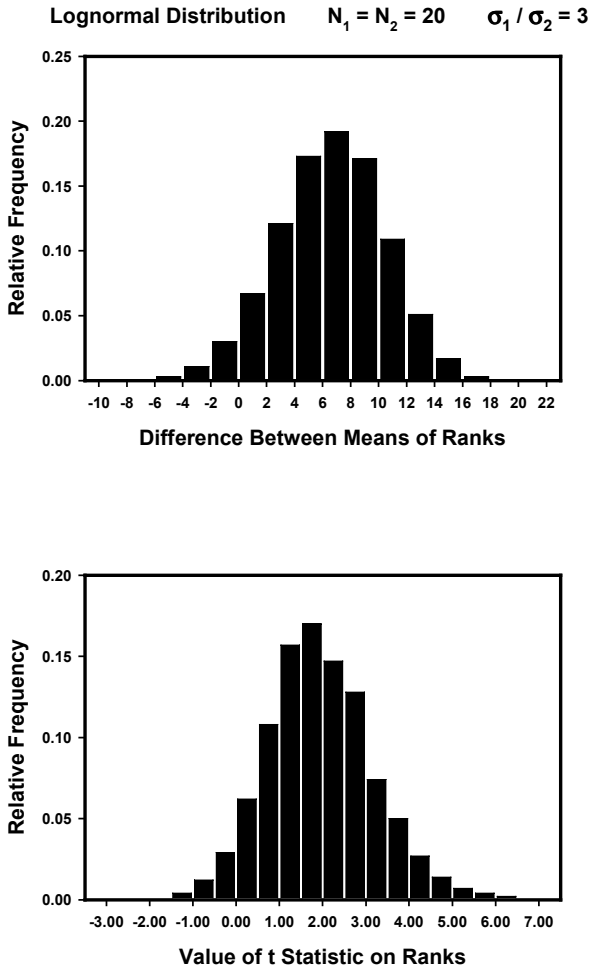

Figure 10. Relative frequency distributions of differences between means of ranks and values of t statistic on ranks—lognormal distribution

FURTHER IMPLICATIONS

It is generally believed that significance tests, both parametric and nonparametric, are robust to heterogeneity of variance when the population distribution is normal and sample sizes are equal. One observes some elevation of the Type I error probability of the Student t test under these conditions. The same conditions produce somewhat larger elevations of the Type I error rates of the t test on ranks and the Wilcoxon-Mann-Whitney test. For the normal case, the t test on scores can be regarded as robust to variance heterogeneity with some justification. However, the t test on ranks, as well as the Wilcoxon-Mann-Whitney test, can be seriously misleading if the ratio σ_1/σ_2 is 1.5 or more and sample sizes are as large as 50.

In the case of non-normal distributions, especially skewed distributions, the picture is quite different. The disruption of the t test on scores still is noticeable but in most cases not extreme. In contrast, the disruption of the t test on ranks is far greater. In the case of a few skewed distributions, such as the lognormal, exponential, and Weibull, both scores and ranks are severely affected, but ranks considerably more than scores. The pattern of results for symmetric non-normal distributions, including the rectangular, triangular, and bimodal, appears to be quite similar to that of the normal distribution. However, for most of the skewed non-normal densities, the Type I error rate is elevated far above the nominal significance level. Another anomalous result appears in the present study: The degree of inflation of the t test under the rank transformation increases with sample size for both normal and many non-normal distributions.

Since the Wilcoxon-Mann-Whitney test is known to be sensitive to any difference in distributions, not just to a difference in means, it is tempting to view the increase in Type I error rates in the present study as rejection of a false null hypothesis having to do with shape rather than location. However, this interpretation requires caution. For one thing, many distributions, can be quite dissimilar in shape, including differences in variances, and still no increase in Type I error rates are observed unless sample sizes are unequal. When both distributions are non-normal and symmetric, the change resulting from variance heterogeneity is minimal; when both distributions are non-normal and skewed, the change is extensive.

Furthermore, the equivalence of the Wilcoxon-Mann-Whitney test and the Student t test under a rank transformation raises another issue. For many distributions, the t test applied to raw scores is not highly sensitive to differences in shape apart from differences in location. However, transformation of raw scores to ranks immediately renders the t test

sensitive to differences in variances, extremely so when sample sizes are unequal. It is difficult to view this outcome as a consequence of the assumptions underlying the Wilcoxon-Mann-Whitney test, even though the Type I error rates of the two tests are nearly identical.

The results of the present study are strikingly different in other ways from the well-established effects of heterogeneous variances on Type I and Type II errors for normal distributions and unequal sample sizes. In the usual case, the modification of probabilities is large, but the change is greater for the parametric t test than for the nonparametric Wilcoxon-Mann-Whitney test. In the present study, where sample sizes are equal, the outcome is the reverse. In both cases, the degree of change is an increasing function of the ratio σ_1/σ_2.

