2009

The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice.

Joseph L. Gastwirth

Yulia R. Gel

Weiwen Miao
Haverford College, wmiao@haverford.edu

Statistical Science
2009, Vol. 24, No. 3, 343–360
DOI: 10.1214/09-STS301

The Impact of Levene's Test of Equality of Variances on Statistical Theory and Practice

Joseph L. Gastwirth, Yulia R. Gel and Weiwen Miao

Abstract. In many applications, the underlying scientific question concerns whether the variances of k samples are equal. There are a substantial number of tests for this problem. Many of them rely on the assumption of normality and are not robust to its violation. In 1960 Professor Howard Levene proposed a new approach to this problem by applying the F-test to the absolute deviations of the observations from their group means. Levene's approach is powerful and robust to nonnormality and became a very popular tool for checking the homogeneity of variances.

This paper reviews the original method proposed by Levene and subsequent robust modifications. A modification of Levene-type tests to increase their power to detect monotonic trends in variances is discussed. This procedure is useful when one is concerned with an alternative of increasing or decreasing variability, for example, increasing volatility of stocks prices or "open or closed gramophones" in regression residual analysis. A major section of the paper is devoted to discussion of various scientific problems where Levene-type tests have been used, for example, economic anthropology, accuracy of medical measurements, volatility of the price of oil, studies of the consistency of jury awards in legal cases and the effect of hurricanes on ecological systems.

Key words and phrases: ANOVA, equality of variances, Levene's test, trend tests, effect of dependence, applied statistics.

INTRODUCTION

Very few statisticians write an article that is still cited forty or fifty years after it is published. Professor Howard Levene, whose research focused on statistical problems arising in biological science, was the sole author of three such classic papers. Not only have they been cited hundreds of times; they continue to be cited today. Professor Levene passed away in July, 2003 and this article is written in recognition of his important contributions to statistical science.

After introducing two earlier well cited articles, Levene (1949) and Levene (1953), the impact of the third article, on a robust test for the equality of

Joseph L. Gastwirth is Professor of Statistics and Economics, George Washington University, Pennsylvania Ave. 2140, Washington, DC 20052, USA e-mail: jlgast@gwu.edu. Yulia R. Gel is Associate Professor, Department of Mathematics and Statistics, University of Waterloo, 200 University Ave. W, Waterloo, Ontario, Canada N2L 3G1 e-mail: ygl@math.uwaterloo.ca. Weiwen Miao is Associate Professor, Department of Mathematics, Haverford College, Haverford, Pennsylvania 19041, USA e-mail: wmiao@haverford.edu.

the variances of k populations, will be emphasized. In particular, both the robustness aspect and the focus on the "spread" or variability of the data in the Levene (1960) article influenced the work of the authors, especially J. L. Gastwirth, who took his first class in Mathematical Statistics from Professor Levene.

The first seminal article of Professor Levene concerned checking that the random mating assumption often used in mathematical models in population genetics holds. This implies that the alleles transmitted by each parent are independent, that is, when there are two possible alleles, A and a at a locus, with frequencies $p(A) = p$ and $p(a) = 1 - p = q$ in the population, the frequencies of the three genotypes (AA, Aa and aa) in the next generation equal p^2, $2pq$ and q^2. Hardy (1908) and Weinberg (1908) showed that in a large randomly mating population these genotype frequencies remain the same from one generation to the next. To test whether the Hardy–Weinberg (HWE) equilibrium holds at a locus, one estimates the frequencies p and q from a sample of n individuals, using $\bar{p} = [2n(AA) + n(Aa)]/2n$, $\bar{q} = 1 - \bar{p}$. Under HWE, the expected genotype frequencies at a particular locus are obtained by substituting these estimates into the equilibrium distribution. Then the standard χ^2-test (Gillespie, 1998, pages 11–15) is conducted. When HWE does not hold, different genetic theories and settings typically predict either a decrease or increase in the number of homozygotes.

An analogous equilibrium distribution holds when there are k possible alleles at a locus and the appropriate χ^2-test is used. In the highly polymorphic (large k) situation, which is of interest in forensic applications (Evett and Weir, 1998), the accuracy of the χ^2-test in moderate sample sizes is questionable; while in studies of rare or endangered species, only small sample sizes are available (Hedrick, 2000, page 74). In the spirit of Fisher's exact test, Levene (1949) obtained an exact test for the number (h) of homozygotes that conditioned on the number of alleles of each of k types. The importance of the problem is reflected by the current literature developing more computer intensive exact procedures (Huber et al., 2006; Maurer, Melchinger and Frisch, 2007); however, Levene's exact test for HWE was the first. The original article also derived the large sample distribution of the statistic and considered the effect of misclassification of a small fraction of heterozygotes as homozygotes. Finally, Levene expressed the problem of finding the distribution of h in terms of card matching; similar analogies between exact tests for HWE and card shuffling problems are still used today (Weir, 1996, page 110).

A few years later, Levene (1953) developed the first theoretical model that examined the effects of spatial variation on fitness (Hedrick, 2000, page 161). During the 1920's Fisher and Haldane asked an important question: How is polymorphism maintained when selection is operating? When there are two alleles at a locus, natural selection should favor the allele (A) most related to survival and mating, so eventually all the entire population should become homozygotes (AA). As described by Pollak (2006), they demonstrated that each of the two alleles can have a substantial equilibrium frequency when heterozygotes are superior in viability to either homozygote and that a deleterious allele, d, can be maintained at a low equilibrium frequency due to recurrent mutation of the favored allele to d. Levene (1953) showed that two alleles could be maintained when a population inhabits K ecological niches, migrates between them, and selection varies among the niches, even if the viabilities of a heterozygote are between those of homozygotes in all K niches. In particular, a stable polymorphism can occur when the harmonic mean fitness of both homozygotes is less than that of the heterozygote. The basic approach taken by Levene (1953) is still used in modern texts (Hedrick, 2000, page 161), where references to developments incorporating genotypic-specific habitat selection, that is, individuals preferentially migrate to niches in which they have higher fitness (viability), are described. Recent developments are surveyed by Hedrick (2006) and Star, Stoffels and Spencer (2007) who investigate the levels of polymorphism in a model incorporating recurrent mutation and selection.

In 1960 Professor Howard Levene proposed a now classic test for the equality of the variances of k populations. The practical importance of Levene's (1960) article is demonstrated by the fact that it has been cited over 1000 times in the scientific literature. The goal of this paper is to discuss the scientific heritage of Professor Levene's contribution on both statistical methodology and its use in a wide variety of disciplines. Other procedures for testing the equality of variances have been surveyed by Boos and Brownie (2004).

Levene's (1960) original article was motivated by the k-sample problem. Before comparing the sample means, one should check that the underlying populations have a common variance. At the time, procedures that were easy to calculate were desired. Section 3 describes the proper use of Levene-type tests as a first stage test to select either the standard or Welch-modified k-sample ANOVA. With modern computers and software, nowadays one can use the Welch method in place of ANOVA, as it incurs only a small loss in power when the variances are equal.

Levene's test, however, remains very useful, as many scientific questions concern the variances of k populations, rather than their means or location parameters (centers). For example, to choose among several ways of delivering the same average dose of a drug, the one with least variability in the measured dose is preferred. When reviewing the applied literature, it became apparent that many alternative hypotheses were best described as a monotonic trend in the variances of the k populations; hence, a modification of Levene-type tests for this situation is proposed. The increased power of a trend test, which is directed at the alternative of interest, is illustrated by reanalyzing data from two published studies.

Levene-type tests have become very popular and are used in a wide variety of applications, for example, clinical data (Grissom, 2000), marine pollution (Johnson, Rice and Moles, 1998), species preservation (Neave et al., 2006), climate change and geology (Henriksen, 2003; Khan, Coulibaly and Dibike, 2006; Coulson and Joyce, 2006), animal science (Waldo and Goering, 1979; Schom and Kit, 1980), food quality (Francois et al., 2006), spherical distributions in astronomy (Fisher, 1986), regional differences of semen quality (Auger and Jouannet, 1997), business (Chang, Jain and Locke, 1995; Christie and Koch, 1997; Plourde and Watkins, 1998), auditing (Davis, 1996), studies of awards in civil cases (Saks et al., 1997; Robbennolt and Studebaker, 1999; Marti and Wissler, 2000; Greene et al., 2001), the analysis of data in actual legal cases (Tyler v. Unocal, 304 F.3d 379, 5th Cir. 2002), genetics and evolution (Mitchell-Olds and Rutledge, 1986; Giraud and Capy, 1996), toxicology (Mayhew, Comer and Stargel, 2003), psychology, education and speech (Flynn and Brockner, 2003; Cattaneo, Postma and Vechi, 2006; O'Neil, Penrod and Bornstein, 2003; Tabain, 2001), sports (Cumming and Hall, 2002) and even sex research (Hicks and Leitenberg, 2001; Hays et al., 2001).

The original tests along with subsequent modifications that improve the robustness of the test to non-normality of the underlying data, for example, Brown and Forsythe (1974), or improve the statistical performance in certain circumstances, for example, unequal sample sizes, are described in Section 1. Section 2 discusses Levene-type tests when the alternative is that the variances of the k-groups follow a monotonic trend. A modification of the statistic along the lines of the Cochran–Armitage trend test, used to analyze dose-response data, is described. The results of a small simulation study illustrate its increased power. Our results are consistent with the detailed investigations of Balakrishnan and Ma (1990) and Lim and Loh (1996) and collectively they provide extensive support for the use of robust Levene-type tests in practice. Section 3 describes the proper use of Levene-type tests as a first stage test to decide whether to analyze the data by the standard or Welch-modified k-sample ANOVA. While the two-stage method, using an appropriate size for a Levene-type preliminary test, remains valid, with modern day statistical software, in most situations one can use the Welch method, as it is only slightly less powerful than the standard test when the variances are equal. The use of Levene-type tests in the analysis of data arising in a wide variety of interesting applications is described in the penultimate section (Section 4). The paper concludes with a summary of recommended methods and a discussion of topics needing further research.

