ABSTRACT

LIU, XIAONI. New Methods Using Levene Type Tests For Hypotheses About Dispersion Differences. (Under the direction of Professors Dennis Boos and Cavell Brownie).

Testing equality of scale arises in many research areas including clinical data analysis. In contrast to procedures for tests on means, tests for variances derived assuming normality of the parent populations are highly non-robust to non-normality. Levene type tests are well known to be robust tests for equality of scale for the one-way design; the current standard test uses the ANOVA F test on absolute deviations from the sample medians. We first develop a new modified version of the standard Levene test that improves its null performance and power. Applying a Box-Andersen correction to the ANOVA F test further improves the performance.

We also extend the robust Levene type tests to the two-way design with one observation per cell, the randomized complete block design (RCB). Currently, the available Levene type tests for RCB designs employ either standard ANOVA F tests on the absolute values of ordinary least squares (OLS) residuals, or weighted least squares (WLS) ANOVA F tests on the OLS residuals. These two tests can be liberal, especially under non-normal distributions. Instead, we use OLS ANOVA F tests on the absolute values of residuals obtained from models fit by least absolute deviation (LAD) estimation and by Huber Proposal 2 M-estimation. We also apply bootstrap

methods to these Levene type tests and compare these tests in terms of robustness and power using simulation.

NEW METHODS USING LEVENE TYPE TESTS FOR HYPOTHESES ABOUT DISPERSION DIFFERENCES

by

Xiaoni Liu

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

STATISTICS

Raleigh

2006

Approved By:

Dennis Boos Chair of Advisory Committee	Cavell Brownie Co-Chair of Advisory Committee
David Dickey	Jason Osborne

To My Parents and Younger Brother

Biography

Xiaoni Liu was born on December 1st, 1978 to Xinhua Liu and Zhongzhen Xie in Shaoyang, China. She graduated from No. 2 High School in Shaoyang in 1996. From September 1996 to July 2001, Xiaoni attended University of Science and Technology of China (USTC), Hefei, Anhui Province, China, where she received a Bachelor of Science degree in Science and English and a Bachelor of Engineering Degree in Computer Science in July 2001. From August 2001 to July 2002, she studied in the Department of Mathematics and Statistics at the University of North Florida. She then joined the graduate program in the Department of Statistics at North Carolina State University in August 2002. In May 2004, she earned her Master of Statistics degree at North Carolina State University. She continued her studies towards a Ph.D degree in statistics at North Carolina State University. Xiaoni was recognized as member of Phi Kappa Phi and Mu Sigma Rho.

Acknowledgements

First, I would like to express my deepest appreciation to both of my advisors: Dr. Dennis Boos and Dr. Cavell Brownie and thank them for their valuable ideas, helpful comments, great support, continuous encouragement and kind patience. Without their help and advising, I can't imagine completing my dissertation. I feel very lucky to have worked with them and have benefited from their valuable experience. In addition, I would like to extend my profound gratitude to Drs. David Dickey and Jason Osborne for their great advice and the time they dedicated to reviewing my thesis.

I would like to thank the entire statistics department for support in the past four years with special thanks to Dr. Sastry Pantula, Dr. William Swallow and Dr. Leonard Stefanski for financial support and supervision, Dr. Peter Bloomfield for his great comments on my research, and Mr. Terry Byron for computing support. My special thanks also go to Drs. Bibhuti Bhattacharyya, Marie Davidian, Jacqueline Hughes-Oliver, Anastasios Tsiatis, Daowen Zhang and Helen Zhang for their wonderful lectures.

I also want to thank my dear friends, Xing Sun, Qiong Wang, Liyun Ma, Yun Chen, Guozhi Gao, Zhaoling Meng, Xiaohui Luo, Xiang Guo, Jiajun Liu, Feng Liu and Xi Chen for their friendship. They make me feel that life is so beautiful.

Finally, my most heartfelt thanks go to my dearest parents and brother, for their selfless love and support!

Contents

List of Figures

List of Tables

Chapter 1

The Modified LevMed Test for a One-Way Design

1.1 Introduction

Testing equality of variances is of interest in many applications including quality control in industry, development of educational methods, and studies on variability in biological populations. Even in clinical research, comparing variability can be as important as comparing averages. Zwinderman and Cleophas (2005) summarized the main applications of variance tests in clinical research and gave some situations where variability is more relevant than means in clinical data analysis. For example, it is important to compare variability in response for different formulations of a drug, especially for drugs with small therapeutic windows. Tests for equality of variances can

also be used to compare variability in patient characteristics for different treatment groups.

Another context in which homogeneity of variances is tested is as a preliminary to some standard statistical procedures. For example, the F test for equality of variances is sometimes suggested in order to decide whether to use the pooled variance t-test or the unequal variance Welch t-test to test equality of two means. In general, it is recommended not to use this preliminary test approach, partly because normal theory tests for means, such as the pooled t test, are robust to non-normality, whereas normal theory tests for variances, such as the F test, are highly sensitive to departures from normality. So typically, it is better to decide whether to use the pooled variance t-test or the unequal variance Welch t-test based on other grounds, such as whether the groups were randomly assigned or not.

1.1.1 Normal Theory Tests

The common normal theory variance tests include the F test for two populations, Bartlett's test (Bartlett, 1937) and Hartley's test (Hartley, 1950b) for k ($k \geq 2$) populations. Let $\{X_{i1}, \cdots, X_{in_i}, i = 1, \cdots, k\}$ represent k independent samples, where for the ith sample, $\{X_{ij}, j = 1, \cdots, n_i\}$, the sample members are iid normally distributed, $N(\mu_i, \sigma_i^2)$. In this situation, $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \overline{X}_i)^2$ is the sample variance for the ith sample and $N = \sum_{i=1}^{k} n_i$.

The F Test for $k = 2$ Populations

The F test is the classical normal theory test to compare variability of two populations. For the alternative hypothesis $H_1 : \sigma_1^2 \neq \sigma_2^2$, we use the F test statistic

$$F = \frac{s_1^2}{s_2^2} \tag{1.1}$$

and reject the null hypothesis when F is greater than the $100(1 - \alpha/2)$th percentile or less than $100(\alpha/2)$th percentile of the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, where n_1 and n_2 are sample sizes for the two groups.

Bartlett's Test for $k \geq 2$ Populations

The normal theory likelihood ratio statistic for testing $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ is

$$B = (N - k) \ln s_p^2 - \sum_{i=1}^{k} (n_i - 1) \ln s_i^2, \tag{1.2}$$

where $s_p^2 = \sum_{i=1}^{k} (n_i - 1) s_i^2 / (N - k)$ and $N = \sum_{i=1}^{k} n_i$. A correction factor is

$$C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^{k} \frac{1}{n_i - 1} - \frac{1}{N - k} \right). \tag{1.3}$$

Bartlett's test statistic is the bias-corrected statistic:

$$B_c = B/C \tag{1.4}$$

The null hypothesis is rejected when B_c is greater than the $100(1-\alpha)$th percentile of the chi-squared distribution with $(k-1)$ degrees of freedom.

Hartley's Test for $k \geq 2$ Populations

Hartley (1950b) proposed the maximum F-ratio as a short-cut test for homogeneity of variances for unbalanced designs under the normal distribution. Suppose $s^2_{\max}$ is the largest sample variance and $s^2_{\min}$ is the smallest sample variance among the k sample variances.

Hartley's test statistic is

$$F = \frac{s^2_{\max}}{s^2_{\min}}. \tag{1.5}$$

The null hypothesis is rejected when F is greater than the $100(1-\alpha)$th percentile of the $F_{\max}$ distribution (Hartley, 1950b) with k and $n-1$ degrees of freedom assuming balanced designs.

The F test and Bartlett's test are the most popular normal theory tests taught in elementary statistical courses. In particular, Bartlett's test has good power under normality, even for unequal sample sizes. However, it is well known to be highly sensitive to non-normality (Box, 1953). Therefore, it is not recommended to be used except when the data are known to be normally distributed.

1.1.2 Robust Variance Tests

The non-robustness of normal theory tests for equality of variances results because test statistics are not asymptotically distribution-free but depend on the kurtosis of the parent distribution. Thus, in the past 50 years, a number of tests have been proposed to solve this problem, that is, to be Type I error robust to non-normality. To paraphrase Boos and Brownie (2004), the three approaches used to construct robust tests are:

1. to use an estimate of kurtosis to adjust the normal-theory test procedures (Box and Andersen, 1955, Shoemaker, 2003)

2. to replace the observations in the original data set by scale variables such as the absolute deviations from the mean or median followed by the ANOVA test on the new data set (Levene, 1960, Miller, 1968, Brown and Forsythe, 1974); a related procedure is to perform ANOVA on the jackknife pseudo-values of a scale variable such as the log of the sample variance (Miller, 1968)

3. to get p values for a given test statistic with a resampling method (Box and Andersen, 1955, Boos and Brownie, 1989).

Among the three approaches, we recommend the second approach in which t or F tests on the scale variables are used to test homogeneity of variances or scale for the original variables. The resulting tests are not only simple to implement but also robust to non-normality, even in small samples. Levene (1960) was the first to

propose procedures using the one-way ANOVA F test on the new variables $Y_{ij} = |X_{ij} - \overline{X}_i|$, or more generally, $Y_{ij} = g(|X_{ij} - \overline{X}_i|)$, where g is monotonically increasing on $(0, \infty)$. Miller (1968) showed that $Y_{ij} = |X_{ij} - \overline{X}_i|$ is asymptotically incorrect for asymmetric populations, but that using the median instead of the mean to center the spread variables is asymptotically correct. Brown and Forsythe (1974) formally studied this modification of Levene's method where the median was used instead of the mean to center the variables. The one-way ANOVA F test on the spread variables, $Z_{ij} = |X_{ij} - M_i|$, where M_i is the sample median of ith group, is referred to in the literature as the Brown-Forsythe test (e.g. SAS PROC GLM) or as Lev1:med (Conover et al., 1981; Boos and Brownie, 1989, 2004). We will refer to it here as the LevMed test. Brown and Forsythe (1974) and Conover et al. (1981) studied the small sample properties of the LevMed test and demonstrated that it had satisfactory Type I and Type II error properties for many distributions. However, Boos and Brownie (1989), Lim and Loh (1996), and Shoemaker (2003) have noted that the LevMed test can be conservative with loss of power. Boos and Brownie (1989) concluded that null performance of the LevMed test will be generally good for sample sizes greater than or equal to 8. However, for small and odd sample sizes, the test is extremely conservative because of zero values of the scale variables inflating the estimate of within-group variance in the denominator of the F statistic.

Layard (1973) developed the k-sample generalization of Miller's two-sample jackknife test based on sample variances. Brown and Forsythe (1974) showed that it was

not as good as the LevMed test in terms of robustness. Boos and Brownie (1989) applied the bootstrap technique to the LevMed test and to other tests for equality of scale. The bootstrap version of the LevMed test performs better than the LevMed test in terms of robustness and power, especially for small sample sizes. However, these methods based on resampling are computationally more complicated than the LevMed test, which can be performed with standard software.

Our aim is to propose a new test for equality of scale that is based on the LevMed test, is simple to compute, and also performs better for small samples than the LevMed test. The chapter is organized as follows. Section 2 defines the model, describes the LevMed test in detail, introduces the modified LevMed test and summarizes the results from preliminary simulations on the LevMed variables and the modified LevMed variables. The asymptotic properties and small sample problems of the LevMed test are introduced briefly. Section 3 describes the simulation results for the performance of the modified LevMed test including the null performance and power. This section also introduces some other tests and compares them to the modified LevMed test. Section 4 gives an example in which some of the robust tests are compared. Section 5 compares the modified LevMed test to the Hines LevMed test (Hines and Hines, 2000). Section 6 presents some conclusions.

1.2 The LevMed Test and the Modified LevMed Test

1.2.1 Model

Let $\{X_{i1}, \cdots, X_{in_i}, i = 1, \cdots, k\}$ represent k independent samples, where for the ith sample, $\{X_{ij}, j = 1, \cdots, n_i\}$, the sample members are iid with distribution function $G_i(x) = G_0((x - \mu_i)/\sigma_i)$. Let $N = \sum_{i=1}^{k} n_i$ be the total sample size. We assume that $G_0(x)$ has mean 0 and variance 1, and finite fourth moment. To test homogeneity of variances, the null hypothesis is $H_0 : \sigma_1^2 = \cdots = \sigma_k^2$ or equivalently $H_0 : \sigma_1 = \cdots = \sigma_k$, where G_0 and μ_i are unknown. Under the location-scale family assumption, any other scale parameter such as the mean absolute deviation from the median is related to the standard deviation by $\mathrm{E}|X_{ij} - \theta_i| = c\sigma_i$, $i = 1, \ldots, k$, for some constant c. For example, if G_0 is the standard normal distribution, $\mathrm{E}|X_{ij} - \theta_i| = \sqrt{2/\pi}\sigma_i = .798\sigma_i$, and here $c = 0.798$. Thus, a test for equality of scale parameters is represented as $H_0 : c\sigma_1 = \cdots = c\sigma_k$, which is equivalent to the null hypothesis of equality of standard deviations or of variances.

1.2.2 The LevMed Test

Definition

Let $Z_{ij} = |X_{ij} - M_i|$, where M_i is the sample median for the ith group. The LevMed test is based on the one-way ANOVA statistic:

$$F = \frac{\sum_{i=1}^{k} n_i(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_{i\cdot})^2/(N-k)} \tag{1.6}$$

where $\overline{Z}_{i\cdot} = \sum_{j=1}^{n_i} Z_{ij}/n_i$ and $\overline{Z}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}/N$. The critical values of F in (1.6) are obtained from the F distribution with $k-1$ and $N-k$ degrees of freedom.

Asymptotic Properties of the Mean Absolute Deviation from the Sample Median

Here we use the M-estimator method to illustrate the asymptotic validity of comparing the LevMed F statistic to an F distribution. Basically, we demonstrate that the denominator of the F statistic is estimating the common variance of the spread variables under the null hypothesis. Miller (1968) was the first to give a related asymptotic argument. To make the notation simpler, we work with a single sample and just one subscript. Let $X_1, \cdots, X_n$ be iid from distribution F, where F has density f, mean μ, median θ_1, and the basic LevMed scale parameter is $\mathrm{E}|X_i - \theta_1| = \theta_2$, and $\mathrm{var}(|X_i - \theta_1|) = \theta_3$. We show that n times the asymptotic variance of the sample mean absolute deviation from the sample median is θ_3.

Following the notation in Stefanski and Boos (2002), the estimators of interest satisfy $\sum_{i=1}^{n} \psi(X_i, \widehat{\theta}) = \mathbf{0}$, where

$$\psi(X_i, \theta) = \begin{pmatrix} \frac{1}{2} - I(X_i \leq \theta_1) \\ |X_i - \theta_1| - \theta_2 \\ (X_i - \theta_1)^2 - \theta_2^2 - \theta_3 \end{pmatrix}. \tag{1.7}$$

Note that $\widehat{\theta}_3$ is the sample variance of the basic spread variables $|X_i - \widehat{\theta}_1|$. After making some calculations, we have

$$\mathbf{A}(\theta_\mathbf{0}) = E[-\psi'(X_1, \theta_\mathbf{0})] = \begin{pmatrix} f(\theta_1) & 0 & 0 \\ 0 & 1 & 0 \\ 2\mu_1 - 2\theta_1 & 2\theta_2 & 1 \end{pmatrix}, \tag{1.8}$$

$$\mathbf{B}(\theta_\mathbf{0}) = E[\psi(X_1, \theta_\mathbf{0})\psi(X_1, \theta_\mathbf{0})^T], \tag{1.9}$$

$$\mathbf{V}(\theta_\mathbf{0}) = \mathbf{A}(\theta_\mathbf{0})^{-1}\mathbf{B}(\theta_\mathbf{0})\{\mathbf{A}(\theta_\mathbf{0})^{-1}\}^T. \tag{1.10}$$

The $(2, 2)$ element of equation (1.10) is the asymptotic variance of $\sqrt{n}\widehat{\theta}_2$:

$$v_{22} = \mu_2' - 2\mu\theta_1 + \theta_1^2 - \theta_2^2 \tag{1.11}$$

where $\mu_2' = E(X_i^2)$. If the LevMed variable $Z_i = |X_i - \widehat{\theta}_1|$ is a correct "spread variable," we need $v_{22} = \theta_3$. On the other hand, expanding the definition of θ_3, we have

$$\theta_3 = \mathrm{E}(X_i - \theta_1)^2 - \theta_2^2 = \mu_2' - 2\mu\theta_1 + \theta_1^2 - \theta_2^2, \tag{1.12}$$

which is the same as v_{22}. Thus, Z_i is an appropriate spread variable to be used in t or F statistics.

Shortcomings of the LevMed Test

Monte Carlo studies carried out by Conover et al. (1981) demonstrated the Type I and Type II error robustness of the LevMed procedure for moderate sample sizes. Boos and Brownie (1989) mentioned that when the sample size is odd and less than 8, the test can be extremely conservative, because zero values of Z_{ij} inflate the estimate of within-group variance in the denominator of the F statistic in equation (1.6). Subtracting the sample median also causes correlation among the Z_{ij} within the same group. To solve these problems due to subtracting the sample median (instead of the true but unknown median), when sample sizes are odd and small, O'Brien (1978) suggested deletion of one observation randomly from each group, and Conover et al. (1981) suggested deletion of the middle observation in each group. O'Brien's suggestion leads to many possible outcomes for the same data set, and the idea of Conover et al. can lead to liberal tests. Instead, we propose in the next section,

deleting the smallest ordered Z_i. A similar but different proposal was given in Hines and Hines (2000).

1.2.3 The Modified LevMed Test

We propose a modification of the LevMed test with the goal of improving performance when sample sizes are small ($n < 10$). For the ith sample, let $Z_{i(1)}, \cdots, Z_{i(n_i)}$ represent the ordered LevMed variables. The conservative behavior of the LevMed test for n_i small and odd, $i = 1, \cdots, k$, is explained in part by noting that $Z_{i(1)}$ must be 0, and consequently the Z_{ij} tend to be negatively correlated and have an inflated variance. Each of these properties will lead to conservative performance of the ANOVA F test. When n_i is even, a consequence of defining M_i as the average of the two middle X_{ij} values is that $Z_{i(1)} = Z_{i(2)}$. In small samples, $Z_{i(1)} = Z_{i(2)}$ results in the within sample variance of Z_{ij} being too small, and liberal behavior of the F test.

Our solution is to delete $Z_{i(1)}$ in each sample and carry out the ANOVA F test on the reduced set of Z_{ij}. The modified LevMed test (MLM) is carried out as follows:

1. Sort the set of the LevMed variables $\{Z_{ij}\}$ within each sample, giving $Z_{i(1)}, \cdots, Z_{i(n_i)}$.

2. Delete $Z_{i(1)}$ from each sample.

3. The remaining scale variables are the set of modified scale variables. We denote them as $\{Z^*_{ij}, j = 1, \cdots, n_i - 1, i = 1, \cdots, k\}$.

Similar to the LevMed test, the modified LevMed test is based on the one-way ANOVA statistic:

$$F^* = \frac{\sum_{i=1}^{k}(n_i - 1)(\overline{Z}^*_{i\cdot} - \overline{Z}^*_{\cdot\cdot})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i-1}(Z^*_{ij} - \overline{Z}^*_{i\cdot})^2/(N-2k)} \tag{1.13}$$

where $\overline{Z}^*_{i\cdot} = \sum_{j=1}^{n_i-1} Z^*_{ij}/(n_i-1)$ and $\overline{Z}^*_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i-1} Z^*_{ij}/(N-k)$. The critical values of F^* are obtained from the F distribution with $k-1$ and $N-2k$ degrees of freedom.

The main step of the Levene type tests is to perform the one-way ANOVA test on a scale variable such as the absolute deviation from the mean or the median. Such scale variables are not normally distributed, which violates the normally-distributed assumption of the ANOVA F test. However, the ANOVA F test has been shown to be robust to non-normality due to the Central Limit Theorem. Box and Andersen (1955) proposed a correction to the degrees of freedom of the ANOVA F test based on permutation theory, which can improve the robustness of the ANOVA F test for means. The main idea of this correction is to match the fourth sample moment of the permutation test statistic to that of an F variable to get a correction factor by which the degrees of freedom in the original ANOVA F test are multiplied.

Therefore, another test procedure is obtained by applying the Box-Andersen degrees of freedom correction to the one-way ANOVA F test on the modified LevMed variables. We call this variation of the modified LevMed test the MLM-BA test which is also based on the ANOVA statistic, F^* in the formula (1.13). When $k = 2$ and

$n_1 = n_2 = 3$, we can't adjust the degrees of freedom of the ANOVA F test. In other cases, we can adjust the degrees of freedom of the one-way ANOVA F test. Therefore, except when $k = 2$, and $n_1 = n_2 = 3$, the critical values of F^* are obtained from the F distribution with $d \times (k-1)$ and $d \times (N-2k)$ degrees of freedom, where d relates to the estimated kurtosis (Box and Andersen, 1955), and it is given by

$$d = 1 + \frac{N+1}{N-1}\frac{c_2}{(N^{-1}+A)^{-1} - c_2}, \text{where } A = \frac{N+1}{2(k-1)(N-k)}\left(\frac{k^2}{N} - \sum_{i=1}^{k}\frac{1}{n_i}\right) \quad (1.14)$$

Here c_2 is the estimated kurtosis for the whole sample, $c_2 = k_4/k_2^2$, where

$$k_4 = \frac{N(N+1)S_4 - 3(N-1)S_2^2}{(N-1)(N-2)(N-3)}, \quad k_2 = \frac{S_2}{N-1},$$

and

$$S_r = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \overline{X})^r$$

where $\overline{X}$ is the overall mean of all N observations. For balanced designs with $n_1 = n_2 = \cdots = n_k$, we can simplify the formula (1.14) to

$$d = 1 + \frac{N+1}{N-1}\frac{c_2}{N - c_2}.$$

1.2.4 Small Sample Properties of the LevMed and Modified LevMed Variables

The LevMed test and the modified LevMed test are based on comparing the one-way ANOVA F statistic to an F distribution. For the usual F statistic based on sample means and variances, the main underlying assumptions are equality of variances, independence of observations within samples, and normality of the observations. Various studies have focused on determining the impact of failures of these assumptions. The effect of inequality of variances is fairly mild if the sample sizes are roughly equal (Box, 1954). The effect of negative correlation within samples is that the sample variance divided by sample size overestimates the variance of the corresponding mean (Box, 1954), causing the test to be liberal. The F statistic is fairly robust to deviations from normality. In general, kurtosis larger than the normal distribution kurtosis causes the resulting test to be conservative (Box and Andersen, 1955). Skewness causes the numerator and denominator of the F statistic to be correlated.

