A Parametric Bootstrap Approach for ANOVA with Unequal Variances: Fixed and Random Models

K. Krishnamoorthy, **Fei Lu**

Department of Mathematics, University of Louisiana at Lafayette
Lafayette, LA 70504, USA

Thomas Mathew[*]

Department of Mathematics and Statistics, University of Maryland
Baltimore, MD 21250, USA

Abstract

This article is about testing the equality of several normal means when the variances are unknown and arbitrary, i.e., the set up of the one-way ANOVA. Even though several tests are available in the literature, none of them perform well in terms of type I error probability under various sample size and parameter combinations. In fact, the type I errors can be highly inflated for some of the commonly used tests; a serious issue that appears to have been overlooked. We propose a parametric bootstrap (PB) approach and compare it with three existing location-scale invariant tests – the Welch test, James test and the generalized F (GF) test. The Type I error rates and powers of the tests are evaluated using Monte Carlo simulation. Our studies show that the PB test is the best among the four tests with respect to Type I error rates. The PB test performs very satisfactorily even for small samples while the Welch test and the GF test exhibit poor Type I error properties when the sample sizes are small and/or the number of means to be compared is moderate to large. The James test performs better than the Welch test and the GF test. It is also noted that the same tests can be used to test the significance of the random effect variance component in a one-way random model under unequal error variances. Such models are widely used to analyze data from inter-laboratory studies. The methods are illustrated using some examples.

Key words: Fixed effects; Generalized F test; Generalized p-value; Inter-laboratory studies; Random effects; Welch test.

[*]Tel: +1-410-455-2418; fax:+1-410-455-1066.

E-mail address: mathew@math.umbc.edu (T. Mathew)

1. Introduction

There has been a continuous interest in the problem of comparing several normal means when the variances are unknown and arbitrary. This problem, when only two normal means are involved, is referred to as the Behrens-Fisher problem, and it has been well addressed in the literature. Among the tests proposed for the Behrens-Fisher problem, Welch's (1947) *approximate degrees of freedom solution* is a popular one. Welch's test is based on Student's t distribution with degrees of freedom (df) depending not only on the sample sizes but also the sample variances. Nevertheless, this approach has been well accepted and commonly used in practical applications because of its simplicity and accuracy. Indeed many introductory level text books (e.g., Moore, 2003, p. 454) recommend the Welch test regardless of the variances being equal or unequal. This is mainly because the conventional approach, regarding the choice between the two-sample t test and the Welch test, is to test first the equality of variances, and if the equality is tenable then use the t test, otherwise use the Welch test. Type I error and power studies of this conventional approach by Moser, Stevens and Watts (1989) showed that the preliminary test about variance homogeneity is superfluous and Welch's approximate test was satisfactory for all parameter configurations. Another criticism about the conventional approach is the appropriateness of the usual variance ratio test; this test is heavily dependent on the normality assumption, and nonnormality and unequal variances cannot be separated with this test.

For the comparison of k normal means under unequal variances, there is no single-stage procedure that performs satisfactorily for all sample sizes and parameter configurations. Bishop and Dudewicz (1978) proposed an exact two-stage sampling procedure; however, this procedure is not well accepted in practice as it is not practical to require additional large samples in the second stage. The usual F test is based on the assumption of equal error variances, and its performance is satisfactory if there is a moderate departure from this assumption. However, the F test is liberal (Type I error rates are appreciably larger than the nominal level) when the sample sizes are negatively correlated with the variances and too conservative if they are positively correlated (Krutchkoff 1988 and Lee and Ahn 2003). Several authors proposed asymptotic solutions when the error variances are unknown and arbitrary. Among them, Welch's (1951) test, which is a generalization of the solution to the Behrens-Fisher problem, appears to be one of the first. Another early paper on the problem is James (1951), who derived a second order approximation to the distribution of a natural test statistic; the resulting test is referred to as James second-order test. Brown and Forsythe (1974), Rice and Gaines (1989), Weerahandi (1995a), Chen and Chen (1998) and Lee and Ahn (2003) proposed tests based on asymptotic or other approaches. Monte Carlo comparison studies by Gamage and Weerahandi (1998), Gerami and Zahedian (2001) and Lee and Ahn (2003) showed that, out of these and other tests, only Welch's test and Weerahandi's (1995a) generalized F test emerged satisfactory provided the sample sizes are moderate or large. Weerahandi (1995a) argued that his generalized F (GF) test is equivalent to the test given by Rice and Gaines (1989). A review study by Coombs, Algina and Oltman

(1996) pointed out that the James second-order test performs much better than the Welch test and the Brown-Forsythe test for small samples.

We have observed in the literature that the asymptotic procedures and other tests are evaluated for their validity for small k and/or moderate to large samples; some of the tests perform satisfactorily in this case. Extensive numerical results reported in Dajani (2002) show that none of the tests perform satisfactorily (in terms of type I error probability) when k is large and the sample sizes are small−a situation of practical importance in the context of the one-way random model, as explained below. In the following section, we describe the Welch test, the James second-order test, the generalized F test due to Weerahandi (1995a), and propose a parametric bootstrap (PB) test, for the one-way ANOVA model with fixed effects and unequal error variances. We chose the Welch test, the GF test and the James test as they are location-scale invariant, and perform better than other asymptotic tests for moderate k and large sample sizes. For $k = 2$, an approximation to the distribution of the PB pivot variable yields the Welch approximate df test. The methods are compared with respect to Type I error rates and power using Monte Carlo simulation. Comparison studies in Section 3 show that the PB test is the best among all the four tests.

Section 4 is on the one-way random model with unequal error variances, and we address the problem of testing the significance of the random effect variance component. In applications dealing with inter-laboratory studies, such a model is used to analyze the data. The primary problem of interest here is inference concerning the common mean; see Rukhin and Vangel (1998) and Vangel and Rukhin (1999) for more details and further references. However, it is of some interest to test the significance of the random effect variance component; if there is strong evidence to conclude that this variance component is insignificant, the data can be analyzed using the much simpler common mean model. It turns out that the tests used in the fixed effects model can also be used to test the significance of the inter-laboratory variance component in the case of the random effects model. This is noted in Section 4. The methods are illustrated using some examples in Section 5. Some concluding remarks are provided in Section 6.

Regarding the parametric bootstrap methodology that we have proposed here, note that the bootstrap can obviously be carried out both parametrically and nonparametrically. However, the problems addressed in this paper are in a strict parametric setting, namely the one-way fixed or random model with the usual normality assumptions, and heterogenous error variances. Thus we have chosen to do the bootstrap parametrically. If the model assumptions are at least approximately correct, Lee (1994) concludes that parametric bootstrap results may be more accurate than their nonparametric versions. Consequently, we have not considered the nonparametric bootstrap in this work.

