Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 25

11-1-2005

Power of the *t* Test for Normal and Mixed Normal Distributions

Marilyn S. Thompson
Arizona State University, m.thompson@asu.edu

Samuel B. Green
Arizona State University, samgreen@asu.edu

Yi-hsin Chen
Arizona State University

Shawn Stockford
Arizona State University

Wen-juo Lo
Arizona State University

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No.2, 591-597

1538 – 9472/05/$95.00

Power of the *t* Test for Normal and Mixed Normal Distributions

Marilyn S. Thompson Samuel B. Green Yi-hsin Chen Shawn Stockford Wen-juo Lo
Division of Psychology in Education
Arizona State University

Previous research suggests that the power of the independent-samples *t* test decreases when population distributions are mixed normal rather than normal, and that robust methods have superior power under these conditions. However, under some conditions, the power for the independent-samples *t* test can be greater when the population distributions for the independent groups are mixed normal rather than normal. The implications of these results are discussed.

Key words: *t* test, mixed normal, power

Introduction

The accepted belief in modern statistical practice is that the assumption of normality for parametric tests, such as the independent-samples t test and the analysis-of-variance F test, seldom, if ever, holds in practice. In psychology and education, Micceri (1989) offered empirical support for this conclusion. He examined over 400 large-sample data sets that included achievement and psychometric measures and found that they had a variety of shapes (e.g., skewed) and generally could not be described as normal.

For a number of years, violation of the normality assumption was not seen as a serious problem in that a number of studies showed that nonnormality, in and of itself, had a minimal effect on Type I error rate unless sample size is quite small (e.g., Boneau, 1960; Glass, Peckham, & Sanders, 1972; Ramsey, 1980; Rogan & Keselman, 1977).

Marilyn Thompson (m.thompson@asu.edu) and Samuel Green (samgreen@asu.edu) are faculty members in the Measurement, Statistics, and Methodological Studies program in the Division of Psychology in Education at Arizona State University. Yi-hsin Chen, Shawn Stockford, and Wen-juo Lo are graduate students in this program.

More recently, researchers have demonstrated that violation of the normality assumption may, however, have a deleterious effect on the power of parametric tests (e.g., MacDonald, 1999; Lix & Keselman, 1998; Wilcox, 1995). Based on these findings and others concerning violation of the homogeneity of variance assumption, Keselman, Wilcox, and Lix (2003) suggested that the application of standard parametric methods should be greatly restricted, and robust methods requiring minimal distributional assumptions should be used in their place. More specifically, they argued that robust methods, such as those using trimmed means and bootstrapping, are superior in terms of Type I and II error rates across a wide number of conditions encountered in practice.

The mixed normal distribution has been used extensively to illustrate the detrimental effect of nonnormality and specifically outliers on parametric tests and, most frequently, on the independent-samples t test (e.g., MacDonald, 1999; Wilcox, 1997, 2001). Based on these presentations, the independent-samples t test shows a dramatic decrease in power when the population distributions for the two independent groups are mixed normal rather than normal. A small-scale simulation may be used to illustrate the decrease in power found in these studies.

Consider the power of the independent-samples t test with 12 observations in each group under normal and mixed normal conditions. For the normal condition, data are generated from normal distributions with means

of 0 and 3 for first and second groups, respectively. The population variances are held constant across groups at 1. Based on 4000 replications, the empirically determined power is 1.00.

For the mixed normal condition, normal data are generated for each group from primary and secondary subpopulations with probabilities of .80 and .20, respectively. The means of the normal distributions for the primary and secondary subpopulations are identical to those under the normal condition: means of 0 for the first group and means of 3 for the second group. As in the normal condition, the variances for the primary distributions are set to 1 in both groups; however, the variances for the secondary distributions are set to 400 in both groups to simulate outliers. Based on 4000 replications, the empirical power is .21 under the mixed normal condition, much lower than the 1.00 found under the normal condition.

The explanation for these results and ones like them is that the standard error of the difference in means is much larger for the mixed normal distribution than for the normal distribution (e.g., Wilcox, 2001). For this example, the within-group variances increased from 1.00 for the normal condition to 80.80 for the mixed normal condition [i.e., combined across the primary and secondary distributions: .80 (1) + .20 (400) = 80.80], as a function of introducing the secondary distribution with a much larger variance (i.e., 400). Because the within-group variances increased for the mixed normal condition, the standard error of the difference in means increased, and the power decreased.

