UNDERSTANDING STATISTICS, 2(4), 267–280

TEACHING ARTICLE

A Warning About the Large-Sample Wilcoxon–Mann–Whitney Test

Donald W. Zimmerman
Department of Psychology
Carleton University

It is known that the Wilcoxon–Mann–Whitney test is strongly influenced by unequal variances of treatment groups combined with unequal sample sizes. This simulation study indicates that, for various continuous and discrete distributions, the discrepancy between the empirical Type I error rate and the nominal significance level is large even when sample sizes are equal. In some cases, it exceeds the similar discrepancy characteristic of the Student *t* test. Furthermore, for some distributions, the discrepancy becomes increasingly more extreme as sample sizes increase. When sample sizes are relatively large, so that the normal-approximation form of the Wilcoxon–Mann–Whitney statistic is appropriate, minor and usually undetected differences in variability of treatment groups can substantially inflate the Type I error rate. For several distributions, including some that occur frequently in psychological research, ratios of population standard deviations as small as 1.1 or 1.2 have sizeable effects.

Keywords: Wilcoxon–Mann–Whitney test, Student *t* test, Type I error, homogeneity of variance, large-sample normal approximation

The validity of widely used significance tests of differences in location, such as the Student *t* test and the analysis of variance *F* test, depends on an assumption of homogeneity of variance of treatment groups. When this assumption is violated and at the same time sample sizes are unequal, Type I error rates are substantially modified. When a larger variance is associated with a larger sample size, the probability of a Type I error declines below the nominal significance level, and when a larger variance is associated with a smaller sample size, the probability increases, sometimes far above the significance level (Hsu, 1938; Overall, Atlas, & Gibson, 1995; Scheffé, 1959).

Requests for reprints should be sent to Donald W. Zimmerman, 1978 134A Street, Surrey, British Columiba V4A 6B6 Canada. E-mail: zimmerma@look.ca

Quite some time ago, Cohen (1965) warned psychologists that nonparametric tests are not tests without assumptions and that some assumptions of these tests are strong. In recent years, it has become apparent that nonparametric tests of location, including the Wilcoxon–Mann–Whitney rank-sum test and the van der Waerden, or normal-scores test, are seriously influenced by unequal variances of treatment groups, although the changes typically are not as large as those of their parametric counterparts (Zimmerman, 1996; Zimmerman & Zumbo, 1993). It is still widely believed that both parametric and nonparametric significance tests are robust to variance heterogeneity when sample sizes are equal. For normal distributions, this belief perhaps is justified in the case of the Student *t* test, although slight modifications of the nominal significance level do occur.

In this study, I disclose that for various non-normal distributions, the outcome for the Wilcoxon–Mann–Whitney test is not the same. For these non-normal distributions, variance heterogeneity inflates the Type I error rate considerably, even when sample sizes are equal. In addition, the magnitude of this inflation increases as sample size increases. When sample sizes are relatively large so that the normal-approximation form of the Wilcoxon–Mann–Whitney test statistic is appropriate, minor differences in the variability of treatment groups, even seemingly insignificant ratios of standard deviations of 1.1 or 1.2, produce sizeable changes in the Type I error rate.

GENERATION OF VARIATES WITH PREDETERMINED DISTRIBUTIONS[1]

The following continuous probability densities were included in this study: normal, exponential, lognormal, gamma, truncated normal (half-normal), Weibull, and power function. In addition, several distributions of the type identified by Micceri (1989), which often occur in psychological research, were examined. A geometric distribution with probability of success .1 and a skewed binomial distribution with 20 trials and probability of success .05 were included. Also, an asymmetric bimodal distribution consisted of a mixture of two normal distributions with different means. Finally, two normal distributions were modified to have "ceilings"—one at .5 σ above the mean and one at 1.0 σ above the mean—in which all scores above a selected cutoff value were replaced by that cutoff value. Exponential and geometric distributions are encountered in measurement of response times in psychological research, and distributions with ceilings often characterize test scores. All these distributions are skewed.