2. Tests for the Fixed Effects Model

The one-way model under fixed effects correspond to k normal populations $N(\mu_i, \sigma_i^2)$, $i = 1, 2, \ldots, k$. Let $X_{i1}, \ldots, X_{in_i}$ be a random sample from $N(\mu_i, \sigma_i^2)$, and let $\bar{X}_i$ and S_i^2 denote the sample mean and sample variance, respectively. That is,

$$\bar{X}_i = \frac{1}{n_i} \sum_{i=1}^{n} X_{ij} \quad \text{and} \quad S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2, \quad i = 1, \ldots, k. \tag{1}$$

Let $\bar{\boldsymbol{X}} = (\bar{X}_1, \ldots, \bar{X}_k)'$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)'$, $\boldsymbol{S} = \text{diag}(S_1^2/n_1, \ldots, S_k^2/n_k)$ and $\Delta = \text{diag}(\sigma_1^2/n_1, \ldots, \sigma_{n_k}^2/n_k)$. The hypotheses of interest are

$$H_0 : \mu_1 = \ldots = \mu_k \quad \text{vs.} \quad H_a : \mu_i \neq \mu_j \quad \text{for some } i \neq j, \tag{2}$$

and we want to carry out a test using $\bar{\boldsymbol{X}}$ and $\boldsymbol{S}$.

If σ_i^2's are known, then a natural statistic for testing (2) is given by

$$
\begin{aligned}
T_N(\bar{X}_1, \ldots, \bar{X}_k; \sigma_1^2, \ldots, \sigma_k^2) &= \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \left[\bar{X}_i - \frac{\sum_{i=1}^{k} n_i \bar{X}_i/\sigma_i^2}{\sum_{i=1}^{k} n_i/\sigma_i^2} \right]^2 \\
&= \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \bar{X}_i^2 - \frac{\left[\sum_{i=1}^{k} n_i \bar{X}_i/\sigma_i^2 \right]^2}{\sum_{i=1}^{k} n_i/\sigma_i^2} \\
&= \bar{\boldsymbol{X}}' \Delta^{-\frac{1}{2}} \left(I_k - \frac{\Delta^{-\frac{1}{2}} \mathbf{1} \mathbf{1}' \Delta^{-\frac{1}{2}}}{\mathbf{1}' \Delta^{-1} \mathbf{1}} \right) \Delta^{-\frac{1}{2}} \bar{\boldsymbol{X}},
\end{aligned} \tag{3}
$$

where I_k is the identity matrix of order k and $\mathbf{1}$ is the $k \times 1$ vector of ones. Since $\Delta^{-\frac{1}{2}} \bar{\boldsymbol{X}} \sim N_k(\Delta^{-\frac{1}{2}} \boldsymbol{\mu}, I_k)$, and $B = \left(I - \frac{\Delta^{-\frac{1}{2}} \mathbf{1} \mathbf{1}' \Delta^{-\frac{1}{2}}}{\mathbf{1}' \Delta^{-1} \mathbf{1}} \right)$ is an idempotent matrix with rank $k - 1$, we have

$$\bar{\boldsymbol{X}}' \Delta^{-\frac{1}{2}} B \Delta^{-\frac{1}{2}} \bar{\boldsymbol{X}} \sim \chi_{k-1}^2 (\boldsymbol{\mu}' \Delta^{-\frac{1}{2}} B \Delta^{-\frac{1}{2}} \boldsymbol{\mu}),$$

where $\chi_m^2(\delta)$ denotes a noncentral chisquare random variable with degrees of freedom m and noncentrality parameter δ. (See Seber 1977, p. 37). The noncentrality parameter

$$\boldsymbol{\mu}' \Delta^{-\frac{1}{2}} B \Delta^{-\frac{1}{2}} \boldsymbol{\mu} = \sum_{i=1}^{k} \frac{n_i}{\sigma_i^2} \left[\mu_i - \frac{\sum_{i=1}^{k} n_i \mu_i/\sigma_i^2}{\sum_{i=1}^{k} n_i/\sigma_i^2} \right]^2,$$

and is equal to zero when $\mu_1 = \ldots = \mu_k$. Let $\bar{\boldsymbol{x}} = (\bar{x}_1, \ldots, \bar{x}_k)$ be the observed value of $\bar{\boldsymbol{X}}$. Then, the test that rejects H_0 in (2) whenever

$$T_N(\bar{x}_1, \ldots, \bar{x}_k; \sigma_1^2, \ldots, \sigma_k^2) > \chi_{k-1,\alpha}^2$$

is a size α test, where $\chi_{m,\alpha}^2$ is the upper αth quantile of a chi-square distribution with df $= m$.

In general, the population variances σ_i^2's are unknown; in this case, a test statistic can be obtained by replacing σ_i^2 in (3) by S_i^2, $i = 1, \ldots, k$, and is given by

$$T_N(\bar{X}_1, \ldots, \bar{X}_k; S_1^2, \ldots, S_k^2) = \sum_{i=1}^{k} \frac{n_i}{S_i^2} \bar{X}_i^2 - \frac{\left[\sum_{i=1}^{k} n_i \bar{X}_i / S_i^2\right]^2}{\sum_{i=1}^{k} n_i / S_i^2}. \tag{4}$$

In the following, we describe Welch's test, James' second-order test, the generalized F test due to Weerahandi (1995a) and the PB test.

2.1 Welch's test

Let $w_j = n_j / S_j^2$, $j = 1, \ldots, k$. Welch (1951) showed that

$$W^* = \frac{T_N(\bar{X}_1, \ldots, \bar{X}_k; S_1^2, \ldots, S_k^2)/(k-1)}{1 + \frac{2(k-2)}{k^2-1} \sum_{i=1}^{k} \frac{1}{n_i-1}\left(1 - \frac{w_i}{\sum w_j}\right)^2} \sim F_{k-1,f} \text{ approximately}, \tag{5}$$

where T_N is given in (4) $F_{r,s}$ denotes a random variable having an $F-$distribution with (r, s) degrees of freedom, and

$$f = \left[\frac{3}{k^2-1} \sum_{i=1}^{k} \frac{1}{n_i-1}\left(1 - \frac{w_i}{\sum w_j}\right)^2\right]^{-1}.$$

For a given level α, and an observed value w^* of W^*, this test rejects the H_0 in (2) whenever the p-value $P(F_{k-1,f_2} > w^*) < \alpha$.

