STATISTICS

Evaluation of Statistical Methods for Determining Differences between Samples from Lognormal Populations

T. B. Parkin*

ABSTRACT

Soil variables often exhibit frequency distributions that are positively skewed. When this situation exists, the assumption of normality associated with analysis-of-variance procedures is violated, and two common recommendations are given: (i) perform a normalizing transformation, or (ii) apply nonparametric statistical methods. Information regarding the relative efficacy of these two procedures with regard to power to detect differences between batches of samples is lacking. This study evaluates five statistical procedures for detecting differences between samples drawn from lognormal populations. The tests evaluated were (i) two-tailed t-test on untransformed data, (ii) two-tailed t-test on natural log-transformed data, (iii) nonparametric Mann-Whitney test, (iv) median confidence interval overlap method, and (v) a mean confidence interval overlap method. The tests were evaluated with regard to Type I error rate by comparing batches of samples drawn from the same lognormal population. Also, the power of the statistical tests to detect differences between two batches of samples drawn from different lognormal populations was evaluated over a range of population variances and sample sizes (n = 4 to 100). It was found that three of the tests (Tests 2, 3, and 4) were sensitive to sample differences when the underlying populations differed with regard to their medians. The other two tests (Tests 1 and 5) were sensitive to differences in population means. Test 2 is recommended when the median is the location parameter of interest; however, Test 5 should be used to detect differences in means.

Environmental and soils data are often not normally distributed, but rather exhibit skewed distributions (Parkin and Robinson, 1992). In many cases, the lognormal distribution adequately describes such data. When lognormality exists, optimum methods for computing summary statistics such as the mean, median, and variance are well defined (Parkin et al., 1988; Parkin et al., 1990; Parkin and Robinson, 1993). However, a typical objective of many studies extends beyond estimation of population parameters from sample data. In many cases, sampling is conducted to evaluate treatment effects.

Application of standard analysis of variance procedures requires several assumptions concerning the underlying error structure of the data, and among these is the assumption of normality. The effects of violations of the assumption of normality on the efficacy of parametric statistical tests, such as the t-test, have long been known (Hey, 1938; Cochran, 1947). Two primary effects result when the normality assumption is violated. Nonnormality will influence the ability of a statistical test to perform at the stated a-level. Cochran (1947) refers to this effect as the validity. Nonnormality will also affect the power of a statistical test to detect differences when real differences in the data actually exist. Two common procedures have been recommended when data are not normally distributed: (i) transform for normality, or (ii) apply nonparametric statistical methods (Snedecor and Cochran, 1967). However, the consequences of implementing these two approaches on the inference base, specifically with regard to the estimand, are not typically considered.

For lognormally distributed data, the population mean and median have different values. Both of these location parameters are indicators of central tendency of the population. The mean is the center of mass of the distribution, while the median is the center of probability of the distribution. In some cases, the median may be a more appropriate indicator of central tendency (Hirano et al., 1982; Loper et al., 1984; Landwehr, 1978); and in other situations, the mean is more appropriate (Parkin, 1991; Gilbert, 1987, p 45–57). The choice of the mean or the median as the summary statistic depends upon the objectives of the experiment and the nature of the sampling (Parkin and Robinson, 1992).

Once a determination has been made regarding the use of the mean or median as the estimator of choice, consideration must be given to the appropriate statistical test for evaluating differences between batches of samples. However, little information exists concerning the relative sensitivities of statistical tests to differences in the mean vs. the median. The purpose of this study was to determine the efficacy of five statistical testing procedures to distinguish differences between samples from lognormal populations. Specifically, Monte Carlo simulation experiments were performed in order to (i) ascertain the actual Type I error rate of common statistical procedures when applied to lognormal data, and (ii) evaluate the statistical tests with regard to power (1 - β, where β is the Type II error rate) in detecting differences in samples drawn from populations with differing means or medians.

METHODS

Populations

This study was designed to address a relatively simple problem in statistics, the two sample problem. Seven different lognormal populations were used in these comparisons. These populations were constructed to test the efficacy of the statistical tests under three conditions: (i) samples drawn from identical populations, (ii) samples drawn from populations with equal means, but differing medians, and (iii) samples drawn from populations with different means but equal medians. The simulations were designed so that comparisons could be made between batches of samples drawn from populations with means or medians which varied by a factor of 2 to 8 (Table 1). Populations A through D had medians of 9 and means ranging from 10 to 80. Populations E, F, and G had means of 10 and medians of 4.5, 2.25, and 1.125, respectively. These evaluations were performed across a wide range of population vari-

*Corresponding author.


