VARIANCE TESTING WITH SIMPLICIAL DATA DEPTH

Karen J. McGaughey
George A. Milliken

Follow this and additional works at: https://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation

This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.
Variance Testing with Simplicial Data Depth

Karen J. McGaughey¹, and George A. Milliken²

¹ Department of Statistics, Cal Poly University, San Luis Obispo, CA 93407
² Department of Statistics, Kansas State University, Manhattan, KS 66505

ABSTRACT

A method is developed and studied for testing equality of variances based on simplicial data depth and Mood’s nonparametric test in the case of two samples. A method for calculating univariate simplicial data depth using a rank transformation is introduced. Type I error rates and power curves are compared for three existing tests for equality of variances and the data depth test using data simulated from the normal distribution and 5 nonnormal distributions. In addition, a new method of aligning two samples with unequal location parameters is proposed. This method shows significant improvement over aligning by either the median or mean in controlling Type I error rates of skewed distributions.

KEY WORDS: Equal Variance Tests, Nonparametric, Mood’s test, Simplicial Data Depth, Two Sample, Univariate Response

1 INTRODUCTION

It is becoming increasingly recognized in agricultural and industrial settings that differences in dispersion of groups is an important issue in its own right. While testing for differences in group means may still be of primary concern, the interest in equality of group variances has gone beyond simply fulfilling the distributional assumptions of hypothesis tests for means. In the baking industry, for example, the quality of the flour is determined by both the protein content and the strength of the gluten of the wheat. Flour milled from wheat with a highly variable protein content or gluten strength will be of lower quality than that milled from less variable wheat. Identifying and controlling process variability was the primary reason for the development of 6-sigma methodology. Used in many major manufacturing sectors, 6-sigma is a well-defined set of steps that a business can take to improve both their product and profitability.

Many tests, both parametric and nonparametric, exist to test for equality of group variances in the case of a single response. Surveys conducted by Duran (1976), Conover, Johnson and Johnson (1981), and Oljenik and Algina (1987) concluded that the Brown-Forsythe test and O’Brien’s test had the best power while consistently holding their Type I error rates at the nominal α level. More recently, other methods of testing for differences in scale for the two sample problem have been considered. Techniques based on the jackknife (Hollander and Wolfe, 1999) and the analysis of means for variances (ANOMV) (Wludyka and Nelson, 1997) have been shown to be competitive with existing tests in terms of Type I error rates and power.

While the single response variable case is interesting, many times a study or manufacturing process involves measurements on multiple response variables. In an animal science study, for example, the researcher may be interested in the effects of different feed additives on several response variables such as average daily gain, feed consumption, and feed efficiency. In precision agriculture there is a need to determine the suitability of different GPS
sources for use in particular fields. Due to the nature of GPS one system may be more variable in locating objects in a field than another system. Locating an object means determining three response variables, the x-, y-, and z-coordinates. In the case of multivariate populations current literature generally addresses three ideas. To test for a difference in the covariance matrices of two populations, one can use the multivariate generalization of Bartlett’s test. Lawley (1963) proposed a test for the equality of k population correlation matrices. Liu, Parelius and Singh (1999) introduced the idea of using convex contours defined by the simplicial data depth of the points in the sample to measure differences in the concentration or dispersion of the populations.

The goal of the research reported here is to introduce a test for the equality of variances for two samples in $\mathbb{R}^1$ based on the ideas of simplicial data depth (Liu, 1990). This procedure is a new look at Mood’s test and is shown to be competitive in terms of Type I error rates and power versus the Brown-Forsythe test, O’Brien’s test and the F-test. The use of simplicial data depth as the vehicle for testing equality of variances enables future extension of the testing process to the more complicated issue of testing for homogeneity of variances in factorial designs and testing for homogeneity of covariance matrices.

2 UNIVARIATE SIMPLICIAL DATA DEPTH, A RANK TRANSFORM

Data depth is a process for measuring the centrality of a multivariate data point with respect to a given data cloud (Liu and Singh, 1993). It is a method for reducing the dimension of a data set from $\mathbb{R}^p$ to $\mathbb{R}^1$. While there is no data reduction per se, the ideas of data depth can be used in the context of a univariate population. Let $\{X_1, X_2, ..., X_m\}$ be a sample data set in $\mathbb{R}^1$.