The random number generator used in this study was introduced by Marsaglia, Zaman, and Tsang (1990) and was described by Pashley (1993, pp. 395–415). Nor-

[1]The computer programs in this study were written in PowerBASIC, Version 3.5, PowerBASIC, Inc., Carmel, California. Listings of the programs can be obtained by writing to Donald W. Zimmerman at 1978 134A Street, Surrey, British Columbia, V4A 6B6, Canada.

mal variates, $N(0,1)$, were generated by the rejection method of Marsaglia and Bray (1964), and these were transformed to have the various distribution shapes mentioned previously. 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 inverse distribution functions given in Table 1. After these transformations, scores were standardized by subtracting the mean and dividing by the standard deviation, so that $\mu = 0$ and $\sigma = 1$. For further discussion of generation of variates, see Evans, Hastings, and Peacock (1993) and Patel, Kapadia, and Owen (1976).

Each replication of the sampling procedure obtained two independent samples of the same size. In successive replications, all scores in one sample were multiplied by a constant so that the ratio σ_1/σ_2 had a predetermined value. The Wilcoxon–Mann–Whitney rank-sum test was performed, and the result was evaluated at the .01, .05, and .10 significance levels. In the small-sample case (Table 2), the critical values of the Wilcoxon statistic were based on exact probabilities obtained from standard tables chosen to be as close as possible to the .01, .05, and .10 levels. As a comparison, for some distributions, the Student t test was evaluated at the same modified significance levels.

In the large-sample case (Table 3 and Figures 1 and 2), the normal-approximation form of the Wilcoxon–Mann–Whitney test was performed, again using the .01, .05, and .10 significance levels. Also, in the case of the data in Figures 1 and 2, an independent samples Student t test was performed, and the result was evaluated at the same significance levels.

The ratio of population standard deviations, σ_1/σ_2, was either 1.0, 1.1, or 1.2. Sensitivity to these small differences is a serious matter. In one part of the study (the data in Tables 2 and 3), sample sizes were 6, 8, or 10. In another part of the

TABLE 1
Generation of Variates With Specified Distributions

Distribution	*Transformation*
Exponential, $\lambda = 1$	$X = -\log(U)$
Lognormal, shape parameter 1, scale parameter 1	$X = \exp(Y)$, where Y is $N(0, 1)$
Gamma, shape parameter 2	
Power function, shape parameter .5	$X = U^2$
Geometric, $p = .1$	$X = [\log(U)/\log(.9)] - 1$
Weibull, shape parameter .5, scale parameter 1	$X = [-\log(U)]^2$
Skewed binomial (20, .05)	Generation of $B(n, p)$, $n = 20$, $p = .05$
Asymmetric bimodal	$N(0, 1)$ with probability .7 and $N(3, 1)$ with probability .3
Normal, ceiling at $\mu + 1.0\ \sigma$	$X = N(0, 1)$, values exceeding $\mu + 1.0\ \sigma$ replaced by $\mu + 1.0\ \sigma$
Normal, ceiling at $\mu + .5\ \sigma$	$X = N(0, 1)$, values exceeding $\mu + .5\ \sigma$ replaced by $\mu + .5\ \sigma$
Truncated normal (half normal)	$X = \text{abs}(Y)$, where Y is $N(0, 1)$

study, using the large-sample normal approximation (the data in Table 3), sample sizes were 20, 30, 60, 90, 120, or 200. For the graphs plotted in Figures 1 and 2, sample sizes varied between 10 and 100 in increments of 10. All significance tests were nondirectional, and there were 50,000 replications of the sampling procedure for each combination of parameters.

RESULTS OF SIMULATIONS

Table 2 provides Type I error rates of the small-sample Wilcoxon–Mann–Whitney test for the 12 population distributions and for ratios of population standard deviations of 1.0, 1.1, and 1.2. In all cases, $N_1 = N_2$, and the mutual sample size was 6, 8, or 10. The significance levels (the columns labeled α) were exact probabilities associated with the Wilcoxon statistic for various sample sizes. Critical values of the statistic were chosen so that these probabilities were close to the conventional .01, .05, and .10 significance levels.