USDA-ARS, National Soil Tilth Laboratory, 2150 Pammel Dr., Ames, IA 50011. Contribution from the USDA-ARS, National Soil Tilth Laboratory. Received 22 Aug. 1991.
Type I Error Rates

with the Monte Carlo techniques for sampling from a lognormal distribution of samples drawn from the same population was determined. Since, in this study, samples are drawn from the same population, the null hypothesis is in fact true, therefore frequency with which differences were detected is the actual Type I error rate ($\alpha$-level). Cochran (1947) refers to the comparison of the actual Type I error rate vs. the nominal $\alpha$-level at which the test is applied as the validity of the statistical test. The Type I error rates of each statistical test were evaluated at $\alpha$-levels of 0.05, 0.1, and 0.2 for each of the seven test populations. Only results for the lognormal populations with the smallest and largest variance (populations A and D, respectively) are presented.

The probability associated with committing a Type I error when comparing batches of samples drawn from the low variance population is presented in Fig. 1. For a given statistical test, the probability of committing a Type I error was generally invariant as a function of sample size; however, differences among the tests are observed. The nonparametric test (Test 3), $t$-test on untransformed data (Test 1), and $t$-test on natural log-transformed data (Test 2) all performed near the target $\alpha$-level for each $\alpha$-level evaluated. The measured Type I error rates associated with the confidence interval overlap methods were similar to one another, yet were lower than the target $\alpha$-levels. In evaluations conducted at the 0.05 level (Fig 1A), the mean and median confidence limit overlap methods were actually performing at the 0.006 level. When evaluated at an $\alpha$-level of 0.10, these methods yielded error rates of approximately 0.017 (Fig 1B), and at the 20% probability level (Fig 1C) the measured $\alpha$-levels of these two tests were approximately 0.06.

Evaluations conducted with the highly skewed population yielded similar results with respect to the nonparametric test and the $t$-test on natural log-transformed data (Fig. 2). The measured Type I error rate of these two tests was at or near the target $\alpha$-level for each level tested (Fig 2A, B, and C). Unlike results obtained from population A, when applied to population D the $t$-test on untransformed data yielded Type I error rates which were lower than the target level. The measured $\alpha$-level of this method was approximately 1.7% for $\alpha = 0.05$ (Fig 2A), 5% for $\alpha = 0.10$ (Fig 2B), and 16% for $\alpha = 0.2$ (Fig 2C). The confidence limit overlap methods yielded Type I error rates that were substantially less than the target $\alpha$-levels, with the CL-MEAN method operating at a slightly lower level than the CL-MEDIAN test.

### Statistical tests

In this study, five statistical tests were evaluated (Table 2). A brief description of each test is presented below.

**Test 1.** This test is the standard two-tailed $t$-test applied to unpaired samples of equal size (Snedecor and Cochran, 1967, p. 123–125). In situations where the samples being compared were drawn from populations with different variances, $n - 1$ degrees of freedom as opposed to $2n - 2$ degrees of freedom was used in obtaining the $t$ statistic (Cochran, 1964).

**Test 2.** This test is the standard $t$-test applied to natural log-transformed data. In comparing samples with unequal variances $n - 1$ degrees of freedom was used (as described for Test 1).

**Test 3.** This test is the nonparametric ranking test developed by Wilcoxon (1945) and is often called the Mann-Whitney test (Snedecor and Cochran, 1967).

**Test 4.** This test is based on computation of confidence limits of the median according to the parametric method described by Parkin and Robinson (1993). In the application of this test, batches of samples are judged not to be significantly different if the confidence intervals of the two batches of samples overlap.

**Test 5.** This test is similar to Test 4 in that significance is based on confidence interval overlap; however, in this case the confidence intervals are constructed about the means of the two batches of samples. Confidence limits of the mean were computed according to the method of Land (1971) as described by Parkin et al. (1990).