Given any two of the data points $X_i$ and $X_j$, a closed line segment can be formed with endpoints $X_i$ and $X_j$, which can be denoted as $(X_i, X_j)$. From the $m$ data points, $\binom{m}{2}$ line segments can be generated. For any point $x$ in $\mathbb{R}^1$, the number of these line segments which contain $x$ can be counted. The proportion of line segments which contain $x$ is the simplicial data depth of this point. If $x$ is near the center of the data set then the proportion of line segments containing $x$ will be large. If $x$ is near the outskirts of the data set, then the proportion of line segments containing $x$ will be small. Thus, the sample simplicial depth of any point $x$ in $\mathbb{R}^1$ can be expressed as

$$D_m(x) = \left(\binom{m}{2}\right)^{-1} \sum \mathbb{I}(x \in (X_i, X_j))$$

$$i, j = 1, 2, ..., m, i \neq j$$

Where $\mathbb{I}(A)=1$ if event $A$ occurs and $\mathbb{I}(A)=0$ otherwise (Liu and Singh, 1990). Equation (2.1) expresses the depth of $x$ as the proportion of line segments which contain the point $x$.

Current algorithms for calculating the sample simplicial depth can be extremely time consuming and computer intensive, especially when the data set is very large. During the development of the equality of variances test to be presented here it was found that the sample simplicial depth of a point, $D_m(x_i)$, can also be expressed as a function of the rank, $R_i$, of that point so that,

$$D_m(x_i) = d(x_i) = \frac{(m-1) + (R_i - 1)(m - R_i)}{\binom{m}{2}}$$

(2.2)
assuming there are no ties in the data set. Where ties exist, let \( R_i \) be the average rank of the tied points and use equation (2.2) to find the approximate sample simplicial depth of each point in the data set.

3 TWO SAMPLE DATA DEPTH TEST OF SCALE CHANGE

Let \( F \) and \( G \) be the distribution functions of two independent populations. Let \( X = \{X_1, X_2, ..., X_m\} \) be a random sample of \( m \) observations from \( F \), and \( Y = \{Y_1, Y_2, ..., Y_n\} \) be a random sample of \( n \) observations from \( G \) where \( n \geq m \). Let \( Z = \{X_1, X_2, ..., X_m, Y_1, Y_2, ..., Y_n\} \) be the combined sample with \( N = m + n \) total observations. The problem is to determine if the two populations, \( F \) and \( G \), are the same in terms of the variance or scale.

The depths of all of the points in the combined sample, \( Z \), are known values \( D = \{d_1, d_2, ..., d_N\} \), calculated using equation 2.2, where \( R_i \) is the rank of each point in \( Z \).

Because the depths of the sample points depend on the locations of the two distributions relative to one another, if the two samples \( X \) and \( Y \) do not have the same location they should be aligned using their respective sample median, or some other respective sample percentile before being combined into one data set. In the following discussion it is assumed that the two samples have been aligned. Let \( \bar{d}_Y \) and \( \bar{d}_X \) be the average depth of \( X \) and \( Y \) respectively, in the combined sample \( Z \). If the variances of the two populations \( G \) and \( F \) are equal, then \( \bar{d}_Y \) and \( \bar{d}_X \) should be approximately equal. If the variance of \( G \) is greater than the variance of \( F \), then \( \bar{d}_Y \geq \bar{d}_X \). If the variance of \( G \) is less than the variance of \( F \), then \( \bar{d}_Y \leq \bar{d}_X \). This leads to the formulation of a two sample test of equal variances based on the average depth of a sample \( X \) or \( Y \) when the two samples are combined into one sample, \( Z \).