1. THE ORIGINAL TEST AND FURTHER ROBUST MODIFICATIONS

A basic problem in ANOVA is to determine whether k populations have a common mean μ. One has k random samples, $x_{i1}, \ldots, x_{in_i}$, of size n_i from each of k populations with respective means, μ_i, and variances σ_i^2, $i = 1, \ldots, k$. The standard F-test assumes that in each of the populations the variable studied has a common variance σ^2 and compares the between group mean square to the within group mean square (s_p^2), that is,

$$F = s_p^{-2} \sum_{i=1}^{k} (\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot})^2/(k-1), \tag{1}$$

where s_p^2 is the pooled variance, $\bar{x}_{i.}$ is the mean of the ith group, $\bar{x}_{..}$ is the grand mean and $N = \sum_{i=1}^{k} n_i$. It has long been known that the actual size of the test based on F may differ noticeably from the nominal size, for example, 0.05, when the groups have different variances (Sheffe, 1959, pages 351–358). This problem is quite serious when the variances are negatively correlated with the sample sizes (Krutchkoff, 1988; Weerhandi, 1995). Hence, it is important to develop methods for checking the validity of the equal variance assumption.

Bartlett (1937) proposed a statistic, M, for testing the equality of k population variances that is a function of the variances (s_i^2) of the ith group. Subsequently, Box (1953) showed that the sampling distribution of Bartlett's M is not robust to violations of the assumed normality of the underlying distributions. Box noted that Bartlett's procedure is more useful as a test of normality than as a test for equality of k group variances. Box and Anderson (1955) showed that the effect of normality depends on the kurtosis, $\gamma_2 = \mu_4/\mu_2^2$, the ratio of the fourth central moment of the underlying distribution to the square of the variance. Assuming the data from the k groups have the same distribution, the natural estimator of γ_2 is

$$\hat{\gamma}_2 = \frac{N\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^4}{[\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_{i\cdot})^2]^2}. \tag{2}$$

Multiplying Bartlett's M by $2/(\hat{\gamma}_2 - 1)$ yields a test statistic, B_3, which has an approximate χ^2-distribution with $(k-1)$ degrees of freedom. Notice that for normal data the expected value of the factor $2/(\hat{\gamma}_2 - 1)$ equals 1.0 and as the kurtosis increases above 3, it becomes *smaller*. The statistic B_3 is the form of the Box–Anderson test discussed by Miller (1986); see also Shorack (1969).

In the small samples often encountered in applications of ANOVA, the higher moments are quite variable, so a test that does not rely on the fourth sample moment is desirable. To appreciate the idea underlying the approach adopted by Levene, assume that the group means μ_i are known. To measure variance or spread, he considered various functions of $x_{ij}-\mu_i$, for example, $|x_{ij}-\mu_i|$ and $(x_{ij}-\mu_i)^2$. The expected value of $(x_{ij}-\mu_i)^2$ is σ_i^2, the variance of the ith group, while the expected value of $|x_{ij}-\mu_i|$ is the mean deviation from the mean, a well-known measure of spread related to a classical measure of income inequality due to Pietra (Gastwirth, 1972). Thus, if one knew the group means, one could apply the standard ANOVA statistic to $|x_{ij}-\mu_i|$ or $(x_{ij}-\mu_i)^2$.

Since the group means, μ_i, are typically unknown, Levene naturally used the sample group means, $\bar{x}_{i\cdot}$, in their places. Then $|x_{ij}-\bar{x}_{i\cdot}|$ or $(x_{ij}-\bar{x}_{i\cdot})^2$ are treated as independent, identically distributed, normal variables, and the usual ANOVA statistic is utilized. While neither $|x_{ij}-\bar{x}_{i\cdot}|$ nor $(x_{ij}-\bar{x}_{i\cdot})^2$ is normally distributed, Levene's approach takes advantage of the fact that classical ANOVA procedures for comparing means are robust to violations of the assumption that the data follow a normal distribution (Miller, 1968, page 80). Of course, Levene realized that $|x_{ij}-\bar{x}_{i\cdot}|$ and $(x_{ij}-\bar{x}_{i\cdot})^2$ are not independent within each group, as they are deviations from the group mean. However, he showed that the correlation is of the order $1/n_i^2$ and had the intuition that this small degree of dependence would not seriously effect the distribution of the F-statistic. After trying different functions of $(x_{ij}-\bar{x}_{i\cdot})$, for example, square, log etc., Levene proposed the final version of the test in the form of the classic ANOVA method applied to the absolute differences between each observation and the mean of its group $d_{ij} = |x_{ij}-\bar{x}_{i\cdot}|$, $i=1,\dots,k$, $j=1,\dots,n_i$. Since the d_{ij} are not normally distributed even when the original x_{ij} are, the resulting F-statistic,

$$F = \frac{N-k}{k-1}\frac{\sum_{i=1}^{k}(\bar{d}_{i\cdot}-\bar{d}_{\cdot\cdot})^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}n_i(d_{ij}-\bar{d}_{i\cdot})^2}, \tag{3}$$

is not exactly distributed as the usual F-statistic with $k-1$ and $N-k$ degrees of freedom. Levene (1960) showed by simulation that the usual F statistic provides a good approximation, especially at the cut-off values corresponding to the commonly used significance levels, $\alpha = 0.01$ and 0.05.

A natural way to increase the robustness of Levene's original statistic is to replace the group means in the definition of d_{ij} by a more robust estimator of location, for example, the median (Brown and Forsythe, 1974) (BFL test). Studies by Conover, Johnson and Johnson (1981) and Lim and Loh (1996) confirm that utilizing the absolute deviations of the observations from their group medians, rather than means, is preferable. Thus, the modern version of Levene's test uses the $z_{ij} = |x_{ij}-\hat{\mu}_i|$ in place of d_{ij} in (3), where $\hat{\mu}_i$ are robust estimators of μ_i.

In small samples, for example, when there are no more than 10 observations in each group, the level of the Levene test can be quite conservative when the group centers are estimated by their medians. The problem arises from the fact that for odd group sizes, one of the absolute deviations from the group median must equal 0; and for even sample sizes, two of the absolute deviations are equal as the group median is estimated by the average of the middle two observations. Thus, a bootstrap version was proposed by Boos and Brownie (1989) and shown to have improved power by Lim and Loh (1996). An alternative modification was suggested by Hines and Hines (2000). When the number of observations n_i in the ith group is odd, they propose to remove a structural zero z_{im} for $m = [n_i/2] + 1$ (here $[y]$ is the floor function of y); when n_i is even, then the two smallest and necessarily equal deviations $z_{i[n_i/2]}$ and $z_{i[n_i/2+1]}$ are replaced by one single value $\sqrt{2}z_{i[n_i/2]}$. The Hines–Hines (2000) procedure increases the variability of z_{ij}, reducing degrees of freedom by one for each group to compensate for the structural zeros as well as decreasing the Error Sum of Squares and Mean Squares in the Levene ANOVA table. As a result, this simple modification provides a test with size closer to the nominal one, especially in small samples. In addition, this usually provides a Levene-type test with increased power.

Several authors, Martin and Games (1977), O'Brien (1979), Keyes and Levy (1997) and O'Neil and Mathews (2000, 2002), examined the effect that unequal sample sizes create when the data follows a normal distribution and proposed appropriate correction factors. In the one-way ANOVA, under H_a, the variances of the observations σ_i^2 differ, implying that the expected values of the d_{ij} are given by

$$E(d_{ij}) = \sigma_i \sqrt{\frac{2}{\pi}\left(1 - \frac{1}{n_i}\right)}. \tag{4}$$

Notice that equation (4) implies that even under H_0, that is, when all groups have a common variance σ^2, the expected group averages differ. Thus, large differences in the sample sizes, n_i, may cause the original Levene test to reject the null hypothesis when it is true.

O'Brien (1979) and Keyes and Levy (1997) remove this design effect by replacing d_{ij} by $u_{ij} = d_{ij}/\sqrt{1 - 1/n_i}$, which have the same expected value and are proportional to the absolute values of the standardized residuals from the original ANOVA. Then one applies OLS ANOVA to the u_{ij}. O'Neill and Mathews (2000) obtained the covariance matrix of u_{ij} and created the appropriate weighted least squares estimates of the within group and between group variances of u_{ij} and obtained the corresponding F-test. When the n_i are equal, to n, they showed that the weighted F-statistic is a factor, m, times the OLS F-test. Furthermore, m tends to 1 as n increases. O'Neill and Mathews (2000) also obtained the corresponding multiplier when deviations from the group medians are used. Manly and Francis (2002) showed that when the significance level of the F-test was determined by randomization of the residuals of deviations from the sample medians, it was very robust to nonnormality and was less affected by modest differences in the n_i.

2. LEVENE-TYPE TESTS FOR A TREND IN THE GROUP VARIANCES

While reviewing the large number of studies applying Levene's test or the Brown–Forsythe modification, we noticed that the alternative hypothesis appropriate to the subject matter often indicated that the variances would follow a decreasing or increasing trend; for example, the groups might correspond to dose levels or could be classified by status on a monotonic scale. It is well known that tests directed at a specific alternative typically are more powerful in detecting a particular alternative (Agresti, 2002; Freidlin and Gastwirth, 2004). Often, under the alternative the k groups can be arranged so that their variances increase, that is, H_a is $\sigma_1 < \sigma_2 < \cdots < \sigma_k$. A number of procedures which employ the idea of regressing the sample variances of each group vs. some preselected scores or considering a particular contrast have been developed for this problem (Vincent, 1961; Chacko, 1963; Fujino, 1979; and Hines and Hines, 2000). Here we follow the simple linear regression approach in which scores $w_1 < w_2 < \cdots < w_k$ are assigned to each observation in the ith group ($i = 1, \ldots, k$). The expected value of the slope $\hat{\beta}$ (5) of the regression line relating the z_{ij} to the w_i is zero under the null hypothesis, but will be positive (negative) under the alternative that there is an increasing (decreasing) trend in the variances. The estimator $\hat{\beta}$ of β is given by

$$\hat{\beta} = \frac{\sum_{i=1}^{k} n_i(w_i - \bar{w})(\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})}{\sum_{i=1}^{k} n_i(w_i - \bar{w})^2}, \tag{5}$$

$$\bar{w} = \sum_{i=1}^{k} n_i w_i / N,$$

where $\bar{z}_{i\cdot}, i = 1, \ldots, k$, are the group means of z_{ij} and $\bar{z}_{\cdot\cdot}$ is the grand mean over $\bar{z}_{i\cdot}, i = 1, \ldots, k$. When the observations in each group come from a normal distribution, the null hypothesis that the group variances are equal implies that the mean deviations from the group means (or medians) also are equal. When the variances or other measure of spread are equal, $\hat{\beta}$ should be centered around zero, while under the alternative that the group variances increase $\hat{\beta}$ should be positive.