### Test Evaluations

The tests were evaluated according to two criteria: (i) the actual or measured Type I error rate ($\alpha$) obtained when two batches of samples from the same distribution were compared, and (ii) the power (1 - $\beta$, where $\beta$ is the Type II error rate) of the statistical tests to detect differences between batches of samples drawn from different distributions. Ten thousand evaluations were performed at each sample size ($n = 4$ to 100) with the Monte Carlo techniques for sampling from lognormal populations as described by Parkin et al. (1988).

### RESULTS

#### Type I Error Rates

The Type I error rate, or $\alpha$-level, of a statistical test is the probability associated with rejecting the null hypothesis when in fact the null hypothesis is true. In these evaluations the frequency at which each of the statistical tests detected a significant difference between two batches of samples drawn from the same population was determined. Since, in this study, samples are drawn from the same population, the null hypothesis is in fact true, therefore frequency with which differences were detected is the actual Type I error rate ($\alpha$-level). Cochran (1947) refers to the comparison of the actual Type I error rate vs. the nominal $\alpha$-level at which the test is applied as the validity of the statistical test. The Type I error rates of each statistical test were evaluated at $\alpha$-levels of 0.05, 0.1, and 0.2 for each of the seven test populations. Only results for the lognormal populations with the smallest and largest variance (populations A and D, respectively) are presented.

The probability associated with committing a Type I error when comparing batches of samples drawn from the low variance population is presented in Fig. 1. For a given statistical test, the probability of committing a Type I error was generally invariant as a function of sample size; however, differences among the tests are observed. The nonparametric test (Test 3), $t$-test on untransformed data (Test 1), and $t$-test on natural log-transformed data (Test 2) all performed near the target $\alpha$-level for each $\alpha$-level evaluated. The measured Type I error rates associated with the confidence interval overlap methods were similar to one another, yet were lower than the target $\alpha$-levels. In evaluations conducted at the 0.05 level (Fig 1A), the mean and median confidence limit overlap methods were actually performing at the 0.006 level. When evaluated at an $\alpha$-level of 0.10, these methods yielded error rates of approximately 0.017 (Fig 1B), and at the 20% probability level (Fig 1C) the measured $\alpha$-levels of these two tests were approximately 0.06.

Evaluations conducted with the highly skewed population yielded similar results with respect to the nonparametric test and the $t$-test on natural log-transformed data (Fig. 2). The measured Type I error rate of these two tests was at or near the target $\alpha$-level for each level tested (Fig 2A, B, and C). Unlike results obtained from population A, when applied to population D the $t$-test on untransformed data yielded Type I error rates which were lower than the target level. The measured $\alpha$-level of this method was approximately 1.7% for $\alpha = 0.05$ (Fig 2A), 5% for $\alpha = 0.10$ (Fig 2B), and 16% for $\alpha = 0.2$ (Fig 2C). The confidence limit overlap methods yielded Type I error rates that were substantially less than the target $\alpha$-levels, with the CL-MEAN method operating at a slightly lower level than the CL-MEDIAN test.
It must be recognized that the efficacy of a statistical test cannot be based solely on how well the measured Type I error rate corresponds to the nominal α-level. In fact, a given statistical test that performs at a lower α-level may be the superior one as long as the power of the test is not compromised.

**Power of Statistical Tests**

The preceding results indicated that the five statistical tests differ with regard to Type I error rate. In addition to Type I error rate, the power of the statistical tests to detect differences in samples drawn from different lognormal populations was evaluated. Power is expressed as the probability associated with detecting a difference between two batches of samples drawn from the lognormal populations which differ with regard to either means or medians. These evaluations not only provide quantitative information of power as a function of degree of mean or median separation, but also yield information on sensitivity of a given test to the magnitude of the differences that exist between the underlying population means or medians.

**Comparisons of Populations with Different Medians**

The five statistical tests applied to samples drawn from populations with equal means, but unequal medians exhibited differences in power (Fig 3 and 4). At the α = 0.05 level, power of the nonparametric test, t-test on natural log-transformed data, and CL-MEDIAN method to detect significant differences between batches of samples increased as a function of sample size (Fig. 3). For small sample sizes, these tests exhibited low power to detect differences, but with increasing sample size, power increased. For any given sample size, power was also related to the degree of median separation of the two underlying populations. For example, with a sample size of n = 8, the t-test on natural log-transformed data detected differences only 30% of the time when the underlying population medians differed by a factor of 2 (Fig 3A). However, when the two population medians differed by a factor of 8 (Fig 3C), the power of this test to detect differences increased to 72% for n = 8 samples.