The present study also reveals that the inflation of the Type I error rate can be different for two distributions from the same family of distributions. For example, consider the Type I error rates of the Weibull distribution with shape parameter 3 and scale parameter 1, compared to the Weibull distribution with shape parameter .5 and scale parameter 1. Note also the difference between the distribution with a ceiling at $\mu + 1.5\sigma$ and the distribution with a ceiling at $\mu + .5\sigma$. Similar differences are evident for the two lognormal distributions. Again, these differences appear to be related to varying degrees of skewness.

The theoretical t distribution is symmetric about zero and approaches the standard normal distribution as degrees of freedom increases. Furthermore, it is known that the distribution of sample values of t obtained from distributions with the same variance are symmetric about zero (see, for example, Cressie and Whitford, 1986; Miller, 1986; Tan, 1982; Wilcox, 2003). The present data suggest this is also at least approximately true for samples from symmetric distributions with different variances. However, the distribution of sample values of t from skewed distributions with unequal variances is not necessarily symmetric about zero, as indicated by Figure 9. Otherwise expressed, for these cases, the statistic calculated by the usual Student t formula does not have a theoretical t distribution. As a consequence, Type I error rates are modified, as indicated by the results in Tables 1 and 2. Moreover, when scores are transformed to ranks before the t test, a procedure equivalent to the Wilcoxon-Mann-Whitney test, alteration of the sample t distribution is more extreme, as shown by Figure 10.

In some instances, the increase in Type I error rates appears to be quite large for relatively small values of σ_1/σ_2, especially when sample sizes are large. This fact has significant practical implications. As an example, suppose a two-sample t test is performed on test scores with a truncated

normal distribution with a ceiling at $\mu + .5\sigma$. Also, suppose that the sample sizes are 100 or more and that $\sigma_1/\sigma_2 = 1.1$, a ratio that many researchers would consider inconsequential. Or, imagine the same sample sizes and ratio of standard deviations characterizing response times having an exponential distribution.

Table 3 provides an indication of what can happen with these combinations of slight, almost insignificant variance heterogeneity and relatively large sample sizes with the above distributions and also with Weibull distribution. Although many researchers would consider the ratio $\sigma_1/\sigma_2 = 1.1$ to be no cause for concern, it is obvious from the table that the inflation of Type I error rates is sizeable for all the above distributions. Even a smaller ratio, $\sigma_1/\sigma_2 = 1.05$, produces considerable inflation.

It should be emphasized again that σ_1/σ_2 is a ratio of population standard deviations. It is quite possible for s_1/s_2 , the ratio of sample standard deviations, to be close to 1.0, although σ_1/σ_2 is 1.1, 1.2, or larger, because of sampling variability. For this reason, a researcher inspecting sample data may not detect inequality of variances when it exists, and, if sample sizes are sufficiently large, a Wilcoxon-Mann-Whitney test or rank-transformation test may lead to an incorrect statistical decision with high probability. As long as there is uncertainty about characteristics of the population distribution in a research study, together with doubt about possible heterogeneity of variance, a rank-based test entails risk, and the risk increases as sample size increases.

SOME PRACTICAL RECOMMENDATIONS

The practice of using t tests and ANOVA F tests when variances are unequal cannot be justified by claims that the tests are "robust," not even if sample sizes are equal. Type I error rates are altered by heterogeneous variances when sample sizes are unequal, and the same is true when sample sizes are equal, if the population distribution is skewed. Moreover, the nonparametric counterparts of these significance tests have the same limitations and often lead to worse results. It is well established that nonparametric methods based on ranks can protect against non-normality, but they are not a solution to variance heterogeneity.

Table 3. Probability of rejecting H_0 for slight variance heterogeneity (σ_1/σ_2 = 1.05 and 1.10) combined with large sample sizes (100 and 200)—distribution with ceiling at $\mu+.5\sigma$, exponential distribution, and Weibull distribution.

			t		t on ranks		W	
			N_1, N_2		N_1, N_2		N_1, N_2	
distribution	σ_1/σ_2	α	100	200	100	200	100	200
ceiling at $\mu+.5\ \sigma$	1.05	.01	.010	.010	.093	.207	.090	.201
		.05	.050	.049	.239	.420	.236	.410
		.10	.100	.100	.347	.540	.345	.534
	1.10	.01	.010	.010	.103	.232	.101	.227
		.05	.049	.050	.256	.447	.256	.444
		.10	.098	.100	.367	.570	.368	.568
exponential	1.05	.01	.010	.010	.013	.016	.014	.016
		.05	.049	.051	.059	.068	.062	.069
		.10	.100	.102	.114	.127	.117	.128
	1.10	.01	.010	.009	.021	.034	.021	.035
		.05	.049	.050	.082	.116	.083	.120
		.10	.100	.099	.148	.194	.149	.198
Weibull	1.05	.01	.007	.007	.120	.268	.116	.270
		.05	.045	.047	.288	.501	.286	.503
		.10	.100	.100	.403	.624	.401	.625
	1.10	.01	.006	.008	.323	.656	.316	.648
		.05	.045	.046	.553	.841	.553	.836
		.10	.099	.098	.672	.906	.671	.900