Under the null hypothesis that the two populations have identical shapes and equal variances (or scales), \( H_0 : \sigma_F^2 = \sigma_G^2 \), the depths of the \( n \) observations in \( Y \) are just a random sample of \( n \) taken without replacement from the \( N \) depths in \( D \). Using some simple random sampling without replacement arguments, the proposed data depth based test statistic using sample \( Y \) is expressed:

\[
dd = \frac{\bar{d}_Y - E(\bar{d}_Y)}{\sqrt{\text{Var}(\bar{d}_Y)}} \quad \text{where} \quad E(\bar{d}_Y) = \frac{1}{N} \left( \frac{N + 4}{3} \right) \quad \text{and} \quad \text{Var}(\bar{d}_Y) = \frac{1}{n} \left( \frac{N - n}{N} \right) \left( \frac{1}{N - 1} \right) \left[ \frac{N^3 + N^2 - 4N - 4}{45N(N - 1)} \right].
\]

The data depth statistic and Mood’s (1954) statistic, \( M = \sum_{i=1}^{n} \left( R_i - \frac{N + 1}{2} \right)^2 \), are linearly related. Because Mood’s score can be expressed as \( R_i^2 - R_i - R_iN + \left( \frac{N + 1}{2} \right)^2 \) and the data depth rank transformation can be expressed as \( -\binom{N}{2} d_i = R_i^2 - R_i - R_iN + 1 \) there is a one to one relationship between the two scores. The data depth test will thus have properties equivalent to
Mood’s test, so that the sampling distribution of the data depth test statistic can be approximated by a N (0, 1). Note that the first and second moments of \( \bar{d}_Y \) are scalar multiples of the first and second moments of \( M_N \). Thus, the decision rule for testing the equality of two population variances using the data depth test (2.3) is to reject the null hypothesis if \( |dd| \geq z_\alpha \), where \( z_\alpha \) is the upper \( \frac{\alpha}{2} \) percentile of the N (0, 1) distribution.

4 SIMULATION STUDY

To explore the performance of the data depth statistic for testing the equality of variances, a simulation study was conducted for 2 groups of sampled data using the SAS® system (Version 8, 1999). Three existing tests were used as comparison tests: the F-test, O’Brien’s test with \( w = 0.5 \), and the Brown-Forsythe test. In simulating the data, the following three factors were varied: (a) sample size, (b) the distributional form of the parent population, and (c) the variance of the parent population. The data were simulated from all possible sample size combinations of 5, 10, 20, 25, 40 and 50 observations. Samples were selected from one normal and 5 nonnormal parent distributions. The nonnormal distributions included a symmetric platykurtic (light-tailed) distribution, a symmetric leptokurtic (heavy-tailed) distribution, a slightly skewed distribution, a moderately skewed distribution, and a skewed-leptokurtic distribution. The data for the nonnormal distributions were simulated using transformations of N(0, 1) random variates suggested by Fleishman (1978). The shapes of the nonnormal distributions can be seen in Figure 1. The data depth test statistic was calculated using sample \( Y \) and the two samples were aligned using their respective sample medians for all simulation studies.

To study the Type I error rates for 0.01, 0.05, 0.10, and 0.25 alpha level tests, two samples of data were generated from the six populations under the assumption of equal variances. To study the power of the test to detect unequal variances for 0.01 and 0.05 alpha level tests, two samples of data were generated from populations with variances in the following ratios: (1:1.25), (1:1.50), (1:1.75), (1:2), (1:2.25), (1:2.50), (1:2.75), (1:3), (1:3.25), (1:3.50), (1:3.75), (1:4). Each of these combinations of parent distribution, sample size, and variance ratio was repeated 5,000 times. The observed frequency of rejection of the null hypothesis for size \( \alpha = 0.05 \) is reported in Table 1 for each parent population. The standard error for the Type I error rate is 0.0031. Selections of the power plots are displayed in Figures 2-7.