With spread variables used in the F statistics, we also have to worry about the bias of the mean estimates and whether the sample variances of the variables divided by sample size overestimate the variances of the estimated means due to the correlation induced within samples by estimating the median. Thus, in the next subsections we focus on these latter two concerns as well as on whether the numerator and denominator of the F statistics are correlated.

Sample Means and Variances of LevMed and Modified LevMed Variables

In the usual one-way ANOVA F test, given independent samples $\{X_{ij}, j = 1, \ldots, n_i\}$, $i = 1, \cdots, k$, the sample means $\overline{X}_i$ and the sample variances s_i^2 are the building blocks of the F statistic. For the usual ANOVA F test, we have $\mathrm{E}(\overline{X}_i) = \mu_i$ (mean unbiasedness), and $\mathrm{var}(\overline{X}_i)$ can be estimated by s_i^2/n_i (unless there is correlation within samples). Actually, we use the pooled estimate of variance s_{P}^2/n_i. In a previous section, we showed that for the ith sample, asymptotically the sample variance of the LevMed Z_{ij} divided by the ith sample size, n_i, estimates the variance of the mean of the Z_{ij}, $\mathrm{var}(\overline{Z}_{i\cdot})$. Here we explore by simulation how well this property holds in small samples. In addition we study the small sample bias of the sample mean of the LevMed and the modified LevMed variables. The bias itself is not the concern, but the fact that the bias changes with sample size is a problem for data sets where the n_i are small and different. Thus, under a null hypothesis of equal scale, if the bias is different for different sample sizes, then given unbalanced data, the numerator of the F statistic will tend to be proportional to a variable more like a noncentral chisquared distribution than the required central chisquared distribution.

Suppose that we draw a small sample $\{X_j, j = 1, \cdots, n\}$ from a population with known distribution. The sample median is $\widehat{\theta}_1$. The LevMed variables are $\{Z_j = |X_j - \widehat{\theta}_1|, j = 1, \cdots, n\}$ and the modified LevMed variables are $\{Z_j^*, j = 1, \cdots, n-1\}$. In the Monte Carlo simulation, we generate $S = 10,000$ such samples. For the ith Monte Carlo sample, let $\widehat{\mu}_i$ be the sample mean of the Z or Z^*, and let s_i^2 be the sample

variance of the Z or Z^*. The average over replicates of $\widehat{\mu}_i$ is $\overline{\widehat{\mu}} = (1/S)\sum_{i=1}^{S}\widehat{\mu}_i$. This estimates $\mathrm{E}\Big((1/n)\sum_{j=1}^{n}|X_j - \widehat{\theta}_1|\Big)$, which in turn should estimate $\theta_2 = \mathrm{E}|X_1 - \theta_1|$, the mean absolute deviation from the true median θ_1.

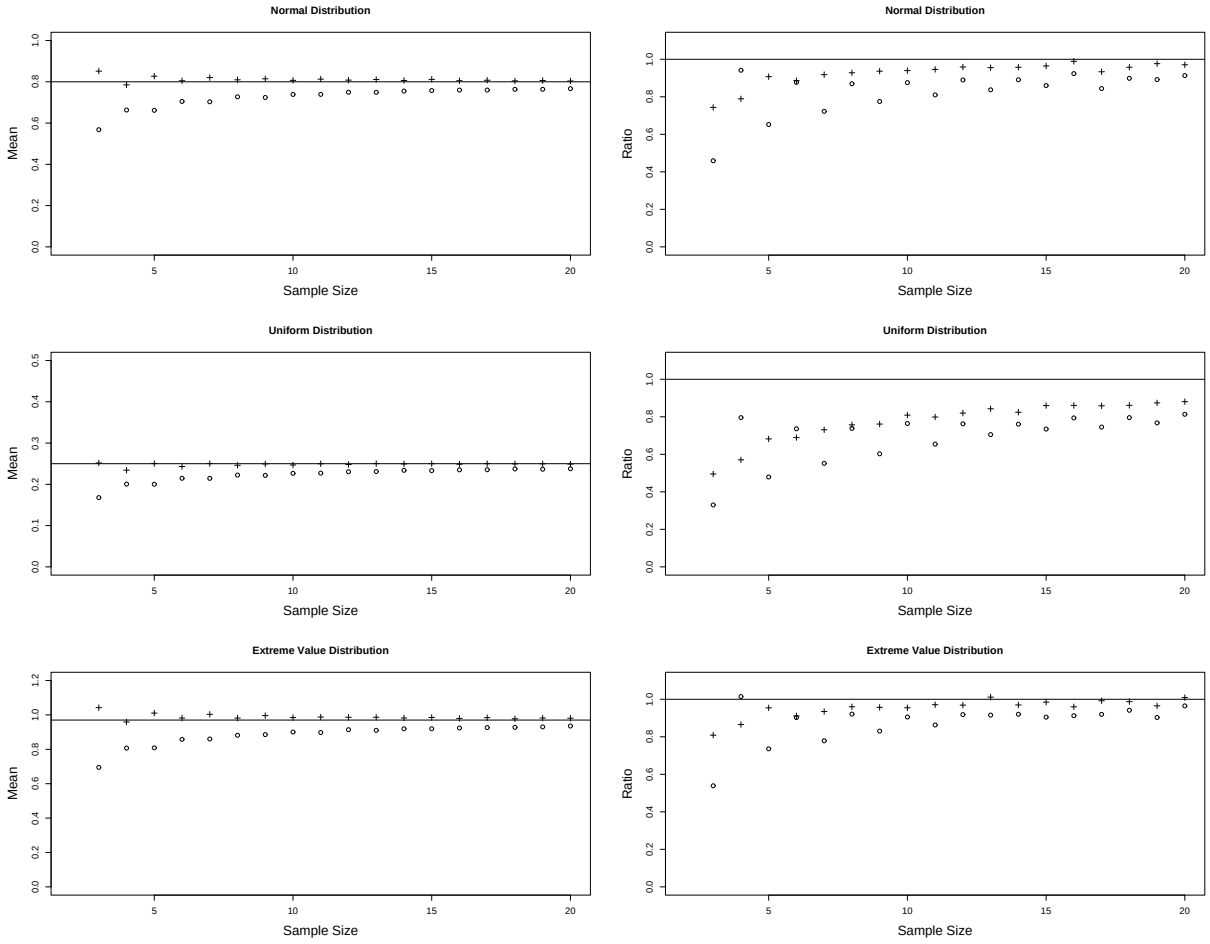

Figure 1.1: Left Side: Monte Carlo Estimates of Expectations of the Sample Means for each of the scale variables, Z and Z^* versus Sample Size. Right Side: Monte Carlo Estimates $(n * s^2_{\widehat{\mu}})/\overline{s}^2$ for Z and Z^* versus Sample Size. $\circ\,\circ\,\circ$ LevMed Scale or Z, $+++$ Modified LevMed Scale or Z^*

The left panels of Figure 1.1 give these average estimates $\overline{\widehat{\mu}}$ versus sample size for three different distributions, the standard normal, uniform(0,1), and extreme

value(0,1) distributions. The horizontal line is for the desired target θ_2. Clearly the "+" symbols for the modified LevMed variable are much more stable as n changes than the LevMed variable represented by open circles. For example, the values estimated by the mean of the LevMed variables at $n = 3$ and $n = 6$ are quite different, whereas the corresponding values for the modified LevMed variables are quite close. By $n = 10$ both sets of means are fairly stable, although the means of the modified variables, Z^* are closer on average to the true values of θ_2, 0.80, 0.25, and 0.97, respectively, for the three distributions.

The average of the sample variances is $\overline{s}^2 = (1/S)\sum_{i=1}^{S} s_i^2$. We also obtain the Monte Carlo sample variance of the $\widehat{\mu}_i$, $s^2_{\widehat{\mu}} = (S-1)^{-1}\sum_{i=1}^{S}(\widehat{\mu}_i - \overline{\widehat{\mu}})^2$. This latter quantity estimates $\text{var}(\widehat{\mu}_i)$, the true variance of the mean estimate. Validity of the F test requires that $\text{var}(\widehat{\mu}_i)$ be estimated by $\overline{s}^2/n$ for each sample. Thus, we want $\overline{s}^2 \approx n * s^2_{\widehat{\mu}}$. To examine whether this relationship holds, we calculated the ratio $(n * s^2_{\widehat{\mu}})/\overline{s}^2$ for each of the two scale variables for each sample size and distribution. If this ratio is less than one, then the sample variances are too large on average, and we might extrapolate that the denominator of the F statistic will be too large on average.

The right panels of Figure 1.1 give these ratios for the three distributions. The horizontal line is for the target value of 1. Again, we see that the modified LevMed variables are more stable as n changes and also closer to the target value than the original LevMed variables. Convergence of both ratios to 1 is slower for the uniform

distribution than for the other two distributions. The LevMed ratio is especially low for $n = 3$. This can be anticipated due to the single 0 value inflating the sample variance.

Correlation within the Samples

In the usual one-way ANOVA F test, one of the assumptions is that the observations within samples are independent. In this part, we focus on the correlation within the samples for the scale variables Z and Z^*.

In the simulation, the distributions are again normal, uniform, and extreme value. The sample sizes configurations are: (I) n=3, (II) n=4, (III) n=5, (IV) n=6, (V) n=20 and (VI) n=100. The number of Monte Carlo replications is S=10,000. The within-sample correlation estimates are summarized in Table 1.1. We compute these entries as follows. First, we generate a sample $\{X_1, \cdots, X_n\}$, then center them to get the LevMed variables, $\{Z_j, j = 1, \cdots, n\}$, or the modified LevMed variables, $\{Z_j^*, j = 1, \cdots, n-1\}$. After that, we calculate the correlation between every pair of scale variables and get the average. For example, when $n = 3$, we generate a random sample, $\{X_1, X_2, X_3\}$. After centering, we get the LevMed variables, $\{Z_1, Z_2, Z_3\}$. After that, we compute estimated correlations across all S samples between Z_1 and Z_2, between Z_1 and Z_3 and between Z_2 and Z_3. The average of the 3 estimated correlations is the entry for LevMed under $n = 3$.

The entries labeled LevMed(θ_1) are results for the LevMed variable using true

medians instead of sample medians to calculate the Z values. As we expect, the correlation estimates for the LevMed(θ_1) are essentially 0 for all the cases. For small sample sizes, both the LevMed variables and the modified LevMed variables lead to negative within-sample correlations. The effect of negative correlation within samples is that the sample variance divided by sample size overestimates the variance of the corresponding mean, causing the test to be conservative. It is apparent that the within-sample correlations are more strongly negative for the LevMed variables than for the modified variables, for n odd and small.

Table 1.1: Estimates of Correlations within Samples for the Scale Variables

		Sample Size					
Distribution	Test	n=3	n=4	n=5	n=6	n=20	n=100
Normal	LevMed(θ_1)	0.00	0.00	0.01	0.00	0.00	0.00
	LevMed	−0.21	−0.01	−0.07	−0.02	0.00	0.00
	MLM	−0.15	−0.07	−0.04	−0.02	0.00	0.00
Uniform	LevMed(θ_1)	0.00	0.00	0.00	0.00	0.00	0.00
	LevMed	−0.29	−0.05	−0.12	−0.04	−0.01	0.00
	MLM	−0.25	−0.16	−0.10	−0.06	−0.01	0.00
Extreme	LevMed(θ_1)	0.00	0.00	0.00	0.00	0.00	0.00
Value	LevMed	−0.18	0.00	−0.06	−0.02	0.00	0.00
	MLM	−0.12	−0.05	−0.03	−0.02	0.00	0.00

LevMed(θ_1) is LevMed using known medians. Entries are based on 10,000 replications and have standard error ≤ 0.01.

Correlation between the Numerator and Denominator of the F Statistics

We also assess the correlation between SSR and SSE based on each of the two scale variables because under the usual ANOVA assumptions, these two sums of squares are independent. Here, SSR is the between-group or numerator sum of squares,

and SSE is the within-group or denominator sum of squares, for the F statistic. For example, for the LevMed variables $\{Z_{ij}, i = 1, \cdots, k; j = 1, \cdots, n_i\}$, SSR is $\sum_{i=1}^{k} n_i(\overline{Z}_{i\cdot} - \overline{Z}_{\cdot\cdot})^2$, and SSE is $\sum_{i=1}^{k}\sum_{j=1}^{n_i}(Z_{ij} - \overline{Z}_{i\cdot})^2$, where $\overline{Z}_{i\cdot} = \sum_{j=1}^{n_i} Z_{ij}/n_i$ and $\overline{Z}_{\cdot\cdot} = \sum_{i=1}^{k}\sum_{j=1}^{n_i} Z_{ij}/N$. Thus the F statistic (1.6) can be written as

$$F = \frac{SSR/df(SSR)}{SSE/df(SSE)}. \tag{1.15}$$

In the simulation, the distributions are again normal, uniform, and extreme value. The sample size configurations are: (I) n=3, (II) n=4, (III) n=5, (IV) n=6, (V) n=20 and (VI) n=100. The group size configurations are: (i) k=2, (ii) k=4 and (iii) k=8. The number of Monte Carlo replications is S=10,000. The correlation estimates between SSR and SSE are summarized in Table 1.2. The entries labeled LevMed(θ_1) are results for the LevMed variable using true medians instead of sample medians. The point is that skewness of the variables will cause some amount of correlation, but the best we might hope for is the correlation present when we use known medians.

For all distributions at $n = 3$, the correlations for the modified LevMed variables are closer to the correlations of the known median case than are the original LevMed variables. This reverses at the normal and uniform for $n = 4$. Generally, the correlations are a little closer for the modified variables, but we have trouble drawing strong conclusions from this table.

Table 1.2: Estimates of Correlation between SSR and SSE

			Sample Size					
Distribution	k	Test	n=3	n=4	n=5	n=6	n=20	n=100
Normal	2	LevMed(θ_1)	0.12	0.11	0.11	0.08	0.05	0.03
		LevMed	0.38	0.10	0.26	0.07	0.06	0.03
		MLM	0.23	0.17	0.18	0.10	0.06	0.03
	4	LevMed(θ_1)	0.15	0.14	0.12	0.13	0.08	0.03
		LevMed	0.47	0.12	0.29	0.14	0.10	0.03
		MLM	0.26	0.22	0.19	0.17	0.10	0.03
	8	LevMed(θ_1)	0.17	0.14	0.14	0.13	0.08	0.03
		LevMed	0.48	0.10	0.29	0.14	0.09	0.03
		MLM	0.27	0.21	0.18	0.17	0.08	0.03
Uniform	2	LevMed(θ_1)	−0.28	−0.25	−0.25	−0.23	−0.15	−0.06
		LevMed	0.12	−0.21	0.03	−0.17	−0.07	−0.04
		MLM	0.09	−0.15	0.00	−0.12	−0.07	−0.04
	4	LevMed(θ_1)	−0.34	−0.32	−0.30	−0.27	−0.16	−0.09
		LevMed	0.15	−0.27	0.03	−0.22	−0.09	−0.08
		MLM	0.11	−0.19	−0.03	−0.16	−0.09	−0.08
	8	LevMed(θ_1)	−0.38	−0.36	−0.33	−0.32	−0.19	−0.09
		LevMed	0.18	−0.29	0.04	−0.23	−0.11	−0.09
		MLM	0.13	−0.21	−0.02	−0.16	−0.10	−0.09
Extreme Value	2	LevMed(θ_1)	0.58	0.51	0.49	0.44	0.29	0.12
		LevMed	0.65	0.35	0.50	0.37	0.24	0.10
		MLM	0.53	0.48	0.45	0.41	0.24	0.10
	4	LevMed(θ_1)	0.63	0.60	0.53	0.51	0.32	0.17
		LevMed	0.74	0.47	0.54	0.43	0.28	0.14
		MLM	0.64	0.59	0.48	0.48	0.28	0.14
	8	LevMed(θ_1)	0.64	0.60	0.55	0.55	0.35	0.17
		LevMed	0.77	0.45	0.59	0.47	0.30	0.14
		MLM	0.65	0.59	0.53	0.51	0.30	0.14

LevMed(θ_1) is LevMed using true medians. Entries are based on 10,000 replications and have standard error from 0.007 to 0.026.

Discussion

The simulation results of the previous subsections suggest that an F statistic calculated on the modified LevMed variables may perform better than that based on

the original LevMed variables. However, because there are a number of competing aspects, we proceed to a direct comparison of the test procedures in the next section.

1.3 Simulations

1.3.1 Other Tests for equality of variances or Scale

Our goal in this section is to compare by simulation the modified LevMed procedures including the MLM test and the MLM-BA test to the classical LevMed procedure and to some other tests for equality of variances or scale. Because it is well known that normal-theory tests such as Bartlett's test are sensitive to non-normality, and these have been reported on extensively elsewhere, we will not consider nonrobust normal-theory tests here. The other robust tests used in our simulation are defined in the next few subsections.

Shoemaker's Test

Shoemaker (2003) proposed a new homogeneity of variances test that is robust to non-normality. Let s_i^2 be the sample variance of the ith sample, $\{X_{ij}, j = 1, \cdots, n_i\}$, and $q_i = \ln(s_i^2)$. Shoemaker (2003, p. 107) suggests

$$\chi^2 = \sum_{i=1}^{k} \frac{(q_i - \overline{q})^2}{\frac{1}{n_i - 1}\left(\frac{\widehat{\mu}_4}{\widehat{\sigma}^4} - \frac{n_i - 3}{n_i}\right)}, \tag{1.16}$$

where $\widehat{\mu}_4 = \sum_i \sum_j (X_{ij} - \overline{X}_i)^4/N$ and $\widehat{\sigma}^2 = \sum_i (n_i - 1)s_i^2/N$. The denominator of (1.16) is an estimate of $\text{var}(q_i)$ designed for small samples. Shoemaker (2003) suggested using the harmonic mean of n_i's, $n_h = \frac{k}{\frac{1}{n_1}+\frac{1}{n_2}+\cdots+\frac{1}{n_k}}$ instead of individual n_i in the formula (1.16). The critical values of χ^2 are obtained from a chi-squared distribution with $k-1$ degrees of freedom.

Procedure Based on Gini's Mean Difference

Miller (1968) proposed a general method of constructing spread variables to be used in the ANOVA F statistic based on jackknife pseudo-values. Here we use the Miller idea with a well-known scale estimator, Gini's mean difference, which is highly efficient across a range of distributions (e.g., see Johnson and Kotz, 1970, p. 67). For two independent observations from the ith group, Gini's mean difference scale parameter is $\theta_i = \text{E}|X_{i1} - X_{i2}|$. It is unbiasedly estimated by the U-statistic

$$\widehat{\theta}_i = \frac{1}{\binom{n_i}{2}} \sum_{j<k} |X_{ij} - X_{ik}|$$

or by the simpler to compute L-statistic version

$$\widehat{\theta}_i = \frac{1}{\binom{n_i}{2}} \sum_{j=1}^{n_i} (2j - n_i - 1) X_{i(j)},$$

where $X_{i(1)} \leq X_{i(2)} \leq \ldots \leq X_{i(n_i)}$ are the ordered values of the ith sample.

To implement the Miller method, define the pseudo-values for the ith sample as

$$u_{ij} = n_i\widehat{\theta}_i - (n_i - 1)\widehat{\theta}_{i,-j},$$

where the notation $\widehat{\theta}_{i,-j}$ refers to the estimator in the ith group with X_{ij} removed from the sample. The Gini test is then based on the one-way ANOVA F statistic of the new variables,

$$F_{\text{GINI}} = \frac{\sum_i n_i(\bar{u}_{i\cdot} - \bar{u}_{\cdot\cdot})^2/(k-1)}{\sum_i \sum_j (u_{ij} - \bar{u}_{i\cdot})^2 / \sum_i (n_i - 1)} \tag{1.17}$$

where $\bar{u}_i = \sum u_{ij}/n_i$ and $\bar{u}_{\cdot\cdot} = \sum\sum u_{ij}/N$. The critical values of F_{GINI} are obtained from the F-distribution with $k-1$ and $N-k$ degrees of freedom.

Bootstrap Version of the LevMed Test

Boos and Brownie (1989) introduced the bootstrap technique for tests of scale equality. Basically, the idea is to resample from the pooled set of residuals obtained by subtracting group trimmed means within each sample. In the simulations of the next subsection, we include a bootstrap version of LevMed. That is, the p-values for the LevMed F statistic are obtained from a bootstrap resample procedure rather than from the F distribution. Details may be found in Boos and Brownie (1989) or Lim and Loh (1996).

1.3.2 Null Simulations

Comparison with Other Tests for Variances

We compare the Type I error robustness of the modified LevMed (MLM) test and the MLM-BA test to the other four tests: the LevMed test (LM), Shoemaker's (SH) test, the F test on pseudo-values based on Gini's mean difference (Gini), and the bootstrap version of LevMed (BLM). The three distributions are: (1) Normal (0,1), (2) Uniform (0,1), and (3) Extreme Value (0,1). The group size configurations are: (i) k=2, (ii) k=4, and (iii) k=8. In the balanced design, the sample size configurations are: (I) $n = 3$, (II) $n = 4$, (III) $n = 5$, (IV) $n = 6$, (V) $n = 7$, (VI) $n = 8$, (VII) $n = 9$, (VIII) $n = 10$, (IX) $n = 20$, and (X) $n = 100$. We use S=1,000 Monte Carlo replications. For BLM, the bootstrap replication size is $B = 499$. The nominal Type I error rate is 0.05.

We summarize the results in Tables 1.3-1.5 for the three distributions, respectively. Note that we only report two decimal places in the tables because the standard error of the entries is approximately $\sqrt{(.95)(.05)/1000} = .007$.