It is apparent that, for various non-normal distributions, Type I error rates were inflated somewhat when σ_1/σ_2 was 1.1 or 1.2, although the change was not extreme. For example, for $N_1 = N_2 = 6$, the rates were slightly inflated for the

TABLE 2
Probability of Rejecting H_0 by Wilcoxon–Mann–Whitney Rank-Sum Test (Small Samples) for Various Sample Sizes, Ratios of Standard Deviations, and Significance Levels

Distribution	σ_1/σ_2	α	$N_1 = 6$, $N_2 = 6$	α	$N_1 = 8$, $N_2 = 8$	α	$N_1 = 10$, $N_2 = 10$
Normal	1.0	.009	.009	.010	.010	.009	.009
		.041	.040	.050	.049	.052	.053
		.093	.092	.105	.103	.105	.106
	1.1	.009	.008	.010	.009	.009	.009
		.041	.041	.050	.049	.052	.053
		.093	.095	.105	.102	.105	.105
	1.2	.009	.009	.010	.010	.009	.010
		.041	.043	.050	.050	.052	.054
		.093	.095	.105	.103	.105	.109
Exponential	1.0	.009	.009	.010	.010	.009	.010
		.041	.041	.050	.052	.052	.054
		.093	.094	.105	.106	.105	.106
	1.1	.009	.009	.010	.011	.009	.009
		.041	.043	.050	.052	.052	.055
		.093	.096	.105	.108	.105	.109
	1.2	.009	.010	.010	.012	.009	.012
		.041	.044	.050	.057	.052	.064
		.093	.100	.105	.118	.105	.121

(continued)

TABLE 2 (Continued)

Distribution	σ_1/σ_2	α	$N_1 = 6$, $N_2 = 6$	α	$N_1 = 8$, $N_2 = 8$	α	$N_1 = 10$, $N_2 = 10$
Lognormal	1.0	.009	.009	.010	.011	.009	009
		.041	.041	.050	.051	.052	.052
		.093	.093	.105	.106	.105	.104
	1.1	.009	.010	.010	.011	.009	.011
		.041	.044	.050	.055	.052	.059
		.093	.099	.105	.112	.105	.114
	1.2	.009	.011	.010	.014	.009	.014
		.041	.050	.050	.063	.052	.071
		.093	.111	.105	.127	.105	.133
Gamma	1.0	.009	.009	.010	.010	.009	.009
		.041	.042	.050	.049	.052	.054
		.093	.093	.105	.104	.105	.107
	1.1	.009	.009	.010	.011	.009	.009
		.041	.042	.050	.050	.052	.052
		.093	.095	.105	.104	.105	.105
	1.2	.009	.009	.010	.011	.009	.010
		.041	.043	.050	.052	.052	.056
		.093	.096	.105	.107	.105	.110
Power function	1.0	.009	.008	.010	.011	.009	.009
		.041	.041	.050	.050	.052	.052
		.093	.094	.105	.105	.105	.104
	1.1	.009	.010	.010	.012	.009	.009
		.041	.045	.050	.054	.052	.056
		.093	.099	.105	.111	.105	.111
	1.2	.009	.011	.010	.014	.009	.011
		.041	.047	.050	.060	.052	.063
		.093	.105	.105	.120	.105	.124
Geometric	1.0	.009	.009	.010	.011	.009	.010
		.041	.042	.050	.051	.052	.052
		.093	.094	.105	.108	.105	.103
	1.1	.009	.009	.010	.010	.009	.010
		.041	.043	.050	.052	.052	.056
		.093	.094	.105	.109	.105	.107
	1.2	.009	.010	.010	.013	.009	.013
		.041	.046	.050	.059	.052	.065
		.093	.104	.105	.121	.105	.125
Weibull	1.0	.009	.008	.010	.010	.009	.009
		.041	.040	.050	.051	.052	.051
		.093	.091	.105	.105	.105	.104
	1.1	.009	.016	.010	.023	.009	.022
		.041	.066	.050	.092	.052	.104
		.093	.137	.105	.169	.105	.181
	1.2	.009	.027	.010	.044	.009	.049
		.041	.095	.050	.137	.052	.168
		.093	.183	.105	.234	.105	.264

(continued)

TABLE 2 (Continued)

