Statistica Sinica **7**(1997), 631-646

TWO-WAY ANOVA WITH UNEQUAL CELL FREQUENCIES AND UNEQUAL VARIANCES

Malwane M. A. Ananda and Samaradasa Weerahandi

University of Nevada and Bell Communications Research

Abstract: In this article we consider the Two-Way ANOVA model with unequal cell frequencies without the assumption of equal error variances. By taking the generalized approach to finding p-values, classical F-tests for no interaction effects and equal main effects are extended under heteroscedasticity. The generalized F-tests developed in this article can be utilized in significance testing or in fixed level testing under the Neyman-Pearson theory. An example is given to illustrate the proposed test and to demonstrate its advantage over the classical F-test. A simulation study is carried out to demonstrate that, despite its increased power under heteroscedasticity, the size of the test does not exceed the intended level.

Key words and phrases: Heteroscedasticity, Two-Way ANOVA, unbalanced models.

1. Introduction

1.1. Motivation

The size of classical F-tests are fairly robust against the assumption of equal variances when the sample sizes are equal. When the sample sizes are different, the size of F-tests can substantially exceed the intended size. Most of all, they suffer from serious lack of power even under moderate heteroscedasticity. As we demonstrate in Section 4 the p-value suggested by a classical F-test can be as large as .3 when actually that data provides sufficient evidence to reject the underlying hypothesis at the .05 level. Lack of power of a test at this magnitude should be considered unacceptable and serious, especially in bio-medical research in which data are vital, expensive, and the data collection is time consuming.

As Krutchkoff (1988) pointed out, transformations cannot rectify the heteroscedasticity problem when the available data are already normal. If one attempts to tackle the problem by performing weighted least squares regression with estimated variances, the size of the test can become much larger than the intended level. The generalized p-value method provides a promising approach to solve such problems with no adverse effect on the size of the test. Until recently there were no Bayesian solutions to ANOVA problems either. Now the generalized p-value approach has lead to Bayesian solutions as well (see Weerahandi and Tsui (1996)). Since the assumption of equal variances is made for mathematical

tractability rather than anything else, then in view of power and size problems of the classical tests, there is a great need for encouraging further research in ANOVA, MANOVA, and ANCOVA problems. The purpose of this article is to undertake one such problem which often arises in practical applications.

1.2. Related work

For the Behrens-Fisher problem of comparing two normal populations, Barnard (1984) described how exact p-values can be computed without fiducial arguments. Tsui and Weerahandi (1989) obtained a numerically equivalent and computationally more efficient formula for the p-value and established that it is indeed the exact probability of a well defined extreme region. In a Bayesian treatment, Meng (1994) obtained a formula of the same form as the posterior predictive p-value under a noninformative prior. The generalized p-values and posterior predictive p-values have some implications in fixed level testing as well. Weerahandi (1995a) and Meng (1994) discuss some of these issues. The generalized inference methods have now been successfully applied to obtain exact tests in a variety of linear models (see, for instance, Weerahandi (1987), Thursby (1992), Griffiths and Judge (1992), and Koschat and Weerahandi (1992) for applications in regression, and Weerahandi (1991), and Zhou and Mathew (1994) for applications in mixed models).

The ANOVA problems pose new difficulties in extending results to complicated designs. According to our empirical studies, the assumption of equal variances is more serious in higher-way ANOVA than in one-way ANOVA. In the former, not only F-tests suffer from lack of power, but also they can lead to serious erroneous conclusions. Therefore, we consider further research in this direction to be very important and practically useful.

1.3. Tests in one-way ANOVA

Rice and Gains (1989) extended the argument given by Barnard (1984) to obtain an exact solution to the one-way ANOVA problem with unequal variances. Weerahandi (1995a) obtained a numerically equivalent form for the p-value which is closer in form to the classical F-test and formally proved that it is the exact probability of an unbiased extreme region (see Tsui and Weerahandi (1989)), a well defined subset of the underlying sample space with the observed sample on its boundary. Welch (1951) gave an approximate solution to the problem which works well with large samples. Krutchkoff (1988) provides a simulation based method of obtaining an approximate solution which is fairly good even with small samples. Krutchkoff (1988) and Weerahandi (1995a) also provide interesting examples to demonstrate the repercussions of applying classical F-tests when the problem of heteroscedasticity is serious.

1.4. Tests in two-way ANOVA

There is a large literature on two-way unbalanced models as well as balanced models. In a recent article Fujikoshi (1993) provided a good survey of previous work on two-way unbalanced models and discussed available solutions in a unified framework. We are aware of only a few attempts to tackle the two-way ANOVA problem under heteroscedasticity. The article of Krutchkoff (1989) which gives a simulation based method of obtaining an approximate test is of particular interest.