Tables 1.3-1.5 show that all six tests except Shoemaker's test and LevMed for $n = 4$ are conservative. Shoemaker's test is liberal for many cases. At the bottom of these three tables we have calculated a column mean and the mean absolute deviation from .05, MAD $= (1/30)\sum |\text{entry} - .05|$. The column mean shows that the new modified LevMed (MLM) is less conservative on average compared to the LevMed test, but it is not as good as the bootstrapped version. The procedure based on Gini's

Table 1.3: Estimated Levels for the Normal Distribution, $\alpha = 0.05$

k	n	LevMed	MLM	MLM-BA	BLM	SH	Gini
2	3	0.00	0.05	0.05	0.01	0.07	0.00
2	4	0.07	0.04	0.04	0.03	0.05	0.03
2	5	0.01	0.04	0.05	0.05	0.07	0.03
2	6	0.05	0.04	0.04	0.04	0.05	0.03
2	7	0.02	0.05	0.05	0.05	0.05	0.05
2	8	0.04	0.04	0.04	0.05	0.06	0.04
2	9	0.03	0.04	0.05	0.05	0.06	0.04
2	10	0.03	0.04	0.04	0.05	0.06	0.04
2	20	0.04	0.04	0.04	0.04	0.05	0.04
2	100	0.06	0.06	0.06	0.06	0.06	0.06
4	3	0.00	0.04	0.04	0.03	0.10	0.00
4	4	0.07	0.03	0.04	0.05	0.07	0.02
4	5	0.00	0.04	0.04	0.04	0.05	0.01
4	6	0.03	0.02	0.03	0.04	0.05	0.02
4	7	0.01	0.03	0.04	0.05	0.05	0.02
4	8	0.03	0.03	0.04	0.04	0.04	0.03
4	9	0.01	0.05	0.05	0.05	0.04	0.03
4	10	0.04	0.05	0.06	0.07	0.06	0.05
4	20	0.03	0.04	0.05	0.05	0.07	0.04
4	100	0.04	0.04	0.04	0.05	0.04	0.04
8	3	0.00	0.03	0.03	0.03	0.12	0.00
8	4	0.06	0.02	0.02	0.05	0.08	0.00
8	5	0.00	0.04	0.04	0.04	0.07	0.02
8	6	0.03	0.03	0.03	0.05	0.07	0.01
8	7	0.01	0.04	0.04	0.05	0.06	0.03
8	8	0.03	0.04	0.04	0.06	0.06	0.03
8	9	0.01	0.03	0.03	0.04	0.04	0.02
8	10	0.04	0.05	0.05	0.06	0.05	0.03
8	20	0.03	0.04	0.04	0.05	0.06	0.03
8	100	0.04	0.04	0.04	0.05	0.06	0.04
Column Mean		0.029	0.039	0.040	0.046	0.061	0.028
MAD from .05		0.025	0.012	0.011	0.007	0.013	0.023

Note: Entries based on 1,000 replications. Standard error of main entries $\leq (.88 * .12/1000)^{1/2} = .01$. "MAD from .05" is the mean absolute deviation from .05. Standard error of summary statistics in the last two rows range from .001 to .004.

Table 1.4: Estimated Levels for the Uniform distribution, $\alpha = 0.05$

k	n	LevMed	MLM	MLM-BA	BLM	SH	Gini
2	3	0.00	0.03	0.03	0.02	0.06	0.00
2	4	0.09	0.05	0.06	0.05	0.05	0.04
2	5	0.01	0.04	0.04	0.04	0.04	0.03
2	6	0.04	0.04	0.03	0.04	0.04	0.03
2	7	0.01	0.04	0.04	0.05	0.04	0.03
2	8	0.04	0.04	0.04	0.05	0.04	0.04
2	9	0.02	0.03	0.03	0.04	0.04	0.04
2	10	0.04	0.04	0.04	0.05	0.04	0.04
2	20	0.04	0.05	0.05	0.06	0.06	0.06
2	100	0.04	0.04	0.04	0.05	0.05	0.05
4	3	0.00	0.02	0.02	0.01	0.10	0.00
4	4	0.09	0.02	0.02	0.06	0.06	0.01
4	5	0.00	0.01	0.01	0.02	0.05	0.01
4	6	0.03	0.02	0.02	0.03	0.05	0.02
4	7	0.00	0.03	0.03	0.04	0.05	0.03
4	8	0.03	0.03	0.03	0.04	0.04	0.03
4	9	0.01	0.03	0.03	0.04	0.04	0.03
4	10	0.02	0.02	0.02	0.04	0.04	0.02
4	20	0.03	0.03	0.03	0.04	0.04	0.03
4	100	0.05	0.06	0.06	0.07	0.06	0.05
8	3	0.00	0.01	0.01	0.01	0.14	0.00
8	4	0.07	0.01	0.02	0.06	0.10	0.01
8	5	0.00	0.01	0.01	0.01	0.08	0.01
8	6	0.02	0.01	0.01	0.03	0.06	0.01
8	7	0.00	0.02	0.02	0.03	0.06	0.02
8	8	0.02	0.01	0.01	0.04	0.04	0.01
8	9	0.00	0.01	0.01	0.02	0.04	0.01
8	10	0.01	0.02	0.02	0.04	0.04	0.02
8	20	0.02	0.03	0.03	0.04	0.06	0.03
8	100	0.03	0.04	0.04	0.05	0.04	0.04
Column Mean		0.025	0.028	0.028	0.039	0.055	0.025
MAD from .05		0.031	0.023	0.023	0.014	0.014	0.026

Note: Entries based on 1,000 replications. Standard error of main entries $\leq (.86 * .14/1000)^{1/2} = .01$. "MAD from .05" is the mean absolute deviation from .05. Standard error of summary statistics in the last two rows range from .001 to .004.

Table 1.5: Estimated Levels for the Extreme Value distribution, $\alpha = 0.05$

k	n	LevMed	MLM	MLM-BA	BLM	SH	Gini
2	3	0.00	0.05	0.05	0.03	0.10	0.00
2	4	0.08	0.03	0.05	0.03	0.08	0.03
2	5	0.01	0.04	0.05	0.04	0.07	0.03
2	6	0.04	0.03	0.05	0.05	0.07	0.04
2	7	0.03	0.06	0.06	0.07	0.08	0.06
2	8	0.04	0.05	0.06	0.05	0.07	0.05
2	9	0.03	0.06	0.07	0.06	0.08	0.05
2	10	0.03	0.03	0.04	0.04	0.06	0.04
2	20	0.04	0.05	0.06	0.05	0.08	0.05
2	100	0.04	0.04	0.04	0.04	0.05	0.04
4	3	0.00	0.03	0.05	0.03	0.11	0.00
4	4	0.07	0.03	0.04	0.05	0.06	0.02
4	5	0.01	0.02	0.03	0.03	0.07	0.02
4	6	0.03	0.03	0.04	0.05	0.08	0.04
4	7	0.01	0.03	0.04	0.04	0.06	0.03
4	8	0.04	0.05	0.06	0.06	0.09	0.05
4	9	0.03	0.05	0.06	0.06	0.06	0.05
4	10	0.04	0.05	0.06	0.06	0.08	0.05
4	20	0.04	0.04	0.04	0.05	0.06	0.05
4	100	0.05	0.05	0.05	0.05	0.05	0.05
8	3	0.00	0.04	0.05	0.04	0.13	0.00
8	4	0.09	0.04	0.04	0.08	0.11	0.02
8	5	0.00	0.03	0.04	0.04	0.08	0.02
8	6	0.04	0.04	0.04	0.06	0.08	0.02
8	7	0.01	0.04	0.05	0.05	0.07	0.04
8	8	0.03	0.03	0.04	0.05	0.06	0.03
8	9	0.01	0.03	0.04	0.05	0.05	0.03
8	10	0.04	0.05	0.05	0.06	0.06	0.04
8	20	0.03	0.05	0.05	0.05	0.06	0.04
8	100	0.04	0.05	0.05	0.05	0.05	0.05
Column Mean		0.032	0.041	0.047	0.049	0.074	0.035
MAD from .05		0.024	0.011	0.007	0.008	0.024	0.016

Note: Entries based on 1,000 replications. Standard error of main entries $\leq (.87 * .13/1000)^{1/2} = .01$. "MAD from .05" is the mean absolute deviation from .05. Standard error of summary statistics in the last two rows range from .001 to .004.

mean difference is the most conservative overall but appears to be slightly better than the LevMed test for the extreme value distribution. An important conclusion from these tables is that the new modified LevMed (MLM) is a definite improvement over the established LevMed procedure for sample sizes $n = 3$ to $n = 5$. That was our original goal for creating the modified procedure. Results from the MLM-BA test are similar to the results from the MLM test for the normal distribution and the uniform distribution. Under the extreme value distribution, the MLM-BA test improves the performance of the MLM test, which shows that the Box-Andersen correction plays an important role for skewed distributions. The good point is that the MLM-BA test performs as well as the bootstrapped version, BLM.

Table 1.6 summarizes some results for unbalanced designs. The results for the unbalanced designs are consistent with those for the balanced design. The modified LevMed test also improves the null performance under the unbalanced design, especially for the normal and extreme value distributions. At the bottom of the table we have also calculated a column mean and the MAD from .05. The column mean shows that the modified LevMed test is less conservative on average compared to the LevMed test, especially for the normal and extreme value distributions. The MAD from .05 shows that the modified LevMed test can lead to estimates of significance levels closer to the nominal rate than the LevMed test. Table 1.6 does not include results for the MLM-BA test because it leads to results similar to the MLM test.

Table 1.6: Estimated Levels of Nominal 0.05 tests for Unbalanced Designs

Sample Size	Distribution					
	Normal		Uniform		Extreme Value	
	LevMed	MLM	LevMed	MLM	LevMed	MLM
(5,15)	0.02	0.05	0.02	0.04	0.02	0.04
(3,4)	0.04	0.05	0.04	0.04	0.04	0.04
(3,8)	0.01	0.04	0.04	0.02	0.02	0.05
(4,7)	0.03	0.03	0.03	0.03	0.04	0.04
(5,20)	0.03	0.04	0.07	0.05	0.02	0.06
(6,19)	0.04	0.04	0.04	0.03	0.03	0.04
(5,5,15,15)	0.01	0.04	0.02	0.02	0.03	0.06
(5,5,5,15)	0.00	0.04	0.01	0.03	0.01	0.05
(5,15,15,15)	0.02	0.05	0.02	0.03	0.03	0.05
(3,3,4,4)	0.03	0.02	0.03	0.03	0.03	0.05
(3,3,3,4)	0.01	0.04	0.01	0.01	0.01	0.04
(3,4,4,4)	0.05	0.03	0.05	0.02	0.05	0.04
(4,4,7,7)	0.03	0.04	0.02	0.02	0.04	0.04
(4,4,4,7)	0.04	0.03	0.04	0.03	0.04	0.03
(4,7,7,7)	0.01	0.03	0.01	0.03	0.02	0.03
(3,3,8,8)	0.02	0.04	0.02	0.01	0.01	0.05
(3,3,3,8)	0.01	0.05	0.01	0.02	0.01	0.05
(3,8,8,8)	0.03	0.04	0.03	0.02	0.03	0.06
(5,5,5,5,15,15,15,15)	0.01	0.04	0.01	0.01	0.02	0.06
(5,5,5,5,5,5,5,15)	0.00	0.03	0.00	0.01	0.01	0.06
(5,15,15,15,15,15,15,15)	0.02	0.04	0.00	0.02	0.02	0.04
Column Mean	0.021	0.037	0.024	0.025	0.025	0.047
MAD from 0.05	0.029	0.013	0.027	0.025	0.025	0.009

Entries are based on 1,000 replications. Standard Error of individual entries $\leq (.93 * .07/1000)^{1/2} = 0.01$

Relationship between the Estimated Type I Error Rates and Sample Sizes and Number of Groups

To further illustrate the comparison between the LevMed test and the new modification, we have increased the Monte Carlo sample size to S=10,000 replications and made a series of plots of the Estimated Type I error versus sample sizes per group

with the LevMed on the left and the new modification on the right. Figures 1.2-1.4 are for the normal distribution, the uniform distribution, and the extreme value distribution, respectively. Figures 1.2-1.4 are at the end of the chapter due to space considerations.

The plots for the modified LevMed test are much more stable than those for the LevMed test. The ranges of the estimated levels (from 0.02 to 0.05) for the modified LevMed test are much smaller than the ranges of the LevMed test (from 0 to 0.07) for small sample sizes. As k increases, both the LevMed test and the modified LevMed test become more conservative for small sample sizes.

Finally, we give two more pages of plots at the end of the chapter to illustrate better how the Type I error rates of the modified LevMed procedure change as k increases. In Figure 1.5 we have taken the data from Figures 1.2-1.4 and put the graphs for a given distribution on the same row with k increasing as we move from left to right. In addition we have used a local smoother to track the trend as sample size increases. In Figure 1.6 we have put the local smoothers for all three group numbers, $k = 2$, $k = 4$, and $k = 8$, on the same plot but without the individual points. In both sets of plots we can see that an increase in k causes the procedure to be more conservative.

1.3.3 Power

We now compare the power of the modified LevMed test (MLM) and the MLM-BA test to power of the other 4 tests. The sample size configurations are n=3, 4, 5, 6, 7, 8, 9, 10, and 20. The variance configurations are (1:4), (1:8), (1:6:11:16), (1:1:1:16), (1:1:1:8:8:8:16:16) and (1:1:1:1:1:1:16:16). We use S=1,000 Monte Carlo replications with the nominal rate equal to 0.05. The results are summarized in Tables 1.11-1.13 at the end of the chapter.

In these tables we see important power gains at $n = 3, 5, 7$, and 9 for the modified LevMed procedure compared to LevMed. When sample sizes reach $n = 20$ all the procedures are similar in power. A very crude summary is to take the mean of the columns for each table. The results are summarized in Table 1.7. The underline emphasizes the most powerful test for every distribution. Thus, the modified procedure has an overall gain in power of about .05 when compared to LevMed, and the bootstrapped LevMed has an average .02 gain in power compared to the modified procedure. Amazingly, the MLM-BA test generally has higher power than the bootstrapped LevMed test and the MLM-BA test is much simpler to perform, which indicates that the Box-Andersen correction can improve the power of the modified LevMed test. The Gini procedure has power in between the LevMed and the modified LevMed procedure. The Shoemaker procedure has good power, but the comparison is unfair due to its inflated Type I error.

Table 1.7: Comparison of Average Power Among the Six Tests

	Distribution		
Test	Normal	Uniform	Extreme Value
LevMed	0.48	0.53	0.43
MLM	0.53	0.57	0.48
MLM-BA	<u>0.56</u>	0.60	<u>0.52</u>
BLM	0.55	0.59	0.50
SH	0.54	<u>0.61</u>	0.51
Gini	0.51	0.59	0.45

Note: Entries based on the results of Table 1.11-1.13.
Standard error of entries ≤ 0.002.

1.4 Example

Phadke et. al (1983) reported on an off-line quality control experiment in the fabrication of integrated circuit chips. We will use part of the data to illustrate the 6 tests used in the simulations.

In order to choose process conditions to minimize variance in contact window sizes of integrated circuit chips, Phadke et. al (1983) conducted an experiment with 18 combinations of levels of factors. For every experimental unit (of 18 experimental units), there are 5 specific measured locations such as "Top," "Bottom," "Left," "Right," and "Center" locations. Most of the experimental units have 10 observations (2 measurements for every location) except units 5, 15 and 18 which have 5 observations (only 1 measurement for every location). We use part of the "pre-etch" window size data to compare the six tests. In our example, if we use the experimental units 1-4, then $k = 4$ and if we use experiments 1-8, then $k = 8$. If we use the first 3 observations for every experimental unit, then $n = 3$, and the experimental

unit is measured at the "Top," "Center," "Bottom," locations. If we use the first 4 observations for every experimental unit, then $n = 4$, and the experimental unit is measured at the "Top," "Center," "Bottom," and "Left" locations. When $n = 5$, the experimental unit is measured at the "Top," "Center," "Bottom," "Left," and "Right" locations.

Table 1.8 summarizes the data including the mean and the sample standard deviation for every case. The standard deviation is a measure of variation within a chip and not between chips produced under the same factor settings.

Table 1.8: Summary Statistics of Data from Off-Line Quality Control Study

Statistic	Mean			SD		
Size	n=3	n=4	n=5	n=3	n=4	n=5
h=1	2.53	2.53	2.52	0.10	0.08	0.07
h=2	2.72	2.69	2.66	0.05	0.07	0.10
h=3	2.77	2.72	2.64	0.06	0.12	0.19
h=4	2.10	2.07	2.08	0.10	0.10	0.09
h=5	1.91	1.88	1.87	0.15	0.13	0.12
h=6	2.54	2.52	2.52	0.03	0.05	0.04
h=7	2.03	2.02	2.02	0.07	0.06	0.05
h=8	3.42	3.34	3.28	0.26	0.27	0.26
h=9	2.99	2.91	2.88	0.08	0.18	0.17
h=10	2.57	2.54	2.51	0.11	0.10	0.11
h=11	3.24	3.23	3.21	0.07	0.06	0.06
h=12	3.29	3.28	3.24	0.07	0.06	0.09
h=13	2.59	2.58	2.58	0.03	0.03	0.03
h=14	2.28	2.26	2.27	0.15	0.13	0.11
h=15	2.49	2.47	2.46	0.03	0.04	0.04
h=16	2.64	2.66	2.64	0.10	0.09	0.08

From Table 1.9, we can see that the MLM test, the MLM-BA test and the BLM test produce similar results. The SH test and the Gini test lack power for small

Table 1.9: P-Values for tests of equality of within chip variance using subsets of data from an Off-Line Quality Control Study

	k=4			k=8			k=16		
Test	n=3	n=4	n=5	n=3	n=4	n=5	n=3	n=4	n=5
LevMed	0.83	0.87	0.32	0.37	0.03	0.18	0.73	0.07	0.13
MLM	0.63	0.90	0.18	0.04	0.12	0.09	0.29	0.24	0.03
MLM-BA	0.63	0.90	0.13	0.04	0.10	0.07	0.29	0.23	0.03
BLM	0.59	0.86	0.12	0.04	0.05	0.06	0.11	0.09	0.01
SH	0.67	0.84	0.41	0.51	0.54	0.17	0.80	0.61	0.18
Gini	0.83	0.92	0.28	0.37	0.19	0.10	0.73	0.39	0.06

sample sizes because they can't detect difference in variances for the case with $k = 8$ and $n = 3$ or the case with $k = 8$ and $n = 4$. The LevMed test does not detect heterogeneity of variances for the case with $k = 8$ and $n = 3$, while the MLM test and the BLM test provide evidence of variance heterogeneity. On the other hand, the MLM test misses the heterogeneity of variances for the case with $k = 8$ and $n = 4$, while the LevMed test and the BLM test can detect it. This is in part because the LevMed test tends to be liberal and consequently has more power than the MLM test for small and even sample sizes. In contrast, the MLM test has more power than the LevMed test for the small and odd sample sizes because it avoids the conservative behavior shown by the LevMed test for $n = 3$ and $n = 5$.

1.5 Comparison with the Hines LevMed Test (Hines and Hines, 2000)

Hines and Hines (2000) proposed a modification of the LevMed test (the Hines LevMed test) by removing linear dependencies (structural zeros) among the LevMed variables, which is similar to our Modified LevMed test. Suppose that $(Z_{i(1)}, \cdots, Z_{i(n_i)})$ are the ordered LevMed variables for the ith group. If the sample size n_i is odd, the Hines LevMed test deletes the smallest value, $Z_{i(1)} = 0$. If the sample size is even, then $Z_{i(1)} = Z_{i(2)}$ and the pair of values $Z_{i(1)}$ and $Z_{i(2)}$ is replaced by the pair $(Z_{i(2)} - Z_{i(1)})/\sqrt{2}$ ($= 0$) and $(Z_{i(2)} + Z_{i(1)})/\sqrt{2}$. The resulting structural zero is then deleted. ANOVA is then applied to the remaining Z values. When the sample size is odd for all k groups, the Hines test is the same as the MLM test. When any of the sample sizes is even, the smallest Z value for that group is $(Z_{i(1)} + Z_{i(2)})/\sqrt{2}$ for Hines LevMed, compared to $(Z_{i(1)} + Z_{i(2)})/2$ for the MLM procedure.

To further illustrate the comparison between the modified LevMed and the Hines test, we have made a series of plots of the estimated Type I error versus sample sizes per group with the Hines test on the left and the MLM on the right. Figures 1.7-1.9 are for the normal distribution, the uniform distribution, and the extreme value distribution, respectively. These three figures are at the end of the chapter due to space considerations. For n odd, the Hines and MLM tests are the same, but for n even and small, estimated levels are noticeably greater for the Hines test. For small

Table 1.10: Comparison of Estimates of Levels between the Modified LevMed test and Hines test for Even Sample Sizes

		Distribution					
k	n	Normal		Uniform		Extreme Value	
		Hines	MLM	Hines	MLM	Hines	MLM
2	4	0.07	0.04	0.09	0.05	0.07	0.03
2	6	0.05	0.04	0.06	0.04	0.05	0.03
2	8	0.05	0.04	0.05	0.04	0.06	0.05
2	10	0.04	0.04	0.06	0.04	0.04	0.03
2	20	0.04	0.04	0.05	0.05	0.05	0.05
2	100	0.06	0.06	0.05	0.04	0.04	0.04
4	4	0.07	0.03	0.08	0.02	0.07	0.03
4	6	0.05	0.02	0.04	0.02	0.06	0.03
4	8	0.04	0.03	0.04	0.03	0.06	0.05
4	10	0.06	0.05	0.03	0.02	0.06	0.05
4	20	0.05	0.04	0.04	0.03	0.04	0.04
4	100	0.04	0.04	0.06	0.06	0.05	0.05
8	4	0.06	0.02	0.06	0.01	0.09	0.04
8	6	0.05	0.03	0.03	0.01	0.06	0.04
8	8	0.06	0.04	0.03	0.01	0.05	0.03
8	10	0.06	0.05	0.02	0.02	0.06	0.05
8	20	0.04	0.04	0.04	0.03	0.05	0.05
8	100	0.04	0.04	0.04	0.04	0.05	0.05
Column Mean		0.052	0.038	0.048	0.031	0.056	0.041
MAD from 0.05		0.006	0.013	0.014	0.020	0.009	0.009

Entries are based on 1,000 replications. Standard Error of individual entries $\leq (.91 * .09/1000)^{1/2} = 0.01$

n, levels of MLM are consistently conservative, whereas levels of the Hines test show an oscillating pattern with peaks when n is even. Although similar to the odd/even pattern for levels of LevMed (Figures 1.2 - 1.4), the Hines modification is clearly an improvement over LevMed.