The expression for the slope $\hat{\beta}$ in (5) is analogous to the classic one degree of freedom test for the strength of linearity (Johnson and Leone, 1964, page 78) or the Cochran–Armitage trend test for binary data (Piegorsch and Bailar, 2005) and its numerator is like a covariance between the group centers $\bar{z}_{i.}$ and scores w_i. Hines and Hines (2000) show that using contrasts that reflect the alternative or suspected trend have higher power than the usual F-statistic (1) for homogeneity applied to the z_{ij}. Abelson and Tukey (1963) showed the linear scores are efficiency robust over a wide range of increasing trends, so they are commonly used. If the alternative hypothesis implies a specific nonlinear trend, one should use the corresponding values for w_i, for example, $w_i = i^2$ or $w_i = \sqrt{i}$. Roth (1983) and Neuhauser and Hothorn (2000) developed trend tests using order-restricted inference. These methods may be more powerful when the trend is monotonic but far from linear, they are not explored here. The increased power of Levene-type trend tests will be seen in Section 4 where we reanalyze data sets from two scientific studies.

Remark. If the true group centers are known, then the standardized Levene-type trend statistic asymptotically follows a standard normal distribution, as do results from Proposition 2.2 of Huber (1973), Theorem 1 of Arnold (1980) and Carroll and Schneider (1985). In practice, however, the "true" group centers are typically unknown and estimated from a sample of observations. In the one-sample setting Miller (1968) showed that Levene's original statistic, using absolute deviations from the group means, is asymptotically distribution-free only when the underlying distribution is symmetric; if the sample group median are employed, then the statistic is asymptotically distribution-free. The corresponding large sample result for k groups was proved by Carroll and Schneider (1985). Using the results of Carroll and Schneider (1985), Bickel (1975) and Carroll and Ruppert (1982), it can be shown that if the "true" group centers are unknown, then the size of Levene's trend statistic determined from its asymptotic distribution is correct only when the group location parameters are estimated by the group medians.

A small simulation study considering samples from normal and heavy-tailed symmetric distributions was conducted where a robust trimmed mean (Crow and Siddiqui, 1967; Gastwirth and Rubin, 1969; Andrews et al., 1972), the average of the middle 50% of the data, was also used to estimate the group centers. Our simulation study[1] indicates that for small and moderate sample sizes, the 25% trimmed versions of Levene's ($L_{0.25}$) trend tests yield the most accurate size for a test at the nominal 5% level for all the distributions (normal, exponential, t- and χ^2-distributions with 3 degrees of freedom) studied. In contrast, the corresponding test statistics using the sample means have levels exceeding the nominal 5%, especially for the heavy tailed and skewed distributions. Using medians, as in the Brown–Forsythe version, substantially underestimates the size of the test for small samples, especially for normal data. Overall, all the three versions of Levene's trend test, that is, the mean, median and 25% trimmed mean based, were more powerful against monotonic trend alternatives than the corresponding homogeneity tests, especially for small sample sizes. This is true even when the scores differ somewhat from the true trend, for example, the linear scores 1, 2, 3 are used when the ratios of the standard deviations are 1 : 3 : 5. As expected, in larger samples the difference in performance between Levene-type homogeneity and trend tests is minor.

3. USING LEVENE'S TEST AS THE FIRST STAGE IN ADAPTIVE ANOVA TESTS

In many applications adaptive procedures that utilize a preliminary test to choose the estimator or test for the final analysis improve the accuracy of the final inference (Hall and Padmanabhan, 1997; O'Gorman, 1997). For example, Hogg (1974) and Hogg, Randles and Fisher (1975) use a measure of tail-weight to select the estimator of the location parameter; Freidlin, Miao and Gastwirth (2003) use the p-value of the Shapiro–Wilk test to select a powerful nonparametric test for the analysis of paired differences. Miao and Gastwirth (2009) use the ratio of two measures of spread to choose the nonparametric test to analyze paired data for the second stage. These methods have been successful in the one-sample problem because heavy-tails can severely

[1] All calculations are performed using the R package *Lawstat* that is freely available from http://cran.r-project.org/.

affect the behavior of the sample mean and an appropriate preliminary test enables one to choose a robust estimator or test that has high efficiency across a class of distributions with tail weight close to that of the sample. Recently, Schucany and Ng (2006) noted that preliminary tests must be used with care, as at the second stage, the analysis is *conditional* on the results of the first-stage test. They demonstrated that graphical diagnostics for normality are preferable to a formal test of normality at the first stage when the objective is to make inferences about the population mean.

For testing the equality of k sample means, when the variances may not be equal, Welch (1951) provided the following modification of the usual ANOVA F-test:

$$F_W = \left(\sum_i w_i(\bar{x}_{i\cdot} - \hat{x})^2/(k-1)\right) \Big/ \left[1 + \frac{2(k-2)}{k^2-1} \times \sum_i \frac{1}{n_i - 1}\left(1 - \frac{w_i}{\sum_j w_j}\right)^2\right], \tag{6}$$

where $w_i = n_i/s_i^2$ and $\hat{x} = \sum w_i x_i / \sum w_i$.

This Welch modification rejects the null hypothesis of equal means if the F statistic (6) is larger than the critical value determined from an F distribution with degrees of freedom f_1^* and f_2^*, where

$$f_1^* = k - 1, \quad f_2^* = \left[\frac{3}{k^2-1}\sum_i \frac{1}{f_i}\left(1 - \frac{w_i}{\sum_j w_j}\right)^2\right]^{-1}. \tag{7}$$

When k is 2, the procedure reduces to the Welch 1938 two-sample t-test. Because the test using (6) allows for unequal variances, one needs to examine whether it incurs a noticeable loss of power when the group variances are equal. This section reports the results of a small simulation study that compares three tests: the usual ANOVA F-test, the Welch modification (6) and an adaptive ANOVA. The adaptive procedure is the following: first use a Levene-type test to see whether the variances are equal or not. If the test concludes that the variances are equal, use the ordinary ANOVA F-test, otherwise, use the Welch modification. The results indicate that just using the Welch method (6), which is now available on statistical packages, is easier to use than the adaptive ANOVA and only incurs a small loss in power when the variances are equal.

The study focused on testing whether the means from three *normal* distributions are equal. Following the recommendations of Bancroft (1964) and Huber (1972) that the level of a preliminary test should be greater than 5%, a level of 15% is used here.

Table 1 shows the observed level of the three tests for different sample sizes and different variance ratios. The nominal level is 5%. Clearly, the Welch adjusted ANOVA test and the adaptive procedure preserve the nominal levels very well for all sample sizes and variance ratios studied. These results are consistent with previous studies of the two-sample situation (Moser, Stevens and Matts, 1989, 1992; Weerhandi, 1995; Zimmerman, 2004 and Vangel, 2005). In contrast, the actual level of the ordinary ANOVA F test is affected when the variances are not equal. In some situations, the actual size of the test can be

TABLE 1
The actual sizes of a nominal 0.05 level test for the three procedures. The results are based on 10,000 simulations

	10, 10, 10			**10, 10, 20**		
$\boldsymbol{\sigma_1 : \sigma_2 : \sigma_3}$	**1 : 1 : 1**	**1 : 2 : 3**	**1 : 3 : 5**	**1 : 1 : 1**	**1 : 2 : 3**	**1 : 3 : 5**
ANOVA	0.0481	0.0665	0.0665	0.0512	0.0264	0.023
Welch ANOVA	0.0485	0.0518	0.053	0.0514	0.0524	0.0529
Adaptive ANOVA	0.0496	0.0572	0.0539	0.0546	0.0514	0.0529
	10, 20, 10			**20, 10, 10**		
$\boldsymbol{\sigma_1 : \sigma_2 : \sigma_3}$	**1 : 1 : 1**	**1 : 2 : 3**	**1 : 3 : 5**	**1 : 1 : 1**	**1 : 2 : 3**	**1 : 3 : 5**
ANOVA	0.0491	0.0714	0.0867	0.0542	0.1212	0.1399
Welch ANOVA	0.0494	0.0495	0.0524	0.0557	0.0506	0.0515
Adaptive ANOVA	0.0523	0.0554	0.0528	0.0572	0.0564	0.052

as large as 0.1399, for example, when $(n_1, n_2, n_3) = (20, 10, 10)$ and $(\sigma_1 : \sigma_2 : \sigma_3) = (1 : 3 : 5)$.

The powers of the adaptive and Welch ANOVA tests were also investigated by simulation. When the variances are equal, the powers of the adaptive procedure are about 2–3% higher than the Welch adjusted ANOVA F-test. When the variances are not equal, the Welch adjusted test has higher power, about 2–3% more than the adaptive one. Overall, the difference in power between the two procedures is quite small, rarely more than 0.02. (Detailed results can be obtained from the authors.) Thus, both the Welch method and the adaptive ANOVA are valid procedures.

The results reported in Table 1 use the group medians to estimate their centers, in the preliminary Levene-type test. Simulation studies, using the 25% trimmed means in place of the medians in the Levene test, yielded similar results. Other simulations explored the role of the size of the preliminary test. The findings indicate that the size of the first-stage test should be in the range 15% to 25% in order for the adaptive procedure to have the nominal size (0.05) and have reasonable power. These results confirm the recommended levels of 25% by Bancroft (1964) or 20% by Huber (1972, 1973) for the size of a preliminary test.

Both the Welch and the adaptive tests are more robust to departures from the equal variance assumption than the usual ANOVA F-test. These two tests are nearly as powerful as the standard F test when the group variances are equal. As the Welch test is simpler, we recommend it for general use. Researchers in areas where the two-stage method is commonly accepted, however, can still rely on it. The size of the Levene-type preliminary test should be between 15% and 25%.