2.2 James' test

James (1951) derived a second order approximation (that is, order of -2 in the df $n_i - 1$) to the distribution of the statistic $T_N(\bar{X}_1, \ldots, \bar{X}_k; S_1^2, \ldots, S_k^2)$. The critical value, which is a function of S_i^2's, based on the second order approximation can expressed as follows. Let

$$Q = \sum_{i=1}^{k} \frac{1}{n_i-1}\left(1 - \frac{w_i}{\sum_{j=1}^{k} w_j}\right)^2, \quad c_s = \frac{\left(\chi_{k-1,\alpha}^2\right)^s}{(k-1)(k+1)\cdots(k+2s-3)},$$

and

$$R_{st} = \sum_{i=1}^{k} \frac{1}{(n_i-1)^s}\left(\frac{w_i}{\sum_{j=1}^{k} w_j}\right)^t.$$

In terms of these quantities, the critical value

$$J_\alpha = \chi_{k-1,\alpha}^2 + \frac{1}{2}(3c_2 + c_1)Q + \left\{\frac{1}{16}(3c_2 + c_1)^2\left(1 - \frac{k-3}{\chi_{k-1,\alpha}^2}\right)Q^2\right.$$

$$+ \quad \frac{1}{2}(3c_2 + c_1)\Big[(8R_{23} - 10R_{22} + 4R_{21} - 6R_{12}^2 + 8R_{12}R_{11} - 4R_{11}^2)$$

$$+ \quad (2R_{23} - 4R_{22} + 2R_{21} - 2R_{12}^2 + 4R_{12}R_{11} - 2R_{11}^2)(c_1 - 1)$$

$$+ \quad \frac{1}{4}(-R_{12}^2 + 4R_{12}R_{11} - 2R_{12}R_{10} - 4R_{11}^2 + 4R_{11}R_{10} - R_{10}^2)(3c_2 - 2c_1 - 1)\Big]$$

$$+ \quad (R_{23} - 3R_{22} + 3R_{21} - R_{20})(5c_3 + 2c_2 + c_1)$$

$$+ \quad \frac{3}{16}(R_{12}^2 - 4R_{23} + 6R_{22} - 4R_{21} + R_{20})(35c_4 + 15c_3 + 9c_2 + 5c_1)$$

$$+ \quad \frac{1}{16}(-2R_{22} + 4R_{21} - R_{20} + 2R_{12}R_{10} - 4R_{11}R_{10} + R_{10}^2)(9c_4 - 3c_3 - 5c_2 - c_1)$$

$$+ \quad \frac{1}{4}(-R_{22} + R_{11}^2)(27c_4 + 3c_3 + c_2 + c_1)$$

$$+ \quad \frac{1}{4}(R_{23} - R_{12}R_{11})(45c_4 + 9c_3 + 7c_2 + 3c_1)\Big\} + O((n_i - 1)^{-3}). \tag{6}$$

This test rejects H_0 in (2) when $T_N(\bar{x}_1, \ldots, \bar{x}_k; s_1^2, \ldots, s_k^2) > J_\alpha$, where $T_N(\bar{x}_1, \ldots, \bar{x}_k; s_1^2, \ldots, s_k^2)$ is the observed value of $T_N(\bar{X}_1, \ldots, \bar{X}_k; S_1^2, \ldots, S_k^2)$.

2.3 The Generalized F (GF) Test

We shall now describe Weerahandi's (1995a) generalized F test. Let $V_i^2 = (n_i - 1)S_i^2$ and v_i^2 be an observed value of V_i^2, $i = 1, \ldots, k$. A *generalized test variable* is given by

$$GV = \frac{T_N(\bar{X}_1, \ldots, \bar{X}_k; \sigma_1^2, \ldots, \sigma_k^2)}{T_N(\bar{x}_1, \ldots, \bar{x}_k; v_1^2/U_1, \ldots, v_k^2/U_k)} = \frac{\sum_{i=1}^{k} \frac{n_i}{\sigma_i^2}\bar{X}_i^2 - \frac{\left[\sum_{i=1}^{k} n_i\bar{X}_i/\sigma_i^2\right]^2}{\sum_{i=1}^{k} n_i/\sigma_i^2}}{\sum_{i=1}^{k} \frac{n_iU_i}{v_i^2}\bar{x}_i^2 - \frac{\left[\sum_{i=1}^{k} \frac{n_iU_i}{v_i^2}\bar{x}_i\right]^2}{\sum_{i=1}^{k} \frac{n_iU_i}{v_i^2}}},$$

where $U_1, \ldots, U_k$ are independent random variables with $U_i \sim \chi_{n_i-1}^2$, $i = 1, \ldots, k$. Furthermore, $T_N(\bar{X}_1, \ldots, \bar{X}_k; \sigma_1^2, \ldots, \sigma_k^2) \sim \chi_{k-1}^2$ independently of $(U_1, \ldots, U_k)$. The "observed value" of GV is defined as the value of GV at $(\bar{X}_1, \ldots, \bar{X}_k; V_1^2, \ldots, V_k^2) = (\bar{x}_1, \ldots, \bar{x}_k; v_1^2, \ldots, v_k^2)$, and this observed value is 1. Therefore, for a given $(\bar{x}_1, \ldots, \bar{x}_k; v_1^2, \ldots, v_k^2)$, the *generalized p-value* is given by

$$P_{\chi_{k-1}^2, U_1, \ldots, U_k}\left(\frac{\chi_{k-1}^2}{T_N(\bar{x}_1, \ldots, \bar{x}_k; v_1^2/U_1, \ldots, v_k^2/U_k)} > 1\right). \tag{7}$$

The GF test rejects the null hypothesis in (2) whenever the generalized p-value in (7) is less than a given nominal level α. Notice that, for a given $(\bar{x}_1, \ldots, \bar{x}_k; v_1^2, \ldots, v_k^2)$, the probability in (7) does not depend on any unknown parameters, so it can be estimated using Monte Carlo simulation or computed using the integral expression given in Weerahandi (1995a).

For further details on the generalized p-value idea, along with a number of examples, we refer to Weerahandi (1995b).

It should be noted that the software XPro (Dataxiom Software Inc., Los Angeles, California, www.dataxiom.com) refers to the GF test as one of the exact parametric methods for the ANOVA, because the generalized p-value can be computed exactly as mentioned above. However, this generalized p-value does not always possess the properties of the usual p-value. In general, the distribution of the generalized p-value may not be uniform(0, 1), and it may depend on unknown parameters; see Weerahandi (1995b). Therefore, the generalized F test is not exact in the classical sense and its properties should be evaluated using Monte Carlo simulation.