5 CONCLUSIONS

The Type I error rates for the data depth test for data simulated from the normal distribution and the 5 nonnormal distributions for \( \alpha = 0.05 \) are presented in Table 1 along with a target confidence interval of 95%. The results are reported for sample size combinations 5/5, 5/10, 5/20, 5/40, 10/10, 10/20, 10/40, 20/20, 20/40, 40/40, 25/50, and 50/50. Results for sample size combinations 10/5, 20/5, 40/5, 20/10, 40/10, 40/20, and 50/25 are not reported because the Type I error rates are similar to those seen in the reported cases. The data depth test statistic was computed using the larger sample when the sample sizes were unequal, however the smaller sample could have been used without loss of generality. For normal and platykurtic samples the data depth test Type I error rate is conservative for approximately 70% of the sample size.
combinations and at the nominal $\alpha$ level for approximately 30% of the sample size combinations across all four $\alpha$-level tests. For leptokurtic and slightly skewed samples the data depth test tends to be right at the specified $\alpha$ level for about 70% of the sample size combinations. The remaining 30% of the sample size combinations have conservative Type I error rates. For the moderately skewed distribution, the Type I error rate is somewhat liberal, within 0.05 of the specified $\alpha$ level, for approximately 50% of the sample size combinations. For the remaining 50% of sample size combinations the Type I error rates are more than 0.10 higher than the specified $\alpha$ level. The same is true for all sample size combinations of the skewed leptokurtic distribution. This problem is addressed in section 6.

Figures 2-7 illustrate selected power curves at $\alpha = 0.05$ for the data depth test as compared to the F-test, O’Brien’s test and the Brown-Forsythe test for all 6 parent populations. For all distributions when the sample sizes are unbalanced the data depth test has better power by approximately 5-10% when the larger sample has the larger variance. When the data are normal or platykurtic the data depth test has as good as or better power than O’Brien’s test and Brown-Forsythe 90% of the time when the sample sizes are small and/or unbalanced. When the sample sizes are balanced, the data depth test power is within 5% of the other tests 60% of the time and within 10% the remaining 60% of the sample size combinations. For leptokurtic samples, the data depth test is the best among the three tests holding their Type I error rates at $\alpha = 0.05$ for all sample size combinations except 20/20, 40/40, and 50/50. For these three sample size combinations, the data depth test power is always greater than O’Brien’s and within 5% of that of the Brown-Forsythe test. When the data are simulated from slightly skewed distributions, the data depth test has power slightly better or within 5% of all three competitors when the sample sizes are unbalanced, and is within 5% of both O’Brien’s test and Brown-Forsythe when the samples sizes are balanced. The data depth test’s power for the moderately skewed distribution tends to be higher for all sample size combinations, and is quite liberal for larger sample sizes. When the data are simulated from a skewed leptokurtic distribution the Type I error rate is extremely liberal, exceeding the nominal Type I error rate by as much as 0.50 in some cases. Due to these extreme Type I error rates, a power study for the skewed leptokurtic distribution is not necessary.

6 ALIGNMENT METHOD FOR SKEWED DISTRIBUTIONS

Intuitively, the data depth test will find a difference in the variances of two distributions when one of the samples has more of its points in the tails of the combined ordered sample than the other. Therefore, the samples being used for the test should be aligned using their respective sample medians before they are combined. However, in the case of moderately to highly skewed distributions, aligning by the sample medians provides a test with an inflated Type I error rate. This excessive Type I error rate is the result of the distribution having only one tail and the large variability of the sample medians. For example, consider sample $X$ and sample $Y$ both sampled from the Exponential(1) distribution. If sample $X$ has a larger median and also a larger range than sample $Y$, when the samples are aligned using their sample medians, the points of sample $Y$ will be contained entirely within the points of sample $X$ in the combined ordered sample. Under these conditions, the data depth test can find that the variances are different when in fact they are not.