4. THE WIDE APPLICABILITY OF LEVENE'S TEST AND ITS MODIFICATIONS

The important role statistical design, methodology and inference have in a wide array of intellectual disciplines is exemplified by the numerous applications of Levene-type tests. This section describes how Levene-type tests were used in a number of interesting studies from a variety of disciplines. In many cases the Levene-type test was used as a preliminary check of the equal variance assumption in classical ANOVA; in others, the scientific issue concerned the equality of the variances of measurements from k populations. The topics described were chosen from hundreds of valuable scientific contributions and illustrate the broad scientific impact of Professor Levene's method.

4.1 Applications in Archeology and Ethnography

Archaeologists are concerned with the effects increasing economic activity has on older civilizations. Economic growth encourages specialization in the production of goods, which led to the "standardization hypothesis," that is, increased production of an item would lead to its becoming more uniform. Kvamme, Stark and Longacre (1996) tested this theory on a type of earthenware, chupa-pots, from three Philippine communities that differ in the way they organize ceramic production. In Dangtalan, pottery is primarily made for household use and restricted exchange. Dalupa has an extensive nonmarket based barter economy, where part-time specialist potters trade their output for other goods. The village of Paradijon is near the Provincial capital; full-time pottery specialists sell their output to shopkeepers, located in the village or in the capital, for sale to the general public. To test the "standardization" hypothesis, these authors took measurements on three characteristics (aperture, circumference and height) of two-chalupa pots from the three areas and used the F-test and Brown–Forsythe version of Levene's test to compare the variation among pots produced in each area. The null hypothesis is that the variance or spread of each characteristic is the same in the three areas, while the alternative is that they differ.

After demonstrating that typically the measurements did not follow a normal distribution and had heavier tails, the authors showed (their Table 5) that the usual F-test can yield substantially different p-values than those obtained from Levene's test. For example, comparing the *circumference* of the 55 pots from Dangtalan with 170 from Dalupa, the standard F-test statistic yielded 1.24, leading to acceptance of the null hypothesis that variances are the same. In contrast, the robust Levene test yields a p-value $= 0.001$. Several other pair-wise comparisons showed that the F-test could yield much lower p-values than the robust Levene method. Here we apply the three Levene type tests for homogeneity of variances described in Section 2 to assess whether the variances of the *apertures* of the two-chalupa pots from the three locations are the same. All three tests, the original Levene's test (L), the Brown and Forsythe version (BFL) and the trimmed version

($L_{0.25}$), conclude that the variation in each of the three measured characteristics of the pots made in the regions are statistically significant. These results provide support for the standardization hypothesis.

The standardization hypothesis predicts that as economies develop, production intensifies, causing products to become more uniform or less variable. A test having high power for this particular alternative hypothesis, that is, the standard deviation of the three characteristics of the pots should *decrease* with *increasing* economic development, is preferable to a general test of homogeneity of the variances. Because the alternative hypothesis predicts that the variances of the three characteristics in pots from Dangtalan should be larger than those produced in Dalupa, which in turn should be larger than pots made in Paradijon, we analyze the data with the trend test (5).

To appreciate the increased power of the directed trend test, we analyzed the aperture data, kindly provided by Professor Kvamme. Using weights 1, 2 and 3 and deviations from the group means, mid-means and medians, respectively, in (5) yielded p-values 0.0001, 0.0004 and 0.0004 respectively. The estimates of the slope $\hat{\beta}$ were similar: -1.77, -1.68 and -1.81. All three p-values are less than one-half those obtained from the corresponding test of homogeneity and provide stronger evidence in favor of the "standardization hypothesis."

4.2 Applications in Environmental Sciences

Even before Katrina, ecologists studied the effect of hurricanes on forests, especially their rejuvenation after a severe storm. The catastrophic uprooting of trees creates mounds, pits and other micro-sites that provide possible locations for a particular species to regenerate. Carlton and Bazzaz (1998) simulated the effect of a hurricane by pulling down selected canopy trees and then measuring several important environmental resources (soil organic matter concentration, nitrogen transformation rates and the amount of CO_2) at five types of micro-sites that are created after a storm. These are as follows: mounds; pits; top sites, which are north facing forest floor surfaces; open sites, which are level and unshaded portions of the forest floor; and level portions of the forest floor that are covered by ferns or similar vegetation, called fern sites. For comparative purposes, measurements of the various resources were taken in a control area. Several questions were addressed, including: what were the residual effects of the disturbance on the average levels of key resources in the disturbed sites three years later? Did the simulated hurricane increase resource heterogeneity among the different micro-sites?

One-way ANOVA was used to test the differences in the average level of a resource among the five types of micro-sites. Samples of size five were taken from eight different micro-sites of each type. The authors applied the original version of Levene's test to check whether the variances of the measurements in the five groups were equal. When it indicated unequal variances, a single degree of freedom contrasts (SDFC) were used in lieu of ANOVA (Milliken and Johnson, 1984). When the homogeneity of variances assumption was satisfied and the ANOVA indicated significantly different effects among the micro-sites, a standard multiple comparison method for contrasts was utilized.

Due to nonhomogeneity of variance, Carlton and Bazzaz (1998) needed to use an SDFC to establish that the top sites were higher in soil organic matter than all other micro-sites, while percent soil water by mass was highest on fern, open and control sites. The standard ANOVA method was applicable to the data on climate factors. The CO_2 concentration was lowest on mounds. A major finding was that photon flux density (PFD), a measure of the amount of light level, on mounds, open sites and pits was *higher* than in the control (undisturbed) area. In contrast, the PFD on fern and top micro-sites was less than in the control area. The results suggest that hurricanes increase light levels immediately, which may encourage the growth of shade-intolerant species, while the change in the availability of various soil resources is more gradual. The authors carefully noted that their simulation cannot replicate all the features, for example, very high winds, of a real hurricane. Presumably, similar studies are underway in the areas most affected by the recent severe storms to assist in the regeneration of plant species.

4.3 Applications in Business and Economics

The problem of comparing k sample variances also arises in business and economics. Here, two applications of Levene's test in this area are briefly described, although there are many other interesting studies (Davis, 1996; Christie and Koch, 1997; Dhillon, Lasser and Watanbe, 1997; Chang, Pinegar and Schacter, 1997; Koissi, Shapiro and Hognas, 2006) that implemented the procedure.

Prior to the 1970s, the price of oil was less variable than that of other commodities; first due to the dominance of the major oil companies and later the formation of OPEC by the main countries producing it. To examine whether the behavior of oil prices changed in the 1980s and became more similar to that of other commodities, which tend to have large price fluctuations, Plourde and Watkins (1998) applied Levene's test to monthly price changes, measured by the logarithm of the ratio of the price in the current month to that of the previous month, in oil and other commodities (tin, zinc, wheat, etc.). After noticing that the monthly price changes of the two oil markets (West Texas and Brent) and the seven other commodities have high kurtosis, the authors realized that the usual assumption that the underlying populations all have the same shape or distribution and differ only in the scale parameter was implausible. Thus, they used both the Brown–Forsythe adaptation of Levene's test and the nonparametric Fligner–Killeen (1976) test in a series of pairwise comparisons to assess the relative dispersions of the price changes. In general, both tests showed that the monthly oil price changes were statistically significantly more dispersed than those of other commodities, except for lead and nickel, during the years 1985–1994. The modified Levene test did detect an increase in the dispersion of the price changes of zinc that the F–K test did not. This is consistent with the findings of Algina, Olejnik and Ocanto (1989), indicating that the O'Brien (1979) and BFL tests have relatively high power and preserve the nominal significance for the family of distributions and sample sizes they studied.

Stock market analysts and investors are interested in deciding whether various actions by companies assist them in predicting the future earnings and market prospects of those firms. Sant and Cowan (1994) studied the impact of an omission of a dividend by a company on the variability of both the forecasts of future earnings and the actual earnings. They compared the earnings and forecasts of companies that omitted a dividend during the period 1963–1984 by comparing the variances of the actual or forecasted earnings per share two years after the omission and two years before. Since the data was not normal, they utilized a robust Levene test (BFL). All comparisons showed that the variability of actual and forecasted earnings were significantly larger after the dividend omission. The authors also were careful to construct a control group of similar firms that did not omit a dividend. In a similar comparison, the earnings of these companies was not significantly greater in the later period. Because the increased earnings variability only occurred in the firms that omitted a dividend, their findings support the hypothesis that managers omit dividends when a firm's earnings become less predictable.

4.4 Applications in Medical Research

Since a cancer patient's probability of survival is increased when the disease is detected at an early stage, screening tests are an essential part of health care. Women over 50 typically have a mammogram every year or two. In many European nations, for example, the UK, mammograms tend to be evaluated at a few central locations, so each radiologist reviews many of them. In contrast, the system in the US is more decentralized, so there are fewer radiologists who assess a large number of mammograms. To study whether the accuracy of the mammogram is related to the volume a radiologist sees, Esserman et al. (2002) obtained a sample of 59 radiologists in the US and 194 high-volume radiologists in the UK The number of US radiologists in each volume category was 19 low (<100 per month), 22 medium (101–300) and 18 high (>300). Each radiologist was given a test set of 60 two-view films that contained 13 cancers.

In the disease screening context (Gastwirth, 1987; Pepe, 2003) accuracy is measured by both sensitivity (the probability a person with cancer is correctly identified) and specificity (the probability a healthy person is correctly classified). One can increase the sensitivity of a screening test by lowering the threshold level for classifying a subject as diseased, which decreases the corresponding specificity. A radiologist's accuracy is evaluated by their sensitivity at a specificity level of 0.90. Therefore, the authors fit an ROC curve (Gastwirth, 2001; Pepe, 2003) to the data for each radiologist using a variant of the binormal model (Dorfman and Berbaum, 2000). For the US radiologists, average sensitivity was 70.3% for those in the low-volume category, 69.7% for the medium volume group and 77% for readers of a high-volume of mammograms. High-volume UK radiologists had an average sensitivity of 79.3%. Because the BFL test indicated that the variances in the sensitivities of the radiologists in the groups were not equal, separate pairwise Welch-type t-tests were performed and showed that the differences among the average sensitivities were statistically significant. The area under the ROC curve (AROC) was used as a second measure of accuracy. The areas under the ROC curve ranged from an average of 0.832 for low-volume readers to 0.902 (0.891) for high volume UK (US) radiologists. Levene's test showed that the variances of the AROC in the four groups were statistically significant. Thus, Bonferroni adjusted pairwise comparisons were carried out and showed that the high volume radiologists were noticeably more accurate than the low and medium volume readers. Several related comparisons were conducted, which confirmed that the percentage of cancers detected by high volume radiologists significantly exceeded the corresponding percentage detected by lower volume radiologists. Their finding that higher volume improves diagnostic performance suggests that the quality and efficiency of screening programs can be improved by reorganizing them into more centralized high-volume centers.