Distribution	σ_1/σ_2	α	$N_1 = 6$, $N_2 = 6$	α	$N_1 = 8$, $N_2 = 8$	α	$N_1 = 10$, $N_2 = 10$
Skewed binomial, $n = 20$, $p = .05$	1.0	.009	.009	.010	.011	.009	.009
		.041	.042	.050	.048	.052	.054
		.093	.094	.105	.102	.105	.108
	1.1	.009	.013	.010	.017	.009	.016
		.041	.052	.050	.068	.052	.076
		.093	.113	.105	.136	.105	.140
	1.2	.009	.013	.010	.017	.009	.016
		.041	.051	.050	.068	.052	.078
		.093	.114	.105	.133	.105	.141
Asymmetric bimodal	1.0	.009	.009	.010	.010	.009	.008
		.041	.042	.050	.050	.052	.052
		.093	.093	.105	.105	.105	.104
	1.1	.009	.009	.010	.010	.009	.009
		.041	.041	.050	.050	.052	.051
		.093	.095	.105	.104	.105	.105
	1.2	.009	.010	.010	.011	.009	.010
		.041	.043	.050	.050	.052	.057
		.093	.096	.105	.104	.105	.110
Normal, ceiling at 1.0 σ	1.0	.009	.008	.010	.011	.009	.009
		.041	.040	.050	.049	.052	.052
		.093	.092	.105	.105	.105	.104
	1.1	.009	.009	.010	.011	.009	.010
		.041	.043	.050	.051	.052	.056
		.093	.097	.105	.108	.105	.110
	1.2	.009	.009	.010	.012	.009	.010
		.041	.042	.050	.053	.052	.058
		.093	.097	.105	.111	.105	.112
Normal, ceiling at .5 σ	1.0	.009	.009	.010	.010	.009	.008
		.041	.041	.050	.049	.052	.051
		.093	.091	.105	.104	.105	.102
	1.1	.009	.012	.010	.016	.009	.015
		.041	.052	.050	.070	.052	.076
		.093	.112	.105	.136	.105	.141
	1.2	.009	.013	.010	.017	.009	.016
		.041	.054	.050	.071	.052	.081
		.093	.118	.105	.139	.105	.146
Truncated normal (half-normal)	1.0	.009	.009	.010	.010	.009	.009
		.041	.043	.050	.050	.052	.052
		.093	.093	.105	.106	.105	.106
	1.1	.009	.009	.010	.010	.009	.010
		.041	.042	.050	.051	.052	.053
		.093	.095	.105	.108	.105	.107
	1.2	.009	.009	.010	.011	.009	.010
		.041	.044	.050	.054	.052	.057
		.093	.097	.105	.111	.105	.112

lognormal distribution and the truncated normal distributions and were inflated to a greater extent for the Weibull distribution. The same pattern held for these distributions when $N_1 = N_2 = 8$. In all these cases, one will see, the change was far greater when sample sizes were larger. For almost all distributions in Table 2, the probability of a Type I error increased systematically from row to row in each section representing a particular distribution (as the ratio σ_1/σ_2 increased) and also increased across columns (as the sample size increased).

For comparison, Table 3 shows results of the Student *t* test using sample sizes of 6, 8, and 10 at significance levels corresponding to the same probabilities associated with the small-sample Wilcoxon–Mann–Whitney test. For $\sigma_1/\sigma_2 = 1.0$, the results for these non-normal distributions was similar to familiar findings: The Type I error probabilities declined slightly below the nominal significance level. For $\sigma_1/\sigma_2 = 1.1$ and $\sigma_1/\sigma_2 = 1.2$, there were only slight increases, less than those of the Wilcoxon–Mann–Whitney test. Furthermore, there were only slight increases attributable to sample size.

Table 4 provides Type I error rates of the large-sample, normal-approximation form of the Wilcoxon–Mann–Whitney test for significance levels of .01, .05, and .10; for the same ratios of standard deviations included in Table 2; and for sample sizes of 20, 30, 60, 90, 120, and 200. For many non-normal distributions, the probability of rejecting H_0 increased far above the nominal significance level. Again, probabilities of rejecting H_0 increased systematically from row to row and across columns in the table—that is, the magnitude of the increase was a function of both the ratio of standard deviations and the sample size. Although the ratios deviated from 1.0 only slightly, to a degree that most likely would be ignored in research, the changes in Type I error rates were sizeable, especially for the larger sample sizes.

Figures 1 and 2 present a more detailed picture of the dependence of the change in Type I error rate on sample size for the exponential, lognormal, and binomial distributions and for the normal distribution with a ceiling. In these graphs, the sample sizes varied from 10 to 100 in increments of 10. The ratio σ_1/σ_2 was 1.1 in all cases. For comparison, the graphs also plot the Type I error rates of the Student *t* test. The *t* test was not influenced by either variance heterogeneity or sample size in the case of the binomial distribution and the normal distribution with a ceiling and was only slightly influenced in the case of the exponential and lognormal distributions. However, it appears that the Type I error rate of the Wilcoxon–Mann–Whitney test is a linear function of sample size over this range of sample sizes.