2.4 The PB test

The parametric bootstrap involves sampling from the estimated models. That is, samples or sample statistics are generated from parametric models with the parameters replaced by their estimates. Recall that under $H_0 : \mu_1 = ... = \mu_k$ all $\bar{X}_i$'s have the same mean. As the test statistic T_N in (4) is location invariant, without loss of generality, we can take this common mean to be zero. Using these facts, the parametric bootstrap *pivot variable* can be developed as follows. Let $\bar{X}_{Bi} \sim N\left(0, \frac{S_i^2}{n_i}\right)$ and $S_{Bi}^2 \sim S_i^2 \chi_{n_i-1}^2/(n_i-1)$, $i = 1, \ldots, k$. Then the PB pivot variable based on the test statistic (4) is given by

$$\sum_{i=1}^{k} \frac{n_i}{S_{Bi}^2} \bar{X}_{Bi}^2 - \frac{\left[\sum_{i=1}^{k} n_i \bar{X}_{Bi}/S_{Bi}^2\right]^2}{\sum_{i=1}^{k} n_i/S_{Bi}^2}. \tag{8}$$

Noticing the fact that $\bar{X}_{Bi}$ is distributed as $Z_i \frac{S_i}{\sqrt{n_i}}$, where Z_i is a standard normal random variable, it can be easily verified that the PB pivot variable in (8) is distributed as

$$T_{NB}(Z_i, \chi_{n_i-1}^2; S_i^2) = \sum_{i=1}^{k} \frac{Z_i^2(n_i-1)}{\chi_{n_i-1}^2} - \frac{\left[\sum_{i=1}^{k} \frac{\sqrt{n_i}Z_i(n_i-1)}{S_i\chi_{n_i-1}^2}\right]^2}{\sum_{i=1}^{k} \frac{n_i(n_i-1)}{S_i^2\chi_{n_i-1}^2}}. \tag{9}$$

For a given $(s_1^2, \ldots, s_k^2)$ of $(S_1^2, \ldots, S_k^2)$ and level α, the PB test rejects H_0 in (2) when

$$P\left(T_{NB}(Z_i, \chi_{n_i-1}^2; s_i^2) > T_{N0}\right) < \alpha, \tag{10}$$

where T_{N0} is an observed value of T_N in (4). For fixed $(s_1, \ldots, s_k)$, the above probability does not depend on any unknown parameters, and so it can be estimated using Monte Carlo simulation given in Algorithm 1.

Algorithm 1

For a given $(n_1, ..., n_k)$, $(\bar{x}_1, ...\bar{x}_k)$ and $(s_1^2, ..., s_k^2)$:
compute T_N in (4) and call it T_{N0}

For $j = 1, m$

    generate $Z_i \sim N(0, 1)$ and $\chi^2_{n_i-1}$, $i = 1, ..., k$

    compute $T_{NB}(Z_i, \chi^2_{n_i-1}; s_i^2)$ using (9)

    if $T_{NB}(Z_i, \chi^2_{n_i-1}; s_i^2) > T_{N0}$, set $Q_j = 1$

(end loop)

$\frac{1}{m} \sum_{j=1}^{m} Q_j$ is a Monte Carlo estimate of the p-value in (10).

Remark 1. In general, it is not easy to find a useful approximation to the distribution of the PB variable in (9). However, for the case of $k = 2$, we can find a convenient approximation; in this case, it is easy to see that the PB variable in (9) is distributed as

$$\frac{Z^2}{p_1 \frac{\chi^2_{n_1-1}}{n_1-1} + p_2 \frac{\chi^2_{n_2-1}}{n_2-1}} = \frac{Z^2}{D(\chi^2_{n_1-1}, \chi^2_{n_2-1} | S_1^2, S_2^2)}, \tag{11}$$

where $p_1 = (S_1^2/n_1)/(S_1^2/n_1 + S_2^2/n_2)$, $p_2 = 1 - p_1$, Z is a standard normal random variable, and

$$D(\chi^2_{n_1-1}, \chi^2_{n_2-1} | S_1^2, S_2^2) = p_1 \frac{\chi^2_{n_1-1}}{n_1-1} + p_2 \frac{\chi^2_{n_2-1}}{n_2-1}.$$

For a fixed (S_1^2, S_2^2), we note that $D(\chi^2_{n_1-1}, \chi^2_{n_2-1} | S_1^2, S_2^2)$ is a linear combination of the two independent chisquares, namely, $\chi^2_{n_1-1}$ and $\chi^2_{n_2-1}$. Hence we approximate the distribution of $D(\chi^2_{n_1-1}, \chi^2_{n_2-1} | S_1^2, S_2^2)$ by the distribution of χ^2_d/d, using the moment matching method. As the first moments of these two variables are the same (equal to one), we solve for d by matching the variances, for fixed (S_1^2, S_2^2). Using the result that $\text{Var}(\chi^2_m) = 2m$, and solving the equation $\text{Var}(D(\chi^2_{n_1-1}, \chi^2_{n_2-1} | S_1^2, S_2^2)) = \text{Var}(\chi^2_d/d)$ for d, we get

$$d = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right)^2}{\frac{S_1^4}{n_1^2(n_1-1)} + \frac{S_2^4}{n_2^2(n_2-1)}}.$$

Notice that the d is the degrees of freedom given in the "approximate degrees of freedom test" for the Behrens-Fisher problem by Welch (1947). Using the chi-square approximation of D, we see that the PB variable in (11) follows an $F_{1,d}$ distribution, and so the PB test rejects the H_0 in (2) whenever

$$T_{N0} > F_{1,d,\alpha},$$

where T_{N0} is an observed value of T_N in (4), and $F_{m,n,\alpha}$ is the upper αth quantile of an $F_{m,n}$ distribution.

3. Type I Error and Power Properties

The Type I error rates of the ANOVA tests are estimated using Monte Carlo simulation. As we already mentioned, the tests that we consider are location-scale invariant, and so we can take, without loss of generality, that $\mu_1 = \ldots = \mu_k = 0$, $\sigma_1^2 = 1$ and $0 < \sigma_i^2 < 1$,

$i = 2, \ldots, k$, in our simulation studies. Thus the sample statistics $\bar{x}_i$ and s_i^2 will be generated independently as $\bar{x}_i \sim N(0, \sigma_i^2/n_i)$ and $s_i^2 \sim \sigma^2 \chi_{n_i-1}^2/(n_i - 1)$, with $0 < \sigma_i^2 < 1$, $i = 2, \ldots, k$.