In contrast to the other three methods, the t-test on untransformed data and the CL-MEAN overlap method exhibited low power to detect differences between the test populations across the range of sample sizes evaluated. For samples sizes n > 4 the CL-MEAN technique was insensitive to sample size and yielded probability levels of 1.2, 1.9, and 2.3% when the underlying population medians differed by a factor of 2, 4, or 8 (Fig 3A, B, and C, respectively). The power curve exhibited by the t-test on untransformed data was different than that exhibited by the other statistical methods. The probability of detecting differences with this method was inversely related to sample size. Decreasing power with increasing sample size indicates that the t-test on untransformed data is insensitive to differences in the medians of the underlying populations.

The relative behavior of the five statistical methods applied at the α = 0.2 level was similar to that observed at the α = 0.05 level with respect to functional dependence upon both sample size and the degree of median separation of the underlying test populations (Fig. 4). The t-test on natural log-transformed data, nonparametric test, and CL-MEDIAN method exhibited increased power with increasing sample size, while the t-test on untransformed data and the CL-MEAN method were insensitive to differences in population medians. Although the relative relationships between the power curves of the different tests was the same, as the α-level is relaxed.

**Fig. 1.** Probability of detecting a significant difference between batches of samples (n = 4 to 100) drawn from the low variance lognormal population as determined by each of the five statistical tests. The statistical methods were applied at α-levels of 0.05 (panel A), 0.1 (panel B), and 0.2 (panel C). The horizontal line in each panel denotes the α-level at which each test was applied.
Comparisons of Populations with Different Means

The power curves for each statistical test when applied to samples drawn from lognormal populations with means that differ by a factor of 2, 4, or 8 are present in Fig. 5A, B, and C, respectively. The nonparametric tech-
Fig. 4. Probability of detecting a significant difference between batches of samples (n = 4 to 100) drawn from lognormal populations with equal means but medians which differed by a factor of 2 (panel A), 4 (panel B) or 8 (panel C). Each of the statistical tests was applied at an α-level of 0.05.

Fig. 5. Probability of detecting a significant difference between batches of samples (n = 4 to 100) drawn from lognormal populations with equal medians but means which differed by a factor of 2 (panel A), 4 (panel B) or 8 (panel C). Each of the statistical tests was applied at an α-level of 0.05.

Nique, t-test on natural log-transformed data, and CL-MEDIAN method were insensitive to differences in samples with different means, as indicated by the low probability levels and the relatively flat responses as a function of sample size. The nonparametric method and the t-test on natural log-transformed data yielded comparable probability levels that approximated the nominal α-level at which the tests were applied (α = 0.05). Across the range of sample sizes, the probability levels associated with the CL-MEAN method was slightly lower than these two methods. In contrast, the mean confidence interval overlap method and the t-test on untransformed data were sensitive differences in population means. However, for a given sample size, the CL-MEAN method had greater power to detect differences than the t-test. The abilities of these two tests to detect differences increased with increasing sample size. Power also increased as a function of mean separation of the underlying
populations. For example, the CL-MEAN method applied to samples of \( n = 12 \) drawn from populations \( A \) and \( B \) detected differences only 14% of the time (Fig. 5A), however with sample sizes of \( n = 12 \) drawn from populations \( A \) and \( D \), differences were detected approximately 59% of the time (Fig. 5C).

These results were consistent when the tests were applied at an \( \alpha \)-level of 0.20 (Fig. 6). The power of the nonparametric method was fairly constant over the range of sample sizes at a level slightly higher than the nominal \( \alpha \)-level of 0.2. The \( t \)-test on untransformed data yielded probability levels approximately equal to the 0.2 level, whereas the CL-MEDIAN technique operated at approximately the 0.1 level. Similar to results obtained at the \( \alpha = 0.05 \) level, the CL-MEAN technique and the \( t \)-test were sensitive to differences in population mean. The power of these two tests increased with increasing sample size and with increasing mean separation of the underlying populations.