Berger et al. (1999) utilized a database of 6026 echocardiograms that were read by one of three similarly qualified readers to assess the differences in frequency of several diagnoses and related measurements. The numbers of echocardiograms read by the readers (1, 2, 3) were 2702, 2101 and 1223, respectively. Levene's test was used to assess the variability in the measurements of several continuous characteristics, of which we discuss two: left atrial dimension (LAD) and left ventricle ejection fraction (LVEF). The median values of LAD for the three readers were as follows: 3.9, 3.9 and 3.8, respectively. The Kruskal–Wallis test (K–W test), however, showed that the three groups were significantly different, but the Median test did not detect any difference. Levene's test indicated statistically significant differences in the variability of LAD measurements made by the three doctors. Like the Wilcoxon test, the null distribution of the K–W test is affected by differences in the scale parameters or variances of the underlying distributions. The investigators may not have been aware of this issue and did not explore whether the differences among the variances of the three distributions would be sufficient to change the inference obtained from the usual K–W test.

The median values of the LVEF measurements made by the three readers were identical, 57.5 and Levene's test found no difference in their variability. A somewhat surprising statistically significant difference in location was found by both the Kruskal–Wallis and the Median tests. This might be due to the large, but unequal, sample sizes and/or the fact that the LVF measurements appear to be left-skewed, as the mean values of all three readers (52.7, 51.5 and 51.6) were less than the corresponding medians. The nonnormality and skewness of both data sets were indicated by Q–Q type plots. In contrast to the LVF data, the LAD measurements appear to be right skewed, with a fairly heavy right-tail.

A major finding was that the prevalence of mitral valve prolapse (MVP) differed in the three groups (5.3%, 3.0% and 4.8%), as did the recognition of clots (1.9% for reader 1 versus about 0.5% for readers 2 and 3). After checking that the individuals in the three groups had similar age and sex compositions, the authors noted that these differences would be difficult to detect in a typical small-scale reproducibility study. The data used in this study, as in many epidemiologic investigations, were observational, and not obtained from a randomized clinical trial. Thus, a sensitivity analysis based on generalizations of Cornfield's inequality (Rosenbaum, 2002) can be used to assess whether an omitted variable could explain the observed differences in the prevalence of heart problems found by the three readers. The article noted that some data was missing in a small proportion of cases but, given the large sample size, the authors decided not to impute those data. In this particular case, they are probably correct, however, from a statistical viewpoint it would be preferable for researchers to report the proportion of missing data. Then readers could assess whether it might affect the results. For example, the Kruskal–Wallis test of equality of the location parameters of the LVF measurements just reached statistical significance at the 0.05 level. If the proportion of missing measurements varied among the three readers, then the data would not be consistent with "missing at random" and the significance of the data might change with the method of imputation adopted.

An interesting study (Rosser, Murdoch and Cousens, 2004) demonstrated that a medical problem, optical defocus, increases the variability of the measurements of visual acuity. When visual acuity is repeatedly measured on the same person, the recorded scores can vary. This test-retest variability (TRV) is measured in units of the logarithm of the minimum angle of resolution (logMAR) and is a form of measurement error. Previous studies yielded estimates of the 95% range of TRV measurements between ± 0.07 to ± 0.19 logMAR. Following up on a conjecture that the length of the 95% TRV range might increase with the amount of defocus, these investigators examined 40 subjects under three conditions: no defocus or full refractive correction, full correction plus 0.50 D and full correction plus 1.00 D. The order of the six measurements given to a participant was randomized and no eye chart was used for consecutive measurements. When the same chart was used the patient was asked to read it forward one time and backward on the other. Thus, memory or learning as well as the potential effect of fatigue were controlled for in the experimental design. Following a common practice in ophthalmology of ignoring the matching, the authors applied the original Levene test of homogeneity of variances and obtained a significant result ($p = 0.00023$). The trend test using the group means yielded a more significant result ($p = 4.16 \times 10^{-5}$). Similarly, the trend test using group medians yielded a lower p-value than the test of homogeneity (0.00024 vs. 0.00124). As expected, the p-values obtained using the 25%-trimmed means of each group as their centers were in between those obtained using the mean and median. The smaller p-value of the trend test, which is directed at the alternative of interest, provides greater support for the conclusion that the variability of measured visual acuity increases with the degree of optical defocus than the test of homogeneity.

4.5 Applications in Legal Studies and Law Cases

In product liability and other tort cases, there is concern that monetary damages are not proportionate to the actual harm. Furthermore, individuals who contract the same illness after exposure to the same toxic product can receive very different monetary compensation from the legal system. Since the deliberations of actual jurors are confidential, researchers (Saks et al., 1997; Goodman, Green and Loftus, 1989; Robbennolt and Studebaker, 1999; Marti and Wissler, 2000) have varied the scenario described or the instructions given to mock jurors to evaluate whether the variability of awards for similar injuries can be reduced.

For example, Saks et al. (1997) explored the effect of giving jurors different types of information to guide their awards. Thus, some jurors were given no guidance (control), some the average award for the type of injury, some a range or interval of values, some both an interval and the average, and some were given some examples of awards in similar cases while some were given a cap or upper limit. These researchers also varied the severity of the injury. For low severity injuries, Levene's original test yielded a highly significant result $F_{(5,114)} = 11.5$, ($p < 0.001$). Significant variation also occurred in the medium and high injury categories. Somewhat unexpectedly, jurors given a cap had the most variable awards for low-level injuries. In the high-level category, the most variable conditions were the ones when no guidance or just the average award was provided to the mock jurors. Robbennolt and Studebaker (1999) explored the effect of varying the cap on punitive damage awards. Levene's test showed that the variability of those awards also increased with the size of the cap the mock jurors were given, however, the variability of the awards the control or no cap mock juries gave was less than those of mock juries given the highest cap ($50 million). These authors also showed that overall variability of jury awards was reduced when the awards for compensatory damages and punitive damages were made in two separate stages of jury deliberation.

The *Tyler v. Union Oil Co. of California* (304 F. 3d 379, 5th Cir. 2002) case concerned age discrimination in layoffs. First, plaintiffs' expert showed that recent job evaluations received by employees and their retention status were not significantly correlated. Then he compared the age distribution of the employees who were terminated to those who were retained in various locations of the firm. Levene's test was used to determine whether the usual t-test, which assumes the variances of the distributions are equal, or the Welch modified t-test is more appropriate. In most comparisons both versions of the t-test were significant. In one location, Ponville, the ages of 36 employees who were placed in a redeployment pool and eventually terminated were compared with the ages of 272 retained employees. Levene's test showed that the standard deviations (9.97 and 6.94) of the age distributions of the two groups were statistically significant. The usual t-test found the difference of three years between the average ages of the two groups significant (two-sided p-value is 0.024), while the modified t-test did not (two-sided p-value is 0.093). Surprisingly, the transcript of the expert testimony does not mention any questions by the defendant about the potential implication of the result that the age distributions of retained and laid-off employees were similar. Comparisons showing that the termination rates of employees aged 50 or more were higher than those of employees under 50, however, were quite significant ($p < 0.001$). This analysis provided very strong evidence supporting the finding of age discrimination.

4.6 Miscellaneous Applications

By the late 1990s researchers had documented geographical differences in semen quality, including sperm concentration, which raised questions about the possible causal roles of genetic differences and environmental factors. Since the criteria for recruiting study subjects, methods of laboratory analysis and experimental design differed among the earlier studies, to eliminate those factors as possible explanations for the basic finding, Auger and Jouannet (1997) conducted a retrospective study of candidate semen donors to sperm banks at University hospitals in eight regions of France during the period 1973–1993. These hospitals adopted the same guidelines for recruiting male semen donors and used similar laboratory methods. The authors analyzed data on seminal volume, sperm concentration, sperm count and the percentage of sperm that were motile. As the data were not normally distributed, they made appropriate transformations for each variable of interest, for example, the square root transform for sperm concentration and total sperm count. Levene's original test indicated that even the transformed data for all four variables had statistically significantly different variances. Hence, the authors used the Welch analog (6) of ANOVA to analyze the data. The results showed statistically significant differences among the eight regions in all four characteristics of semen quality (all p-values are less than 0.0001). While these small p-values arose in part because the total sample size was large (4710), varying from 226 in Caen to 1396 in Paris, the differences appear to be quite meaningful. For instance, the mean total sperm count varied from 284 per million in Toulouse to 409 per-million in Caen. The authors showed that these regional differences remained statistically significant after controlling for age, year of semen donation and number of days the subject abstained from sex prior to sample collection.

Sexual fantasies and their content can provide insight into the process of sexual arousal as well as gender differences in what people find exciting. As previous research indicated that men have more fantasies than women, Hicks and Leitenberg (2001) studied whether men and women differ in their likelihood of having sexual fantasies about their current partner as compared to extra-dyadic fantasies (about someone else) after controlling for the overall difference in number of fantasies. Using an anonymous questionnaire, they obtained 317 surveys from students (94% response rate) and 273 completed surveys (24% response rate) from faculty and staff at a mid-sized University. Eliminating a few cases with missing data, six outliers and 188 forms from individuals not currently in a relationship, they analyzed 349 responses (215 females, 134 males); apparently females had a higher response rate than males. Levene's test showed a significant gender difference in the variance of the number of fantasies, so the Welch modified t-test was used to compare the means. Men had a statistically significantly higher number of fantasies per month than women (76.7 vs. 34.1, $t_{192} = -4.77$). To control for this gender difference in total number of fantasies, the researchers calculated the percentage of each respondent's fantasies that were extra-dyadic. Since the variances of these percentages again differed by gender, the Welch t-test showed that men reported a greater number of sexual fantasies with an outsider than women (54% vs. 36%, $t_{311} = -5.1$). While only a modest percentage of extra-dyadic fantasies concerned former partners, on average, women had significantly more of them than men (34% vs. 22%, $p = 0.004$).