The purpose of this article is to develop an exact test of significance for the two-way unbalanced model with unequal variances when the unbalancedness arises due to unequal samples available from different factor combinations. As in Weerahandi (1987), we accomplish this by taking a generalized approach to constructing extreme regions. The proposed tests will be referred to as *generalized F-tests* as they are based on the form of the classical F-test for the case of known variances.

1.5. Exactness and unbiasedness of proposed tests

In significance testing of hypotheses in the two-way ANOVA model, each generalized F-test proposed in this article is exact in the sense that it is based on a p-value which is the exact probability of a well defined subset of the sample space (extreme region). The test is unbiased in the sense that the probability of the extreme region increases for any deviation from the null hypothesis. This means that if the probability of the extreme region is computed assuming that the null hypothesis is true it tends to be lower than it is supposed to be, thus resulting in smaller probabilities when the null hypothesis is not true.

It should be emphasized that, as in the case of testing parameters of discrete distributions, these assertions are not valid under the Neyman-Pearson fixed level testing. More precisely, probability of rejecting a null hypothesis if the p-value is less than α is not necessarily equal to α. As a matter of fact, exact fixed level tests based on minimal sufficient statistics do not exist for this type of problem. Nevertheless, the generalized F-tests developed in this paper can be utilized in fixed level testing as well. Our limited simulation study has suggested that rejecting a null hypothesis when the generalized p-value is less than α provides an excellent approximate α level test. Moreover, according to our simulation study, the actual size of the generalized F-test presented in this article does not exceed the intended level under homoscedasticity as well as under serious heteroscedasticity. According to other simulation studies reported in the literature (see, for instance, Thursby (1992), Weerahandi and Johnson (1992), and Zhou and Mathew (1994)), in many linear models, approximate tests based on generalized p-values often outperform more complicated approximate tests available in the

literature. Robinson (1976) established that the Behrens-Fisher solution to the two sample problem does not exceed the intended level. The result also applies to the Bayesian solution under the noninformative prior and to the generalized t-test of Tsui and Weerahandi (1989) because all three solutions are numerically equivalent. In view of that result it is expected that sizes of generalized F-tests in balanced linear models also do not exceed the intended level in most situations. Therefore, the generalized F-tests can be utilized as fixed-level tests with size not exceeding the intended level or as exact significance tests. It is conjectured that they are also Bayesian tests under appropriate discrepancy measures (see Meng (1994)) and noninformative priors. For a complete coverage and applications of the generalized p-values the reader is referred to Weerahandi (1995b).

2. Generalized F-test for No Interaction Effect

Consider the two-way ANOVA model with factors A and B, with factor levels $A_1, \ldots, A_I$ and $B_1, \ldots, B_J$, respectively giving a total of IJ factorial combinations or treatments. Suppose a random sample of size n_{ij} is available from ijth treatment, $i = 1, \ldots, I;\ j = 1, \ldots, J$. Let X_{ijk}, $i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, n_{ij}$ represent these random variables and x_{ijk} represent their observed (sample) values. Assume that $n_{ij} > 1$ so that sample variances can be computed for each cell of the design. Sample mean and the sample variance of the ijth treatment are denoted by $\bar{X}_{ij}$ and S^2_{ij}, $i = 1, \ldots, I;\ j = 1, \ldots, J$ respectively, that is,

$$\bar{X}_{ij} = \sum_{k=1}^{n_{ij}} X_{ijk}/n_{ij} \qquad \text{and} \qquad S^2_{ij} = \sum_{k=1}^{n_{ij}} \left(X_{ijk} - \bar{X}_{ij}\right)^2 /n_{ij}.$$

Their observed sample values are denoted by $\bar{x}_{ij}$ and s^2_{ij}, $i = 1, \ldots, I$; $j = 1, \ldots, J$ respectively. Consider the Two-Way ANOVA model with unequal variances:

$$X_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}, \tag{1}$$

$$\epsilon_{ijk} \sim N\left(0, \sigma^2_{ij}\right),\ i = 1, \ldots, I;\ j = 1, \ldots, J;\ k = 1, \ldots, n_{ij},$$

where μ is the general mean, α_i is an effect due to the ith level of the factor A, β_j is an effect due to the jth level of the factor B, and γ_{ij} represents an effect due to the interaction of the factor level A_i and the factor level B_j.