In the current Monte Carlo study, unexpected results were found when investigating the comparative power of the independent-samples t test under normal and mixed normal conditions. Conditions were included that were similar to those in previous research: the variances for the normal distributions were set equal to the variances of the primary distributions of the mixed normal distributions. In these conditions, the combined variances for the mixed normal distributions were greater due to the larger variances of the secondary distributions. However, different from previous studies, control conditions were included in which normal distributions had variances set equal to the combined variances in the mixed normal conditions. Presumably, the power of the independent-samples t test would be equivalent for the normal and mixed normal conditions if the population variances for the two conditions were equal and, thus, the standard errors of the difference in means were equal. However, the results of this study demonstrate the counterintuitive result that the power may be greater under the mixed normal condition.

Methodology

Data were generated using the normal pseudorandom number generator available in the IML procedure in SAS 8.2. Fifty-four conditions were created by manipulating four factors: the form of the population distribution, variances of these distributions, sample size, and mean differences.

Form of distributions. Data were generated for two independent groups from populations with normal or mixed normal distributions.

Variance. When the distributions were normal, the variances were equal to 1 for both groups or 80.8 for both groups. When the distributions were mixed normal, the variances for both groups were 1 for the normal distribution with a probability of .80 and 400 for the normal distribution with a probability of .20; therefore, the mixed normal distributions had a combined variance of 80.8.

Sample size. The total sample size (N) consisted of 24, 48, or 96 cases, with an equal number of cases in each of the two independent groups.

Mean differences. To evaluate the Type I error rates of the test statistics, data were generated such that the differences in population means were equal to zero. To assess power, data were generated so that the population mean for one group was zero, and the population mean for the second group was one of five values: 0.5, 1.0, 1.5, 3.0, or 4.5. For mixed normal distributions, the means of the primary and secondary distributions for any one group were always the same.

Data Analysis

Two-tailed independent-samples t tests were conducted using the ttest procedure within SAS 8.2 and evaluated at the .05 level. Four-thousand replications were generated for each of the 54 conditions. Empirical alphas were computed for the conditions in which the means were equivalent. Empirical powers were calculated as proportions of rejections of a false null hypothesis in the correct direction for conditions in which the means differed between groups.

In addition, empirical Type III error rates–proportions of rejections of a false null hypothesis in the wrong direction–were computed. However, Type III error rates were excluded from the discussion because they were strongly inversely related to power and were uniformly very low; Type III error rates were less than .01 for 87% of the conditions and never exceeded .02.

Results

Empirical Alphas

For the six conditions with normal distributions and equal population means, the empirical alphas were very close to .05, ranging from .046 to .054. These results were expected in that all assumptions of the independent-samples t test were met under these conditions. On the other hand, the empirical alphas were somewhat conservative when the distributions were mixed normal, particularly for smaller sample sizes. The alphas were .025, .042, and .048 with Ns of 24, 48, and 96, respectively. Given these results, any power advantage observed under mixed normal conditions cannot be attributed to inflated alphas.

Empirical Powers

Figure 1 shows the power of the t test as a function of the difference in means and sample size for three population distributions: mixed normal with a variance of 80.8, normal with a variance of 80.8, and normal with a variance of 1.0. As expected, the power was greater for conditions with a normal distribution and a variance of 1 than for conditions with a mixed normal distribution and a variance of 80.8. The differential power was substantial across most sample sizes and mean differences.

The more provocative findings were the power comparisons between the mixed normal and the normal distributions when both distributions had within-group variances of 80.8. For these comparisons, the power tended to be greater when distributions were mixed normal, particularly for the smaller sample sizes (N of 24 or 48). This power differential became larger as the difference in means increased. In contrast, the power differential was minimal for the largest sample size (N = 96).

Exploration of the Power Differential

The results indicate that the power for an independent-samples t test is greater when samples are drawn from mixed normal distributions rather than normal distributions, given both distributions have comparable variances. To better understand these results, it is useful to examine relevant population and sampling distributions.