PRACTICAL IMPLICATIONS FOR STATISTICAL SIGNIFICANCE TESTING

For all the non-normal distributions included in this study, the Type I error rates of the Wilcoxon–Mann–Whitney test were close to the .01, .05, and .10 significance levels when σ_1/σ_2 was 1.0. This outcome is hardly surprising in light of the fact that

TABLE 3

Probability of Rejecting H_0 by Student t Test (Small Samples), for Various Sample Sizes, Ratios of Standard Deviations, and Significance Levels

Distribution	σ_1/σ_2	α	$N_1 = 6$, $N_2 = 6$	α	$N_1 = 8$, $N_2 = 8$	α	$N_1 = 10$, $N_2 = 10$
Weibull	1.0	.009	.002	.010	.002	.009	.002
		.041	.021	.050	.024	.052	.026
		.093	.068	.105	.076	.105	.080
	1.1	.009	.002	.010	.002	.009	.002
		.041	.025	.050	.027	.052	.028
		.093	.075	.105	.080	.105	.084
	1.2	.009	.006	.010	.005	.009	.005
		.041	.036	.050	.034	.052	.035
		.093	.089	.105	.092	.105	.093
Exponential	1.0	.009	.006	.010	.007	.009	.006
		.041	.039	.050	.042	.052	.042
		.093	.095	.105	.101	.105	.099
	1.1	.009	.006	.010	.006	.009	.007
		.041	.041	.050	.043	.052	.046
		.093	.097	.105	.100	.105	.106
	1.2	.009	.007	.010	.008	.009	.008
		.041	.043	.050	.045	.052	.046
		.093	.101	.105	.101	.105	.106
Lognormal	1.0	.009	.004	.010	.004	.009	.004
		.041	.031	.050	.032	.052	.034
		.093	.084	.105	.088	.105	.090
	1.1	.009	.005	.010	.005	.009	.004
		.041	.033	.050	.034	.052	.034
		.093	.086	.105	.088	.105	.092
	1.2	.009	.005	.010	.006	.009	.006
		.041	.037	.050	.037	.052	.040
		.093	.092	.105	.091	.105	.098
Normal, ceiling at .5 σ	1.0	.009	.010	.010	.010	.009	.009
		.041	.048	.050	.049	.052	.048
		.093	.102	.105	.104	.105	.104
	1.1	.009	.010	.010	.010	.009	.010
		.041	.049	.050	.049	.052	.049
		.093	.105	.105	.104	.105	.106
	1.2	.009	.011	.010	.011	.009	.011
		.041	.052	.050	.050	.052	.051
		.093	.106	.105	.106	.105	.105

TABLE 4
Probability of Rejecting H_0 by Wilcoxon–Mann–Whitney Rank-Sum Test Using Large-Sample Normal Approximation, for Various Sample Sizes, Ratios of Standard Deviations, and Significance Levels

Distribution	σ_1/σ_2	α	$N_1 = 20$, $N_2 = 20$	$N_1 = 30$, $N_2 = 30$	$N_1 = 60$, $N_2 = 60$	$N_1 = 90$, $N_2 = 90$	$N_1 = 120$, $N_2 = 120$	$N_1 = 200$, $N_2 = 200$
Normal	1.0	.010	.009	.009	.010	.010	.010	.010
		.050	.048	.050	.051	.050	.050	.050
		.100	.101	.101	.101	.101	.101	.100
	1.1	.010	.009	.010	.010	.011	.010	.010
		.050	.049	.051	.050	.050	.048	.051
		.100	.100	.101	.101	.102	.100	.101
	1.2	.010	.009	.009	.010	.010	.010	.010
		.050	.050	.051	.050	.051	.051	.051
		.100	.104	.100	.101	.101	.102	.102
Exponential	1.0	.010	.009	.009	.011	.010	.010	.010
		.050	.049	.049	.050	.051	.050	.050
		.100	.102	.100	.099	.101	.100	.100
	1.1	.010	.011	.012	.017	.020	.023	.035
		.050	.056	.060	.070	.081	.090	.118
		.100	.111	.115	.129	.145	.158	.196
	1.2	.010	.015	.019	.031	.046	.059	.106
		.050	.069	.085	.112	.148	.177	.266
		.100	.132	.149	.190	.237	.273	.379
Lognormal	1.0	.010	.009	.008	.010	.010	.010	.010
		.050	.049	.048	.050	.049	.049	.050
		.100	.102	.098	.101	.099	.101	.101
	1.1	.010	.012	.015	.022	.030	.036	.060
		.050	.059	.068	.085	.105	.122	.174
		.100	.119	.126	.153	.179	.202	.271
	1.2	.010	.020	.029	.059	.090	.125	.231
		.050	.088	.110	.173	.236	.299	.452
		.100	.160	.184	.270	.345	.417	.578
Gamma	1.0	.010	.009	.009	.009	.010	.011	.010
		.050	.049	.049	.049	.050	.051	.051
		.100	.101	.100	.100	.100	.102	.101
	1.1	.010	.009	.010	.012	.014	.013	.017
		.050	.051	.052	.055	.059	.062	.070
		.100	.106	.104	.109	.113	.119	.129
	1.2	.010	.012	.012	.016	.020	.023	.036
		.050	.057	.060	.071	.080	.089	.120
		.100	.115	.116	.133	.144	.156	.197
Power function	1.0	.010	.009	.010	.009	.010	.010	.010
		.050	.049	.051	.050	.049	.050	.050
		.100	.103	.099	.100	.100	.101	.098
	1.1	.010	.011	.012	.016	.021	.026	.038
		.050	.057	.061	.070	.084	.093	.124
		.100	.115	.115	.132	.149	.163	.207