In order to have μ, α_i, β_j, and γ_{ij} uniquely defined, we need to have additional constraints. Let $w_1, \ldots, w_I$ and $v_1, \ldots, v_J$ be nonnegative weights (to be defined later to have additional properties) such that $\sum_{i=1}^{I} w_i > 0$ and $\sum_{j=1}^{J} v_j > 0$. We use the constraints

$$\sum_i w_i \alpha_i = 0, \sum_j v_j \beta_j = 0, \sum_{i=1}^{I} w_i \gamma_{ij} = 0, \sum_{j=1}^{J} v_j \gamma_{ij} = 0. \tag{2}$$

We are mainly interested in testing the following hypotheses

$$H_{0AB} : \gamma_{ij} = 0, i = 1, \ldots, I, j = 1, \ldots, J \tag{3}$$

$$H_{0A} : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \tag{4}$$

$$H_{0B} : \beta_1 = \beta_2 = \cdots = \beta_J = 0 \tag{5}$$

against their natural alternative hypotheses. Testing procedures for equal main effects will be considered in Section 3.

First consider testing the hypothesis H_{0AB} given in (3) for no interaction effect. Define the standardized interaction sum of squares and the error sum of squares

$$\tilde{S}_I\left(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}}{\sigma_{ij}^2}\left(\bar{X}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j\right)^2 \tag{6}$$

and

$$\tilde{S}_E\left(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \frac{1}{\sigma_{ij}^2}\left(X_{ijk} - \bar{X}_{ij}\right)^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} S_{ij}^2/\sigma_{ij}^2, \tag{7}$$

where $\hat{\mu}, \hat{\alpha}_i$, and $\hat{\beta}_j$ are solutions of μ, α_i, and β_j that minimize the quadratic equation

$$\tilde{S}\left(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \frac{1}{\sigma_{ij}^2}\left(X_{ijk} - \mu - \alpha_i - \beta_j\right)^2 \tag{8}$$

subject to the constraints given in equation (2). Also let us denote the observed value of $\tilde{S}_I$ as $\tilde{s}_I$. When variances are equal or when they are known parameters, testing procedure for two-way unbalanced design does not depend on chosen weights and the conventional F-test (Arnold (1981), Chap. 7) is based on the p-value

$$p = 1 - H_{(I-1)(J-1),(N-IJ)}\left[\frac{(N-IJ)\tilde{s}_I(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2)}{(I-1)(J-1)\tilde{s}_E}\right], \tag{9}$$

where $H_{k,l}$ is the cdf of the F-distribution with k and l degrees of freedom.

In order to present the formula for computing the p-value when the variances are unequal and unknown, define $IJ-1$ independent beta random variables B_{ij} as

$$B_{ij} \sim \text{Beta}\Big(\sum_{l=1}^{i-1}\sum_{m=1}^{J}(n_{lm}-1)/2 + \sum_{m=1}^{j}(n_{im}-1)/2, (n_{i,j+1}-1)/2\Big) \tag{10}$$
$$\text{if} \quad i = 1, \ldots, I; j = 1, \ldots, J-1$$

$$B_{ij} \sim \text{Beta}\Big(\sum_{l=1}^{i}\sum_{m=1}^{J}(n_{lm}-1)/2, (n_{i+1,1}-1)/2\Big)$$
$$\text{if} \quad i = 1, \ldots, I-1, \text{ and } j = J.$$

and a series of $Y_{ij},\ i = 1,\ldots,I,\ j = 1,\ldots,J$ random variables obtained by changing beta variables as

$$Y_{ij} = (1 - B_{i,j-1}) \prod_{j'=j}^{J} B_{ij'} \prod_{i'=i+1}^{I} \prod_{j'=1}^{J} B_{i'j'} \tag{11}$$

with $B_{1,0} = 0, B_{I,J} = 1$, and $B_{i,0} = B_{i-1,J}$ for $i > 0$. Then, the hypothesis H_{0AB} is tested on the basis of the p-value

$$p = 1 - E\Big\{H_{(I-1)(J-1),(N-IJ)}\Big[\frac{(N-IJ)}{(I-1)(J-1)}\tilde{s}_I\Big(\frac{n_{11}s_{11}^2}{Y_{11}}, \frac{n_{12}s_{12}^2}{Y_{12}}, \ldots, \frac{n_{IJ}s_{IJ}^2}{Y_{IJ}}\Big)\Big]\Big\}, \tag{12}$$

where $H_{k,l}$ is the cdf of the F-distribution with k and l degrees of freedom and the expectation is taken with respect to $Y_{ij}, i = 1,\ldots,I, j = 1,\ldots,J$ random variables.