The plots for the modified LevMed test are more stable than those for the Hines test. The ranges of the estimated levels (from 0.02 to 0.05) for the modified LevMed

test are much smaller than the ranges of the Hines Levene test (from 0.02 to 0.09) for the small sample sizes.

Because the Hines test is the same with the MLM test for odd sample sizes, Table 1.10 presents levels for the MLM and Hines tests only for even sample sizes. When $n = 4$, the Hines test is liberal under all the three distributions. Under the normal distribution and the uniform distribution, as k, the number of groups, increases, the Hines test holds its levels better than MLM which becomes more conservative. Under the extreme value distribution, the Hines test tends to be liberal while the MLM test is conservative. Except at the extreme value distribution, based on the MAD summary, the Hines test appears to hold its level better than the MLM test for n small and even.

We also compare the Hines test to the other tests in terms of power by simulation. Using the same seeds, Monte Carlo samples were generated for the 30 non-null cases summarized in Table 1.7 for the six other tests. Power was obtained for each case for the Hines test and average power was calculated. The resulting average powers under the normal distribution, the uniform and the extreme value for the Hines test are 0.55, 0.59 and 0.49, respectively. Compared to the MLM test, the Hines has slightly greater power for the normal distribution and the uniform distribution due to its liberal performance for n even and small. However, it does not perform as well as the MLM-BA test in terms of power.

1.6 Conclusion

This chapter demonstrates that the modified LevMed (MLM) test can yield valid levels and good power for most configurations studied. The MLM test performs well for small and odd sample sizes, where the shortcomings of the LevMed test are most pronounced. MLM, which differs from the Hines test only if some n_i are even, performs better than the Hines test under skewed distributions, in this case. Although the modified LevMed test is inferior to the BLM test in terms of level and power, it is much simpler than the BLM test. The Box-Andersen correction can improve the power of the Modified LevMed test, especially for skewed distributions. In general we believe the modified LevMed procedure is a good test for homogeneity of scale.

1.7 Appendix

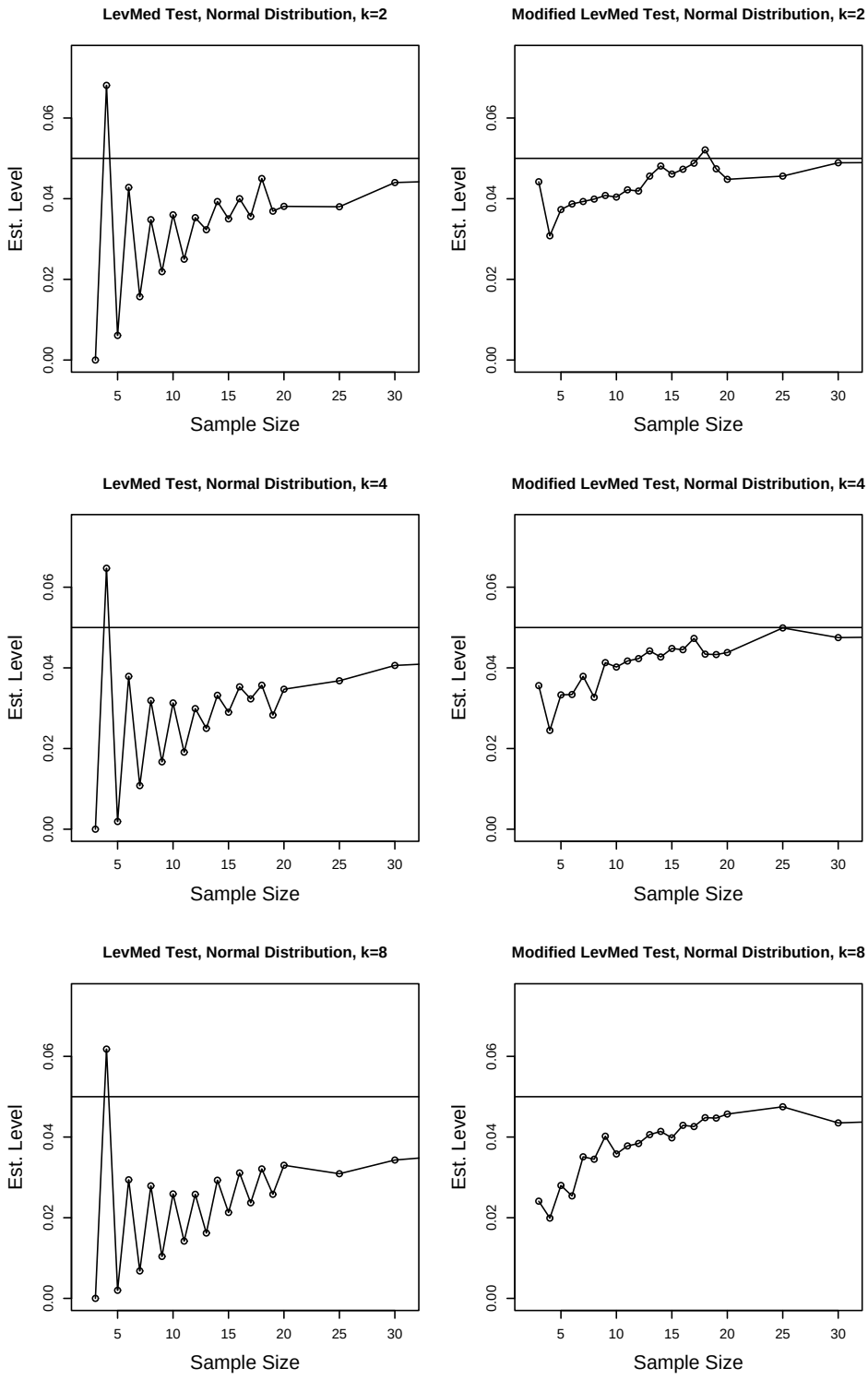

Figure 1.2: Estimated levels versus sample sizes for the normal distribution. Standard deviations of plotted values are bounded by $(40000)^{-1/2} = .005$.

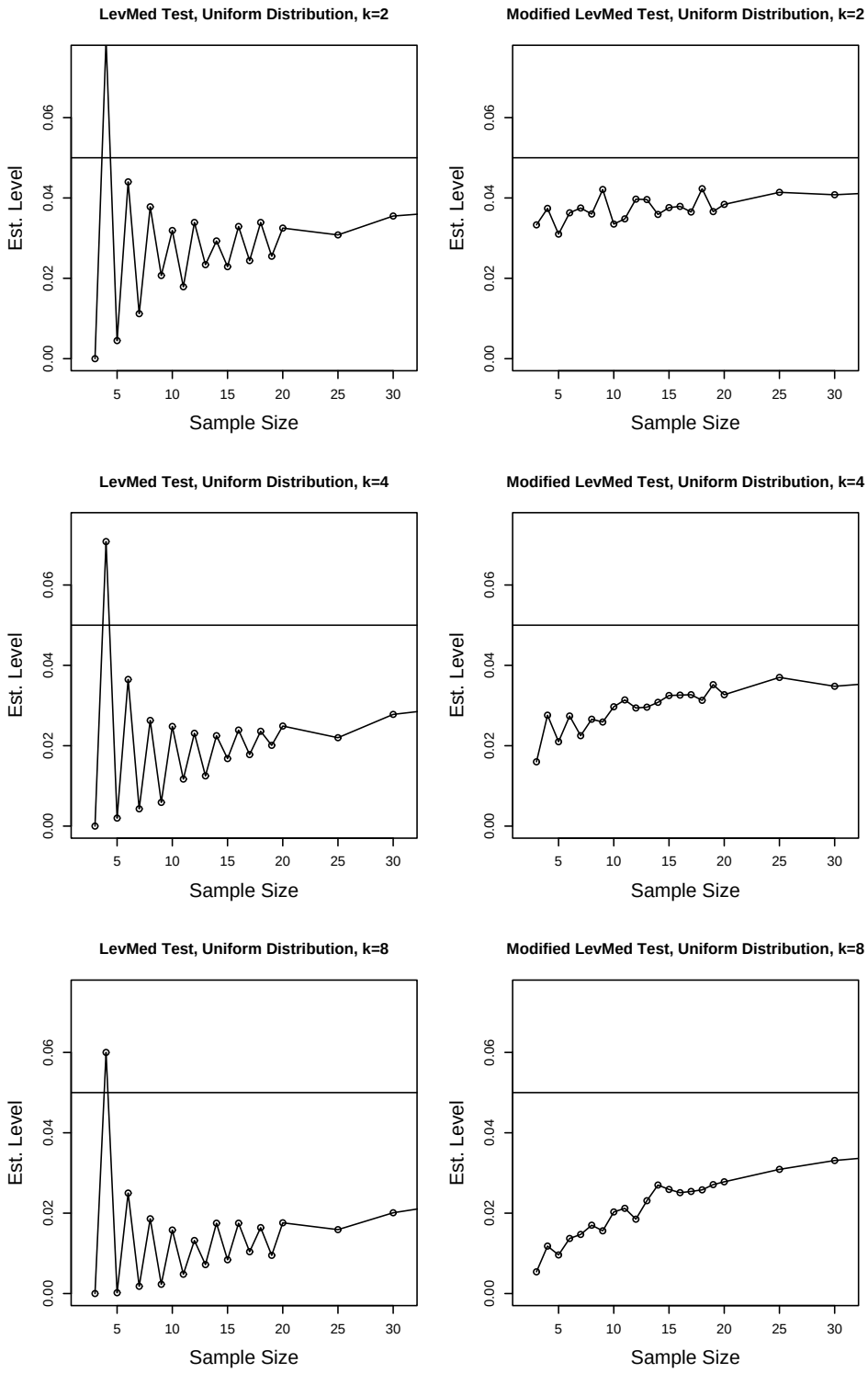

Figure 1.3: Estimated levels versus sample sizes for the uniform distribution. Standard deviations of plotted values are bounded by $(40000)^{-1/2} = .005$.

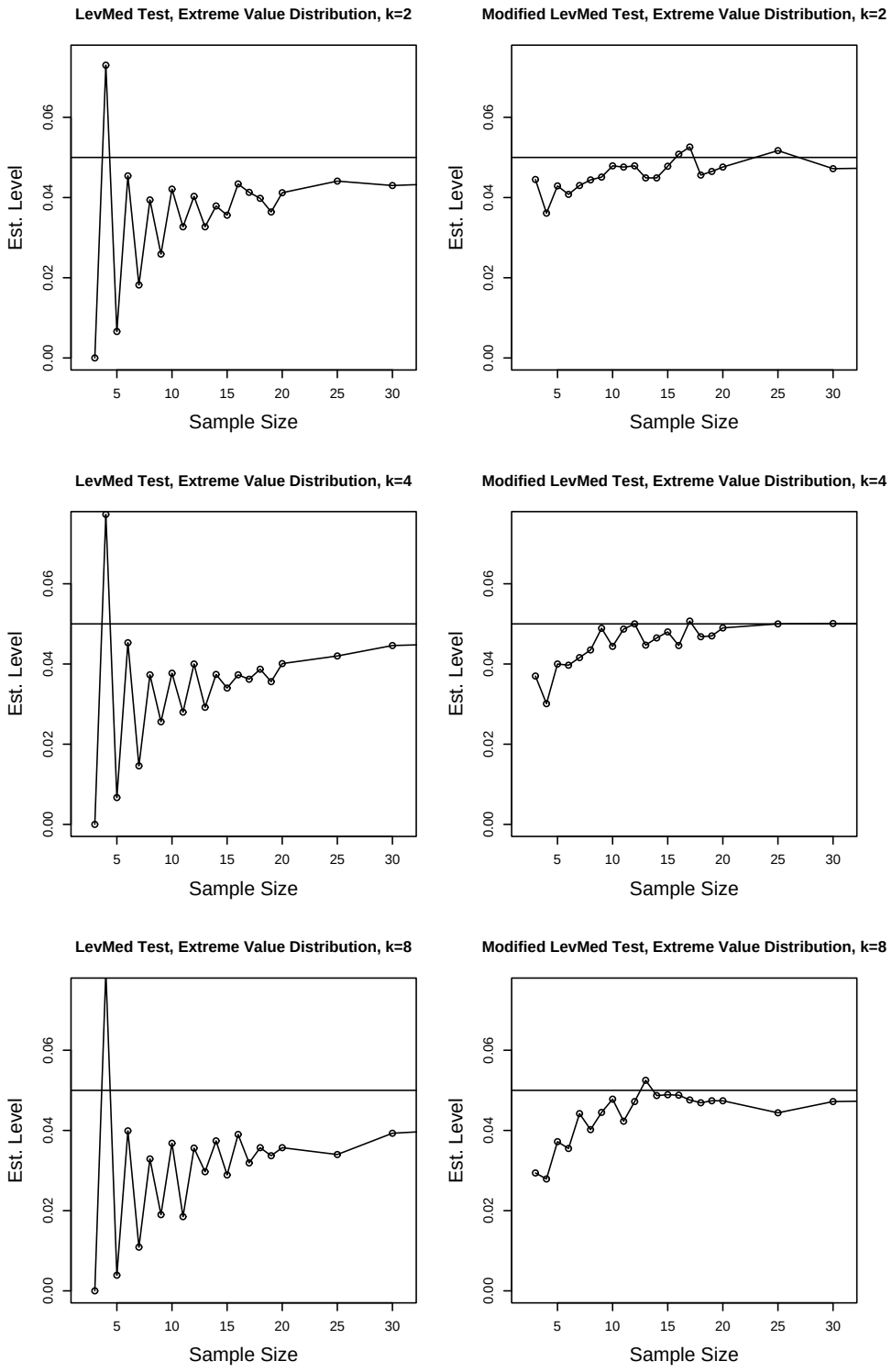

Figure 1.4: Estimated levels versus sample sizes for the extreme value distribution. Standard deviations of plotted values are bounded by $(40000)^{-1/2} = .005$.

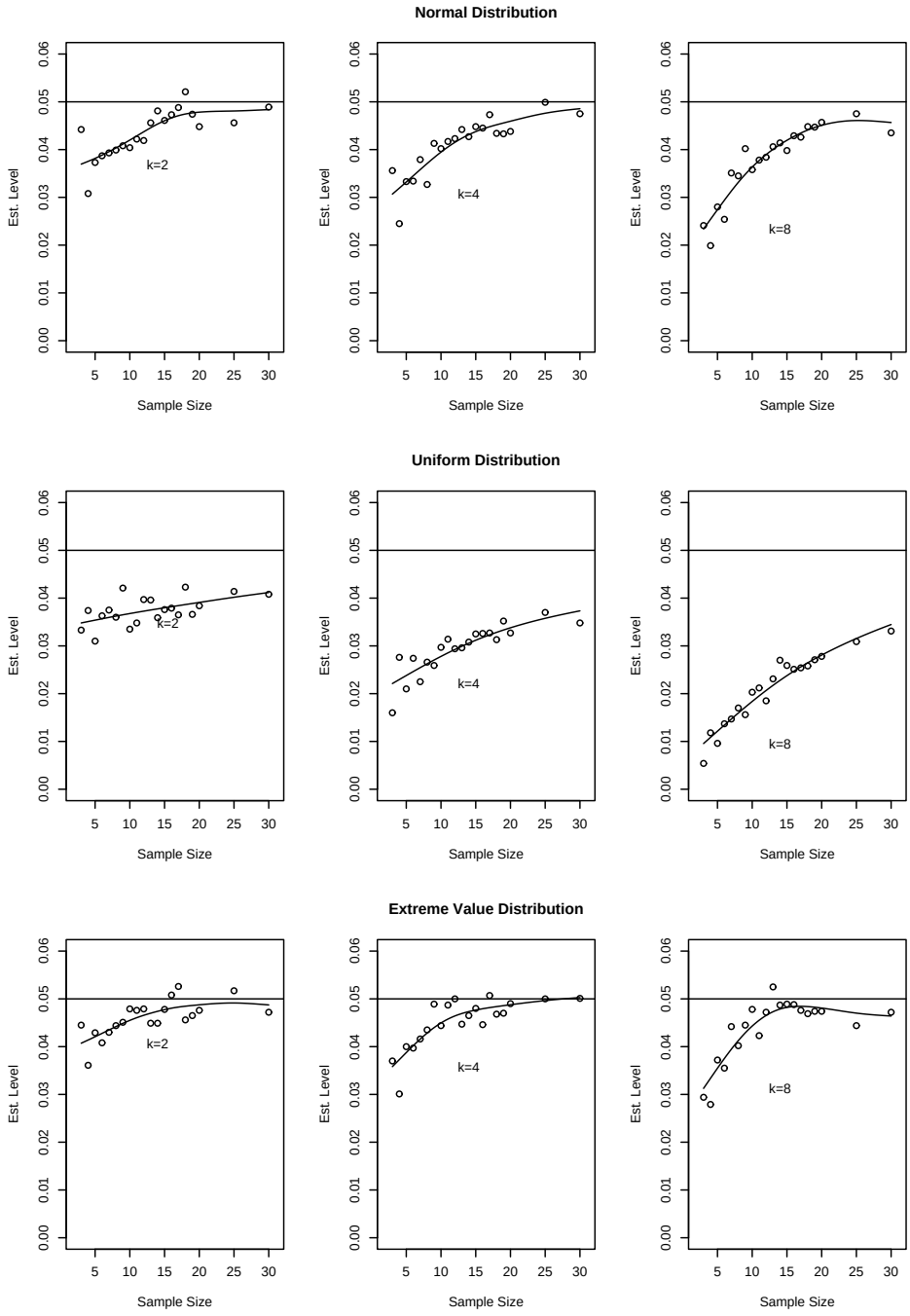

Figure 1.5: Plots of estimated Type I error rates versus sample size for the modified LevMed procedure local smoother added.

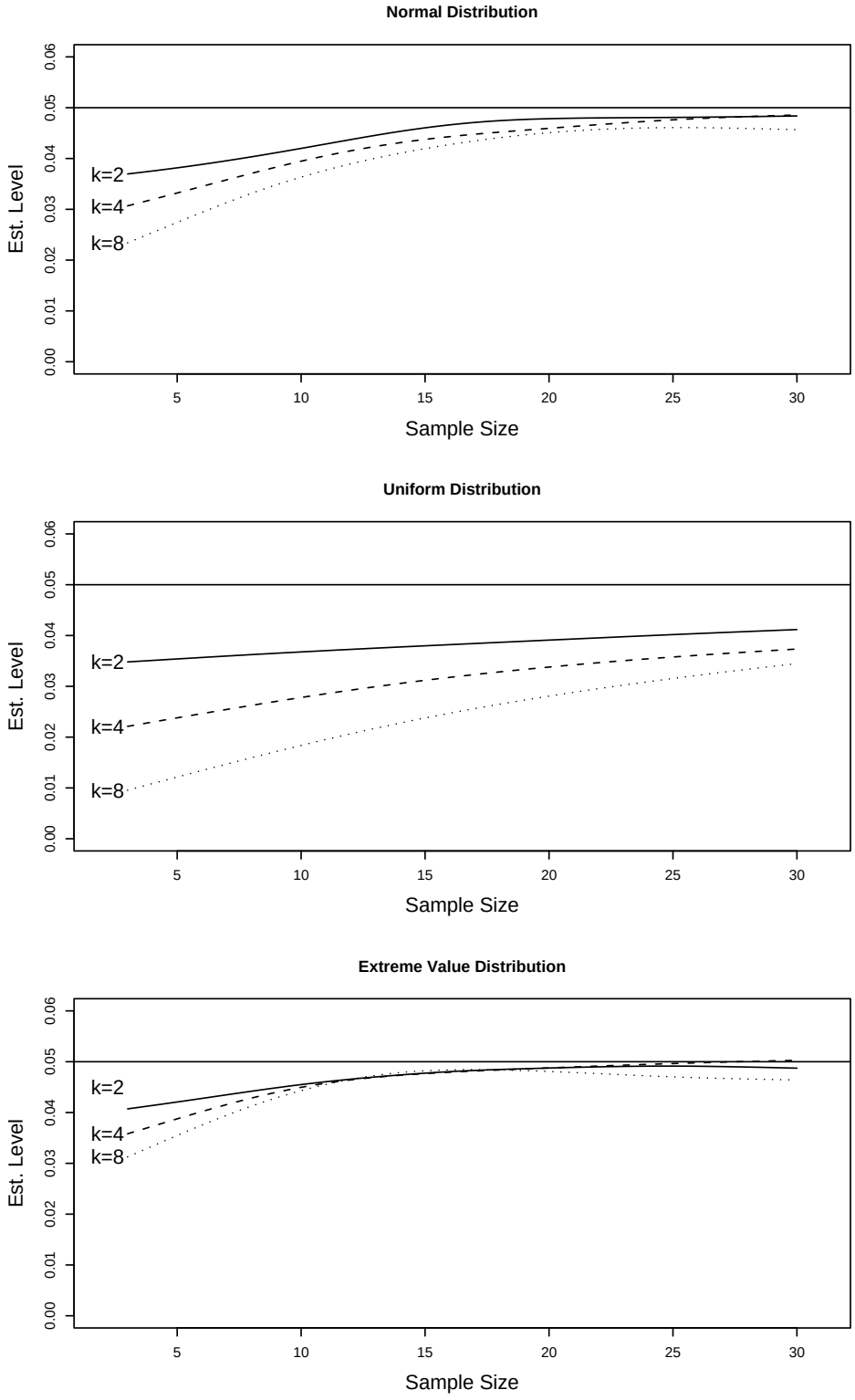

Figure 1.6: Local smoother of estimated Type I error rate versus sample size for the modified LevMed procedure.