To estimate the Type I error rates of the Welch test, we used simulation consisting of 100,000 runs for each of the sample size and parameter configurations. That is, for a given $(n_1, \ldots, n_k)$ and $(\sigma_1^2, \ldots, \sigma_k^2)$, we generated 100,000 W^*'s given in (5), and estimated the Type I error rates by the proportion of times W^* exceeded $F_{k-1, f_2, \alpha}$, where $F_{a, b, \alpha}$ denotes the upper αth quantile of an F distribution with degrees of freedoms a and b. Type I error rates of the James test are also estimated similarly. To estimate the Type I error rates of the GF and PB tests, we have used a two-step simulation. The Monte Carlo method used for estimating the Type I error rates of the PB test is as follows. For a given sample size and parameter configuration, we generated 2,500 observed vectors $(\bar{x}_1, \ldots, \bar{x}_k, s_1^2, \ldots, s_k^2)$, and the observed value T_{N0} in (10) was computed for each of the generated vectors. For each of the generated T_{N0}'s, we used 5,000 runs to estimate the p-value in (10). Finally, the Type I error rate of the PB test was estimated by the proportion of the 2,500 p-values that are less than the nominal level α. The Type I error rates of the GF test were similarly estimated.

In Table 1, we present the estimates of Type I error rates for $k = 2, 3, 6, 10$ and 20, and sample sizes ranging from very small to moderate. We observe the following from the numerical results in Table 1.

1. For $k = 2$, the Welch test, the James test and the PB test have similar Type I error rates, except in some cases ($n_1 = n_2 = 3$), where the Welch test appears to be very conservative. In the worst cases, the Type I error rates of both tests are around 0.06 when the nominal level is 0.05. The GF test seems to be very conservative for small samples.

2. Type I error rates of the Welch test, the James test and the PB test are similar for the $k = 3$ case. We again note that, even for small samples ($n_1 = 2, n_2 = 3, n_3 = 2$) the Type I error rates of these three tests never exceeded 0.065 whereas the Type I error rates of the James test goes as high as 0.10 when the nominal level is 0.05.

3. We see from the reported Type I error rates for $k = 6, 10$ and 20 that the PB test is the only test that controls the Type I errors satisfactorily. In particular, we see that the Type I errors of the GF test can be as large as 0.28 when $\alpha = 0.05$; this test, in general, appears to be liberal for moderate values of k. The Type I errors of the Welch test also exceed the nominal level considerably but not to the extent of the GF test. The James test performs superior to the Welch test and the GF test, and it is the second best among all four tests.

In Table 2, we provide the powers of the four tests for $k = 3$ and 10. We once again observe from this table that the PB test, James' test and Welch's tests control the Type I errors very well. All four tests exhibit similar power properties provided the Type I error rates are

close to each other. In one case, where $n_1 = n_2 = n_3 = 10$ and $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, .3, .9)$, the James test appears to be less powerful than other tests. In some cases, the GF test appears to be more powerful than other tests because of its inflated Type I error rates exceeding the intended level 0.05.

4. Tests for the Random Effects Model

As already pointed out, the one-way random model with heteroscedastic error variances is important in the modeling and analysis of inter-laboratory data. Let X_{ij} denote the jth measurement at the ith lab, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, k$. The model is

$$X_{ij} = \mu + \tau_i + e_{ij},$$

where $\tau_i \sim N(0, \sigma_\tau^2)$, $e_{ij} \sim N(0, \sigma_i^2)$ are all independent random variables. The quantities $\bar{X}_i$ and S_i^2 defined earlier form a set of sufficient statistics, having the distributions

$$\bar{X}_i \sim N\left(\mu, \sigma_\tau^2 + \frac{\sigma_i^2}{n_i}\right) \text{ and } S_i^2 \sim \sigma_i^2 \chi_{n_i-1}^2/(n_i - 1),$$

and all the random variables are independently distributed. The problem we shall address is that of testing

$$H_0 : \sigma_\tau^2 = 0 \text{ vs. } H_a : \sigma_\tau^2 > 0.$$

If H_0 is not rejected, we conclude that the lab effect is not significant.

We shall first note that the tests in Section 2, in the context of the fixed effects model, are also appropriate for testing the significance of σ_τ^2. For this, we shall use an observation in Dajani and Mathew (2003), which states that any nonnegative definite quadratic form in $\bar{\mathbf{X}} = (\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_k)'$ has a distribution that is stochastically increasing in σ_τ^2. It is easily seen that the test statistics corresponding to the Welch test, generalized F test, as well the PB test described in Section 2, are all positive definite quadratic forms in $\bar{\mathbf{X}}$, conditionally given the S_i^2's. In other words, conditionally given the S_i^2's, these test statistics are stochastically increasing in σ_τ^2; obviously, the same stochastic monotonicity holds unconditionally as well. So, the same tests can be used to test $H_0 : \sigma_\tau^2 = 0$. Furthermore, the null distributions (and hence the type I error probabilities of the corresponding tests) of the test statistics are the same as under the fixed effects model. Thus the parametric bootstrap test continues to be a satisfactory test in the random effects model as well.

5. Illustrative Examples

We shall illustrate the four tests using two examples. The summary statistics for the first example are taken from Example 2 of Weerahandi (1995) so that we can compare our results with those of Weerahandi, and understand the behavior of these tests for a small number of groups. In the second example, we illustrate the tests using inter-laboratory data on the

fiber content in apples obtained by nine laboratories; the summary statistics are taken from Vangel and Rukhin (1999).

Example 1. The data were generated by Weerahandi (1995) for comparing five treatments A, B, C, D and E, corresponding to five normal distributions with the common mean 10 and variances as provided in Table 3. Thus the null hypothesis $H_0 : \mu_1 = \ldots = \mu_5$ is true in this case. The sample sizes were chosen so that they are negatively correlated with the variances – the situation where the usual F test produces inflated type I error probabilities. The summary statistics in Table 3 of Weerahandi (1995) are reproduced here in our Table 3. We note that the values of s_i given in Table 3 of Weerahandi (1995) are computed using $s_i^2 = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2/n_i$ which is the maximum likelihood estimate of σ_i^2. The s_i^2 given in our Table 3 is the value of the usual unbiased estimate of the σ_i^2 as defined in (1).

The p-values are computed for comparing three treatments at a time, four treatments at a time, and finally for comparing all five treatments. Notice that we can not compute the p-values for the James test, and so we reported the values of the test statistic and the corresponding critical values $J_{.05}$ in (6) (recall that the critical values depend on s_i^2's). We observe from Table 3 that the p-values of the PB test and Welch's test are very close except for the case where all the five treatments are compared. But the p-values of the GF test are smaller than those of other two tests except for the case where the treatments A, B and C are compared. All the tests made correct decision of accepting the equality of the means which is true.