In Figure 2, three sets of population distributions with means of 0 and 4.5 (and equal variances) are presented: mixed normal distributions with within-group variances of 80.8; normal distributions with within-group variances of 1.0; and normal distributions with within-group variances of 80.8. Examination of these population distributions suggests that some sample distributions from the mixed normal may be more similar to those from the normal with variances of 1.0 than those from the normal with variances of 80.8, particularly for smaller samples. In these samples from mixed normal distributions, there should be a greater likelihood of rejecting the null hypothesis than in samples drawn from the normal distribution with a variance of 80.8. However, sampling distributions are next examined to gain a deeper insight into the differential power of t test under normal and mixed normal conditions.

Table 1 shows the sampling distributions of the *t* statistic, the difference in means, and the pooled within-group variance for 30,000 samples drawn from normal and mixed normal distributions with a difference in means equal to 4.5, within-group variances of 80.8, and *N*s of 24 (with equal sample sizes).

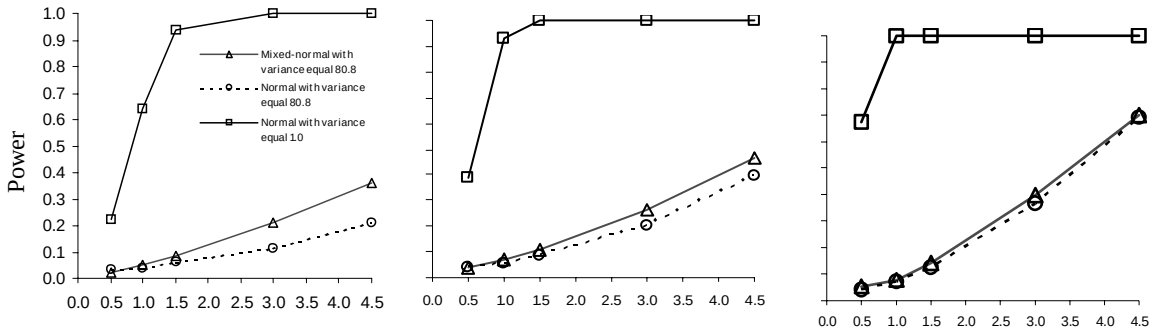


Figure 1. Power of the *t* test as a function of the difference in means and sample size for three population distributions: mixed normal with $\sigma^2 = 80.8$, normal with $\sigma^2 = 80.8$, and normal with $\sigma^2 = 1.0$. From left to right, $N = 24$; $N = 48$; and $N = 96$ (with equal cases across group).

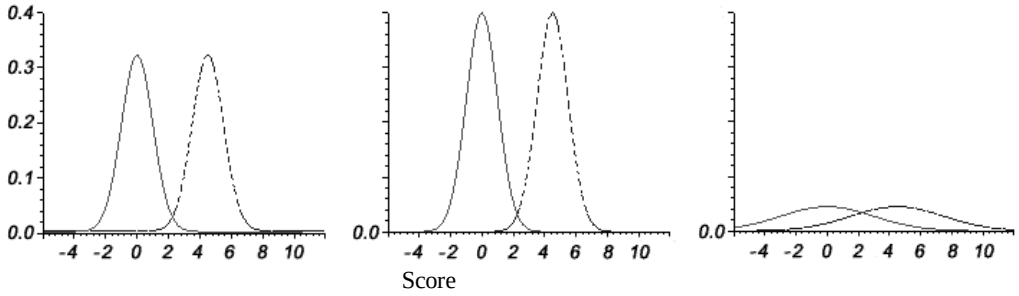


Figure 2. Group population distributions for three conditions where variances are equal across groups and the difference in means is 4.5. From left to right, mixed normal distributions with $\sigma^2 = 80.8$; normal distributions with $\sigma^2 = 1.0$; and normal distributions with$\sigma^2 = 80.8$.

As shown in the first row, the t distribution for the mixed normal condition was quite skewed and thick tailed (i.e., skewness = 2.37 and kurtosis = 11.23) compared to the t distribution for the normal condition (i.e., skewness = 0.19 and kurtosis = 0.37). Given |tcritical(22)| = 2.07, the empirical power of the t test was .34 for the mixed normal distribution, which was considerably larger than the empirical power of .21 for the normal condition.