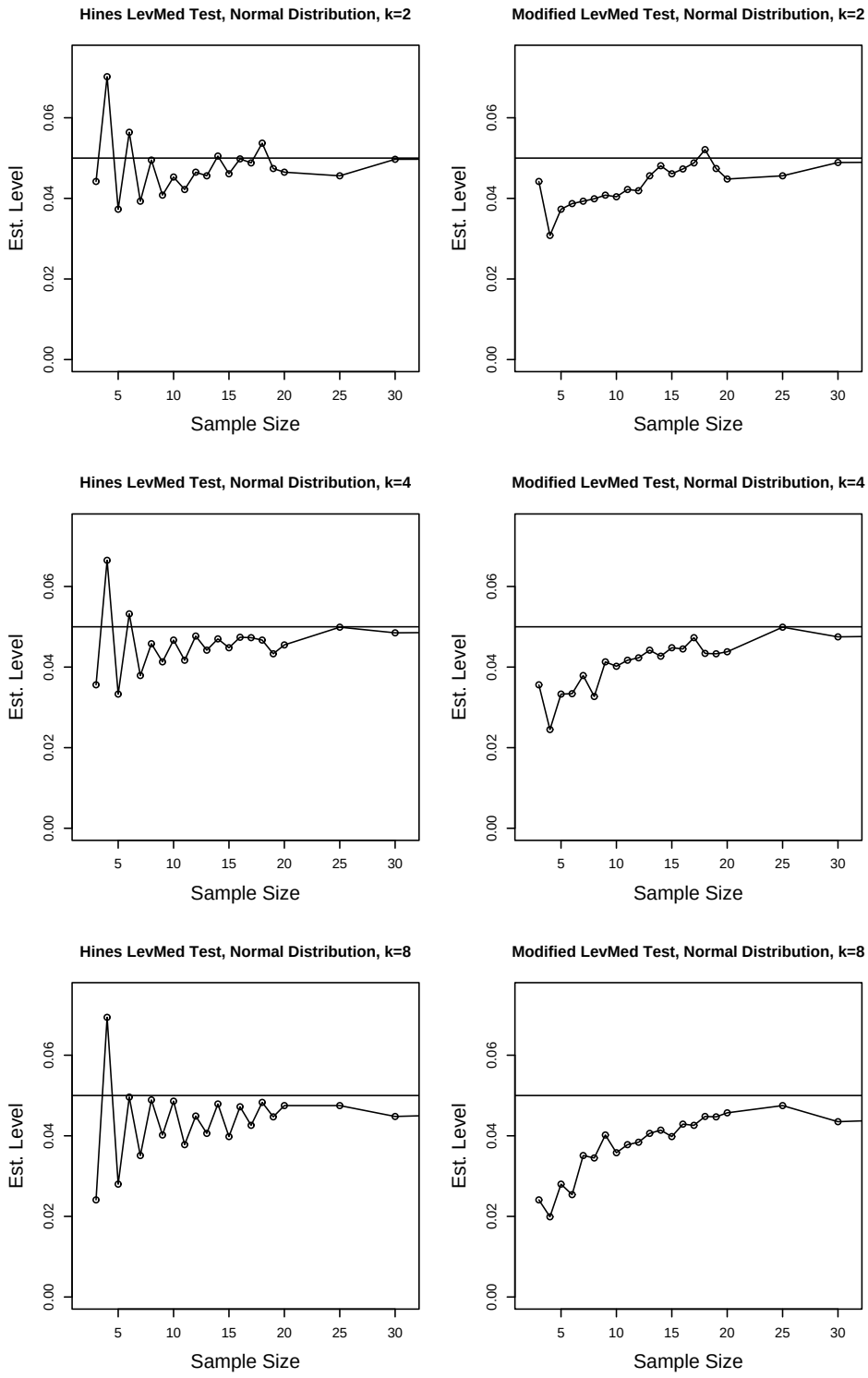

Figure 1.7: Estimated levels versus sample sizes for the normal distribution. Standard deviations of plotted values are bounded by $(40000)^{-1/2} = .005$.

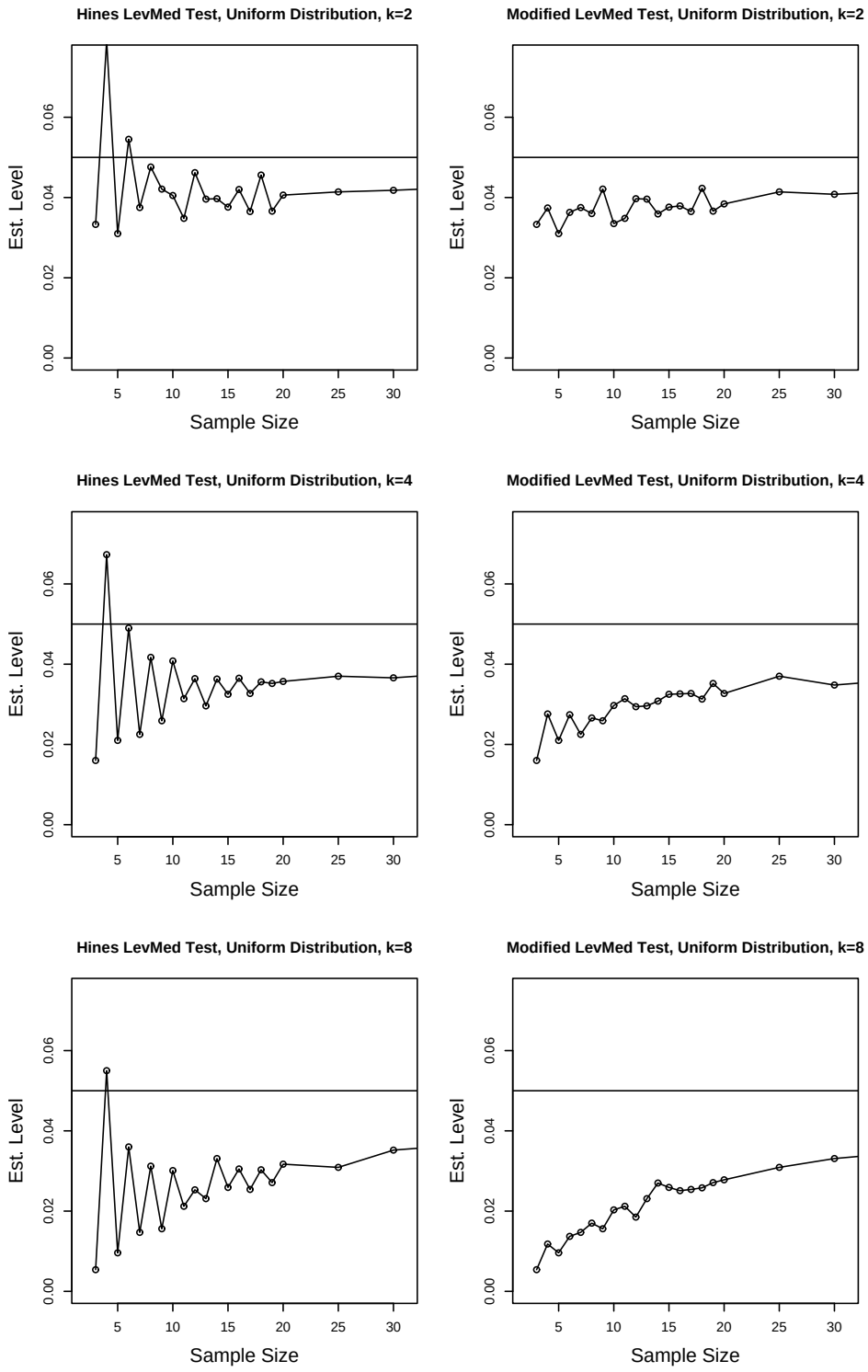

Figure 1.8: Estimated levels versus sample sizes for the uniform distribution. Standard deviations of plotted values are bounded by $(40000)^{-1/2} = .005$.

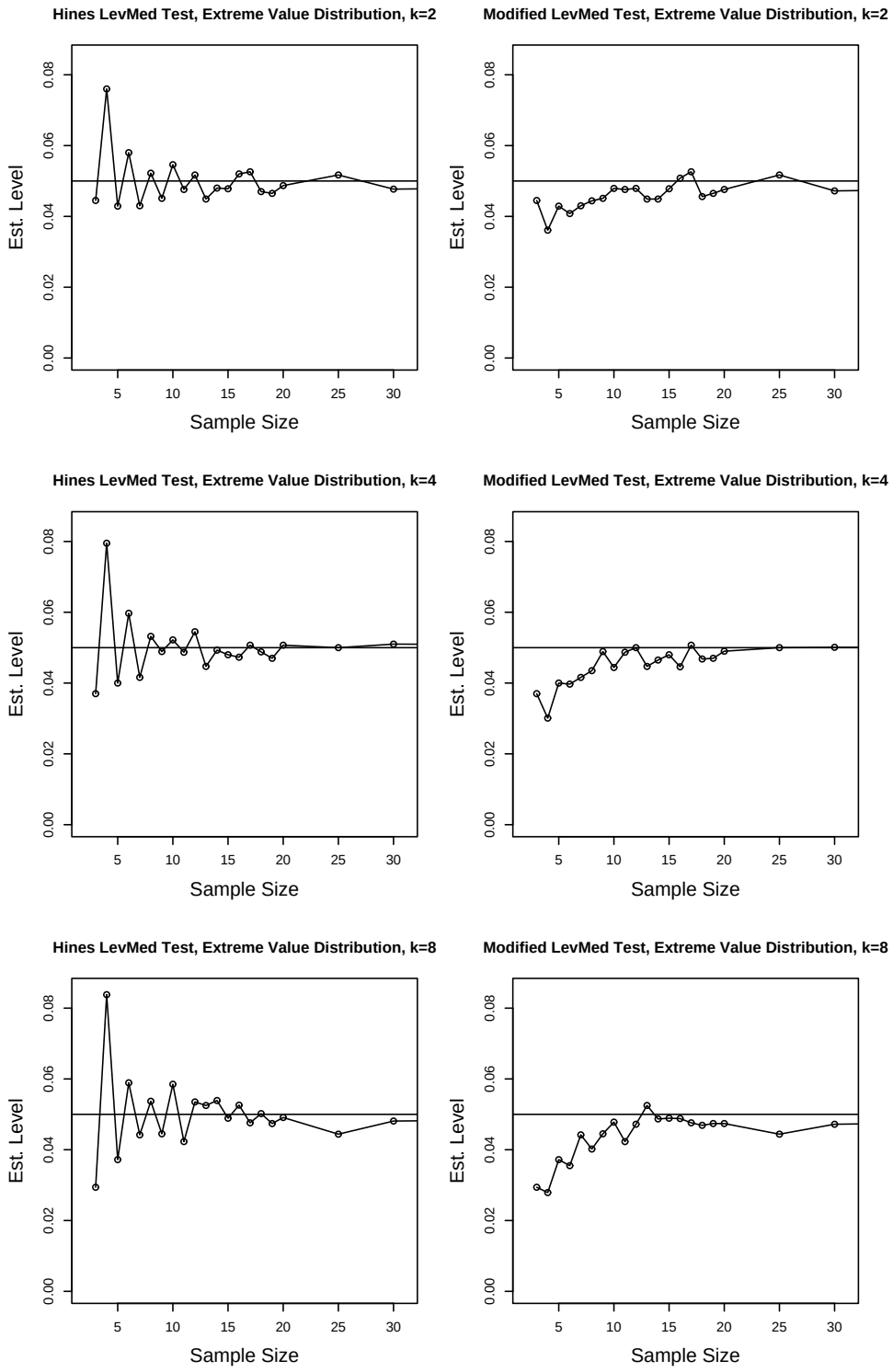

Figure 1.9: Estimated levels versus sample sizes for the extreme value distribution. Standard deviations of plotted values are bounded by $(40000)^{-1/2} = .005$.

Table 1.11: Estimated Power for the Normal Distribution, $\alpha = 0.05$.

k	n	Variance	LevMed	MLM	MLM-BA	BLM	SH	Gini
2	3	(1,4)	0.00	0.08	0.08	0.04	0.14	0.00
2	4	(1,4)	0.12	0.07	0.08	0.07	0.17	0.06
2	5	(1,4)	0.04	0.13	0.15	0.14	0.22	0.11
2	6	(1,4)	0.16	0.17	0.18	0.16	0.26	0.17
2	7	(1,4)	0.11	0.19	0.22	0.23	0.28	0.19
2	8	(1,4)	0.24	0.26	0.28	0.27	0.36	0.27
2	9	(1,4)	0.21	0.30	0.32	0.31	0.37	0.30
2	10	(1,4)	0.30	0.33	0.35	0.33	0.42	0.35
2	20	(1,4)	0.72	0.74	0.76	0.74	0.84	0.79
2	3	(1,8)	0.00	0.08	0.08	0.07	0.21	0.00
2	4	(1,8)	0.21	0.15	0.16	0.13	0.30	0.13
2	5	(1,8)	0.06	0.18	0.23	0.24	0.41	0.18
2	6	(1,8)	0.28	0.29	0.33	0.29	0.49	0.30
2	7	(1,8)	0.24	0.36	0.41	0.41	0.56	0.37
2	8	(1,8)	0.43	0.47	0.53	0.48	0.65	0.49
2	9	(1,8)	0.47	0.57	0.62	0.61	0.73	0.58
2	10	(1,8)	0.61	0.65	0.69	0.64	0.76	0.68
2	20	(1,8)	0.95	0.96	0.97	0.96	0.98	0.97
4	3	(1,6,11,16)	0.00	0.08	0.10	0.10	0.26	0.00
4	4	(1,6,11,16)	0.19	0.10	0.12	0.14	0.36	0.08
4	5	(1,6,11,16)	0.06	0.20	0.21	0.25	0.48	0.17
4	6	(1,6,11,16)	0.26	0.26	0.28	0.29	0.59	0.23
4	7	(1,6,11,16)	0.21	0.36	0.39	0.41	0.70	0.35
4	8	(1,6,11,16)	0.43	0.48	0.50	0.50	0.77	0.48
4	9	(1,6,11,16)	0.47	0.61	0.62	0.64	0.88	0.64
4	10	(1,6,11,16)	0.62	0.69	0.72	0.69	0.92	0.70
4	20	(1,6,11,16)	1.00	1.00	1.00	1.00	1.00	1.00
4	3	(1,1,1,16)	0.00	0.19	0.29	0.23	0.17	0.00
4	4	(1,1,1,16)	0.38	0.27	0.42	0.28	0.16	0.23
4	5	(1,1,1,16)	0.29	0.46	0.59	0.57	0.27	0.45
4	6	(1,1,1,16)	0.60	0.62	0.72	0.64	0.33	0.61
4	7	(1,1,1,16)	0.68	0.76	0.83	0.81	0.45	0.77
4	8	(1,1,1,16)	0.82	0.85	0.88	0.86	0.53	0.85
4	9	(1,1,1,16)	0.84	0.89	0.92	0.91	0.58	0.89
4	10	(1,1,1,16)	0.92	0.93	0.95	0.94	0.71	0.94
4	20	(1,1,1,16)	1.00	1.00	1.00	1.00	0.99	1.00

Note: Entries based on 1,000 replications. Standard deviation $\leq (4000)^{-1/2} = .016$.

Table 1.11 Continued

k	n	Variance	LevMed	MLM	MLM-BA	BLM	SH	Gini
8	3	(1,1,1,8,8,8,16,16)	0.00	0.11	0.13	0.18	0.32	0.00
8	4	(1,1,1,8,8,8,16,16)	0.33	0.19	0.22	0.25	0.40	0.11
8	5	(1,1,1,8,8,8,16,16)	0.13	0.39	0.42	0.45	0.62	0.30
8	6	(1,1,1,8,8,8,16,16)	0.57	0.62	0.65	0.62	0.75	0.56
8	7	(1,1,1,8,8,8,16,16)	0.58	0.78	0.79	0.79	0.85	0.73
8	8	(1,1,1,8,8,8,16,16)	0.83	0.86	0.87	0.86	0.92	0.85
8	9	(1,1,1,8,8,8,16,16)	0.87	0.95	0.95	0.95	0.96	0.93
8	10	(1,1,1,8,8,8,16,16)	0.95	0.97	0.97	0.97	0.98	0.97
8	20	(1,1,1,8,8,8,16,16)	1.00	1.00	1.00	1.00	1.00	1.00
8	3	(1,1,1,1,1,1,16,16)	0.01	0.22	0.29	0.30	0.12	0.01
8	4	(1,1,1,1,1,1,16,16)	0.54	0.45	0.53	0.45	0.13	0.35
8	5	(1,1,1,1,1,1,16,16)	0.52	0.72	0.78	0.76	0.21	0.66
8	6	(1,1,1,1,1,1,16,16)	0.79	0.82	0.86	0.81	0.32	0.81
8	7	(1,1,1,1,1,1,16,16)	0.88	0.93	0.94	0.94	0.45	0.93
8	8	(1,1,1,1,1,1,16,16)	0.96	0.97	0.98	0.97	0.56	0.97
8	9	(1,1,1,1,1,1,16,16)	0.99	0.99	0.99	0.99	0.67	0.99
8	10	(1,1,1,1,1,1,16,16)	0.99	0.99	0.99	0.99	0.74	0.99
8	20	(1,1,1,1,1,1,16,16)	1.00	1.00	1.00	1.00	1.00	1.00

Note: Entries based on 1,000 replications. Standard deviation $\leq (4000)^{-1/2} = .016$.

Table 1.12: Estimated Power for the Uniform Distribution, $\alpha = 0.05$.

k	n	Variance	LevMed	MLM	MLM-BA	BLM	SH	Gini
2	3	(1,4)	0.00	0.06	0.06	0.04	0.14	0.00
2	4	(1,4)	0.18	0.10	0.12	0.10	0.16	0.10
2	5	(1,4)	0.03	0.13	0.14	0.14	0.20	0.15
2	6	(1,4)	0.19	0.19	0.20	0.20	0.27	0.20
2	7	(1,4)	0.14	0.25	0.26	0.28	0.34	0.29
2	8	(1,4)	0.30	0.33	0.35	0.33	0.45	0.42
2	9	(1,4)	0.28	0.37	0.38	0.41	0.52	0.50
2	10	(1,4)	0.39	0.43	0.45	0.43	0.58	0.55
2	20	(1,4)	0.85	0.87	0.87	0.87	0.96	0.94
2	3	(1,8)	0.00	0.07	0.07	0.04	0.22	0.00
2	4	(1,8)	0.25	0.17	0.19	0.16	0.31	0.16
2	5	(1,8)	0.08	0.21	0.25	0.27	0.43	0.26
2	6	(1,8)	0.38	0.38	0.41	0.38	0.58	0.45
2	7	(1,8)	0.32	0.45	0.50	0.50	0.67	0.55
2	8	(1,8)	0.56	0.60	0.64	0.58	0.77	0.69
2	9	(1,8)	0.57	0.66	0.69	0.69	0.85	0.76
2	10	(1,8)	0.74	0.78	0.81	0.78	0.91	0.87
2	20	(1,8)	0.99	0.99	0.99	1.00	1.00	1.00
4	3	(1,6,11,16)	0.00	0.05	0.05	0.05	0.22	0.00
4	4	(1,6,11,16)	0.25	0.14	0.15	0.20	0.38	0.11
4	5	(1,6,11,16)	0.03	0.18	0.19	0.24	0.51	0.18
4	6	(1,6,11,16)	0.32	0.34	0.35	0.38	0.74	0.38
4	7	(1,6,11,16)	0.25	0.44	0.44	0.51	0.82	0.53
4	8	(1,6,11,16)	0.56	0.61	0.63	0.62	0.93	0.73
4	9	(1,6,11,16)	0.56	0.73	0.74	0.78	0.97	0.83
4	10	(1,6,11,16)	0.77	0.83	0.84	0.82	0.99	0.91
4	20	(1,6,11,16)	1.00	1.00	1.00	1.00	1.00	1.00
4	3	(1,1,1,16)	0.00	0.15	0.25	0.20	0.12	0.00
4	4	(1,1,1,16)	0.41	0.33	0.45	0.35	0.17	0.31
4	5	(1,1,1,16)	0.34	0.52	0.65	0.61	0.25	0.58
4	6	(1,1,1,16)	0.67	0.71	0.80	0.70	0.37	0.75
4	7	(1,1,1,16)	0.76	0.83	0.88	0.88	0.51	0.87
4	8	(1,1,1,16)	0.87	0.89	0.92	0.90	0.67	0.93
4	9	(1,1,1,16)	0.92	0.94	0.96	0.96	0.77	0.96
4	10	(1,1,1,16)	0.97	0.97	0.98	0.98	0.87	0.98
4	20	(1,1,1,16)	1.00	1.00	1.00	1.00	1.00	1.00

Note: Entries based on 1,000 replications. Standard deviation $\leq (4000)^{-1/2} = .016$.

Table 1.12 Continued

k	n	Variance	LevMed	MLM	MLM-BA	BLM	SH	Gini
8	3	(1,1,1,8,8,8,16,16)	0.00	0.06	0.06	0.10	0.30	0.00
8	4	(1,1,1,8,8,8,16,16)	0.39	0.22	0.24	0.33	0.48	0.15
8	5	(1,1,1,8,8,8,16,16)	0.11	0.41	0.42	0.49	0.70	0.39
8	6	(1,1,1,8,8,8,16,16)	0.63	0.63	0.65	0.67	0.89	0.68
8	7	(1,1,1,8,8,8,16,16)	0.67	0.84	0.85	0.89	0.98	0.90
8	8	(1,1,1,8,8,8,16,16)	0.91	0.94	0.95	0.95	0.99	0.97
8	9	(1,1,1,8,8,8,16,16)	0.95	0.98	0.98	0.98	1.00	0.99
8	10	(1,1,1,8,8,8,16,16)	0.99	0.99	1.00	0.99	1.00	1.00
8	20	(1,1,1,8,8,8,16,16)	1.00	1.00	1.00	1.00	1.00	1.00
8	3	(1,1,1,1,1,1,16,16)	0.00	0.18	0.24	0.25	0.10	0.00
8	4	(1,1,1,1,1,1,16,16)	0.58	0.45	0.53	0.49	0.14	0.38
8	5	(1,1,1,1,1,1,16,16)	0.51	0.70	0.75	0.74	0.22	0.72
8	6	(1,1,1,1,1,1,16,16)	0.87	0.89	0.92	0.89	0.40	0.90
8	7	(1,1,1,1,1,1,16,16)	0.91	0.96	0.97	0.96	0.58	0.97
8	8	(1,1,1,1,1,1,16,16)	0.99	0.99	0.99	0.99	0.77	1.00
8	9	(1,1,1,1,1,1,16,16)	1.00	1.00	1.00	1.00	0.89	1.00
8	10	(1,1,1,1,1,1,16,16)	1.00	1.00	1.00	1.00	0.95	1.00
8	20	(1,1,1,1,1,1,16,16)	1.00	1.00	1.00	1.00	1.00	1.00

Note: Entries based on 1,000 replications. Standard deviation $\leq (4000)^{-1/2} = .016$.