We shall show in Section 5 that this is the exact probability of a well defined extreme region, an unbiased subset of the sample space with observed data on its boundary; that is a p-value the way Fisher treated problems of significance testing. The p-value serves to measure the evidence in favor of H_{0AB}. The proof of this test is given in Section 5. Using the extended definition of the p-values given in Tsui and Weerahandi (1989), this test is derived from the F-test when σ_{ij} values are known. In view of this fact, this test is referred to as the generalized F-test for two-way ANOVA. Practitioners who prefer to take the Neyman-Pearson approach and perform tests at a nominal level α can also find a good approximate test (with size not exceeding the intended level, according to our simulation study) by the following rule:

$$\text{reject} \quad H_0 \text{ if } p < \alpha.$$

This p-value can be either computed by numerical integration exact up to a desired level of accuracy or can be well approximated by a Monte Carlo method. When there are a large number of factor combinations the latter method is more desirable and computationally more efficient. In this method the expected value appearing in (12) is approximated by using a large number of data sets simulated from underlying beta random variables. Each of these data sets consist of a total of $IJ - 1$ independent beta random variates defined in (10). For each data set, the beta random numbers are transformed to Y_{ij} using formula (11), $\hat{\mu}, \hat{\alpha}$ are computed, $\hat{\beta}$ are evaluated, and then the cdf of $H_{(I-1)(J-1),(N-IJ)}$ is evaluated. Finally, the expected value is estimated by using their sample mean. The accuracy of this approximation can also be assessed. The Monte Carlo variance of the estimate of p is σ_h^2/L, where L is the number of Monte Carlo samples used to estimate p and σ_h is the sample (simulated) standard deviation

of H values. For example, with a probability of .999, the estimated p-value is accurate up to about $3\sigma_h/L^{1/2}$. This integrand is well behaved so that, when the number of factor levels is small, this p-value can also be computed by numerical integration. For instance, this method can easily be implemented in FORTRAN program using IMSL subroutine QAND to evaluate the integral. The p-value based on the Monte Carlo method has now been integrated into the XPro software package.

To calculate $\tilde{s}_I$, noting that it does not depend on the selected weights, let us pick the particular set of weights $w_i = u_{i.} = \sum_{j=1}^{J} u_{ij}$ and $v_j = u_{.j} = \sum_{i=1}^{I} u_{ij}$, with restrictions $\sum u_{i.}\alpha_i = 0$ and $\sum u_{.j}\beta_j = 0$, where $u_{ij} = n_{ij}/\sigma_{ij}^2$ or $u_{ij} = y_{ij}/s_{ij}^2$ according as the variances are known or unknown. Then $\hat{\mu}$, $\hat{\alpha}_i$ and $\hat{\beta}_j$ can be found by solving the system of equations

$$\begin{aligned} \sum_j u_{ij}(\bar{x}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0, \quad & i = 1, \ldots, I \\ \sum_i u_{ij}(\bar{x}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j) = 0, \quad & j = 1, \ldots, J. \end{aligned} \tag{13}$$

Then we have $\hat{\mu} = \sum\sum u_{ij}\bar{x}_{ij} / \sum\sum u_{ij}$. Moreover, the problem of solving the above system of equations can be somewhat simplified by eliminating $\hat{\beta}_j$'s from the first equation:

$$u_{i.}\hat{\alpha}_i - \sum_{k=1}^{I} \hat{\alpha}_k \left[\sum_{j=1}^{J} \frac{u_{ij}u_{kj}}{u_{.j}} \right] = b_i, \quad i = 1, \ldots, I,$$

where

$$b_i = \sum_{j=1}^{J} u_{ij}\bar{x}_{ij} - \sum_{j=1}^{J} \frac{u_{ij}}{u_{.j}} \sum_{k=1}^{I} u_{kj}\bar{x}_{kj}, \quad i = 1, \ldots, I.$$

This can be written in matrix form as $\hat{\mathbf{A}}\alpha = \mathbf{b} \Rightarrow \hat{\alpha} = \mathbf{A}^{-1}\mathbf{b}$ where $\mathbf{A} = \text{Diag}(u_{1.}, \ldots, u_{I.}) - \mathbf{C}$ and ikth element of $\mathbf{C}$ is $c_{ik} = \sum_{j=1}^{J} u_{ij}u_{kj}/u_{.j}$ and then replace the last row of the matrix $\mathbf{A}$ by $(u_{1.}, \ldots, u_{I.})$. Similarly $\hat{\beta} = \mathbf{B}^{-1}\mathbf{d}$ where $\mathbf{B} = \text{Diag}(u_{.1}, \ldots, u_{.J}) - \mathbf{C}$; jkth element of $\mathbf{C}$ is $c_{jk} = \sum_{i=1}^{I} u_{ij}u_{ik}/u_{i.}$ and then replace the last row of the matrix $\mathbf{B}$ by $(u_{.1}, \ldots, u_{.J})$; and the jth element of $\mathbf{d}$ is $d_j = \sum_{i=1}^{I} u_{ij}\bar{x}_{ij} - \sum_{i=1}^{I} u_{ij}/u_{i.} \sum_{k=1}^{J} u_{ik}\bar{x}_{ik}$.