The t statistic is a function of three quantities: the difference in means, the pooled variance, and sample size–and the latter was held constant. As shown in the second row of Table 1, the sampling distributions for the difference in means were symmetric and quite similar, except that the sampling distribution for the mixed normal was somewhat kurtotic (kurtosis = .45). As presented in the third row of Table 1, the sampling distributions for the pooled variance were very different for the two types of distributions. Although the means of the variances were nearly equal (normal: 80.76; mixed normal: 80.65), the variance of the pooled variance was 6.56 times larger for the mixed normal than for the normal condition. Further, the sampling distribution of the pooled variances was more skewed and had thicker tails for the mixed normal condition compared to the normal

condition (normal condition: skewness = 0.59 and kurtosis = 0.52; mixed normal: skewness = 1.38 and kurtosis = 2.90). Most importantly, a much larger proportion of replications had small variances for the mixed normal distribution than for the normal distribution. For example, approximately 11% of the pooled variances were less than 16 for the mixed normal condition, while none were less than 16 for the normal condition.

A greater percentage of small pooled variances are obtained with the mixed normal in comparison with the normal distribution in that the secondary distribution (with the large population variance of 400) for the mixed normal may have no or minimal effect on the pooled variance in some samples.

For example, some samples may contain no scores from the secondary distribution, and others may contain one score from the secondary distribution, but not an extreme score. The smaller pooled variances produce larger t values and, thus, greater power for the mixed normal distribution in comparison with the normal distribution with the equal population variances.

Conclusion

The results do not contradict the primary conclusions of previous research on the mixed normal distribution and the independent-samples t test. To the extent that the population distributions have outliers, the power of the t test is diminished. In the context of the mixed normal distribution, the power of the independent-samples t test decreases dramatically as the probability of a secondary distribution with a large variance increases from .00 to .20. In the presence of extreme scores, robust methods such as trimmed means become advantageous.

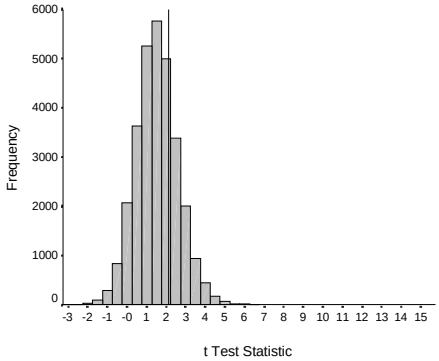
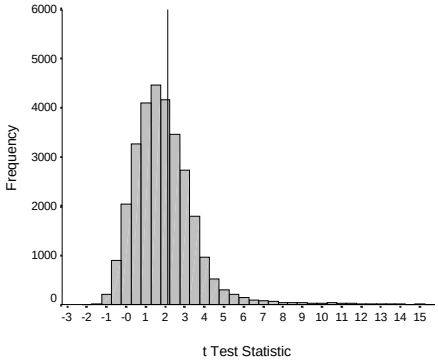
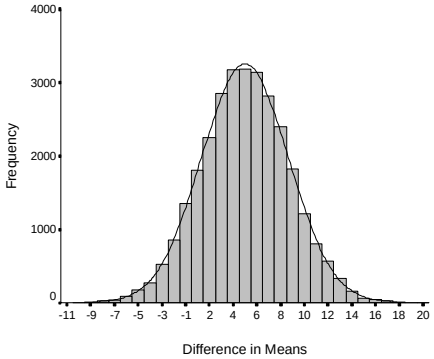
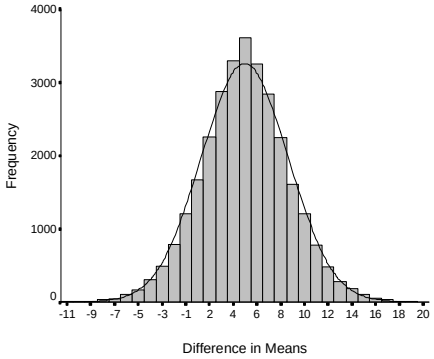
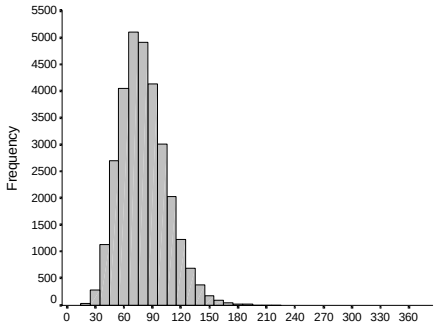
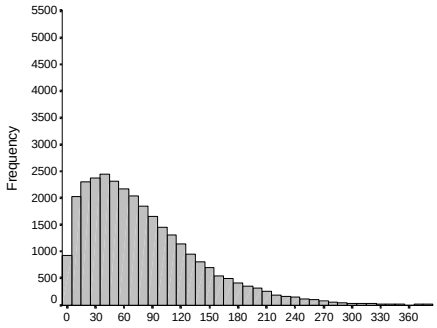
The results, however, contradict the hypothesis that the power of the test for normal and mixed normal conditions would be equal if the within-group variances were held constant or, comparably, if the effect sizes (difference in means divided by the within-group standard deviation) were held constant. Under these conditions, the power, in fact, was greater for the mixed normal distribution in that some samples produce relatively small pooled variance as a function of having few, if any, outliers drawn from the secondary distributions. The superior power was achieved despite the conservative Type I error rate for the mixed normal.