(continued)

TABLE 4 (Continued)

Distribution	σ_1/σ_2	α	$N_1 = 20$, $N_2 = 20$	$N_1 = 30$, $N_2 = 30$	$N_1 = 60$, $N_2 = 60$	$N_1 = 90$, $N_2 = 90$	$N_1 = 120$, $N_2 = 120$	$N_1 = 200$, $N_2 = 200$
	1.2	.010	.014	.017	.028	.037	.049	.086
		.050	.067	.077	.104	.128	.155	.224
		.100	.130	.140	.174	.210	.242	.328
Geometric	1.0	.010	.009	.009	.010	.009	.010	.010
		.050	.049	.049	.050	.049	.049	.050
		.100	.102	.099	.101	.100	.099	.100
	1.1	.010	.011	.012	.014	.016	.018	.027
		.050	.055	.058	.063	.071	.076	.097
		.100	.111	.111	.121	.133	.137	.168
	1.2	.010	.016	.019	.035	.049	.069	.120
		.050	.072	.085	.121	.155	.193	.289
		.100	.139	.151	.199	.247	.295	.406
Weibull	1.0	.010	.009	.008	.009	.010	.009	.010
		.050	.048	.047	.048	.048	.048	.048
		.100	.101	.097	.099	.099	.100	.098
	1.1	.010	.046	.073	.175	.278	.393	.650
		.050	.156	.210	.372	.508	.631	.842
		.100	.253	.315	.494	.634	.740	.907
	1.2	.010	.106	.173	.399	.605	.757	.947
		.050	.275	.377	.632	.809	.904	.986
		.100	.396	.495	.740	.880	.945	.994
Skewed binomial, $n = 20$, $p = .05$	1.0	.010	.008	.009	.010	.010	.010	.010
		.050	.049	.050	.051	.050	.049	.050
		.100	.103	.100	.101	.100	.100	.100
	1.1	.010	.020	.027	.045	.068	.091	.159
		.050	.085	.101	.144	.192	.233	.343
		.100	.153	.172	.227	.291	.338	.462
	1.2	.010	.020	.027	.047	.067	.091	.159
		.050	.085	.102	.144	.187	.231	.345
		.100	.156	.173	.232	.285	.335	.461
Asymmetric bimodal	1.0	.010	.009	.010	.009	.010	.010	.010
		.050	.049	.050	.050	.050	.050	.049
		.100	.101	.100	.101	.099	.100	.100
	1.1	.010	.009	.010	.010	.011	.010	.012
		.050	.052	.052	.053	.051	.052	.056
		.100	.106	.102	.105	.102	.103	.109
	1.2	.010	.009	.011	.011	.012	.013	.016
		.050	.052	.054	.055	.059	.062	.071
		.100	.106	.105	.108	.111	.118	.131
Normal, ceiling at 1.0 σ	1.0	.010	.008	.009	.009	.010	.010	.011
		.050	.049	.050	.050	.050	.051	.052
		.100	.102	.100	.099	.101	.101	.102