**DISCUSSION**

This study was predicated on the observation that many soil variables are not normally distributed, but rather are skewed and often better described by the lognormal family of distributions (Warrick and Nielsen, 1980, p. 385; Gilbert, 1987; Parkin and Robinson, 1992). The consequences of applying analysis of variance, specifically the two-tailed \( t \)-test, to data that do not conform to the assumption of normality are known; both the Type I error rate and the power of the test are affected. If deviations from normality are slight, or if sample size is large, these consequences are insignificant (Cochran, 1947). However, precise guidelines relating sample size and population skewness to the validity and power of statistical tests are lacking, or at best, not readily available to soil scientists. Similarly, there are, scattered throughout the literature, inferences regarding the target estimand of different statistical tests, but again no specific guidelines are accessible. The present study provides comparative data on the efficacy of five statistical tests as a function of sample size and underlying variances of the lognormal populations being compared.

The nonparametric test, \( t \)-test on natural log-transformed data, and the median confidence interval overlap method were sensitive to detecting differences between samples from populations with differing medians, but not means. The behavior of the median-CL-overlap method in this regard is not surprising since implementation of this method involves comparisons of confidence intervals constructed about the sample medians. Similarly, the fact that the nonparametric method employed in this study is a ranking procedure, where the actual magnitudes of the sample values do not contribute to the value of the Wilcoxon statistic, suggests that the center of probability of the distribution (i.e., the median) is the target location estimator (Miller, 1986, p. 40–66). The results obtained on the sensitivity of the \( t \)-test on natural log-transformed data to differences in population median were expected. A logarithmic transformation on data from a lognormal distribution will normalize the data, and a \( t \)-test performed on this normalized data will discern differences between the means of the natural log-transformed data. Since the median of a lognormal distribution is defined by:

\[ \exp(\mu) \]  

where \( \mu \) is the mean of the natural log-transformed variable. The \( t \)-test on natural log-transformed data is, therefore, sensitive to differences in samples drawn from populations that differ with regard to \( \mu \). This test will, therefore, be sensitive to differences in \( \exp(\mu) \) which is an estimator of the population median. Of these three
median separation tests, the method of choice is the $t$-test on natural log-transformed data when it is assured that the data are conform to a lognormal distribution. However, little is lost with regard to either power or validity when the nonparametric test is used.

The $t$-test on untransformed data and mean confidence limit method were insensitive to differences in population median but were sensitive to differences in population means. At any given sample size, the $t$-test on untransformed data detected differences at a lower frequency than the mean confidence interval overlap method. This latter test was also operating at a Type I error rate that was substantially less than the nominal $\alpha$-level at which it was applied. Thus, the mean confidence interval overlap method is a very conservative test.

Results of this study indicate that caution must be used in the application of the common recommendations given for the analysis of nonnormal data. For the lognormal case described here, transformation for normality and applying $t$-test is a test of differences in medians. Such a procedure is insensitive to differences that may exist between population means. A similar result is obtained when parametric approaches are applied. If, indeed, the median is the location estimator of interest, then this problem is not a consideration. However, if the mean is the location estimator of interest, then neither of these recommendations is sufficient. The mean-CL-overlap method has good power to detect differences in means when applied at moderate sample sizes ($n > 12$). This test is also conservative, in that the actual $\alpha$-level is well below the nominal $\alpha$-level at which this test is applied.

These data also indicate the dramatic effect of sample size on the power to detect differences. Regardless of whether the mean or median is the estimator of interest, at sample sizes of $n = 4$, very poor power is available. When lower sample numbers are available, the only way to increase power is to apply the tests at higher $\alpha$-levels.

It must be recognized that the results reported in this study were obtained from Monte Carlo analysis of true lognormal populations. The absolute power of a given test will likely be influenced by the extent to which the underlying population deviates from true lognormality. However, it is speculated that the relative sensitivities of the tests to differences in either the mean or median will be consistent with the results of this study.

The techniques described in this paper, along with some knowledge of the properties of the underlying distribution derived from the sample data, can be used to evaluate the adequacy of mean or median separation methods for a given sample size, on a case-by-case basis. The availability of inexpensive computational power provided by desktop computers offer opportunities for a more detailed analysis of the adequacy of statistical procedures applied to a given data set.

REFERENCES