Example 2. This example is on an inter-laboratory study involving nine laboratories carried out by the Nutrient Composition Laboratory of the US Department of Agriculture (Li and Cardozo 1994). The objective was to validate a proposed simple nonenzymatic-gravimetric method for determining total dietary fiber in some foods. Six samples (apple, apricots, cabbage, carrots, onions and soy fiber) were sent in blind duplicates to the participating laboratories. The data on fiber in apples were analyzed by Vangel and Rukhin (1999), and the summary statistics are reproduced here in Table 4. We note that for this example, $k = 9$ and the number of measurements n_i made by the ith laboratory is 2, $i = 1, \ldots, 9$.

We applied all the four tests for testing equality of the laboratory means. The p-values of the PB test and the generalized F test were computed using simulations consisting of 100,000 runs. For the purpose of demonstration, we computed the p-values for comparing the first i laboratory means, $i = 2, \ldots, 9$, and the results are presented in Table 4. We observe first that the GF test produced the smallest p-values when $k \geq 4$. This is consistent with the simulation findings reported earlier that the GF test is too liberal for moderate k and small samples. At 5% level, the PB test rejects the null hypothesis of equality of means for $i = 8$ and $i = 9$, the Welch test and the James test reject the null hypothesis for $i = 7, 8$ and 9, and the GF test rejects the null hypothesis for $i \geq 5$.

6. Concluding Remarks

The available tests for the one-way ANOVA model with heteroscedastic error variances have serious type I error problems that have been overlooked; this has been pointed out by Dajani (2002). In this article, we have suggested the parametric bootstrap (PB) approach in order to arrive at a test procedure, and have compared the PB test with some of the existing tests−the Welch test, the generalized F−test, and the James (1951) second order test. For a range of choices of the sample size and parameter configurations, we have investigated the performance of the above tests using Monte Carlo simulation. In terms of controlling the Type I error rate, the overall conclusion is that the PB test is the only procedure that performs satisfactorily, regardless of the sample sizes, values of the error variances, and the number of means being compared. The James second order test is a close second. The tests developed for the one-way fixed model with heteroscedastic error variances are also applicable to the one-way random model with heteroscedastic error variances, when the problem is that of testing the significance of the random effect variance component. We would like to emphasize that care should be taken regarding the choice of the test, since the different tests can produce different conclusions, in terms of accepting or rejecting the null hypothesis−a point emphasized by Dajani (2002). One of our examples also demonstrate this point.

Acknowledgments

References

Bishop, D.J., Dudewicz, E.J., 1978. Exact analysis of variance with unequal variances: test procedures and tables. Technometrics 20, 419–430.

Brown, M.B., Forsythe, A.B., 1974. Small sample behavior of some statistics which test the equality of several means. Technometrics 16, 129–132.

Coombs, W.T., Algina, J., Oltman, D.O., 1996. Univariate and multivariate omnibus hypothesis tests selected to control Type I error rates when population variances are not necessarily equal. Rev. Educational Res. 66, 137–179.

Chen, S.Y., Chen, H., 1998. Single-stage analysis of variance under heteroscedasticity. Comm. Statist. Simulation and Computation 27, 641–666.

Dajani, A.N., 2002. Contributions to Statistical Inference for Some Fixed and Random Models. Ph.D. Dissertation submitted to the Department of Mathematics and Statistics, University of Maryland - Baltimore County.

Dajani, A.N., Mathew, T., 2003. Comparison of some tests in the one-way ANOVA with unequal error variances. ASA Proceedings of the Joint Statistical Meetings, 1149-1155.

Gamage, J., Weerahandi, S., 1998. Size performance of some tests in one-way ANOVA. Comm. Statist. Simulation and Computation 27, 625–640.

Gerami, A., Zahedian, A., 2001. Comparing the means of normal populations with unequal variances. Proc. 53rd Session of International Statistical Institute, Seoul, Korea.

James, G.S., 1951. The comparison of several groups of observations when the ratios of population variances are unknown. Biometrika 38, 324-329.

Kesselman, H.J., Wilcox, R.R., 1999. The improved Brown a nd Forsythe test for mean equality: some things can't be fixed. Comm. Statist. Simulation and computation 28, 687–698.

Krutchkoff, R. G. 1988. One-way fixed effects analysis of variance when the error variances may be unequal. Journal of Statistical Computation and Simulation 30, 259-271.

Lee, S., Ahn, C.H., 2003. Modified ANOVA for unequal variances. Comm. Statist. Simulation and Computation 32, 987–1004.

Lee, S.M.N., 1994. Optimal choice between parametric and nonparametric bootstrap estimates. Math. Proc. Cambridge Philosophical Society 115, 335-363.

Li, B.W., Cardozo, M.S., 1994. Determination of total dietary fiber in foods or products with little or no starch, nonenzymatic-gravimetric method: Collaborative study. J. Assoc. Analytical Chemists International 77, 687–689.

Mehrotra, D.V., 1997. Improving the Brown and Forsythe solution to the generalized Behrens-Fisher problem. Comm. Statist. Computation and Simulation 26, 1139–1145.

Moore, D.S., 2003. The Basic Practice of Statistics. W. H. Freeman, NY.

Moser, B.K., Stevens, G.R., Watts, C.L., 1989. The two-sample t test versus the Satterthwaite's approximate F test. Comm. Statist. Theory and Methods 18, 3963-3975.

Rice, W.R., Gaines, S.D., 1989. One-way analysis of variance with unequal variances. Proc. National Academy of Science 86, 8183–8184.

Rukhin, A.L., Vangel, M.G., 1998. Estimation of a common mean and weighted means statistics. J. Amer. Statist. Assoc. 93, 303-308.

Seber, G.A.F., 1977. Linear Regression Analysis. Wiley, NY.

Vangel, M.G., Rukhin, A.L., 1999. Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. Biometrics 55, 129-136.

Weerahandi, S., 1995a. ANOVA under unequal error variances. Biometrics 51, 589–599.

Weerahandi, S., 1995b. Exact Statistical Methods for Data Analysis. Springer-Verlag, NY.

Welch, B.L., 1947. The generalization of Student's problem when several different population variances are involved. Biometrika 34, 28–35.

Welch, B.L., 1951. On the comparison of several mean values: An alternative approach. Biometrika 38, 330–336.