In the case of normal distributions and some non-normal distributions, substitution of a separate-variances significance test, such as the Welch (1938) or Satterthwaite (1946) versions of the t test, often has favorable results. However, a decision to substitute an alternative test cannot reasonably be made on the basis of a preliminary test of equality of variances. If a separate-variance test is to be used, it is more efficient to perform it unconditionally whenever sample sizes are unequal.

Of course, it is prudent to make certain that sample sizes are equal in research designs whenever possible. However, this course of action does not unfailingly protect the significance level when the population distribution is not known. One cannot confidently infer population parameters like variance and skewness from sample data alone, especially if samples are

small. If one is using a well-established measurement technique to obtain large samples from a known distribution, uncertainly about parameters may be minimal. In other cases, like those identified in the present study, researchers should recognize that familiar, widely used statistical methods may not be suitable. Alternative procedures with improved properties have been described by Keselman, Cribbie, and Zumbo (1997), Keselman, Wilcox, Othman, and Fradette (2001), Wei-ming (1999), and Wilcox (2001, 2003). These computer-intensive bootstrap, trimming, and related methods are promising, but still unfamiliar to many researchers and not yet routinely included in textbooks and statistical software packages.

REFERENCES

Albers, W., Boon, P.C., & Kallenberg, W.C.M. (2000). The asymptotic behavior of tests for normal means based on a variance pre-test. *Journal of Statistical Planning and Inference, 88,* 47-57.

Alexander, R.A., & Govern, D.M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics, 19,* 91-101.

Byrk, A.S., & Raudenbush, S.J. (1987). Heterogeneity of variance in experimental studies: A challenge to conventional interpretations. *Psychological Bulletin, 104,* 396-404.

Chen, S.Y., & Chen, H.J. (1998). Single-stage analysis of variance under heteroscedasticity. *Communications in Statistics—Simulation and Computation, 27,* 641-666.

Cressie, N.A.C., & Whitford, H.J. (1986). How to use the two sample t-test. *Biomedical Journal, 28,* 131-148.

Evans, M., Hastings, N., & Peacock, B. (1993). *Statistical distributions* (2nd ed).

Harwell, M.R. (1990). Summarizing Monte Carlo results in methodological research. *Journal of Educational Statistics, 17,* 297-313.

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.

Keselman, H.J., Cribbie, R., & Zumbo, B.D. (1997). Specialized tests for detecting treatment effects in the two-sample problem. *Journal of Experimental Education, 65,* 355-366.

Keselman, H.J., Wilcox, R.R., Othman, A.R., & Fradette, K. (2002). Trimming, transforming statistics, and bootstrapping: Circumventing the biasing effects of heteroscedasticity and nonnormality, *Journal of Modern Applied Statistical Methods, 1,* 288-309.

Marsaglia, G., & Bray, T.A. (1964). A convenient method for generating normal variables. *SIAM Review, 6,* 260-264.

Marsaglia, G., Zaman, A., & Tsang, W.W. (1990). Toward a universal random number generator. *Statistics & Probability Letters, 8*, 35-39.

Mendes, M. (2003). The comparison of Levene, Bartlett, Neyman-Pearson, and Bartlett 2 tests in terms of actual Type I error rates. *Journal of Agricultural Sciences, 9 (2),* 143-146.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 155-166.

Miller, R.C. (1986). *Beyond ANOVA: Basics of applied statistics.* Boca Raton, FL: Chapman & Hall.

Nelson, L.S. (2000). Comparing two variances from normal populations. *Journal of Quality Technology, 32,* 79-80.

Overall, J.E., Atlas, R.S., & Gibson, J.M. (1995). Tests that are robust against variance heterogeneity in k 2 designs with unequal cell frequencies. *Psychological Reports, 76,* 1011-1017.

Pashley, P.J. (1993). On generating random sequences. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 395-415). Hillsdale, NJ: Lawrence Erlbaum Associates.

Patel, J.K., Kapadia, C.H., & Owen, D.B. (1976). *Handbook of statistical distributions.* New York: Marcel Dekker.

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.

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

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

Schneider, P.J., & Penfield, D.A. (1997). Alexander and Govern's approximation: Providing an alternative to ANOVA under variance heterogeneity. *Journal of Experimental Education, 65,* 271-286.