Regardless of the distributional form of the parent population, but especially if the sample data appear to come from moderately to highly skewed distributions, bootstrapping can be used
to select the best percentile by which to align the two samples. So that the assumption of equal variances is met, using sample \( X \) generate two new samples with replacement. Denote these two samples as \( X^* \) and \( Y^* \). The sizes of \( X^* \) and \( Y^* \) should be the same as the original samples \( X \) and \( Y \). Align \( X^* \) and \( Y^* \) using their respective tenth percentiles. Once aligned, combine \( X^* \) and \( Y^* \) into one data set, \( Z^* \), rank the data, and calculate the sample data depths, \( d^*_i \), using equation 2.2. These data depths will be approximate due to the large number of ties which are possible when sampling with replacement. After finding the sample data depths, use equation 3.2 to calculate the data depth test statistic, \( dd^* \), and compare this value to the \( \frac{\alpha}{2} \) percentile of the \( \text{N}(0, 1) \) distribution for a size \( \alpha \)-level test. Repeat the above process 1,000 times to determine the bootstrap Type I error rate, \( T \alpha \), for the data depth test aligning with the tenth percentile. Similarly, determine the bootstrap Type 1 error rates, \( T \alpha_{20}, T \alpha_{30}, \ldots, T \alpha_{90} \), for the data depth test aligning with the 20th, 30th, 40th, \ldots, 90th percentiles, respectively. Plot the \( T \alpha \)'s against the percentiles. Select the percentile with which to align the data, \( P^*_X \), as the percentile at or below the desired alpha level of the test. The entire process above should be repeated using sample \( Y \). At the conclusion of this bootstrapping procedure there will be two selected percentiles by which the data may be aligned, \( P^*_X \) and \( P^*_Y \).

Figure 8 illustrates the actual bootstrap Type I error rates for skewed leptokurtic samples of size 20/20. The bootstrap selection method suggests aligning the samples with the 30th or 35th percentiles. Figure 9 shows the simulation results for the 20/20 sample size combination when aligning with the 30th percentile instead of the median. The data depth test is more robust for this distribution when aligning by the 30th percentile rather than the median, and is now competes much better against O'Brien's test and Brown-Forsythe in terms of power.

7 DATA DEPTH VARIANCE TESTING PROCEDURE

Based on the simulation and bootstrap alignment results, to test for equality of variances in two samples in \( \mathcal{H}^1 \), the following procedure should be used. If the sample sizes are both less than 10, there is not enough data to use the bootstrap alignment method. Thus, aligning by the sample medians will yield the best results. If the sample sizes are larger than 10, regardless of the suspected distributional form of the parent population, use the bootstrap alignment method to determine the best percentile by which to align the data. Align the two samples using the bootstrap selected percentile \( P^*_X \), combine the two samples into one, rank the aligned data, and use the rank transform (2.2) to calculate the data depths of each point in the combined sample. Calculate the data depth test statistic (3.2), \( dd^*_X \), using either of the two samples and compare this value to the appropriate percentile from the standard normal distribution. The calculation of the data depths using the rank transform (2.2), and the calculation of the test statistics (3.1) should be repeated after aligning the data using \( P^*_Y \) providing a second data depth test statistic, \( dd^*_Y \). If \( P^*_X \neq P^*_Y \), then \( |dd^*_X| \neq |dd^*_Y| \). Use \( dd_\cdot = \max(|dd^*_X|,|dd^*_Y|) \) as the test statistic to test the equality of variances hypothesis.

8 SUMMARY
If the data come from parent distributions which are not highly skewed, the data depth test is a viable alternative to the most reliable and most widely used two sample tests of equality of variances available. The ease of computation of the sample simplicial depths using equation (2.2), makes the test quick and easy to implement and the test statistic itself is intuitively appealing. Simulation studies indicate that the data depth test holds its own against the best competitors in terms of Type I error rates and power when the samples are balanced. When the sample sizes are small and/or unbalanced, in general the data depth test has more power than the three competitors.

For moderately to highly skewed distributions, the data depth test is an unreliable test to use due to extremely large Type I error rates when the samples are aligned with their respective medians. Further simulation studies indicate that aligning using a percentile with less variability than the median will help to control the large number of rejections under the null hypothesis. The exact percentile by which to align depends on the degree and direction of the skewness of the parent population. The more right skewed the distribution the smaller the percentile, and the more left skewed the distribution the larger the percentile will be to align the samples. A bootstrapping method can be used to select the appropriate percentile and using this method results in a more stable Type I error rate.