A regression analysis, adjusting for length of the relationship and whether one cheated on their partner, showed that the number of prior partners a person had was significantly more highly related to the percentage of extra-dyadic fantasies of women than men. The percentages of fantasies that involved someone other than their current partner was nearly identical for men and women who had cheated on their partner (55% vs. 53%), implying that the major difference between the genders in extra-dyadic fantasies occurs in faithful partners. Since the percentages of male and female respondents who admitted to having cheated on their current partner were nearly identical (28% vs. 29%), the previous finding is not likely to have been affected by nonresponse. For both sexes, the percentage of fantasies that were extra-dyadic increased with the length of the relationship. As most of the individuals in long-term relationships were faculty and staff rather than students, the subjects with a high degree of nonresponse, this last finding might require further confirmation. Since the overall regression had an R^2 of only 0.25, more research is needed to determine other explanatory factors as well as improving the accuracy of the recall data collected in similar studies.

5. DISCUSSION AND OPEN QUESTIONS

Levene's original article and the statistical procedures that developed and refined his original test enabled researchers in many intellectual disciplines to check the validity of an important assumption underlying the analysis of data obtained from studies using an ANOVA design. With modern day computer programs for calculation of statistical tests and estimators, the results in Section 3 show that today there is less need for a Levene-type test as a *preliminary* step to decide whether a standard or Welch-modified ANOVA test statistic should be applied, as the Welch procedure does not lose much power when the variances are equal. With an appropriate choice for the size of the Levene-type preliminary test, the two-stage procedure is valid and can be reliably used in disciplines where it has become a standard technique.

Levene's article and the subsequent literature have properly focused users of statistics on the need to examine whether their data "fit" the assumptions underlying the methods they apply. If one observes a "borderline" result, a Levene-type test may be used as one of the diagnostic tools to assess the sensitivity of the inference to potential violations of the basic assumptions. In particular, an analog of the Sprott and Farewell (1993) use of a confidence interval for the ratio, ρ^2, of both sample variances in the Behrens–Fisher problem to assess the sensitivity of inferences on the difference of the two means should be developed for the k-group setting. Using the ratios of the mean absolute deviations from a robust estimate of the group centers in place of the ratio of the sample variances may increase the applicability of this technique to data from heavier tailed distributions.

The Welch-modified t-test now appears in some standard textbooks and statistical packages. Since that procedure has been shown to be nearly as powerful as the standard one used in the equal variance setting and has much superior control of the Type I error when the group variances differ, authors of statistical textbooks should consider including it in their discussion of ANOVA. The main extra complications are the calculation of the denominator of the statistic (6) and the degrees of freedom (7), which are now readily carried out in statistical software. Since Levene-type tests for equal variance or a trend in variances are easy to describe and nearly as powerful as more complicated alternative procedures (Pan, 2002), these methods can now be included in statistics curriculum.

Reviewing the applied literature showed that comparing the variability of data from several groups frequently is the scientific question of interest. In particular, analysis of the variability of the measurements of medical characteristics obtained from different devices or techniques should lead to more reliable diagnosis. Quite often the problem of interest was whether there was a decreasing or increasing trend in the variability of the characteristic of interest that is associated with a covariate. This was the focus of articles from a variety of fields: the study relating characteristics of pots to the degree of economic development, the investigations of the relationship between the amount of information given to juries and the variability of the monetary damages they award, or the variability of eye examination measurements.

The simple test described in Section 2, along with related references, should be useful to researchers concerned with similar trend alternatives. For example, Kutner, Nachtsheim and Neter (2004) describe the use of the BFL two-sample test for checking the equality of variances of residuals from a time series regression against a time-trend alternative. It is likely that the power of such a test would be increased if more than two groups were formed and the trend test was applied. Further research is needed, as the appropriate number of groups is likely to depend on the total sample size as well as the magnitude of the trend.

The increased power of the test will also enable researchers to use smaller samples in those studies. Graubard and Korn (1987) noted that the choice of scores used in the Cochran–Armitage (CA) trend test in proportions is an important topic, as they can have a noticeable effect on the p-value of the test. Their point also applies to the trend test for variances. When there are several scientifically plausible choices for the weights, analogs of the efficiency robust methods (Zheng et al., 2003) developed for the CA test can be obtained, as the correlations of the test statistics based on each set of weights can be estimated from the data. These correlations are used in creating a suitable test statistic that has high power over the family of scientifically plausible models of the trend.

Although there exist several methods based on Levene-type statistics for studying differences in variability or the scale parameter of two variables measured on paired data (Wilcox, 1989; Grambsch, 1994), the visual acuity study (Rosser, Murdoch and Cousens, 2004) indicates that appropriate k-sample versions should be developed. A related problem occurs when the same technician assesses the same sample with several devices. This topic is related to tests for the equality of variance in randomized block designs. The survey of Schaalje and Despain (1996) found that when the block effect is mild, the method of Wilcox (1989) performs well. When the block effect is strong and the distributions are symmetric, a variant of Levene's test due to Yitnosumarto and O'Neill (1986) is recommended. Further research is needed for the situation of asymmetric or very heavy-tailed distributions.

Textbook discussions of ANOVA focus on comparing a relatively small number of treatments (groups) and the large sample theory is derived assuming that the numbers of observations in each group increase at the same rate. In some situations the number of treatments can also be large (Boos and Brownie, 1995). Bathke (2002, 2004) examines the effect of unequal variances in the multi-factor situation. In the commonly occurring two-factor design, when the number of levels of the first factor, A1, increases but the number of levels of the second, A2, remains finite, as long as the inequality in the error variances is not related to the level of factor A1, the F-test for the main effect of the first factor is almost unaffected by differences in the variances at the levels of the other factor. The tests for the main effect of factor A2 and interaction, however, are affected. A thorough analysis of tests of equality of variance when there are many treatments with a modest sized sample for each one remains to be done.

In most of the applications discussed here the observations in each group are independent random samples. It is well known (van Belle, 2002) that dependence can have a major effect on the distribution of many standard statistics. Thus, researchers will need to design their experiments and studies carefully to ensure that the observations in each group are independent of each other and those in other groups. This may not be a routine problem in studies where the same individuals and devices are used to make the measurements. More statistical procedures that model the dependence appropriately and incorporate it in the analysis need to be developed.

In several large studies we reviewed there was some nonresponse or missing data. In general, the potential effect of missing data on the conclusions of a study should be examined, as in English, Armstrong and Kricker (1998). In the study by Berger et al. (1999), only a small proportion of data was missing, which was unlikely to affect the conclusions. Nevertheless, researchers should be encouraged to report the pattern of missing data and any methods of imputation they adopted in the statistical analysis.

In contrast, the probability of nonresponse in the study of sexual fantasies (Hicks and Leitenberg, 2001) was highly correlated with age, a characteristic that is related to two independent variables in the regression predicting percentage of fantasies that were extradyadic. Thus, a study population containing a greater proportion of older respondents might yield different estimates of the effects of the number of prior partners and the length of current relationship, respectively. Since the slope of the regression relating the proportion of extradyadic fantasies to number of prior partners was stronger for women than for men, whether the nonresponse rates of older males and females differed should also be investigated. Given the recent development of imputation and other techniques for handling missing data (Little and Rubin, 2002; Molenberghs and Kenward, 2007), it would be useful to explore how they can be used in these applications to realistically assess the affect of missing data on the results of Levene-type tests, both for homogeneity and trend.

The number of observational, rather than designed, studies we encountered in the area of quality control or accuracy of medical measurements indicates the importance of developing methods for assessing the sensitivity of inferences based on tests of the equality of variance to an unobserved variable. Hopefully, this review will stimulate the development of methods analogous to those used to assess the potential impact of omitted variables on the comparison of the means or proportions from two samples (Rosenbaum, 2002) or in regression analysis (Dempster, 1988).

For cost-effectiveness many government sponsored surveys have a complex design based on stratified multistage probability cluster sampling, which produces estimates of population means and proportions with larger standard errors than would be obtained from a purely random sample of the same size (Nygard and Sandstrom, 1989; Korn and Graubard, 1999). Appropriate modifications of Levene-type tests for variance or measures of relative variability should be useful when the status of several sub-groups of the population is studied.

ACKNOWLEDGMENTS

The research of Professor Gastwirth was supported in part by NSF Grant SES-0317956. The research of Professor Gel was in part supported by a Grant from NSERC of Canada and was made possible by the facilities of SHARCNET.

REFERENCES

Abelson, R. P. and Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Ann. Math. Statist.* **34** 1347–1369. MR0156411

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, New York. MR1914507

Algina, J., Olejnik, S. and Ocanto, R. (1989). Type I error rates and power estimates for selected two-sample tests of scale. *Journal of Educational Statistics* **14** 373–383.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rodgers, W. H. and Tukey, J.W. (1972). *Robust Estimates of Location: Survey and Advances*. Princeton Univ. Press, Princeton, NJ. MR0331595

Arnold, S. F. (1980). Asymptotic validity of F tests for the ordinary linear model and the multiple correlation model. *J. Amer. Statist. Assoc.* **75** 890–894. MR0600972

Auger, J. and Jouannet, P. (1997). Evidence for regional differences of semen quality among fertile french men. *Human Reproduction* **12** 740–745.

Balakrishnan N. and Ma, C. W. (1990). A comparative study of various tests for the equality of two population variances. *J. Stat. Comput. Simul.* **35** 41–89.

Bancroft, T. A. (1964). Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance. *Biometrics* **20** 427–442. MR0181066

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proc. Roy. Soc. Ser. A* **160** 268–282.