3. Generalized F-test for the Main Effect

The literature on two-way unbalanced models provide several procedures available for testing the main effect (in the presence of possible interactions). There is no common agreement about the circumstances under which these alternative testing procedure should be used. The controversy is not about the

derivation of testing procedures, but about the appropriate weights. In many situations there are no natural weights to justify a particular procedure. Arnold (1981), p. 96-99 gives an excellent coverage of this problem and the controversies behind it. He provides five testing procedures in an attempt to tackle this problem. Probably the widely used method (see, for instance, Lindman (1992), Chap. 5) is the one discussed by Fujikoshi (1993) in detail. We will consider all six methods and give testing procedures for unbalanced design under heteroscedasticity using the generalized p-value approach.

Most of the procedures can be easily described by first considering the case of analyst specified weights to be used in the constraints. Let $w_1, \ldots, w_I$ and $v_1, \ldots, v_J$ be the prespecified weights. It is important to note that this testing procedure depends on the chosen weights and therefore the weights must be chosen prior to performing the experiment. Define the standardized sum of squares due to factor A

$$\tilde{S}_A\left(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{n_{ij}}{\sigma_{ij}^2}\left(\bar{X}_{ij} - \hat{\mu} - \hat{\beta}_j - \hat{\gamma}_{ij}\right)^2, \tag{14}$$

and it's observed value by $\tilde{s}_A\left(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2\right)$. Here $\hat{\mu}, \hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are the estimates of μ, β_j, and γ_{ij} obtained by minimizing

$$\tilde{S}_1\left(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2\right) = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{n_{ij}} \frac{1}{\sigma_{ij}^2}\left(X_{ijk} - \mu - \beta_j - \gamma_{ij}\right)^2 \tag{15}$$

subject to the constraints given in equation (2). In the case of known variances and known weights, testing of H_{0A} in the presence of interactions is based on the above chi-squared quantity (Arnold (1981), Chap. 7) and, is given by

$$p = 1 - H_{(I-1),(N-IJ)}\left[\frac{(N-IJ)\tilde{s}_A(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{IJ}^2)}{(I-1)\tilde{s}_E}\right]. \tag{16}$$

In the case of unknown variances (unequal) and unequal cell frequencies, the p-value of the hypothesis H_{0A} is given by

$$p = 1 - E\Big\{H_{(I-1),(N-IJ)}\Big[\frac{(N-IJ)}{(I-1)}\tilde{s}_A\Big(\frac{n_{11}s_{11}^2}{Y_{11}}, \frac{n_{12}s_{12}^2}{Y_{12}}, \ldots, \frac{n_{IJ}s_{IJ}^2}{Y_{IJ}}\Big)\Big]\Big\}, \tag{17}$$

where the expectation is taken with respect to the Y_{ij} random variables given in equation (12); $\hat{\mu}, \hat{\beta}_j$, and $\hat{\gamma}_{ij}$ are the solutions of μ, β_j, and γ_{ij} by minimizing the quantity $\tilde{S}_1\left(n_{11}s_{11}^2/Y_{11}, n_{12}s_{12}^2/Y_{12}, \ldots, n_{IJ}s_{IJ}^2/Y_{IJ}\right)$ given in (15) subject to the weights given in (2). The sketch of this proof is also given in Section 5. The p-value can be computed using the XPro software package.

Equation (17) enables us to compute p-values with any desired set of weights. The $\tilde{S}_1()$ function is determined by point estimates appearing in (14) which in turn depend on the weights. With particular weights this function can be determined explicitly and (17) can be written in more compact form. To do this, first consider the widely used procedure discussed by Fujikoshi (1993). This procedure is based on the nonnegative weights u_{ij} which satisfies the identifiability condition, $\sum_i u_{i.}\alpha_i = \sum_j u_{.j}\beta_j = \sum_i u_{ij}\gamma_{ij} = \sum_j u_{ij}\gamma_{ij} = 0$. In the presence of interactions when variances are equal, using $u_{ij} = n_{ij}$ he showed that the main effect can be tested using the sum of squares

$$s_a = \sum_i \sum_j n_{ij}(\bar{x}_{ij} - \bar{x}_{..})^2 - s_I - \sum_j n_{.j}(\bar{x}_{.j} - \bar{x}_{..})^2, \tag{18}$$

where s_I is calculated as in (6) with variances equal to 1, as appropriate in the case of equal variance assumption considered by Fujikoshi (1993). Here s_a^2 was referred to as the sum of squares due to the levels of factor A eliminating the levels of factor B. The last term of the expression (18) was referred to as the sum of squares due to levels of factor B ignoring the levels of factor A.