In practice most industrial applications of variance testing will involve large, multivariate data sets. While the data depth test is not the best test to use in all two sample univariate cases, it affords the opportunity for extension to multiple samples of univariate data and also multiple samples of multivariate data.

Table 1. Type I Error Rates of Data Depth Statistic for All Distributions at $\alpha = 0.05$

<table>
<thead>
<tr>
<th>Size of Sample X</th>
<th>Size of Sample Y</th>
<th>Normal</th>
<th>Platykurtic</th>
<th>Leptokurtic</th>
<th>Slightly Skewed</th>
<th>Moderately Skewed</th>
<th>Skewed Leptokurtic</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>5</td>
<td>0.0500</td>
<td>0.0506</td>
<td>0.0528</td>
<td>0.0514</td>
<td>0.0628</td>
<td>0.1094</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>0.0322</td>
<td>0.0366</td>
<td>0.0358</td>
<td>0.0352</td>
<td>0.0510</td>
<td>0.1060</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>0.0392</td>
<td>0.0392</td>
<td>0.0396</td>
<td>0.0408</td>
<td>0.0534</td>
<td>0.1254</td>
</tr>
<tr>
<td>5</td>
<td>40</td>
<td>0.0384</td>
<td>0.0418</td>
<td>0.0384</td>
<td>0.0460</td>
<td>0.0564</td>
<td>0.1372</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.0410</td>
<td>0.0358</td>
<td>0.0474</td>
<td>0.0412</td>
<td>0.0634</td>
<td>0.1726</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0.0386</td>
<td>0.0370</td>
<td>0.0478</td>
<td>0.0480</td>
<td>0.0704</td>
<td>0.2170</td>
</tr>
<tr>
<td>10</td>
<td>40</td>
<td>0.0432</td>
<td>0.0402</td>
<td>0.0476</td>
<td>0.0444</td>
<td>0.0870</td>
<td>0.2378</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>0.0436</td>
<td>0.0414</td>
<td>0.0458</td>
<td>0.0476</td>
<td>0.0954</td>
<td>0.2878</td>
</tr>
<tr>
<td>20</td>
<td>40</td>
<td>0.0486</td>
<td>0.0436</td>
<td>0.0536</td>
<td>0.0490</td>
<td>0.1036</td>
<td>0.3506</td>
</tr>
<tr>
<td>40</td>
<td>40</td>
<td>0.0452</td>
<td>0.0438</td>
<td>0.0468</td>
<td>0.0556</td>
<td>0.1242</td>
<td>0.4276</td>
</tr>
<tr>
<td>25</td>
<td>50</td>
<td>0.0504</td>
<td>0.0418</td>
<td>0.0424</td>
<td>0.0536</td>
<td>0.1160</td>
<td>0.4034</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>0.0498</td>
<td>0.0452</td>
<td>0.0480</td>
<td>0.0648</td>
<td>0.1294</td>
<td>0.4684</td>
</tr>
</tbody>
</table>
Figure 1. Nonnormal distributions.

Platykurtic Distribution

Leptokurtic Distribution

Slightly Skewed Distribution

Moderately Skewed Distribution

Skewed Leptokurtic Distribution
Figure 2. Selected power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Normal data at $\alpha=0.05$.

Figure 3. Selected power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Platykurtic data at $\alpha=0.05$.

Figure 4. Selected power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Leptokurtic data at $\alpha=0.05$.

Figure 5. Selected power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Slightly Skewed data at $\alpha=0.05$. 
Figure 6. Selected power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Moderately Skewed data at $\alpha=0.05$.

Figure 7. Selected power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Skewed Leptokurtic data at $\alpha=0.05$.

Figure 8. Bootstrap Type I error rates for the data depth test. Horizontal reference line is at $\alpha = 0.05$.

Figure 9. Power curves for data depth test, F-test, O’Brien’s test, and Brown-Forsythe test for Skewed Leptokurtic data at $\alpha=0.05$. The samples were aligned using their respective 30th percentiles.
REFERENCES