Table 1.13: Estimated Power for the Extreme Value Distribution, $\alpha = 0.05$.

k	n	Variance	LevMed	MLM	MLM-BA	BLM	SH	Gini
2	3	(1,4)	0.00	0.06	0.06	0.04	0.16	0.00
2	4	(1,4)	0.13	0.08	0.09	0.08	0.20	0.07
2	5	(1,4)	0.02	0.09	0.12	0.12	0.22	0.09
2	6	(1,4)	0.15	0.15	0.19	0.16	0.27	0.14
2	7	(1,4)	0.10	0.17	0.19	0.21	0.29	0.17
2	8	(1,4)	0.18	0.20	0.24	0.21	0.32	0.21
2	9	(1,4)	0.19	0.26	0.31	0.28	0.36	0.27
2	10	(1,4)	0.28	0.30	0.34	0.31	0.42	0.31
2	20	(1,4)	0.57	0.59	0.62	0.60	0.65	0.61
2	3	(1,8)	0.00	0.09	0.09	0.07	0.28	0.00
2	4	(1,8)	0.18	0.11	0.14	0.10	0.34	0.10
2	5	(1,8)	0.06	0.18	0.22	0.21	0.41	0.18
2	6	(1,8)	0.25	0.25	0.31	0.25	0.48	0.26
2	7	(1,8)	0.22	0.31	0.40	0.37	0.52	0.32
2	8	(1,8)	0.36	0.38	0.47	0.40	0.60	0.40
2	9	(1,8)	0.43	0.49	0.58	0.53	0.66	0.51
2	10	(1,8)	0.51	0.55	0.62	0.55	0.68	0.55
2	20	(1,8)	0.89	0.89	0.91	0.90	0.92	0.90
4	3	(1,6,11,16)	0.00	0.09	0.11	0.10	0.29	0.00
4	4	(1,6,11,16)	0.19	0.11	0.13	0.14	0.38	0.07
4	5	(1,6,11,16)	0.05	0.15	0.19	0.20	0.45	0.14
4	6	(1,6,11,16)	0.24	0.24	0.27	0.26	0.55	0.22
4	7	(1,6,11,16)	0.20	0.34	0.38	0.39	0.64	0.33
4	8	(1,6,11,16)	0.33	0.37	0.42	0.39	0.68	0.37
4	9	(1,6,11,16)	0.36	0.48	0.52	0.52	0.73	0.47
4	10	(1,6,11,16)	0.49	0.54	0.58	0.55	0.79	0.54
4	20	(1,6,11,16)	0.95	0.96	0.97	0.96	0.97	0.95
4	3	(1,1,1,16)	0.00	0.15	0.27	0.23	0.18	0.00
4	4	(1,1,1,16)	0.32	0.23	0.38	0.24	0.21	0.20
4	5	(1,1,1,16)	0.26	0.42	0.57	0.52	0.30	0.39
4	6	(1,1,1,16)	0.55	0.55	0.69	0.60	0.36	0.54
4	7	(1,1,1,16)	0.58	0.66	0.77	0.73	0.44	0.68
4	8	(1,1,1,16)	0.72	0.74	0.82	0.77	0.55	0.75
4	9	(1,1,1,16)	0.79	0.84	0.90	0.87	0.58	0.83
4	10	(1,1,1,16)	0.87	0.88	0.92	0.90	0.67	0.88
4	20	(1,1,1,16)	1.00	1.00	1.00	1.00	0.93	1.00

Note: Entries based on 1,000 replications. Standard deviation $\leq (4000)^{-1/2} = .016$.

Table 1.13 Continued

k	n	Variance	LevMed	MLM	MLM-BA	BLM	SH	Gini
8	3	(1,1,1,8,8,8,16,16)	0.00	0.10	0.12	0.18	0.33	0.00
8	4	(1,1,1,8,8,8,16,16)	0.30	0.19	0.21	0.23	0.42	0.12
8	5	(1,1,1,8,8,8,16,16)	0.15	0.35	0.38	0.40	0.55	0.26
8	6	(1,1,1,8,8,8,16,16)	0.45	0.46	0.51	0.47	0.69	0.43
8	7	(1,1,1,8,8,8,16,16)	0.45	0.63	0.67	0.67	0.73	0.59
8	8	(1,1,1,8,8,8,16,16)	0.68	0.72	0.75	0.72	0.81	0.68
8	9	(1,1,1,8,8,8,16,16)	0.73	0.83	0.86	0.85	0.85	0.79
8	10	(1,1,1,8,8,8,16,16)	0.85	0.88	0.91	0.88	0.88	0.85
8	20	(1,1,1,8,8,8,16,16)	1.00	1.00	1.00	1.00	0.98	1.00
8	3	(1,1,1,1,1,1,16,16)	0.01	0.20	0.26	0.29	0.16	0.01
8	4	(1,1,1,1,1,1,16,16)	0.47	0.37	0.47	0.38	0.18	0.30
8	5	(1,1,1,1,1,1,16,16)	0.40	0.59	0.67	0.64	0.24	0.53
8	6	(1,1,1,1,1,1,16,16)	0.75	0.77	0.84	0.78	0.35	0.73
8	7	(1,1,1,1,1,1,16,16)	0.77	0.85	0.90	0.88	0.40	0.81
8	8	(1,1,1,1,1,1,16,16)	0.90	0.91	0.94	0.93	0.50	0.89
8	9	(1,1,1,1,1,1,16,16)	0.94	0.96	0.98	0.98	0.59	0.95
8	10	(1,1,1,1,1,1,16,16)	0.97	0.98	0.98	0.98	0.64	0.97
8	20	(1,1,1,1,1,1,16,16)	1.00	1.00	1.00	1.00	0.94	1.00

Note: Entries based on 1,000 replications. Standard deviation $\leq (4000)^{-1/2} = .016$.

Chapter 2

New Methods Using Levene Type Tests for RCB Design

2.1 Introducation

Chapter 1 develops new methods using Levene type tests to test equality of dispersion for the one-way design. It is easy to extend these tests to the two-way design with more than one observation per cell. However, in the randomized complete block design (RCB), a two-way design with only one observation per cell, the situation is more complicated. Suppose we have independent observations from t treatments in b blocks. Let Y_{ij} be the observation in the ith block from the jth treatment. Table 2.1 shows the format of the data array for the RCB design.

Table 2.1: Data Array for the RCB Design

	Trt 1	Trt 2	...	Trt t
Block 1	Y_{11}	Y_{12}	...	Y_{1t}
Block 2	Y_{21}	Y_{22}	...	Y_{2t}
⋮	⋮	⋮	⋮	⋮
Block b	Y_{b1}	Y_{b2}	...	Y_{bt}

2.1.1 Model

Assuming a fixed effects model and homogeneity of variance, the RCB model is given by:

$$Y_{ij} = \mu + \alpha_i + \tau_j + \sigma e_{ij}, \quad i = 1, \cdots, b, \quad j = 1, \cdots, t, \tag{2.1}$$

where α_i represents the effect of the ith block under the constraint $\sum_{i=1}^{b} \alpha_i = 0$, τ_j the effect of the jth treatment under the constraint $\sum_{j=1}^{t} \tau_j = 0$, $\sigma > 0$ and the e_{ij}'s are iid with mean 0 and variance 1. The ANOVA for this model assumes that the variance of Y_{ij} is constant across blocks and treatments. When variances are not constant across blocks and treatments, by replacing σ in (2.1) by σ_{ij}, the new RCB model can be formularized as:

$$Y_{ij} = \mu + \alpha_i + \tau_j + \sigma_{ij} e_{ij}, \quad i = 1, \cdots, b, \quad j = 1, \cdots, t, \tag{2.2}$$

where α_i represents the ith block effect with $\sum_{i=1}^{b} \alpha_i = 0$, τ_j the jth treatment effect and $\sum_{j=1}^{t} \tau_j = 0$, and e_{ij}'s are iid with mean 0 and variance 1.

Under (2.2), the expectation of Y_{ij} satisfies:

$$E(Y_{ij}) = \mu + \alpha_i + \tau_j, \quad i = 1, \cdots, b, \quad j = 1, \cdots, t. \tag{2.3}$$

Note that (2.3) is an additive model. Similarly, we can also propose a model for the standard deviation of Y_{ij} that parallels the additive model for means (2.3),

$$\sigma_{ij} = \sigma + \phi_i + \theta_j, \ i = 1, \cdots, b, \ j = 1, \cdots, t, \tag{2.4}$$

where ϕ_i represents the ith block effect on standard deviation with the constraint $\sum_{i=1}^{b} \phi_i = 0$, θ_j represents the jth treatment effect on standard deviation with $\sum_{j=1}^{t} \theta_j = 0$, and $\sigma_{ij} > 0$ for any $i = 1, \cdots, b$, $j = 1, \cdots, t$.

If $\theta_j \neq 0$ for some j, then the standard deviation of Y_{ij} is not constant across treatments. If $\phi_i \neq 0$ for some i, then the standard deviation of Y_{ij} is not constant across blocks. Therefore, we have two sets of hypotheses:

$$H_{T0}: \ \theta_j = 0, j = 1, \cdots, t, \text{ vs. } H_{T1}: \ \theta_j \neq 0 \text{ for some } j;$$

$$H_{B0}: \ \phi_i = 0, i = 1, \cdots, b, \text{ vs. } H_{B1}: \ \phi_i \neq 0 \text{ for some } i.$$

2.1.2 Normal-Theory Tests

All of the normal-theory tests introduced in this section are only for testing the homogeneity of column variances (treatment effects on variance) assuming equality of row variances (no block effects on variance). In other words, these normal-theory tests are used to test H_{T0} assuming H_{B0} is true.

Box (1954b), Graybill (1954), Hartley (1950a) and Russell and Bradley (1958) discussed tests of the heterogeneity of variances for the RCB design with the assumption of normality. Russell and Bradley (1958) derived the likelihood ratio test for the case of only three treatments. Han (1969) proposed two test procedures to test the homogeneity of column variances in a two-way design, the multiple correlation test and the F_{max} test. The multiple correlation test is an exact test under normality for designs with $b > t$, where t is the number of treatments and b is the number of blocks. The procedure can be described as follows. Replace the t observations in the ith block by the ith block mean, $\overline{Y}_{i.} = \frac{\sum_{j=1}^{t} Y_{ij}}{t}$. The equality of variability across treatments can be concluded if and only if R, the multiple correlation coefficient between the block means $\overline{Y}_{i.}$ and any $(t-1)$ mean-adjusted observations, $Y_{ij} - \overline{Y}_{i.}$, for $j = 2, \cdots, t$, is zero. The test statistic is:

$$F_{Han} = \left(\frac{b-t}{t-1}\right)\left(\frac{\widehat{R}^2}{1-\widehat{R}^2}\right), \tag{2.5}$$

where $\widehat{R}$ is the empirical estimation of R by using b blocks as individuals. The equality

of variability across treatments will be rejected at level α if the test statistic is larger than the $100(1-\alpha)$ percentile of the F distribution with $t-1$ and $b-t$ degrees of freedom. O'Neil and Mathews (2002) noted that Han's test for H_{T0} requires H_{B0} to be true.

Han's test performs well for testing H_{T0} assuming H_{B0} is true under normality. However, Han's test is dependent on the parameters α_i in (2.2) because the sample multiple correlation $\widehat{R}$ is related to the values of the α_i. Shukla (1972) developed a new test based on Han's test that is invariant to the parameters α_i.

Han's test (Han, 1969) and Shukla's test (Shukla, 1972) are valid under normality, but very sensitive to non-normality. In addition, when applying Han's test and Shukla's test for equality of variances across treatments, we need to assume no block effects on variance.

O'Neil and Mathews (2002) developed a weighted least squares modification of the Levene type test (WLS Levene test) for the RCB design that is more robust than the OLS Levene test. These two methods using Levene type tests are not only invariant to the mean parameters α_i and τ_j, but also able to test equality of variances across treatments without the assumption of equality of variability across blocks. These two methods will be described in detail in the following sections.

2.2 Levene Type Tests for the Two-Way RCB Design

From (2.2), we have $E|Y_{ij} - \mu - \alpha_i - \tau_j| = \sigma_{ij}E|e_{ij}| = c\sigma_{ij}$ where c is free of σ_{ij}. With this relationship, an RCB ANOVA on the absolute values of the residuals, $|r_{ij}| = |Y_{ij} - \widehat{Y}_{ij}|$ where $\widehat{Y}_{ij}$'s are the fitted values from a particular fitting method, will provide tests of H_{T0} and H_{B0}. We now discuss how to fit the model in order to obtain the residuals, r_{ij}.

2.2.1 Existing Methods

Two current methods for testing H_{T0} and H_{B0}, the OLS Levene test and the WLS Levene test, are based on residuals r_{ij}, where the $\widehat{Y}_{ij}$ are obtained by fitting (2.2) using ordinary least squares. Standard ANOVA on the absolute value of the OLS residuals is referred to as the OLS Levene test, whereas a weighted ANOVA on the OLS residuals (O'Neil and Mathews, 2002) is referred to as the WLS Levene test.

OLS Levene Test

The ordinary least squares Levene test (OLS Levene test) is performed as follows:

1. Use the ordinary least squares method (OLS) to fit the model (2.2) and get the fitted values, $\widehat{Y}_{ij} = \widehat{\mu} + \widehat{\alpha}_i + \widehat{\tau}_j$.

2. Get the absolute values of the OLS residuals, $|r_{ij}| = |Y_{ij} - \widehat{Y}_{ij}|$.

3. Perform the two-way ANOVA F test on the absolute values of the OLS residuals, $|r_{ij}|$.

Actually, there are two F tests, one for blocks and one for treatments. Since they are interchangeable with the same properties, we just refer to the test for treatments as the "OLS Levene test".

The simulation results from the next section show that the OLS Levene test for treatments performs well for the normal distribution except when b is small. Similarly, the test for blocks perform well except when t is small. The OLS Levene test is however, highly sensitive to non-normality due to its non-robust estimation method. In fact, the OLS Levene test is a generalization of Levene's test (1960) using absolute deviations from the mean in the one-way model. Miller (1968) showed that this Levene test (Levene, 1960) is not asymptotically distribution-free. Therefore, it makes sense that the OLS Levene test is not fully appropriate here either.

WLS Levene Test

O'Neill and Mathews (2002) developed a WLS Levene test for two-way designs based on the OLS residuals, $r_{ij} = Y_{ij} - \widehat{Y}_{ij}$. Compared to the OLS Levene test, the only difference is in the third step in which WLS F tests are performed instead of the two OLS F tests on the absolute residuals. The WLS F-value for testing H_{T0} or H_{B0} in an RCB design is simply $F_{WLS} = m \times F_{OLS}$, where m can be obtained from Table 1 in O'Neill and Mathews (2002). Consider, for example, a study with

$b = 3$ and $t = 8$. In this situation, the appropriate WLS F-value for testing H_{T0} is $F_{T;WLS} = m \times F_{T;OLS}$, where $m = 0.537$. Similarly, the appropriate WLS F-value for testing H_{B0} is $F_{B;WLS} = m \times F_{B;OLS}$, where $m = 0.876$ can be obtained from Table 1 by switching the values of b and t.

O'Neill and Mathews (2002) gave the formula to get the multiplier m for the WLS Levene test. The weights are based on the covariance matrix for the $|r_{ij}|$, assuming normality of r_{ij}. There are three distinct correlations in the correlation matrix of the standardized residuals, r_{ij}. They are $\rho_1 = -1/(t-1)$ for the correlation between any two residuals in the same block, $\rho_2 = -1/(b-1)$ for the correlation between any two residuals in the same treatment, and $\rho_3 = 1/[(b-1)(t-1)]$ for the correlation between any two residuals in different blocks and different treatments. Based on the correlation matrix of residuals, we can calculate the variance-covariance matrix, $\mathbf{V}$, for the absolute values of residuals, $|r_{ij}|$. The diagonal elements of $\mathbf{V}$ are the same, $\omega_0 = 1 - 2/\pi$, and there are three distinct off-diagonal elements, $\omega_i = (2/\pi)[(1-\rho_i^2)^{1/2} + \rho_i \sin^{-1}(\rho_i) - 1]$, $i = 1, 2, 3$.

With some calculation, the appropriate WLS F-value for testing H_{T0} in the randomized block design is simply $F_{T;WLS} = m \times F_{T;OLS}$, where

$$m = \frac{\omega_0 - \omega_1 - \omega_2 + \omega_3}{\omega_0 - \omega_1 + (b-1)(\omega_2 - \omega_3)}. \tag{2.6}$$

Also, the multiplier m to test T_{B0} is

$$m = \frac{\omega_0 - \omega_2 - \omega_1 + \omega_3}{\omega_0 - \omega_2 + (t-1)(\omega_1 - \omega_3)}. \tag{2.7}$$

The WLS Levene test is more robust than the OLS Levene test, but it is still not sufficiently robust to non-normality.

2.2.2 New Methods

The OLS Levene test and the WLS Levene test are not Type I error robust enough due to their using OLS for fitting the mean parameters. We develop new methods based on Levene type tests for the two-way RCB design by applying robust estimation methods in fitting the model (2.2). The least absolute deviation method (LAD) and Huber M-estimation method are good choices.

LAD Levene

The least absolute deviation method is suggested because it is the natural generalization to the linear model of using medians for location estimation. The LAD estimation is a mathematical optimization technique similar to OLS. However, the OLS is not as robust as the LAD, where the parameters are estimated through minimization of the sum of the absolute values of residuals, $\sum_{i=1}^{b}\sum_{j=1}^{t}\rho(Y_{ij}-\mu-\alpha_i-\tau_j)$, where $\rho(z) = |z|$. The difference between OLS and LAD is that $\rho(z) = |z|$ in LAD instead of $\rho = \frac{1}{2}z^2$ in OLS. As pointed out by Huber (1981), a robust regression

estimation requires $\psi(z) = \rho'(z)$ to be bounded. Obviously, LAD is more robust than OLS.

The LAD Levene test is performed as follows:

1. Use the least absolute deviation method (LAD) to fit (2.2) and get the fitted values, $\widehat{Y}_{ij}^{L} = \widehat{\mu}^{L} + \widehat{\alpha}_i + \widehat{\tau}_j^{L}$.

2. Get the absolute values of the LAD residuals, $|r_{ij}^{L}| = |Y_{ij} - \widehat{Y}_{ij}^{L}|$.

3. Perform the two-way ANOVA F tests on the absolute values of the residuals, $|r_{ij}^{L}|$.

The LAD estimation is robust to outliers but the solutions are not unique. LAD estimators are not stable because the ρ-function $|z|$ is not strictly convex in z. In the simulations, we use the **quantreg** package of R (Koenker, 2004) to perform the LAD regression. The LAD Levene test is a robust test but is conservative in situations with small b or t. Good performance of the MLM test in Chapter 1 suggests deletion of zero values from the set of $|r_{ij}^{H}|$. Unfortunately, it is not obvious how to delete the zero values in r_{ij}^{L} since different solutions result in different $r_{ij}^{L} = 0$. In addition, deleting multiple zeros will cause unbalanced problem with F tests.

Huber Levene

Huber (1973) introduced robust M-estimation in regression. Here, we use Huber's Proposal 2 M-estimation method (Huber, 1964, 1973; Carroll, 1980) to estimate the

mean parameters in Model (2.2).

It's well known that the OLS estimates minimize $\sum_{i=1}^{b}\sum_{j=1}^{t}\rho(Y_{ij}-\mu-\alpha_i-\tau_j)$, where $\rho(z)=\frac{1}{2}z^2$ and its influence function $\psi(z)=z$ is unbounded. The quadratic form of ρ and unboundedness of the influence function result in OLS's non-robustness to outliers and non-normality. Huber M-estimation just generalizes OLS by replacing the quadratic function $\rho(z)=\frac{1}{2}z^2$ by

$$\rho(z)=\begin{cases} z^2/2 & \text{if } |z|\le k \\ |z| & \text{if } z>k \end{cases}$$

Huber's influence function is $\psi(z)=\rho'(z)=\ \max(-k,\ \min(z,k))$, which is a bounded function. Huber Proposal 2 M-estimation solves the equations (2.8) and (2.9) simultaneously subject to the constraints, $\sum_{i=1}^{b}\alpha_i=0$, and $\sum_{j=1}^{t}\tau_j=0$:

$$\sum_{i=1}^{b}\sum_{j=1}^{t}\psi((Y_{ij}-\mathbf{X_{ij}}^{\mathbf{T}}\beta)/\sigma)\mathbf{X_{ij}}=\mathbf{0} \tag{2.8}$$

$$(bt-(b+t-1))^{-1}\sum_{i=1}^{b}\sum_{j=1}^{t}\psi^2((Y_{ij}-\mathbf{X_{ij}}^{\mathbf{T}}\beta)/\sigma)=\mathrm{E}_{\mathbf{\Phi}}\psi^{\mathbf{2}}(\mathbf{Z}) \tag{2.9}$$

where $\beta=(\mu,\alpha_{\mathbf{1}},\cdots,\alpha_{\mathbf{b}},\tau_{\mathbf{1}},\cdots,\tau_{\mathbf{t}})^{\mathbf{T}}$ and $\mathbf{X_{ij}}$ is the design vector such that $\mathbf{X_{ij}}^{\mathbf{T}}\beta=\mu+\alpha_{\mathbf{i}}+\tau_{\mathbf{j}}$. Z in (2.9) is a standard normal random variable.