When variances are known and unequal, using $u_{ij} = n_{ij}/\sigma_{ij}^2$ sum of squares due to factor A given in the above equation is equivalent to

$$\tilde{s}_a(\sigma_{11}^2, \ldots, \sigma_{IJ}^2) = \sum_i \sum_j u_{ij}(\bar{x}_{ij} - \hat{\mu})^2 - \tilde{s}_I - \sum_j u_{.j}(\bar{x}_{.j} - \hat{\mu})^2,$$

where $\bar{x}_{.j} = \sum_{i=1}^I u_{ij}\bar{x}_{ij}/u_{.j}$ and $\tilde{s}_I$ is as given in (6).

When variances are unknown and unequal, the expression for the generalized p-value (4) is the same as in (17) except that $\tilde{s}_A$ must be replaced by $\tilde{s}_a$ and $\hat{\mu}, \hat{\alpha}_i$ and $\hat{\beta}_j$ are solutions of the system of equations in (13) with u_{ij} replaced by y_{ij}/s_{ij}^2. Therefore, the p-value of (4) under heteroscedasticity can be written as

$$p = 1 - E\Big\{H_{(I-1),(N-IJ)}\Big[\frac{(N-IJ)}{(I-1)}\sum_{j=1}^J\sum_{i=1}^I \frac{Y_{ij}}{s_{ij}^2}((\bar{x}_{ij} - \hat{\mu})^2 - (\bar{x}_{.j} - \hat{\mu})^2 - (\bar{x}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2)\Big]\Big\}, \tag{19}$$

where the expectation is taken with respect to the Y_{ij} random variables. It should be noted that Y_{ij} enter into the formula as a result of σ_{ij} in the weights being tackled by the random variables whose distribution depend on such parameters. The proof of this is very similar to the proof of (10) given in Section 5.

Another popular solution to the problem (Searle (1971), Chap. 7, Arnold (1981), Chap. 7) is not quite a test for $\alpha_i = 0$ in the presence of interactions, but rather to test the null hypothesis

$$H_{0A*}: \quad \alpha_i + \gamma_{ij} = 0 \qquad \text{for all } i, j \tag{20}$$

subject to the constraint $\sum_j \sum_i n_{ij}\beta_j/\sigma_{ij}^2 = 0$. Here the testing procedure does not depend on the chosen weights. When variances are known, the conventional F–test is based on the chi-squared statistic $\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\sigma_{ij}^{-2}(\bar{x}_{ij} - \bar{x}_{.j})^2$. When variances are unknown and unequal, the p-value of the hypothesis H_{0A*} in (20) is given by

$$p = 1 - E\Big\{H_{(I-1),(N-IJ)}\Big[\frac{(N-IJ)}{(I-1)}\sum_{i=1}^{I}\sum_{j=1}^{J}\frac{Y_{ij}}{s_{ij}^2}(\bar{x}_{ij} - \bar{x}_{.j})^2\Big]\Big\}, \tag{21}$$

where the expectation is taken with respect to the random variables defined in (12).

The third specific approach (Seber (1977), Chap. 9) is to assume $\gamma_{ij} = 0$ for all i, j if the no interaction effect hypothesis H_{0AB} in (3) is not rejected. Then the experiment is equivalent to two one-way experiments, one for factor A and the other one for factor B. Then the main effect for an unbalanced heteroscedastic design can be tested using the generalized p-value given by Weerahandi (1995a).

The fourth approach (Arnold (1981), Chap. 7) is to use uniform weights $w_i = v_j = 1$, considering all treatment levels are equally important. Then the p-value of (4) is as same as in equation (17) except the weights $w_i = v_j = 1$ must be used in the minimization process.

The fifth approach (Arnold (1981), Chap. 7, Scheffe (1959), Chap. 4) is to use weights $w_i = n_{i.}$ and $v_j = n_{.j}$. When variances are equal, these weights produce nice closed form solutions for the model with proportional sampling. Again here, the p-value of (4) is as same as in equation (17) except the above weights must be used in the process.

4. Examples

The objective of this section is to illustrate the proposed test and to demonstrate with the aid of two examples that it is worth while to resort to a numerically extensive testing procedure when the problem of heteroscedasticity is serious. The first example attempts to show how one can come to misleading conclusions or fail to detect significant differences in parameters as a result of ignoring the possibility of unequal variances implied by sample variances. This in turn can lead to very serious repercussion in many applications. Data collected in many applications are so vital and designed experiments can be so very expensive that such repercussions are unacceptable. The second example examines the size performance of the generalized F-test to demonstrate that under serious heteroscedasticity as in Example 1, one can draw conclusions based on generalized p-values in fixed level testing as well.