Tan, W.Y. (1982). Sampling distributions and robustness of t, F, and variance-ratio in two samples and ANOVA models with respect to departure from normality. *Communications in Statistics, A11,* 2485-2511.

Weerahandi, S. (1995). ANOVA under unequal error variances. *Biometrics, 51,* 589-599.

Wei-ming, L. (1999). Developing trimmed mean test statistics for two-way fixed effects ANOVA models under variance heterogeneity and nonnormality. *The Journal of Experimental Education, 67,* 243-264.

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

Westfall, P.H., & Young, S.S. (1993). *Resampling based multiple testing.* New York: Wiley.

Wilcox, R.R. (2001). *Fundamentals of modern statistical methods: Substantially increasing power and accuracy.* New York: Springer-Verlag.

Wilcox, R.R. (2003). *Applying contemporary statistical techniques.* San Diego: Academic Press.

Wludyka, P., & Nelson, P.R. (1999). Two nonparametric analysis-of-means type tests for homogeneity of variances. *Journal of Applied Statistics, 26,* 243-256.

Zimmerman, D.W., & Zumbo, B.D. (1993a). Rank transformations and the power of the Student t test and Welch $\underline{t}'$ test for non-normal populations with unequal variances. *Canadian Journal of Experimental Psychology, 47*, 523-539.

Zimmerman, D.W., & Zumbo, B.D. (1993b). The relative power of parametric and nonparametric statistical methods. *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 481-517). Hillsdale, NJ: Lawrence Erlbaum Associates.

Zimmerman, D.W. (in press). Conditional probabilities of rejecting H_0 by pooled and separate variances $\underline{t}$ tests given heterogeneity of sample variances. *Communications in Statistics—Simulation and Computation.*

Zimmerman, D.W. (in press). A note on preliminary tests of equality of variances. *British Journal of Mathematical and Statistical Psychology.*

APPENDIX

Generation of Variates with Specified Continuous and Discrete Distributions.

Let U be a unit rectangular variate and Y a unit normal variate. The various continuous and discrete distributions in the study were obtained by the transformations in Table A1. After these transformations, X was normalized by subtracting the mean and dividing by the standard deviation, so that all variates had $\mu = 0$ and $\sigma = 1$. For further discussion of generation of variates, see Evans, Hastings, & Peacock (1993) and Patel, Kapadia, & Owen (1976).

Table A1.

Distribution	Transformation
exponential $\lambda = 1$	$X = -\log(U)$
lognormal shape parameter 1 scale parameter 1	$X = \exp(Y)$
lognormal shape parameter 1 scale parameter .6	$X = \exp(.6Y)$
mixed-normal (.05, 20)	X is N(0,1) with probability .95 and N(0, 20) with probability .05.
mixed-normal (.02, 10)	X is N(0,1) with probability .98 and N(0,10) with probability .02.
mixed-normal (.01, 20)	X is N(0,1) with probability .99 and N(0,20) with probability .01.
gamma shape parameter 2	$X = \sum_{i=1}^{2} -\log(U_i)$
gamma shape parameter 5	$X = \sum_{i=1}^{5} -\log(U_i)$

Table A1 (continued).

Distribution	Transformation
Laplace (double-exponential) location parameter 0 scale parameter 1	$X = \log(U_1/U_2)$, where U_1 and U_2 are unit rectangular.
logistic location parameter 0 scale parameter 1	$X = \log[U/(1 - U)]$
logistic location parameter 0 scale parameter .5	$X = .5 \log[U/(1 - U)]$
logistic location parameter 0 scale parameter 2	$X = 2 \log[U/(1 - U)]$
Gumbel (extreme value) location parameter 0 scale parameter 1	$X = -\log(-\log U)$
power function shape parameter .5 scale parameter 1	$X = U^2$
power function shape parameter 3 scale parameter 1	$X = U^{1/3}$
bimodal, 1 σ between modes	$N(0,1)$ with probability .5 and $N(1,1)$ with probability .5
bimodal, 2 σ between modes	$N(0,1)$ with probability .5 and $N(2,1)$ with probability .5

Table A1 (continued).

Distribution	Transformation
Weibull shape parameter .5 scale parameter 1	$X = [-\log(U)]^2$
Weibull shape parameter 3 scale parameter 1	$X = [-\log(U)]^{1/3}$
Weibull shape parameter 2 scale parameter 1	$X = [-\log(U)]^{1/2}$
ceiling at $\mu + 1.5\ \sigma$	$X = N(0,1)$, scores exceeding $\mu + 1.5\sigma$ replaced by $\mu + 1.5\sigma$.
ceiling at $\mu + .5\ \sigma$	$X = N(0,1)$ scores exceeding $\mu + .5\sigma$ replaced by $\mu + .5\sigma$.
triangular, range (0,2)	$X = (U_1 + U_2)/2$

(Manuscript received: 3 July 2003; accepted: 27 October 2003)