In this paper, we use Huber's Proposal 2 method with $k=1$. The Huber Levene

test is performed as follows:

1. Use the Huber Proposal 2 M-estimation method (Huber, 1964) to fit the model (2.2) and get the fitted values, $\widehat{Y}_{ij}^{H} = \widehat{\mu}^{H} + \widehat{\alpha}_{i}^{H} + \widehat{\tau}_{j}^{H}$.

2. Compute the absolute values of the Huber residuals, $|r_{ij}^{H}| = |Y_{ij} - \widehat{Y}_{ij}^{H}|$.

3. Perform the two-way ANOVA F test on the absolute values of the residuals, $|r_{ij}^{H}|$.

Bootstrap Versions

We also apply the bootstrap technique to the four tests in the two-way design. The bootstrap procedure here can be illustrated with the bootstrap version of the LAD Levene test (BLAD). The BLAD test is carried out as follows.

1. Apply the LAD Levene test to the data $\{Y_{ij}\}$ to get the test statistic for blocks, T_1^L and the test statistic for treatments, T_2^L.

2. Use Huber Proposal 2 M-estimation to fit model (2.2) and get the residuals $r_{ij}^{H} = Y_{ij} - \widehat{Y}_{ij}^{H}$, $i = 1, \cdots, b,\ j = 1, \cdots, t$, where $\widehat{Y}_{ij}^{H}$ are the fitted values.

3. Initialize $A_1 = A_2 = 0$.

4. Draw $N = bt$ bootstrap errors r_{ij}^{*} with replacement from the set of residuals $\varphi = \{r_{ij}^{H} : i = 1, \cdots, b, j = 1, \cdots, t\}$. Construct a bootstrap data set, $Y_{ij}^{*} = \widehat{Y}_{ij}^{H} + r_{ij}^{**}$, where $r_{ij}^{**} = \frac{r_{ij}^{*}}{\sqrt{(t-1)\times(b-1)/(bt)}}$

5. Apply the LAD Levene test to the data set $\{Y_{ij}^*\}$ to get the bootstrap statistics T_1^{L*} for blocks, and T_2^{L*} for treatments. If $T_1^{L*} > T_1^L$, increment A_1 to $A_1 + 1$. If $T_2^{L*} > T_2^L$, increment A_2 to $A_2 + 1$.

6. Repeat steps 4 and 5 a total of $B = 499$ times. The bootstrap p-value for block is given by A_1/B, and the p-value for treatments is given by A_2/B.

In the same way, we obtain the bootstrap version of the OLS Levene test (BOLS test). The only differences are in Step 1 and Step 5. In these two steps, we compute the OLS Levene test statistic instead of the LAD Levene test statistic. Similarly, for the bootstrap version of the Huber Levene test (BH), we compute the Huber Levene test statistic in steps 1 and 5. Note that we need not consider the bootstrap version of the WLS Levene test because it is identical to the BOLS test.

2.3 Simulation

The goal of this section is to compare by simulation the four Levene type tests, the LAD Levene test, the OLS Levene test, the WLS Levene test and the Huber Levene test, in terms of robustness and power. In the simulation, we also include the bootstrap versions of these tests, namely the bootstrap version of the LAD Levene test (BLAD), the bootstrap version of the OLS Levene test (BOLS) and the bootstrap version of the Huber Levene test (BH). We consider three distributions for the errors e_{ij} in (2.2): (1) Normal (0,1), (2) t_3 distribution, and (3) Extreme Value. Distributions

(2) and (3) have been transformed so that all three distributions have mean 0 and variance 1. We use S=10,000 Monte Carlo replications for the four non-bootstrap tests. For the three bootstrap tests, we use S=1,000 Monte Carlo replications. The bootstrap replication size is $B = 499$. The nominal Type I error rate is 0.05.

For the simulation, we need to give the parameter configurations for (2.2). However, because the test results are invariant to the mean parameters α_i and τ_j, and we list the b and t combination without specifying the mean parameters.

(1) $b = 5,\ t = 3$; (2) $b = 5,\ t = 4$; (3) $b = 8,\ t = 3$;

(4) $b = 8,\ t = 4$; (5) $b = 15,\ t = 3$; (6) $b = 15,\ t = 4$.

For a given data set, there are two tests of hypothesis: H_{T0} vs H_{T1}, which tests for treatment effects on variance, and H_{B0} vs H_{B1}, which tests for block effects on variance. We assess null performance and power of the different test procedures in the four situations as follows:

1. H_{T0} and H_{B0} are both true.

   Null performance is assessed for testing H_{T0} when H_{B0} is true, and for testing H_{B0} when H_{T0} is true.

2. H_{T0} is true and H_{B0} is false.

   Null Performance is assessed for testing H_{T0} when H_{B1} is true.

   Power under an alternative H_{B1} is assessed when H_{T0} is true.

3. H_{T0} is false and H_{B0} is true.

Power under an alternative H_{T1} is assessed when H_{B0} is true.

Null Performance is assessed for testing H_{B0} when H_{T1} is true.

4. H_{T0} and H_{B0} are both false.

   Power under an alternative H_{T1} is assessed when H_{B0} is false, and power under an alternative H_{B1} is assessed when H_{T0} is false.

As we have chosen combinations of $(b,\ t)$ that are not symmetric, the tests of H_{T0} vs H_{T1} and H_{B0} vs H_{B1} in the different scenarios do not represent duplication but provide information about situations where the values of b and t are switched.

2.3.1 Situation I: H_{T0} and H_{B0} Both True

In this situation, we estimate the Type I error rate for each hypothesis test given that the other null is true. In Table 2.2, we summarize results for the four tests that obtain p-values from the F distribution. The left half summarizes levels for testing H_{B0} given absence of treatment effects on variance, and the right half summarizes levels for testing H_{T0} when there are no block effects on variance. From Table 2.2, the LAD Levene test is conservative, especially for extreme situations with a large number of blocks, b, and small number of treatments, t when testing H_{B0}. For example, the Type I error rates are much less than .05 for tests of H_{B0} when t is small. The OLS Levene test is liberal with unsatisfactory levels except in the case of test of H_{T0} when b is large ($b = 15$). The WLS Levene test is somewhat liberal under the non-normal distributions, particularly when testing H_{B0} for b large and t small. The Huber Levene

Table 2.2: Estimated Levels of Variance Tests in the RCB Design.

Distri-			H_{B0} vs H_{B1}				H_{T0} vs H_{T1}			
bution	b	t	LAD	OLS	WLS	Huber	LAD	OLS	WLS	Huber
Normal	5	3	0.000	0.126	0.021	0.007	0.022	0.089	0.058	0.061
	5	4	0.005	0.105	0.043	0.074	0.012	0.081	0.039	0.054
	8	3	0.000	0.194	0.017	0.002	0.036	0.071	0.054	0.052
	8	4	0.002	0.140	0.047	0.082	0.027	0.068	0.044	0.047
	15	3	0.000	0.362	0.013	0.000	0.047	0.059	0.049	0.052
	15	4	0.000	0.202	0.049	0.094	0.039	0.061	0.052	0.052
Column Mean			0.001	0.188	0.032	0.043	0.031	0.072	0.049	0.053
MAD from 0.05			0.049	0.138	0.018	0.040	0.020	0.022	0.005	0.004
t_3	5	3	0.000	0.230	0.056	0.005	0.023	0.099	0.059	0.053
	5	4	0.010	0.170	0.068	0.069	0.015	0.112	0.059	0.050
	8	3	0.001	0.412	0.113	0.002	0.033	0.074	0.053	0.040
	8	4	0.009	0.280	0.111	0.075	0.026	0.090	0.058	0.042
	15	3	0.001	0.740	0.241	0.005	0.039	0.065	0.057	0.042
	15	4	0.010	0.529	0.228	0.093	0.039	0.075	0.061	0.042
Column Mean			0.005	0.394	0.136	0.042	0.029	0.086	0.058	0.045
MAD from 0.05			0.045	0.344	0.086	0.038	0.021	0.036	0.008	0.006
Extreme	5	3	0.000	0.169	0.027	0.006	0.025	0.095	0.060	0.063
Value	5	4	0.009	0.141	0.057	0.089	0.018	0.100	0.054	0.061
	8	3	0.000	0.319	0.045	0.001	0.039	0.076	0.057	0.052
	8	4	0.004	0.218	0.074	0.096	0.031	0.091	0.061	0.055
	15	3	0.000	0.603	0.093	0.001	0.045	0.070	0.059	0.054
	15	4	0.003	0.389	0.125	0.118	0.037	0.083	0.069	0.050
Column Mean			0.003	0.307	0.070	0.052	0.033	0.086	0.060	0.056
MAD from 0.05			0.047	0.257	0.030	0.049	0.018	0.036	0.010	0.006

Note: Individual entries are based on 10,000 replications and have standard error ≤ 0.005.
OLS: OLS mean-based Levene Test; LAD: LAD Levene Test;
WLS: WLS mean-based Levene Test; Huber: Huber Levene Test

test has better null performance than LAD, OLS and WLS, and but it displays the oscillation in level seen with LevMed for even and odd sample sizes in the one-way design. Note that for testing H_{T0}, the number of blocks, b, can be thought of as the "sample size," while for the test of H_{B0}, $t = 3$ and $t = 4$ are analogous to the n odd

and n even sample sizes in the one-way layout. In this table, we also calculate the column mean and MAD from .05 for every distribution. The column mean and the MAD from .05 show that the Huber Levene test performs better than the other three tests.

Table 2.3: Estimated Levels of Variance Tests (Bootstrap Versions) in the RCB Design

Distri-			H_{B0} vs H_{B1}			H_{T0} vs H_{T1}		
bution	b	t	BLAD	BOLS	BH	BLAD	BOLS	BH
Normal	5	3	0.034	0.023	0.044	0.039	0.048	0.046
	5	4	0.043	0.040	0.049	0.052	0.043	0.048
	8	3	0.033	0.003	0.030	0.047	0.054	0.055
	8	4	0.038	0.032	0.034	0.049	0.044	0.051
	15	3	0.018	0.001	0.020	0.054	0.044	0.057
	15	4	0.033	0.051	0.058	0.047	0.054	0.062
Column Mean			0.033	0.025	0.039	0.048	0.048	0.053
MAD from 0.05			0.017	0.025	0.014	0.004	0.005	0.005
t_3	5	3	0.046	0.025	0.041	0.046	0.042	0.042
	5	4	0.053	0.034	0.039	0.065	0.048	0.045
	8	3	0.048	0.011	0.043	0.057	0.051	0.056
	8	4	0.068	0.036	0.048	0.048	0.056	0.056
	15	3	0.034	0.002	0.052	0.038	0.039	0.045
	15	4	0.054	0.043	0.042	0.043	0.031	0.033
Column Mean			0.051	0.025	0.044	0.050	0.045	0.046
MAD from 0.05			0.008	0.025	0.007	0.008	0.008	0.008
Extreme	5	3	0.031	0.022	0.038	0.042	0.037	0.038
Value	5	4	0.044	0.032	0.037	0.057	0.045	0.047
	8	3	0.033	0.007	0.039	0.037	0.053	0.051
	8	4	0.055	0.052	0.057	0.054	0.043	0.045
	15	3	0.023	0.001	0.032	0.053	0.052	0.054
	15	4	0.048	0.061	0.066	0.056	0.048	0.049
Column Mean			0.039	0.029	0.045	0.050	0.046	0.047
MAD from 0.05			0.013	0.025	0.013	0.007	0.005	0.004

Note: Individual entries are based on 1,000 replications and B=499 Bootstrap resamples.
BLAD: Bootstrap version of LAD Levene Test
BOLS: Bootstrap version of OLS mean-based Levene Test
BH: Bootstrap version of Huber Levene Test

In Situation I, we also compare by simulation the bootstrap versions of the tests, namely the BLAD test, the BOLS test and the BH test. The results are summarized in Table 2.3. In the extreme situations, the BOLS test is conservative especially when "sample size," is small and odd, i.e., the test of H_{B0} when $t = 3$. In contrast, the BLAD test and the BH test perform well in this situation. In terms of the column mean and the MAD from .05, the BH test outperforms the other two tests. Comparing this table to Table 2.2, we see that the bootstrap technique results in substantial improvement in the null performance of the original tests.

2.3.2 Situation II: H_{T0} True, H_{B0} False

This section aims to estimate levels of the test for treatment effects on variance in the presence of block effects on variance. In addition, it aims to estimate power of the test of H_{B0} vs H_{B1} given that H_{T0} is true.

We use the following parameter configurations for Model (2.4), each with $\theta_i = 0,\ i = 1, \cdots, t,$:

(1) $b = 5, t = 3,\ \phi = (0.6, 0.4, -0.3, -0.5, -0.2)$;

(2) $b = 5, t = 4,\ \phi = (-0.5, 0.5, 0.4, 0, -0.4)$;

(3) $b = 8, t = 3,\ \phi = (-0.2, 0.3, 0.8, -0.9, 0.5, -0.8, 0.6, -0.3)$;

(4) $b = 8, t = 4,\ \phi = (-0.5, 0.5, 0.4, 0.2, 0.6, -0.8, -0.3, -0.1)$;

(5) $b = 15, t = 3,\ \phi = (0, 0.9, -0.7, 0.6, -0.3, -0.7, 0.5, -0.45, 0.45, -0.79, 0.79,$

$0.4, -0.2, -0.5, 0)$;

Table 2.4: Estimated Levels for Tests across Treatments and Power for Tests across Blocks in the RCB Design

Distri-			H_{B0} vs H_{B1}				H_{T0} vs H_{T1}			
bution	b	t	LAD	OLS	WLS	Huber	LAD	OLS	WLS	Huber
Norm	5	3	0.000	0.220	0.040	0.004	0.022	0.091	0.055	0.052
	5	4	0.024	0.220	0.101	0.147	0.014	0.079	0.038	0.044
	8	3	0.001	0.549	0.114	0.004	0.044	0.088	0.066	0.044
	8	4	0.028	0.429	0.196	0.244	0.031	0.070	0.045	0.042
	15	3	0.002	0.891	0.253	0.009	0.043	0.070	0.060	0.047
	15	4	0.079	0.854	0.486	0.499	0.037	0.071	0.056	0.044
Column Mean			0.022	0.527	0.198	0.151	0.032	0.078	0.053	0.046
MAD from 0.05							0.018	0.028	0.009	0.005
t_3	5	3	0.000	0.300	0.084	0.004	0.021	0.112	0.063	0.046
	5	4	0.026	0.263	0.120	0.117	0.014	0.131	0.066	0.050
	8	3	0.005	0.647	0.223	0.010	0.033	0.100	0.075	0.037
	8	4	0.040	0.498	0.244	0.182	0.025	0.109	0.073	0.040
	15	3	0.013	0.939	0.473	0.027	0.038	0.080	0.067	0.037
	15	4	0.107	0.904	0.603	0.348	0.038	0.101	0.082	0.041
Column Mean			0.032	0.592	0.291	0.115	0.028	0.106	0.071	0.042
MAD from 0.05							0.022	0.056	0.021	0.008
Extreme	5	3	0.000	0.264	0.058	0.006	0.024	0.097	0.058	0.051
Value	5	4	0.027	0.254	0.114	0.151	0.016	0.110	0.057	0.060
	8	3	0.001	0.612	0.162	0.005	0.036	0.100	0.076	0.045
	8	4	0.033	0.499	0.238	0.235	0.032	0.101	0.068	0.049
	15	3	0.003	0.933	0.388	0.013	0.040	0.084	0.071	0.044
	15	4	0.067	0.907	0.595	0.440	0.037	0.097	0.082	0.048
Column Mean			0.022	0.578	0.259	0.142	0.031	0.098	0.069	0.050
MAD from 0.05							0.019	0.048	0.019	0.004

Note: Individual entries are based on 10,000 replications and have standard error ≤ 0.005.
OLS: OLS mean-based Levene Test; LAD: LAD Levene Test;
WLS: WLS mean-based Levene Test; Huber: Huber Levene Test

(6) $b = 15, t = 4$, $\phi = (0.1, 0.3, 0.7, 0.1, 0.4, -0.9, 0.8, 0.23, 0.4, -0.23, 0.6,$

$-0.6, -0.5, -0.7, -0.7)$.

We summarize the results for the four Levene type tests in Table 2.4. We have calculated the column mean for levels and power under different distributions. We

Table 2.5: Estimated Levels for Bootstrap Tests across Treatments and Power for Bootstrap Tests across Blocks in the RCB Design

Distri-			H_{B0} vs H_{B1}			H_{T0} vs H_{T1}		
bution	b	t	BLAD	BOLS	BH	BLAD	BOLS	BH
Norm	5	3	0.075	0.028	0.044	0.056	0.053	0.043
	5	4	0.105	0.091	0.097	0.060	0.049	0.058
	8	3	0.142	0.013	0.097	0.053	0.046	0.043
	8	4	0.171	0.126	0.136	0.043	0.037	0.034
	15	3	0.139	0.001	0.100	0.060	0.058	0.057
	15	4	0.351	0.267	0.293	0.051	0.056	0.062
Column Mean			0.164	0.088	0.128	0.054	0.050	0.050
MAD from 0.05						0.006	0.006	0.010
t_3	5	3	0.083	0.024	0.041	0.039	0.035	0.025
	5	4	0.085	0.057	0.059	0.055	0.041	0.037
	8	3	0.151	0.020	0.105	0.055	0.060	0.053
	8	4	0.168	0.085	0.104	0.044	0.043	0.034
	15	3	0.159	0.001	0.144	0.046	0.040	0.045
	15	4	0.293	0.154	0.188	0.043	0.045	0.038
Column Mean			0.157	0.057	0.107	0.047	0.044	0.039
MAD from 0.05						0.006	0.009	0.012
Extreme	5	3	0.082	0.019	0.045	0.040	0.039	0.040
Value	5	4	0.086	0.078	0.084	0.050	0.055	0.045
	8	3	0.121	0.018	0.097	0.048	0.044	0.042
	8	4	0.167	0.164	0.153	0.050	0.041	0.040
	15	3	0.137	0.002	0.114	0.049	0.061	0.054
	15	4	0.284	0.255	0.268	0.060	0.054	0.042
Column Mean			0.146	0.089	0.127	0.050	0.049	0.044
MAD from 0.05						0.004	0.008	0.008

Note: Individual entries are based on 1,000 replications and B=499 Bootstrap resamples.
BLAD: Bootstrap version of LAD Levene Test
BOLS: Bootstrap version of OLS mean-based Levene Test
BH: Bootstrap version of Huber Levene Test

only calculate the MAD from .05 for estimated levels. In the presence of block effects on variance, estimated levels for testing H_{T0} in Table 2.4 are similar to those in Table 2.2. This shows that inequality of variances across blocks has little effct on the

estimated levels for the test of treatment effects on variance. As for power, the OLS and the WLS tests have higher power than the other tests, but the comparison is unfair due to their inflated type I error rates. The LAD Levene test is conservative with loss of power. The Huber Levene test has higher power than the LAD Levene test.

Table 2.5 summarizes the results for the bootstrap tests, BLAD, BOLS and BH. Estimated levels for all three tests are similar to those in Table 2.3 for the test of H_{T0}. The BLAD test has a little higher power than the BH test. Power is low for BOLS, particularly when $t = 3$. In fact, the BOLS test has less power than the Huber Levene test which can achieve power close to BLAD.

2.3.3 Situation III: H_{T0} False, H_{B0} True

In this section, we estimate levels for the test of H_{B0} vs H_{B1} in the situation where treatments affect variance. Also, we assess power for the test of H_{T0} vs H_{T1} given that H_{B0} is true. We consider the following parameter configurations, for Model (2.4), all of which include $\phi_i = 0,\ i = 1, \cdots, b,$:

(1) $b = 5, t = 3,\ \theta = (0, -0.4, 0.4)$;

(2) $b = 5, t = 4,\ \theta = (-0.4, -0.1, 0.3, 0.2)$;

(3) $b = 8, t = 3,\ \theta = (-0.7, 0.1, 0.6)$;

(4) $b = 8, t = 4,\ \theta = (0, 0.5, -0.8, 0.3)$;

(5) $b = 15, t = 3,\ \theta = (-0.4, 0, 0.4)$;

Table 2.6: Estimated Levels for Tests across Blocks and Power for Tests across Treatments for the RCB Design

Distri-			H_{B0} vs H_{B1}				H_{T0} vs H_{T1}			
bution	b	t	LAD	OLS	WLS	Huber	LAD	OLS	WLS	Huber
Norm	5	3	0.000	0.180	0.036	0.010	0.038	0.133	0.088	0.090
	5	4	0.006	0.109	0.041	0.068	0.028	0.126	0.069	0.087
	8	3	0.001	0.376	0.065	0.006	0.148	0.227	0.184	0.181
	8	4	0.006	0.162	0.044	0.067	0.117	0.290	0.223	0.206
	15	3	0.000	0.531	0.052	0.002	0.210	0.246	0.220	0.245
	15	4	0.002	0.410	0.094	0.060	0.614	0.782	0.750	0.767
Column Mean			0.003	0.295	0.055	0.036	0.193	0.301	0.256	0.263
MAD from 0.05			0.048	0.245	0.015	0.030				
t_3	5	3	0.000	0.187	0.075	0.008	0.033	0.127	0.078	0.065
	5	4	0.012	0.279	0.084	0.068	0.022	0.146	0.080	0.074
	8	3	0.001	0.588	0.224	0.005	0.117	0.225	0.173	0.124
	8	4	0.010	0.372	0.157	0.069	0.089	0.250	0.184	0.137
	15	3	0.001	0.813	0.338	0.006	0.151	0.195	0.171	0.147
	15	4	0.011	0.763	0.432	0.071	0.481	0.644	0.603	0.538
Column Mean			0.006	0.500	0.218	0.038	0.149	0.265	0.215	0.181
MAD from 0.05			0.044	0.450	0.168	0.032				
Extreme	5	3	0.000	0.219	0.046	0.009	0.038	0.130	0.078	0.083
Value	5	4	0.009	0.146	0.052	0.076	0.025	0.152	0.085	0.097
	8	3	0.000	0.462	0.122	0.005	0.136	0.241	0.194	0.160
	8	4	0.008	0.265	0.087	0.083	0.099	0.293	0.223	0.183
	15	3	0.000	0.690	0.169	0.002	0.184	0.247	0.221	0.198
	15	4	0.008	0.593	0.247	0.081	0.548	0.754	0.721	0.660
Column Mean			0.004	0.396	0.121	0.043	0.172	0.303	0.254	0.230
MAD from 0.05			0.046	0.346	0.072	0.037				

Note: Individual entries are based on 10,000 replications and have standard error ≤ 0.005.
OLS: OLS mean-based Levene Test; LAD: LAD Levene Test;
WLS: WLS mean-based Levene Test; Huber: Huber Levene Test

(6) $b = 15, t = 4$, $\theta = (0.8, -0.9, 0.3, -0.2)$.