Example 1. First consider an example of a situation where the interaction between two factors is negligible. Suppose the first factor has 4 levels and the

second factor has 5 levels. Let the factors be A and B with factor levels A1, A2, A3, A4 and B1, B2, B3, B4, B5 respectively. Suppose that the design is balanced and that 7 observations are available from each factor combination. To demonstrate what can happen in a typical application with moderate heteroscedasticity, consider the sample means and sample variances given by the following table:

Table 1. Sample means and sample variances

	B1		B2		B3		B4		B5	
	$\bar{x}_{i1}$	s^2_{i1}	$\bar{x}_{i2}$	s^2_{i2}	$\bar{x}_{i3}$	s^2_{i3}	$\bar{x}_{i4}$	s^2_{i4}	$\bar{x}_{i5}$	s^2_{i5}
A1	5.1	0.61	4.9	1.7	4.2	1.5	3.7	3.4	3.8	1.7
A2	5.0	2.9	4.1	0.31	4.3	1.2	4.0	1.1	4.1	0.30
A3	4.9	1.1	5.1	0.31	5.0	1.2	5.0	1.0	3.9	3.4
A4	4.8	1.8	4.8	4.1	4.0	1.8	3.8	1.7	3.7	2.1

The null hypotheses of no different main effects and no interactions can be easily performed by classical F-tests under the assumption that cell variances are all equal. The elements of the F-tests are displayed in the following ANOVA table; the p-values appearing in the table are computed under the assumption that all cell variances are equal.

Table 2. ANOVA and p-values under the assumption of equal variances

Source	d.f.	SS	MS	F-Value	p-value
Inter. AxB	12	9.73	0.8108	0.4183	0.9539
Factor A	3	6.65	2.2167	1.1435	0.3344
Factor B	4	21.26	5.3165	2.7427	0.0317

With these p-values one would conclude that the differences between the levels of factor A as well as the interactions are not statistically significant. Moreover, the null hypothesis of equal effects of factor B levels would be rejected. Now let us drop the assumption of equal variances and retest the hypotheses. The p-value for testing the interactions is computed using Equation 10 and the p-values for testing the main effects are computed by applying the formula given by Equation 21. The p-values computed by these methods are as follows:

Table 3. P-values without the assumption of equal variances

Source	p-value
Interaction AxB	0.815
Factor A	0.033
Factor B	0.123

Now while we come to the same conclusion as far as the interactions are concerned, these p-values suggest that it is the levels of factor A, and not factor

B, that are significantly different. Since the significance of factor A differences are detected with milder assumptions the latter conclusion is more credible. The misleading conclusions made by applying the classical F-test in this example is a consequence of trying explain all statistical variations through the means. In many applications, this is considered more serious than the assumption of normal distributions. This example demonstrates the danger of resorting to simple procedures based on unreasonable assumptions. The reader is referred to Krutchkoff (1988, 1989) for further discussion of these issues.

Example 2. The purpose of this example is to investigate the size performance of the generalized p-value based on a simulation study. Since the test concerning H_{0AB} does not depend on chosen weights we confine our study to the problem of testing the interaction terms. In our limited study we carried out the simulation for a number of combinations of sample sizes n_{ij} and standard deviations σ_{ij} with population parameters: $\mu = 10.0$, $\alpha_1 = 0.0$, $\alpha_2 = 5.0$, $\beta_1 = 0.0$, $\beta_2 = 5.0$, $\beta_3 = 10.0$. When H_{0AB} is true, we also have $r_{11} = r_{12} = r_{13} = r_{21} = r_{22} = r_{23} = 0$. For each combination of sample sizes and variances, a simulated sample of size 10,000 was generated from normal populations with appropriate means and variances. The generalized p-value given by (10) was computed using each simulated sample and then the rejection rate (fraction of times the p-value is less than the nominal level) of the null hypothesis H_{0AB} was calculated. Two values of nominal levels, namely $\alpha = .1$ and $\alpha = .05$ are studied based on 10000 simulated samples. A simulation study based on about 100000 samples may provide sufficient data to assess the size performance at .01 level, which is beyond the scope of this paper. The findings of the simulated study are summarized in Table 4.