Bathke, A. (2002). ANOVA for a large number of treatments. *Math. Methods Statist.* **11** 118–132. MR1900976

Bathke, A. (2004). The ANOVA F-test can still be used in some balanced designs with unequal variances and non-normal data. *J. Statist. Plann. Inference* **2** 413–422. MR2088750

Berger, A. K., Gottdiener, J. S., Yohe, M. A. and Guerro, J. L. (1999). Epidemiologic approach to quality assessment in echocardiographic diagnosis. *Journal of the American College of Cardiology* **34** 1831–1836.

Bickel, P. J (1975). One-step Huber estimates in the linear model. *J. Amer. Statist. Assoc.* **70** 428–434. MR0386168

Boos, D. D. and Brownie, C. (1989). Bootstrap methods for testing homogeneity of variances. *Technometrics* **31** 69–82. MR0997671

Boos, D. D. and Brownie, C. (1995). ANOVA and ranks test when the number of treatments is large. *Statist. Probab. Lett.* **23** 183–191.

Boos, D. D. and Brownie C. (2004). Comparing variances and other measures of dispersion. *Statist. Sci.* **19** 571–578. MR2185578

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika* **40** 318–335. MR0058937

Box, G. E. P. and Andersen, S. L. (1955). Permutation theory in the derivation of robust criteria and the study of departures from assumption. *J. Roy. Statist. Soc. Ser. B* **17** 1–26.

Brown, M. B. and Forsythe, A. B. (1974). Robust tests for equality of variances. *J. Amer. Statist. Assoc.* **69** 364–367.

Carlton, G. C. and Bazzaz, F. A. (1998). Resource congruence and forest regeneration following an experimental hurricane blowdown. *Ecology* **79** 1305–1319.

Carroll, R. J. and Ruppert, D. (1982). Robust estimation in heteroscedastic linear models. *Ann. Statist.* **10** 429–441. MR0653518

Carroll, R. J. and Schneider, H. (1985). A note on Levene's tests for equality of variances. *Statist. Probab. Lett.* **3** 191–194.

Cattaneo, Z., Postma, A and Vecchi, T. (2006). Gender differences in memory for object and word. *Quarterly Journal of Experimental Psychology* **59** 904–919.

Chacko, V. J. (1963). Testing homogeneity against ordered alternatives. *Ann. Math. Statist.* **34** 945–956. MR0150882

Chang, E. C., Jain, P. C. and Locke, P. R.(1995). Standard and Poors 500 index futrues volatility and price changes around the New York stock exchange close. *Journal of Business* **68** 61–84.

Chang, E. C., Pinegar, J. M. and Schacter B. (1997). Interday variations in volume, variance and participation of large speculators. *Journal of Banking and Finance* **21** 797–810.

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.

Crow, E. L. and Siddiqui, M. M. (1967). Robust estimation of location. *J. Amer. Statist. Assoc.* **62** 353–389. MR0212953

Coulson, D. and Joyce, L. (2006). Indexing variability: A case study with climate change impacts on ecosystems. *Ecological Indicators* **6** 749–769.

Christie, D. R. and Koch, T. W. (1997). The impact of market-specific public information on return variance in an illiquid market. *Journal of Futures Markets* **17** 887–908.

Cumming, J. and Hall, C. (2002). Athlete's use of imagery in the off-season. *Sport Pshychologist* **16** 160–172.

Davis, J. T. (1996). Experience and auditors' selection of relevant information for preliminary control risk assessments. *Auditing* **15** 16–37.

Dempster, A. P. (1988). Employment discrimination and statistical science. *Statist. Sci.* **3** 149–161. MR0968389

Dhillon, U. S., Lasser, D. J. and Watanbe, T. (1997). Volatility, information and double versus walrasian auction pricing in US and Japanese futures markets. *Journal of Banking and Finance* **21** 1045–1061.

Dorfman, D. D. and Berbaum, K. S. (2000). A contaminated binormal model for ROC data-part III: Initial evaluation with detection ROC data. *Academic Radiology* **7** 438–447.

English, D. R., Armstrong, B. K. and Kricker, A. (1998). Reproducibility of reported measurements of sun exposure in a case-control study. *Cancer, Epidemiology, Biomarkers and Prevention* **7** 857–863.

Esserman, L., Cowley, H., Eberle, C., Kirkpatrick, A., Chang S., Berbaum, K. and Gale, A. (2002). Improving the accuracy of mammography: Volume and outcome relationships. *Journal of the National Cancer Institute* **94** 369–375.

Evett, I. W. and Weir, B. S. (1998). *Interpreting DNA Evidence*. Sinauer, Sunderland, MA.

Fisher, N. I. (1986). Robust-tests for comparing the dispersions of several Fisher or Watson distributions on the sphere. *Geophysical Journal of the Royal Astronomical Society* **85** 563–572.

Fligner, M. A. and Killeen, T. J. (1976). Distribution-free two-sample tests for scale. *J. Amer. Statist. Assoc.* **71** 210–213. MR0400532

Flynn, F. J. and Brockner, J. (2003). It is different to give than to receive: Predictors of givers' and receivers' reactions to favor exchange. *Journal of Applied Psychology* **88** 1034–1045.

Francois, N., Guydot-Declerck, C., Hug, B., Callemien, D., Govaerts, B. and Collin, S. (2006). Beer astringency assessed by time-intensity and quantitative descriptive analysis: Influence of pH and accelerated aging. *Food Quality and Preference* **17** 445–452.

Freidlin, B. and Gastwirth, J. L. (2004). A note on the use of tests of mutation rates on ordered groups. *Genetic Testing* **8** 437–440.

Freidlin, B., Miao M. and Gastwirth, J. L. (2003). On the use of the Shapiro–Wilk test in two-stage adaptive inference for paired data from moderate to very heavy tailed distributions. *Biometrical Journal* **45** 887–900. MR2012347

Fujino, Y. (1979). Tests for the homogeneity of variances for ordered aternatives. *Biometrika* **66** 133–139. MR0529157

Gastwirth, J. L. and Rubin, H. (1969). On robust linear estimators. *Ann. Math. Statist.* **40** 24–39. MR0242329

Gastwirth, J. L. (1972). Robust estimation of the Lorenz curve and Gini index. *Rev. Econom. Statist.* **54** 306–316. MR0314429

Gastwirth, J. L. (1987), The statistical precision of medical screening procedures: Application to polygraph and AIDS antibodies test data. *Statist. Sci.* **2** 213–238. MR0920139

Gastwirth, J. L. (2001). Screening and selection. In *International Encyclopedia of Social Sciences* (N. J. Smelser and P. B. Bates, eds.). Elsevier, Oxford, U.K. 13755–13767.

Gillespie, J. H. (1998). *Population Genetics: A Concise Guide*. Johns Hopkins Univ. Press, Baltimore, MD.

Giraud, T. and Capy, P. (1996). Somatic activity of the mariner trasposable element in natural populations of Drosophila simulans. *Proceedings: Biological Sciences* **263** 1481–1486.

Grambsch, P. M. (1994). Simple robust tests for scale differences in paired data. *Biometrika* **81** 359–372. MR1294897

Goodman, J., Green, E. and Loftus, E. F. (1989). Runaway verdicts or reasonable determination: Mock juror strategies in awarding damages. *Jurimetrics Journal* **29** 285–309.

Grissom, R. J. (2000). Heterogeneity of variance in clinical data. *Journal of Consulting and Clinical Psychology* **68** 155–165.

Graubard, B. I. and Korn, E. L. (1987). Choice of column scores for testing independence in ordered $2 \times k$ contingency tables. *Biometrics* **43** 471–476. MR0897415

Greene, E., Coon, D. and Boornstein, B. (2001). The effects of limiting punitive damage awards. *Law and Human Behavior* **25** 217–234.

Hall, P. and Padmanabhan, A. R. (1997). Adaptive inference for the two-sample scale problem. *Technometrics* **39** 412–422. MR1482518

Hardy, G. H. (1908). Mendelian proportions in a mixed population. *Science* **28** 40–50.

Hedrick, P. W. (2000). *Genetics of Populations*, 2nd ed. Jones and Bartlett, Sudbury, MA.

Hedrick, P. W. (2006). Genetic polymorphism in heterogeneous environments: The age of genomics. *Ann. Rev. Ecol. Systems* **37** 67–93.

Henriksen, H. (2003). The role of some regional factors in the assessment of well yields from hard-rock aquifers of Fennoscandia. *Hydrogeology Journal* **11** 628–645.

Hays, M. A., Irsula, B., McMullen, S. L. and Feldblum, P. J. (2001). A comparison of three daily coital diary designs and a phone-in regimen. *Contraception* **63** 159–166.

Hicks, T. V. and Leitenberg, H. (2001). Sexual fantasies about one's partner versus someone else: Gender differences in incidence and frequency. *The Journal of Sex Research* **38** 43–50.

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.

Hogg, R. V. (1974). Adaptive robust procedures: A partial review and some suggestions for future applications and theory. *J. Amer. Statist. Assoc.* **69** 909–923. MR0461779

Hogg, R. V., Fisher, V. M. and Randles, R. H. (1975). A two-sample adaptive distribution-free test. *J. Amer. Statist. Assoc.* **70** 656–661.

Huber, P. J. (1972). Robust statistics: A review. *Ann. Math. Statist.* **43** 1041–1067. MR0314180

Huber, P. J. (1973). Robust regression: Asymptotic, conjectures and Monte Carlo. *Ann. Statist.* **1** 799–821. MR0356373

Huber, M, Chen, Y. G., Dinwoodie, I., Dobra, A. and Nicholas, M. (2006). Monte Carlo algorithms for Hardy–Weinberg proportions. *Biometrics* **62** 49–53. MR2226555

Johnson, S. W., Rice, S. D. and Moles, D. A. (1998). Effects of submarine mine tailings disposal on juvenile yellowfin sole (Pleuronectes asper): A laboratory study. *Marine Pollution Bulletin* **36** 278–287.

Johnson, N. L. and Leone, F. C. (1964). *Statistics and Experimental Design in Engineering and Physical Sciences*, 2nd ed. Wiley, New York. MR0172362

Kahn, M. S., Coulibaly, P. and Dibike, Y. (2006). Uncertainty analysis of statistical downscaling methods using canadian global climate predictors. *Hydrological Processes* **20** 3085–3104.