We summarize the results for the four non-bootstrap tests in Table 2.6. We have calculated column mean for levels and power under different distributions. We calculate the MAD from .05 only for estimated levels. Estimated levels in Table 2.6 for

Table 2.7: Estimated Levels for Bootstrap Tests across Blocks and Power for Bootstrap Tests across Treatments in the RCB Design

Distri-			H_{B0} vs H_{B1}			H_{T0} vs H_{T1}		
bution	b	t	BLAD	BOLS	BH	BLAD	BOLS	BH
Norm	5	3	0.053	0.042	0.079	0.073	0.073	0.070
	5	4	0.051	0.034	0.039	0.078	0.072	0.064
	8	3	0.082	0.038	0.103	0.198	0.175	0.197
	8	4	0.060	0.024	0.029	0.154	0.192	0.188
	15	3	0.044	0.006	0.076	0.235	0.218	0.254
	15	4	0.049	0.031	0.029	0.656	0.737	0.791
Column Mean			0.057	0.029	0.059	0.232	0.245	0.261
MAD from 0.05			0.009	0.021	0.027			
t_3	5	3	0.063	0.039	0.062	0.067	0.067	0.060
	5	4	0.053	0.029	0.036	0.081	0.061	0.065
	8	3	0.071	0.046	0.076	0.152	0.123	0.124
	8	4	0.065	0.034	0.038	0.135	0.130	0.149
	15	3	0.052	0.011	0.059	0.166	0.144	0.159
	15	4	0.054	0.038	0.026	0.545	0.518	0.583
Column Mean			0.060	0.033	0.050	0.191	0.174	0.190
MAD from 0.05			0.010	0.017	0.016			
Extreme	5	3	0.052	0.033	0.065	0.079	0.072	0.074
Value	5	4	0.041	0.031	0.035	0.076	0.076	0.079
	8	3	0.066	0.031	0.083	0.176	0.165	0.162
	8	4	0.063	0.044	0.045	0.158	0.191	0.182
	15	3	0.033	0.009	0.066	0.204	0.209	0.222
	15	4	0.054	0.035	0.037	0.567	0.651	0.682
Column Mean			0.052	0.031	0.055	0.210	0.227	0.234
MAD from 0.05			0.010	0.020	0.016			

Note: Individual entries are based on 1,000 replications and B=499 Bootstrap resamples.
BLAD: Bootstrap version of LAD Levene Test
BOLS: Bootstrap version of OLS mean-based Levene Test
BH: Bootstrap version of Huber Levene Test

testing H_{B0} are similar to those in Table 2.2. It appears that inequality of variances across treatments results in slightly higher estimated levels for the test of equality of variances across blocks except for Huber Levene. As for power, the OLS and the

WLS test have higher power than the other tests, but the comparison is unfair due to their inflated type I error rates. The Huber Levene test achieves a little higher power than the LAD Levene test.

Table 2.7 summarizes the results for the bootstrap tests, BLAD, BOLS and BH. These tests have estimated levels similar to those in Table 2.3. It also shows that inequality of variances across treatments increases estimated levels of tests for a block effect on variance for all the boostrap tests. The BH test has a little higher power than the BLAD test.

2.3.4 Situation IV: H_{T0} and H_{B0} Both False

In this section, we estimate the power of the test for block effects on variance given that treatment effects on variance are present, and vice versa.

We use the following parameter configurations for Model (2.4):

(1) $b = 5, t = 3$, $\theta = (0, -0.4, 0.4)$, $\phi = (0.6, 0.4, -0.3, -0.5, -0.2)$;

(2) $b = 5, t = 4$, $\theta = (-0.4, -0.1, 0.3, 0.2)$, $\phi = (-0.5, 0.5, 0.4, 0, -0.4)$, ;

(3) $b = 8, t = 3, \theta = (0.6, -0.2, -0.4)$, $\phi = (-0.2, -0.3, -0.3, 0.8, -0.5, 0.2, 0.6, -0.3)$;

(4) $b = 8, t = 4, \theta = (0, 0.8, -0.5, -0.3)$, $\phi = (-0.4, 0.5, -0.4, -0.2, -0.2, 0.8, -0.1, 0)$;

(5) $b = 15, t = 3$, $\theta = (-0.4, 0, 0.4)$,

$\phi = (0, 0.9, -0.4, 0.1, -0.3, -0.5, 0.5, -0.45, 0.45, -0.39, 0.39, 0.4, -0.2, -0.5, 0)$;

(6) $b = 15, t = 4$, $\theta = (0.7, 0, -0.4, -0.3)$,

$\phi = (0.1, 0.47, 0, 0.4, 0.8, -0.3, 0.23, 0.4, -0.23, -0.47, -0.1, -0.5, -0.3, -0.5)$.

Table 2.8: Estimated Power of Tests in the RCB Design

Distri-			H_{B0} vs H_{B1}				H_{T0} vs H_{T1}			
bution	b	t	LAD	OLS	WLS	Huber	LAD	OLS	WLS	Huber
Norm	5	3	0.000	0.258	0.058	0.009	0.037	0.129	0.082	0.075
	5	4	0.024	0.218	0.094	0.134	0.027	0.124	0.066	0.073
	8	3	0.001	0.499	0.106	0.006	0.093	0.148	0.113	0.093
	8	4	0.022	0.347	0.128	0.165	0.087	0.207	0.150	0.136
	15	3	0.002	0.885	0.267	0.012	0.164	0.190	0.168	0.145
	15	4	0.031	0.690	0.287	0.244	0.377	0.513	0.477	0.473
Column Mean			0.013	0.483	0.157	0.095	0.131	0.219	0.176	0.166
t_3	5	3	0.000	0.341	0.108	0.008	0.032	0.140	0.086	0.061
	5	4	0.026	0.271	0.122	0.107	0.023	0.166	0.084	0.065
	8	3	0.003	0.631	0.235	0.008	0.071	0.165	0.124	0.070
	8	4	0.029	0.476	0.224	0.132	0.068	0.203	0.148	0.097
	15	3	0.011	0.941	0.510	0.026	0.127	0.175	0.155	0.104
	15	4	0.041	0.844	0.528	0.179	0.269	0.411	0.374	0.300
Column Mean			0.018	0.584	0.288	0.077	0.098	0.210	0.162	0.116
Extreme	5	3	0.000	0.296	0.072	0.009	0.035	0.131	0.081	0.070
Value	5	4	0.024	0.249	0.106	0.134	0.026	0.149	0.078	0.082
	8	3	0.001	0.556	0.157	0.006	0.086	0.160	0.124	0.088
	8	4	0.024	0.419	0.181	0.158	0.073	0.209	0.154	0.120
	15	3	0.004	0.924	0.401	0.014	0.146	0.195	0.173	0.127
	15	4	0.029	0.774	0.415	0.218	0.318	0.496	0.459	0.395
Column Mean			0.014	0.536	0.222	0.090	0.114	0.223	0.178	0.147

Note: Individual entries are based on 10,000 replications and have standard error ≤ 0.005.
OLS: OLS mean-based Levene Test; LAD: LAD Levene Test;
WLS: WLS mean-based Levene Test; Huber: Huber Levene Test

We summarize the results for the four tests in Table 2.8 with column means listed. Assuming difference in variances due to treatments, the OLS Levene test and the WLS Levene test achieve much higher power than the other two tests. However, the comparison is unfair due to their inflated type I error rates. The Huber Levene test has higher power than the LAD Levene test. Assuming inequality of variances across blocks, the OLS Levene test achieves the highest power. The Huber Levene

test and the WLS Levene test perform equally well and are more powerful than the LAD Levene test.

Table 2.9 summarizes the results for the bootstrap version of variance tests, BLAD, BOLS test, and BH. From Table 2.9, the BLAD test performs better than the other two tests, especially in the extreme situations (testing H_{B0} with small t and large b). The BOLS test performs poorly in the extreme situations. Overall, the BH test is a little less powerful than BLAD.

2.3.5 Summary of Simulation Results

In general, the LAD Levene test is a very conservative test with loss of power. The OLS Levene test is liberal especially for non-normality or extreme situations (with small t for testing H_{B0} or with small b for testing H_{T0}). At the same time, the OLS Levene test has very high power, but this is not fair due to its highly inflated type I error rates. In terms of null performance, the WLS Levene performs well under normality and is a little liberal under non-normality, but it performs very poorly in extreme situations (with small t for testing H_{B0} or with small b for testing H_{T0}). The Huber Levene test performs better than the other three tests. The Huber Levene test is not as good as the WLS Levene test in terms of power, but the WLS Levene has inflated type I error rates. The Huber Levene test is more powerful than the LAD Levene. The bootstrap can greatly improve the performance of all of the tests in terms of null performance and power. However, the bootstrap versions of the tests

Table 2.9: Estimated Power of Bootstrap Tests in the RCB Design

Distri-			H_{B0} vs H_{B1}			H_{T0} vs H_{T1}		
bution	b	t	BLAD	BOLS	BH	BLAD	BOLS	BH
Norm	5	3	0.078	0.040	0.077	0.061	0.073	0.060
	5	4	0.106	0.085	0.098	0.075	0.073	0.075
	8	3	0.137	0.062	0.135	0.203	0.152	0.163
	8	4	0.150	0.101	0.099	0.297	0.252	0.311
	15	3	0.096	0.009	0.118	0.213	0.174	0.182
	15	4	0.173	0.147	0.144	0.515	0.514	0.576
Column Mean			0.123	0.074	0.112	0.227	0.206	0.228
t_3	5	3	0.097	0.041	0.054	0.057	0.041	0.034
	5	4	0.093	0.069	0.071	0.065	0.053	0.049
	8	3	0.112	0.073	0.105	0.163	0.126	0.123
	8	4	0.148	0.085	0.076	0.257	0.166	0.212
	15	3	0.114	0.015	0.116	0.160	0.128	0.135
	15	4	0.147	0.090	0.086	0.382	0.319	0.393
Column Mean			0.119	0.062	0.085	0.181	0.139	0.158
Extreme	5	3	0.087	0.030	0.053	0.052	0.062	0.055
Value	5	4	0.095	0.065	0.070	0.059	0.068	0.056
	8	3	0.094	0.058	0.112	0.197	0.158	0.156
	8	4	0.153	0.113	0.109	0.272	0.265	0.287
	15	3	0.118	0.010	0.121	0.198	0.179	0.188
	15	4	0.168	0.151	0.146	0.422	0.435	0.476
Column Mean			0.119	0.071	0.102	0.200	0.195	0.203

Note: Individual entries are based on 1,000 replications and B=499 Bootstrap resamples.
BLAD: Bootstrap version of LAD Levene Test
BOLS: Bootstrap version of OLS mean-based Levene Test
BH: Bootstrap version of Huber Levene Test

are much more complicated. The BLAD test and the BH perform much better than the BOLS test. From these tables, we also see that inequality of one set of variances has only a small effect on tests of the other set of variances.

2.4 Example

Swinderman and Cleophas (2005) proposed that it is important to compare variability among subjects in drug response for different formulations. Sometimes, comparing variability is more important than comparing means, especially for drugs with small therapeutic window sizes. The goal of a bioequivalence study in pharmaceutical research is to demonstrate that a new formulation is equivalent to an existing formulation. Usually, the mean response is regarded as the main criterion for comparison formulations.

Table 2.10: Part of Dataset GN24 on the FDA Website

		CMAX				Subject	Huber
Subject	Sequence	A	B	C	D	Mean	Mean
4	ABDC	1.571	1.393	1.199	1.682	1.461	0.172
5	ABDC	1.491	1.791	1.747	1.695	1.681	0.097
6	ABDC	1.643	2.136	1.708	1.958	1.861	0.195
10	CABD	1.984	1.284	2.123	2.200	1.898	0.265
11	CABD	2.004	2.077	1.778	1.712	1.893	0.146
12	CABD	1.983	1.445	2.646	1.895	1.992	0.314
16	DCAB	1.339	1.578	1.228	1.479	1.406	0.132
17	DCAB	1.389	1.348	2.038	1.467	1.561	0.194
18	DCAB	1.436	1.359	1.400	1.080	1.319	0.111
22	BDCA	1.883	1.716	1.713	1.792	1.776	0.068
23	BDCA	2.375	2.144	2.615	2.199	2.333	0.153
24	BDCA	1.507	1.562	1.264	1.199	1.383	0.150
Treatment Mean		1.717	1.653	1.788	1.697		
Huber Mean		0.103	0.222	0.224	0.116		

Data source: Data Set GN24 from
http://www.fda.gov/cder/bioequivdata/
Huber Mean: Average of the absolute values of Huber residuals.

The FDA website (http://www.fda.gov/cder/bioequivdata/) lists data sets from

a number of bioequivalence studies. We use part of the data set GN24 to illustrate the seven tests compared by simulation in earlier sections. The GN24 study compared anticonvulsant drugs on 24 subjects, including 12 males and 12 females. The experimental design is a four-period crossover design with 4 treatments including two reference treatments, A and D, and two test treatments, B and C. There are four different sequences: ABDC, CABD, DCAB and BDCA. Data set GN24 includes three pharmacokinetic (PK) parameters, the maximum plasma concentration (CMAX), the area under the plasma concentration-time curve from time zero to time infinity (AUCINF) and the area under the plasma concentration-time curve from time zero to time of last measurable concentration (AUCLAST). For illustration, we use a subset of the data consisting of the values for the PK parameter, CMAX, from only the male subjects (SEX=1). Table 2.10 lists all the data in our example. For the crossover design, the full model includes effects for period, treatment and subject. For illustration, we ignore the period and sequence effects. The subject effect is regarded as a block effect, leading to the two-way RCB design considered in this chapter. The null hypothesis of interest is that there are no treatment effects on variability in drug response, or H_{T0}: $\theta_j = 0, \quad j = 1, \cdots, 4$. We can also test for the presence of subject effects on variability in drug response, or H_{B0}: vs H_{B1}, though this test is of less interest.

In this example, we consider the following two scenarios:

Scenario 1:

It includes 10 male subjects with $b = 10$ and $t = 4$. The two male subjects excluded are subject 18 and subject 22 randomly in Table 2.10

Scenario 2:

It includes all male subjects with $b = 12$ and $t = 4$. Table 2.10 lists all the data used in this scenario.

Table 2.10 also summarizes the mean of CMAX for every treatment (Treatment Mean) and the mean of CMAX for every subject (Subject Mean). "Huber Mean" in Table 2.10 shows the average of the absolute values of Huber residuals across subjects or treatments.

From Table 2.11, we can see that none of the tests detect subject effects on variance. In Scenario 2, there is no evidence that formulations differ with respect to their effect on variability in drug response. However, in Scenario 1, p-values for the OLS Levene test, the WLS Levene test, and the BOLS test are each $< .05$, and are $\leq .10$ for the other tests. There are two explanations for this result. One is that the OLS Levene test and the WLS Levene test are more powerful than the other tests. The other is that the OLS Levene and the WLS Levene test are liberal tests. Based on the results from the simulations and comparing the results between the two scenarios, the second explanation seems more reasonable.

Table 2.11: P-Values for Variance Equality across Blocks and Variance Equality across Treatments

	t=4			
	b=10		b=12	
	BL	TR	BL	TR
OLS	0.54	0.02	0.35	0.08
WLS	0.76	0.03	0.63	0.10
LAD	0.87	0.10	0.77	0.26
Huber	0.70	0.05	0.63	0.11
BLAD	0.79	0.08	0.63	0.25
BOLS	0.90	0.03	0.88	0.08
BH	0.76	0.05	0.74	0.10

LAD: LAD Levene Test
OLS: OLS mean-based Levene Test
WLS: WLS mean-based Levene Test
Huber: Huber Levene Test
BLAD: Bootstrap version of LAD Levene Test
BOLS: Bootstrap version of OLS Levene Test
BH: Bootstrap version of Huber Levene Test

2.5 Conclusion

This chapter introduces some new variance tests for the two-way RCB design including the LAD Levene test, the Huber Levene test, the BLAD test, the BOLS test and the BH test. Among the simple tests, the Huber Levene test performs best in terms of null performance and power. The LAD Levene test is a very conservative test. The bootstrap technique can greatly improve the performance of these tests. The BLAD test and the BH test perform very well, compared to the BOLS test. The BOLS test is not recommended due to its highly conservative performance. The Huber Levene test does not perform as well as the BH test in extreme situations, but

it is much simpler to compute. In general, the Huber Levene test is recommended for use in the two-way RCB design.

Bibliography

[1] Bartlett, M. S. (1937). Properties of Sufficiency and Statistical tests. *Proc. Royal Society., Ser. A*. 160, 268-282.

[2] Boos, D. D., and Brownie, C. (1989). Bootstrap Methods for Testing Homogeneity of Variances. *Technometrics*, 31, 69-82.

[3] Boos, D. D., and Brownie, C. (2004). Comparing Variances and Other Measures of Dispersion. *Statistical Science*, 19, 571-578.

[4] Box, G. E. P. (1953). Non-Normality and Tests on Variances. *Biometrika*, 40, 318-335.

[5] Box, G. E. P. (1954a). Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, I. Effects of Inequality of Variance and of Correlation Between Errors in the Two-Way Classification. *Annals of Mathematical Statistics*, 25, 484-498.

[6] Box, G. E. P. (1954b). Some Theorems on Quadratic Forms Applied in the Study

of Analysis of Variance Problems, II. Effect of Inequality of Variance in the One-Way Classification. *Annals of Mathematical Statistics*, 25, 290-302.

[7] Box, G. E. P., and Andersen, S. L. (1955). Permutation Theory in the Derivation of Robust Criteria and the Study of Departure from Assumption. *Journal of the Royal Statistical Society, Ser. B*, 17, 1-26.

[8] Brown, M. B., and Forsythe, A. B. (1974). Robust Tests for the Equality of Variances. *Journal of the American Statistical Association*, 69, 364-367.

[9] Carroll, R. J. (1980). Robust Methods for Factorial Experiments with Outliers. *Applied Statistics*, 29, 246-251.

[10] Conover, W. J., Johnson, M. E., and Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, With Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, 23, 351-361.

[11] Graybill, F. (1954). Variance Heterogeneity in a Randomized Block Design. *Biometrics*, 10, 516-520.

[12] Han, C. P. (1969). Testing the Homogeneity of Variances in a Two-Way Classification. *Biometrics*, 25, 153-158.

[13] Hartley, H. O. (1950a). The Use of Range in Analysis of Variance. *Biometrika*, 37, 271-280.

[14] Hartley, H. O. (1950b). The Maximum F-Ratio as a Short-Cut Test for Heterogeneity of Variance. *Biometrika*, 37, 308-312.

[15] Hines, W. G. S., and Hines, R. J. O. (2000). Increased Power with Modified Forms of the Levene (Med) Test for Heterogeneity of Variance. *Biometrics*, 56, 451-454.

[16] Huber, P. J. (1964). Robust Estimation of location parameter. *Annals of Mathematical Statistics*, 35, 73-101.

[17] Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics*, 1, 799-821.

[18] Huber, P. J. (1981). *Robust Statistics*. Wiley.

[19] Johnson, N. L., and Kotz, S. (1970). *Distributions in statistics. Continuous Univariate Distributions - 2*. New York: Wiley.

[20] Koenker, R. (2004). quantreg: Quantile Regression, R package version 3.70. (http://www.econ.uiuc.edu/roger/research/rq/rq.html).

[21] Layard, M. W. J. (1973). Robust Large-Sample Tests for Homogeneity of Variances. *Journal of the American Statitical Association*, 68, 195-198.

[22] Levene, H. (1960). Robust Tests for Equality of Variances. *Contributions to Probability and Statistics*, ed, I. Olkin, Palo Alto, Calif.: Stanford University Press, 278-292.

[23] Lim, T.-S., and Loh, W.-Y. (1996). A Comparison of Tests of Equality of Variances. *Computational Statitics and Data Analysis*, 22, 287-301.

New York: Wiley.

[24] Miller, R. G. (1986). Jackknifing Variances. *The Annals of Mathematical Statistics*, 39, 567-582.

[25] O'Brian, R. G. (1978). Robust Techniques for Testing Heterogeneity of Variance Effects in Factorial Designs. *Psychometrika*, 43, 327-342.

[26] O'Neil, M. E., and Mathews, K. L. (2002). Levene Tests of Homogeneity of Variance for General Block and Treatment Designs. *Biometrics*, 58, 216-224.

[27] Phadke, M. S., Kackar, R. N., Speeney, D. V., and Grieco, M. J. (1983). Off-Line Quality Control in Integrated Circuit Fabrication Using Experimental Design. *The Bell System technical Journal*, 62, 1273-1309.

[28] Russell, T. S., and Bradley, R. A. (1958). One-way Variances in a Two-Way Classification. *Biometrika*, 45, 111-129.

[29] Shoemaker, L. H. (2003). Fixing the F test for Equal Variances. *The American Statistician*. 57, 105-114.

[30] Shukla, G. K. (1972). An Invariant Test for the Homogeneity of Variances in a Two-Way Classification *Biometrics*, 28, 1063-1072.

[31] Stefanski, L. A., and Boos, D. D. (2002). The Calculus of M-Estimation. *The American Statistician*. 56, 29-38.

[32] Zwinderman, A. H., and Cleophas, T. J. (2005). Variability in Clinical Data Is often More Useful than the Mean: Illustration of Concept and Simple Methods of Assessment *International Journal of Clinical Pharmacology and Therapeutics*. 43, 536-542.