Table 1
Monte Carlo estimates of type I error rates
$k = 2, \sigma_1^2 = 1$

σ_2^2	n = (3,3)				n = (5,5)				n = (8,8)				n = (4,8)			
	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J
.01	.07	.04	.06	.07	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.05	.06
.05	.06	.03	.06	.07	.06	.05	.05	.06	.05	.04	.05	.05	.05	.04	.06	.06
.10	.06	.02	.05	.06	.05	.04	.05	.05	.05	.04	.05	.05	.06	.04	.06	.06
.20	.05	.02	.05	.06	.05	.03	.05	.05	.05	.03	.05	.05	.06	.04	.06	.06
.30	.05	.01	.04	.05	.05	.03	.05	.05	.05	.04	.05	.05	.05	.04	.06	.06
.40	.04	.02	.04	.05	.05	.03	.05	.05	.05	.03	.05	.05	.06	.03	.06	.06
.50	.04	.01	.04	.05	.05	.03	.04	.05	.05	.03	.05	.05	.06	.03	.06	.06
.60	.04	.01	.04	.05	.04	.03	.05	.05	.05	.03	.05	.05	.06	.03	.06	.06
.70	.04	.01	.04	.05	.05	.02	.04	.05	.05	.03	.05	.05	.06	.04	.05	.06
.80	.04	.01	.04	.05	.05	.02	.04	.05	.05	.04	.05	.05	.05	.03	.05	.05
.90	.04	.01	.04	.05	.05	.03	.04	.05	.05	.04	.05	.05	.05	.03	.05	.05
1.00	.04	.01	.03	.04	.05	.02	.04	.05	.05	.04	.05	.05	.05	.03	.05	.05

$k = 3, \sigma_1^2 = 1$

(σ_2^2, σ_3^2)	n = (5,5,5)				n = (10,10,10)				n = (4,6,20)				n = (2,3,2)			
	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J
(1,1)	.05	.03	.05	.05	.05	.05	.05	.05	.05	.03	.06	.05	.03	.05	.04	.06
(1,.5)	.04	.04	.05	.05	.04	.04	.05	.05	.05	.02	.06	.06	.03	.06	.04	.06
(1,.1)	.05	.05	.05	.06	.05	.06	.05	.05	.06	.04	.06	.06	.04	.07	.05	.08
(.5,.5)	.05	.04	.05	.05	.05	.04	.05	.05	.06	.02	.06	.06	.04	.05	.04	.06
(.5,.7)	.04	.04	.05	.05	.05	.04	.05	.05	.06	.02	.06	.06	.04	.05	.04	.06
(.1,.1)	.05	.04	.05	.05	.05	.04	.05	.05	.06	.04	.06	.06	.05	.06	.06	.09
(.1,.9)	.05	.05	.05	.06	.05	.06	.05	.05	.06	.05	.06	.06	.05	.04	.06	.10
(.5,.9)	.05	.04	.05	.05	.05	.05	.05	.05	.06	.03	.06	.06	.03	.05	.04	.06
(.3,.9)	.04	.04	.05	.05	.05	.05	.05	.05	.06	.03	.06	.06	.04	.05	.04	.07
(.3,.6)	.05	.04	.05	.05	.05	.05	.05	.05	.05	.03	.06	.06	.04	.05	.04	.07
(.1,.3)	.05	.04	.05	.05	.05	.04	.05	.05	.06	.05	.06	.06	.04	.06	.06	.09
(.05,.05)	.04	.04	.05	.05	.05	.04	.05	.05	.06	.04	.06	.06	.06	.06	.07	.10

Table 1 (continued).
$k = 6$ and $\sigma_1^2 = 1$

$(\sigma_2^2,\ldots,\sigma_6^2)$	a				b				c				d			
	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J
$(1,1,1,1,1)$	.05	.08	.06	.05	.05	.07	.06	.05	.04	.08	.08	.06	.05	.08	.07	.06
$(.1,.1,.5,.5,.5)$	.05	.09	.07	.06	.05	.07	.05	.05	.05	.09	.08	.06	.05	.08	.07	.06
$(.1,.2,.3,.4,.5)$	.05	.09	.06	.06	.05	.06	.05	.05	.05	.09	.08	.06	.05	.07	.07	.06
$(.1,1.1,1,1)$	.05	.08	.07	.06	.05	.06	.06	.05	.05	.11	.08	.07	.06	.08	.07	.06
$(.2,.4,.4,.2, .1)$	.05	.08	.06	.06	.05	.06	.06	.05	.05	.10	.09	.07	.05	.07	.08	.06
$(.5,.5,.5,.5,1)$	.05	.08	.06	.06	.05	.06	.06	.05	.05	.10	.08	.07	.06	.09	.07	.06
$(.3,.9,.4,.7,.1)$	.05	.09	.07	.06	.05	.08	.06	.05	.05	.11	.09	.07	.05	.08	.07	.06
$(.01,.01,.06,.1,.1)$	.05	.10	.07	.06	.05	.07	.06	.05	.05	.09	.09	.07	.05	.07	.07	.06

a. $\mathbf{n} = (5,\ldots,5)$; b. $\mathbf{n} = (10,\ldots,10)$; c. $\mathbf{n} = (3,3,4,5,6,6)$; d. $\mathbf{n} = (4,8,12,24,30,40)$

$k = 10$ and $\sigma_1^2 = 1$

$(\sigma_2^2,\ldots,\sigma_{10}^2)$	e				f				g				h			
	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J	PB	GF	W	J
$(1,1,1,1,$ $1, 1,1,1,1)$	.05	.15	.08	.06	.05	.08	.05	.05	.04	.18	.11	.08	.05	.10	.09	.06
$(.1,.2,.3,.4,$ $.5,.6,.7,.8,.9)$	.04	.13	.09	.06	.05	.07	.06	.05	.04	.17	.11	.07	.05	.09	.08	.06
$(.1,.1,.2,.2,$ $.3,.3,.4,.4,.5)$	.05	.14	.08	.06	.05	.08	.05	.05	.05	.18	.11	.08	.05	.11	.09	.06
$(.1,.1,.1,.1,$ $.1,.2,.2,.2,.2)$	.05	.14	.08	.06	.05	.08	.05	.05	.04	.17	.12	.08	.05	.11	.09	.07
$(.1,1,.1,1,$ $.1,1,.1,1,.1)$	.04	.14	.09	.06	.05	.07	.06	.05	.04	.17	.13	.09	.05	.12	.10	.06
$(.3,.3,.3,.6,$ $.6,.6,.9,.9,.9)$	.05	.14	.08	.06	.05	.07	.05	.05	.05	.16	.11	.07	.05	.10	.09	.07
$(.1,.1,.1,.1,$ $.1,.1,.1,.1,.1)$	.04	.13	.08	.06	.05	.07	.05	.05	.04	.18	.12	.08	.06	.11	.09	.07

e. $\mathbf{n} = (5,\ldots,5)$; f. $\mathbf{n} = (15,\ldots,15)$;
g. $\mathbf{n} = (3,3,3,4,4,4,5,5,5,5)$; h. $\mathbf{n} = (4,4,4,12,12,12,15,15,15,15)$