These results support a number of conceptual points. First, care should be used in discussing the diminished power of the independent-samples t test when population distributions are mixed normal rather than normal. An accurate statement is that the independent-samples t test has diminished power with a mixed normal distribution in comparison with the normal distribution to the extent that the secondary normal distribution has a much larger variance than the primary distribution and the probability of the secondary distribution is relatively large.

Second, although the independent-samples t test is the most powerful method for comparing two means if the assumptions, including normality, are met, variations of this statement may not be true. In particular, it is not true that the independent-samples t test has greater power if the population distributions are normal in comparison with other distributions, holding all other conditions constant. As demonstrated in this study, the independent-samples t test can have greater power when the population distributions are mixed normal rather than normal, given the variances of these two types of distributions are held constant.

Third, these results may be used to speculate about trimming strategies for the independent-samples t test. Some samples may include no outliers, even though the population distributions have outliers. For these samples, robust methods relying on trimming lower the likelihood of rejecting the null hypothesis by reducing the effective sample size without decreasing the pooled variance. Adaptive trimming methods–ones that trim based on the outliers present in the sample data–should produce greater power in these circumstances than those that use a fixed proportion of trimming (e.g., trim 20% from both tails of sample distributions). Future Monte Carlo studies are required to investigate whether adaptive trimming methods under these conditions maintain proper control of Type I error while increasing power.

Table 1. Sampling distributions based on independent samples of equal size (N = 24) drawn from two population distributions that are both either normal or mixed normal with a difference in population means of 4.5 and a common population variance of 80.8

Sampling distribution	Population distributions: Normal	Population distributions: Mixed normal[c]
t test statistic[a]		
Difference in means[b]		
Pooled variance		

[a]The vertical reference line indicates the critical value for rejecting the null hypothesis in the correct direction: $t(22)=2.07$.

[b]A normal curve is superimposed on the plots of the difference in means.

[c]The abscissas for the distributions based on the mixed normal were not extended to include all possible values of statistics if the frequencies for intervals including these values were sufficiently small (< .04 % of samples) that they could not be observed on the graphs. The most extreme values not shown were for the pooled variance, with six values being greater than 500.

References

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

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

Keselman, H. J., Wilcox, R. R., & Lix, L. M. (2003). A generally robust approach to hypothesis testing in independent and correlated groups designs. Psychophysiology, 40, 586-596.

Lix, L. M., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticity and nonnormality. Educational and Psychological Measurement, 58, 409-429.

MacDonald, P. (1999). Power, Type I, and Type III error rates of parametric and nonparametric statistical tests. The Journal of Experimental Education, 67, 367-379.

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

Ramsey, P.H. (1980). Exact type I error rates for robustness of Student's T test with unequal variances. Journal of Educational Statistics, 5, 337-349.

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.

Wilcox, R.R. (1995). ANOVA: The practical importance of heteroscedastic method, using trimmed means versus means, and designing simulation studies. British Journal of Mathematical and Statistical Psychology, 48, 99-114.

Wilcox, R. R. (1997). Introduction to robust estimation and hypothesis testing. San Diego, CA: Academic Press.

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