(continued)

TABLE 4 (Continued)

Distribution	σ_1/σ_2	α	$N_1 = 20$, $N_2 = 20$	$N_1 = 30$, $N_2 = 30$	$N_1 = 60$, $N_2 = 60$	$N_1 = 90$, $N_2 = 90$	$N_1 = 120$, $N_2 = 120$	$N_1 = 200$, $N_2 = 200$
	1.1	.010	.010	.010	.013	.015	.017	.022
		.050	.053	.054	.060	.067	.071	.085
		.100	.108	.107	.116	.125	.132	.151
	1.2	.010	.011	.012	.015	.017	.021	.028
		.050	.054	.058	.066	.072	.079	.101
		.100	.110	.113	.125	.134	.143	.172
Normal, ceiling at .5 σ	1.0	.010	.009	.009	.010	.010	.010	.010
		.050	.050	.050	.051	.050	.051	.049
		.100	.104	.101	.101	.100	.101	.100
	1.1	.010	.021	.032	.058	.091	.125	.228
		.050	.091	.113	.175	.234	.295	.447
		.100	.167	.189	.269	.342	.407	.572
	1.2	.010	.026	.034	.069	.106	.150	.268
		.050	.101	.124	.194	.263	.334	.494
		.100	.177	.205	.296	.376	.453	.613
Truncated normal (half-normal)	1.0	.010	.010	.009	.010	.010	.010	.010
		.050	.051	.050	.052	.049	.050	.048
		.100	.105	.100	.102	.099	.100	.098
	1.1	.010	.010	.010	.012	.013	.013	.016
		.050	.051	.051	.058	.059	.063	.068
		.100	.105	.104	.111	.114	.118	.127
	1.2	.010	.011	.013	.015	.019	.023	.033
		.050	.056	.060	.068	.076	.087	.114
		.100	.114	.115	.128	.140	.153	.191

the Wilcoxon statistic is nonparametric distribution free. This well-established protection of the significance level by the test under violation of normality is not related to sample size and in fact holds when sample sizes of treatment groups are unequal.

On the other hand, it is apparent from this data that the Wilcoxon–Mann–Whitney test does not protect the significance level when variances are unequal. It is widely believed that the test is robust to variance heterogeneity when sample sizes are equal. For various skewed non-normal distributions, however, variance heterogeneity distorts the Type I error rates of the test even though sample sizes are equal. In this study, the effect is obvious despite relatively small differences in variability that typically would be regarded as unimportant. It is difficult to say how much deviation from the nominal significance level would be too large for the test to be considered robust, but for the parameters considered in this study, many deviations are far above the conservative cutoff values recommended by Bradley (1978).

It is notable that the magnitude of this distortion is a function of sample size. In circumstances where the large-sample, normal-approximation form of the Wilcoxon–Mann–Whitney test is recommended, the change in the Type I error

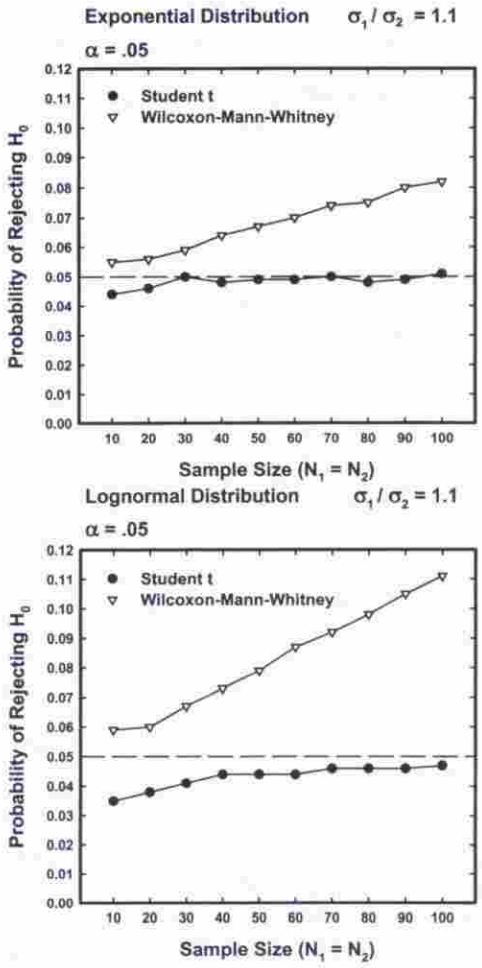

FIGURE 1 Probability of rejecting H_0 by the large-sample Wilcoxon–Mann–Whitney rank-sum test and the two-sample Student *t* test as a function of sample size (exponential and lognormal distributions).

rate can be extreme. For several non-normal distributions, there is extensive inflation if sample sizes are as large as 20 or 30, and ratios of standard deviations are 1.1 or 1.2. For these same distributions, the inflation exceeds all reasonable bounds when sample sizes are still larger.