$k = 20$, $\sigma_1^2 = 1$ and $n_1 = ... = n_{20} = 5$				
$(\sigma_2^2, \ldots, \sigma_{20}^2)$	PB	GF	W	J
(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)	.05	.26	.13	.07
(.1,.1,.2,.2,.3,.3,.4,.4,.5,.5,.6,.6,.7,.7,.8,.8,.9,.9,1)	.05	.28	.12	.08
(.1,.2,.3,.4,.5,.1,.2,.3,.4,.5,.1,.2,.3,.4,.5,.1,.2,.3,.4)	.05	.28	.13	.08
(.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1,.1)	.05	.27	.13	.08
(.2,.2,.2,.2,.4,.4,.4,.4,.6,.6,.6,.6,.8,.8,.8,.8,1,1,1)	.04	.28	.13	.08
(.9,.8,.7,.6,.5,.4,.3,.2,.1,.9,.8,.7,.6,.5,.4,.3,.2,.1,1,)	.05	.28	.13	.08
(.01,.01,.01,.05,.05,.05,.1,.1,.1,.5,.5,.5,.6,.6,.6,.8,.8,.8,.8)	.05	.28	.14	.08

PB – Parametric Bootstrap; GF – Generalized F test; W – Welch's test; J – James' test

Table 2
Powers of the tests
$k = 3, \sigma_1^2 = 1$ and $\mu_1 = 0$

n $= (10, 10, 10)$

(σ_2^2, σ_3^2)	Tests	(μ_2, μ_3)						
		$(0,0)$	$(0,.2)$	$(0,.5)$	$(0,.7)$	$(.5,1)$	$(0,1)$	$(1.5,1)$
$(.3,.9)$	PB	.05	.08	.25	.52	.45	.76	.94
	GF	.04	.08	.24	.51	.43	.75	.93
	W	.05	.08	.25	.51	.45	.76	.94
	J	.05	.07	.21	.37	.46	.67	.93
$(.1,.5)$	PB	.05	.10	.36	.64	.59	.91	.98
	GF	.05	.10	.35	.63	.60	.91	.98
	W	.05	.09	.36	.63	.59	.92	.98
	J	.05	.09	.36	.63	.58	.91	.98

n $= (10, 5, 15)$

(σ_2^2, σ_3^2)	Tests							
$(.3,.9)$	PB	.05	.08	.22	.42	.51	.73	.86
	GF	.05	.07	.21	.41	.52	.73	.84
	W	.05	.08	.23	.42	.51	.74	.86
	J	.05	.07	.23	.42	.51	.74	.86
$(.1,.5)$	PB	.05	.10	.41	.73	.69	.96	.95
	GF	.05	.09	.39	.71	.67	.95	.94
	W	.05	.10	.42	.71	.68	.96	.95
	J	.05	.10	.42	.71	.68	.96	.95

Table 2 (continued)

$k = 10$ and $(\mu_1, ..., \mu_8) = \mathbf{0}$,

$\mathbf{n} = (15, 15, 15, 20, 20, 20, 25, 25, 25, 25)$

$\boldsymbol{\sigma}^2$	Tests	$(0,0)$	$(0,.2)$	$(0,.5)$	$(0,.7)$	$(.5,1)$	$(0,1)$	$(1.5,1)$
							(μ_9, μ_{10})	
a	PB	.05	.08	.33	.63	.98	.93	1
	GF	.06	.11	.37	.67	.98	.94	1
	W	.05	.08	.33	.62	.98	.94	1
	J	.05	.08	.33	.62	.98	.93	1
b	PB	.05	.09	.41	.77	.98	1	1
	GF	.08	.12	.46	.81	.98	1	1
	W	.05	.09	.43	.77	.98	1	1
	J	.05	.09	.43	.76	.98	.98	1

$\mathbf{n} = (15,17,19,21,23,25,27,29,31,33)$

$\boldsymbol{\sigma}^2$	Tests	$(0,0)$	$(0,.2)$	$(0,.5)$	$(0,.7)$	$(.5,1)$	$(0,1)$	$(1.5,1)$
a	PB	.05	.08	.41	.77	1	.99	1
	GF	.08	.11	.47	.81	1	.99	1
	W	.05	.09	.43	.78	1	.99	1
	J	.05	.09	.43	.77	1	.99	1
b	PB	.05	.12	.54	.89	1	1	1
	GF	.06	.15	.59	.90	1	1	1
	W	.05	.11	.56	.89	1	1	1
	J	.05	.11	.56	.89	1	1	1

a. $(\sigma_1^2, ..., \sigma_{10}^2) = (1, .1, .2, .3, .4, .5, .6, .7, .8, .9)$; **b.** $(\sigma_1^2, ..., \sigma_{10}^2) = (1, .1, .1, .1, .3, .3, .3, .7, .7, .7)$

Table 3
Summary statistics and p-values of the tests for comparison of different
sets of treatments

Treat-ments	σ_i^2	n_i	$\bar{x}_i$	s_i^2	Treatments compared	PB	GF	W	James T_N	James $J_{.05}$
A	1	16	10.03	1.24	A, B, C	.380	.376	.378	2.18	7.91
B	4	12	9.57	3.97	A, C, E	.232	.184	.229	4.23	11.64
C	9	8	8.70	6.92	A, C, D, E	.252	.163	.239	6.07	15.12
D	16	6	7.92	13.39	A, B, D, E	.326	.270	.320	4.71	14.49
E	16	4	12.96	15.41	A, B, C, D, E	.326	.239	.310	6.36	16.97

Table 4
Dietary fibers in apples and p-values of the tests

Laboratory	$\bar{x}_i$	s_i	p-values*			James	
			PB	GF	W	T_N	$J_{.05}$
1	12.460	.028	—-	—-	—-		
2	13.035	.233	.154	.196	.173	12.00	29.39
3	12.440	.325	.310	.289	.268	12.02	55.90
4	12.870	.071	.147	.079	.085	67.67	71.26
5	13.420	.339	.189	.038	.078	81.66	92.53
6	12.080	.325	.243	.026	.080	85.40	112.3
7	13.180	.099	.182	.008	.035	167.8	127.6
8	14.335	.064	.037	.000	.001	1497	139.6
9	12.230	.212	.044	.000	.001	1512	156.4

* p-values in the ith row are for comparing
the first i laboratory means