Keyes, T. K. and Levy, M. S. (1997). Analysis of Levene's test under design imbalance. *Journal of Educational and Behavioral Statistics* **22** 845–858.

Korn, E. L. and Graubard, B. I. (1999). *The Analysis of Health Surveys*. Wiley, New York.

Koissi, M. C., Shapiro, A. R. and Hognas, G. (2006). Evaluating and extending the Lee–Carter model for mortality forecasting: Bootstrap confidence interval. *Insurance Math. Econom.* **38** 1–20. MR2197300

Krutchkoff, R. G. (1988). One-way fixed effects analysis of variance when the variances may be unequal. *J. Stat. Comput. Simul.* **30** 259–183.

Kutner, M. H., Nachtsheim, C. J. and Neter, J. (2004). *Applied Regression Analysis*. McGraw-Hill/Irwin, Boston.

Kvamme, K. L., Stark, M. T. and Longacre, M. A. (1996). Alternative procedures for assessing standardization in ceramic assemblages. *American Antiquity* **61** 116–126.

Levene, H. (1949). On a matching problem arising in genetics. *Ann. Math. Statist.* **20** 91–94. MR0029149

Levene, H. (1953). Genetic equilibrium when more than one ecological niche is available. *American Naturalist* **87** 331–333.

Levene, H. (1960). Robust testes for equality of variances. In *Contributions to Probability and Statistics* (I. Olkin, ed.) 278–292. Stanford Univ. Press, Palo Alto, CA. MR0120709

Lim, T. S. and Loh, W. Y. (1996). A comparison of tests of equality of variances. *Comput. Statist. Data Anal.* **22** 287–301. MR1410388

Little, R. J. A. and Rubin, D. A. (2002). *Statistical Analysis with Missing Data*. Wiley, New York. MR1925014

Manly, B. F. J. and Francis, R. I. C. C. (2002). Testing for mean and variance differences with samples from distributions that may be non-normal with unequal variances. *J. Stat. Comput. Simul.* **72** 633–646. MR1930485

Marti, M. W. and Wissler, R. L. (2000). Be careful what you ask for: The effect of anchors on personal injury damages awards. *Journal of Experimental Psychology-Applied* **6** 91–103.

Martin, C. G. and Games, P. A. (1977). Tests for homogeneity of variance: Non-normality and unequal samples. *Journal of Educational Statistics* **2** 187–206.

Maurer, H. P., Melchinger, A. E. and Frisch, M. (2007). An incomplete enumeration algorithm for an exact test of Hardy–Weinberg proportions with multiple alleles. *Theoretical and Applied Genetics* **115** 393–398.

Mayhew, D. A., Comer, C. P. and Stargel, W. W. (2003). Food consumption and body weight changes with neotame, a new sweetener with intense taste: Differentiating effects of palatability from toxicity in dietary safety studies. *Regulatory Toxicology and Pharmacology* **38** 124–143.

Miao, W. and Gastwirth, J. L (2009). A new two stage adaptive nonparametric test for paired difference. *Statistics and Its Interface* **2** 213–221. MR2516072

Miller, R. G., Jr. (1968). Jacknifing variances. *Ann. Math. Statist.* **39** 567–582. MR0223001

Miller, R. G., Jr. (1986). *Beyond ANOVA: Basics of Applied Statistics*. Wiley, New York. MR0838087

Milliken, G. A. and Johnson, D. E. (1984). *Analysis of Messy Data*, Vol.1. Van Nostrand Reinhold, New York.

Mitchell-Olds, T. and Rutledge, J. J. (1986). Quantitative genetics in natural populations: A review of the theory. *The American Naturalist* **127** 379–402.

Molenberghs, G. and Kenward, M. G. (2007). *Missing Data in Clinical Studies*. Wiley, Chichester, UK.

Moser, B. K., Stevens, G. R. and Watts, C. L. (1989). The two-sample T-test versus Satterthwaite's approximation F-test. *Communication in Statistics—Theory and Methods* **18** 3963–3975. MR1058922

Moser, B. K., Stevens, G. R. and Watts, C. L. (1992). Homogeneity of variances in the two-sample means test. *Amer. Statist.* **46** 19–21.

Neave, F. B., Mandrak, N. E., Docker, M. F. and Noakes, D. L. (2006). Effects of preservation on pigmentation and length measurements in larval lampreys. *Journal of Fish Biology* **68** 991–1001.

Neuhauser, M. and Hothorn, L. A. (2000). Parametric location-scale and scale trend tests based on Levene's transformation. *Comput. Statist. Data Anal.* **33** 189–200.

Nygard, F. and Sandstrom, A. (1989). Income inequality measures based on sample surveys. *J. Econometrics* **42** 81–95.

O'Brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *J. Amer. Statist. Assoc.* **74** 877–880. MR0556482

O'Gorman, T. (1997). A comparison of an adaptive two-sample test to the t-test and the rank sum test. *Commun. Statist. Simulation and Comput.* **26** 1393–1411.

O'Neil, K. M., Penrod, S. D. and Bornstein, B. H. (2003). Web-based research: Methodological variables' effects on dropout and sample characteristics. *Behavior Research Methods Instruments and Computers* **35** 217–226.

O'Neil, M. E. and Mathews, K. L. (2000). A weighted least squares approach to Levene's test of homogeneity of variance. *Aust. N. Z. J. Stat.* **42** 81–100. MR1747464

O'Neil, M. E and Mathews, K. L. (2002). Levene tests of homogeneity of variance for general block and treatment designs. *Biometrics* **58** 216–2224. MR1891382

Pan, G. (2002). Confidence intervals for comparing two scale parameters based on levene statistics. *J. Nonparametr. Stat.* **14** 459–476. MR1919050

Piegorsch, W. W. and Bailer, A. J. (2005). *Analyzing Environmental Data*. Wiley, Chichester, UK.

Pepe, M. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Wiley, Chichester, UK. MR2260483

Plourdes, A. and Watkins, G. C. (1998). Crude oil prices between 1985 and 1994: How volatile in relation to other commodities? *Resource and Energy Economics* **20** 245–262.

Pollak, E. (2006). The influence of Levene's paper on polymorphism in subdivided populations. In *Proceedings of the Joint Statistical Meetings, August, 2006*. Amer. Statist. Assoc., Alexandria, VA.

Robbennolt, J. K. and Studebaker, C. A. (1999). Anchoring in the courtroom: The effect of caps on punitive damages. *Law and Human Behavior* **23** 353–373.

Rosenbaum, P. R. (2002). *Observational Studies*. Springer, New York. MR1899138

Rosser, D. A. Murdoch, I. E. and Cousens, S. N. (2004). The effect of optical defocus on the test-retest variability of visual acuity measurements. *Investigative Ophththalmology and Visual Science* **45** 1076–1079.

ROTH, A. J. (1983). Robust trend tests derived and simulated: Analogs of the Welch and Brown–Forsythe tests. *J. Amer. Statist. Assoc.* **78** 1972–1980. MR0727583

SAKS, M. J., HOLLINGER, L. A., WISSLER, R. L., EVANS, D. L. and HART, A. (1997). Reducing variability in civil jury awards. *Law and Human Behavior* **21** 243–256.

SANT, R. and COWAN, A. R. (1994). Do dividends signal earnings—the case of omitted dividends. *Journal of Banking and Finance* **18** 1113–1133.

SCHAALE, G. B. and DESPAIN, D. J. (1996). Robustness of variance tests for randomized complete block data. *Commun. Statist. Simulation* **25** 961–977.

SCHEFFE, H. (1959). *The Analysis of Variance*. Wiley, New York. MR0116429

SCHOM, C. B. and KIT, J. M. (1980). Genetic and environmental-control of avian embryos response to a teratogen. *Poultry Science* **59** 473–478.

SCHUCANY, W. R. and NG, H. K. T. (2006). Preliminary goodness of fit tests for normality do not validate the one-sample student *t*. *Comm. Statist.* **5** 2275–2286. MR2338931

SHORACK, G. R. (1969). Testing and estimating ratios of scale parameters. *J. Amer. Statist. Assoc.* **64** 999–1013.

SPROTT, D. A. and FAREWELL, V. T. (1993). The difference between two normal means. *Amer. Statist.* **47** 126–128.

STAR, B., STOFFELS, R. J. and SPENCER, H. G. (2007). Evolution of fitness and allele frequencies in a population with spatially heterogeneous selection pressures. *Genetics* **177** 1743–1751.

TABAIN, M. (2001). Variability in frictive production and spectra: Implications for the hyper- and hypo- and quantal theories of speech production. *Language and Speech* **44** 57–94.

VAN BELLE, G. (2002). *Statistical Rules of Thumb*. Wiley, New York. MR1886359

VANGEL, M. G. (2005). A numerical approach to the Behrens–Fisher problem. *J. Statist. Plann. Inference* **130** 341–350. MR2128012

VINCENT, S. E. (1961). A test of homogeneity for ordered variances. *J. Roy. Statist. Soc. Ser. B* **23** 195–206. MR0141190

WALDO, D. R. and GOERING, H. K. (1979). Insolubility of proteins in ruminant feeds by 4 methods. *Journal of Animal Science* **49** 1560–1568.

WEERHANDI, S. (1995). ANOVA under unequal error variances. *Biometrics* **51** 589–599.

WEINBERG, W. (1908). Uber den Nachweis der Vererbung beim Menschen. *Jaresh. Verein f. Vaterl. Naturk. In Wuttemberg* **64** 364–382.

WEIR, B. (1996). *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

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

WELCH, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika* **38** 330–336. MR0046617

WILCOX, R. R. (1989). Comparing the variances of dependent groups. *Psychometrika* **54** 305–315.

YITNOSUMARTO, S. and O'NEILL, M. E. (1986). On Levene's tests of variance homogeneity. *Aust. J. Statist.* **28** 230–241. MR0860468

ZHENG, G., FREIDLIN, B., LI, Z. and GASTWIRTH, J. L. (2003). Choice of scores in trend tests for case-control studies of candidate-gene associations. *Biometrical Journal* **45** 335–348. MR1973305

ZIMMERMAN, D. W. (2004). A note on preliminary tests of variances. *British J. Math. Statist. Psych.* **57** 173–181. MR2087822