In statistical significance testing, it is generally believed that large samples are desirable because the power of significance tests, both parametric and nonparametric, is an increasing function of sample size. However, these results indicate that increasing sample size can be counterproductive when assumptions of normality and homogeneity of variance are violated together. As a result of sampling error, it is possible for sample variances to be nearly equal or only slightly different when population variances are substantially different. From sample data

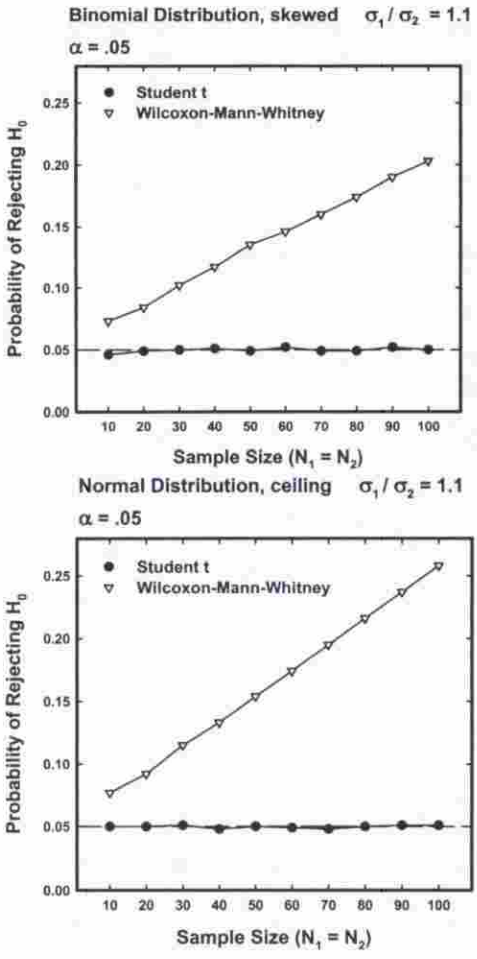

FIGURE 2 Probability of rejecting H_0 by the large-sample Wilcoxon–Mann–Whitney rank-sum test and the two-sample Student *t* test as a function of sample size (skewed binomial and truncated normal distributions).

alone, it would be impossible to rule out slight differences in population variability like those examined in this study. A researcher may be tempted to employ the Wilcoxon–Mann–Whitney test to counteract non-normality, erroneously believing that population variances are equal. In that scenario, a decision to employ larger samples in hopes of increasing the power of a significance test could lead to gross inflation of Type I error rates.

It is well known that neither the *t* test nor the Wilcoxon–Mann–Whitney test is robust to unequal variances combined with unequal sample sizes. The main practical message of this study is that the nonparametric Wilcoxon test is vulnerable to unequal variances even with equal sample sizes. Furthermore, slight differences can be crucial, and the distortion becomes greater as sample size increases. For this reason, the Wilcoxon test cannot be relied on as a nonparametric solution to

non-normality when at the same time there is a suspicion of unequal variances. In that case, other methods, such as the Welch (1938) test applied to the ranks of scores (Zimmerman & Zumbo, 1993) or the robust rank test of Fligner and Policello (1981), are recommended.

REFERENCES

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology, 31,* 144–152.

Cohen, J. (1965). Some statistical issues in psychological research. In B. B. Wolman (Ed.), *Handbook of clinical psychology* (pp. 95–121). New York: McGraw-Hill.

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

Fligner, M. A., & Policello, G. E., III. (1981). Robust rank procedures for the Behrens–Fisher problem. *Journal of the American Statistical Association, 76,* 162–168.

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.

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.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105,* 156–166.

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, Inc.

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

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

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

Zimmerman, D. W. (1996). A note on homogeneity of variance of scores and ranks. *Journal of Experimental Education, 64,* 351–362.

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