Table 4. Size performance of generalized p-values: Rejection rate of null hypothesis when it is true

Standard deviations and sample sizes												Sizes when level	
σ_{11}	σ_{21}	σ_{12}	σ_{22}	σ_{13}	σ_{23}	n_{11}	n_{21}	n_{12}	n_{22}	n_{13}	n_{23}	$\alpha = .1$	$\alpha = .05$
1.0	1.5	2.0	2.5	3.0	2.0	5	5	5	5	5	5	.064	.031
1.0	1.5	2.0	2.5	3.0	2.0	10	10	10	10	10	10	.080	.037
1.0	1.5	2.0	2.5	3.0	2.0	10	10	10	15	15	15	.083	.039
1.0	1.5	2.0	2.5	3.0	3.5	10	10	10	15	15	15	.087	.042
1.0	1.5	2.0	2.5	3.0	3.5	15	15	15	15	15	15	.089	.045
1.0	1.0	1.0	1.0	1.0	1.0	5	5	5	5	5	5	.067	.034
1.0	1.0	1.0	1.0	1.0	1.0	10	10	10	15	15	15	.084	.044
1.0	1.0	1.0	1.0	1.0	1.0	15	15	15	15	15	15	.086	.043

Our study covers a wide variety of possible situations ranging from the case of equal variances and equal sample sizes to extreme heteroscedasticity where the

largest variance is as large as 12 times the smallest variance. As expected, in all cases the actual type I error (rejection rate of the null hypothesis) was found to be less than the intended type I error (nominal level).

5. Derivation of Generalized F-tests

The purpose of this section is to derive the p-values of generalized F-tests presented in Section 2 and Section 3. This is a generalization of classical F-tests to the case of unequal variances and it is accomplished by taking the generalized p-value approach introduced by Tsui and Weerahandi (1989).

To describe the approach briefly, consider the problem of testing a point null hypothesis of the form $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$ based on an observable random vector $\mathbf{X}$. Let $\zeta = (\theta, \delta)$ be the vector of unknown parameters, where δ are the nuisance parameters and $\mathbf{x}$ be the observed (sample) value of $\mathbf{X}$. Traditionally, testing of a hypothesis is performed by ordering the sample space according to the possible values of θ by means of a test statistic and then by establishing a extreme region, a subset of the sample space whose boundaries depend on $\mathbf{x}$. This method will provide a solution to the problem only if the probability of the resulting extreme region can be computed without any knowledge of the nuisance parameters. In many statistical problems, for instance Behrens-Fisher problem, a test statistic satisfying these requirements does not exist. Motivated by an exact p-value given by Weerahandi (1987) for regression comparisons, Tsui and Weerahandi (1989) extended the definition of an extreme region that can be defined any number of statistics. The idea is to define subsets in higher dimension and for the purpose of facilitating this they also defined a test variable of the form $T(\mathbf{X}; \mathbf{x}, \zeta)$ and extreme regions which may depend on the nuisance parameters ζ, and the observed data $\mathbf{x}$ as well as on θ and $\mathbf{X}$. By means of a test variable, Weerahandi (1995a) gave the exact test for the unbalanced heteroscedastic one-way ANOVA. Following their notations and definitions, a test variable $T(\mathbf{X}; \mathbf{x}, \zeta)$ is a real valued function defined on the sample space with the following properties.

1. The distribution function of $T(\mathbf{X}; \mathbf{x}, \zeta_0)$ and $t_{obs} = T(\mathbf{x}; \mathbf{x}, \zeta)$ both do not depend on nuisance parameters δ, where $\zeta_0 = (\theta_0, \delta)$;
2. $\Pr(T(\mathbf{X}; \mathbf{x}, \zeta) \geq t) \geq \Pr(T(\mathbf{X}; \mathbf{x}, \zeta_0) \geq t)$ for all θ and given any fixed $t, \mathbf{x}$ and δ.

Then, the generalized p-value is defined as $p = \Pr(T(\mathbf{X}; \mathbf{x}, \zeta_0) \geq t_{obs})$. Requirement 1 is imposed to ensure that the p-value is computable and Requirement 2 ensures that tests based on this p-value are unbiased. Just like conventional p-values, generalized p-values serve to measure the evidence in support or against hypotheses. It should be also noted that although this approach generalizes the method of specifying an extreme region, a well defined subset of the underlying

sample space, there is no difference between classical p-values and generalized p-values the way Fisher treated problems of significance testing.

First let us derive the test for no interaction effect given in (10). To do this, consider the standardized error sum of squares $\tilde{S}_E$ defined in (7) and the standardized interaction sum of squares $\tilde{S}_I$ defined in (6). Since $n_{ij}S_{ij}^2/\sigma_{ij}^2$ has a chi-squared distribution with $n_{ij}-1$ degrees of freedom for all ij, the standardized error sum of squares $\tilde{S}_E$ has a chi-squared distribution with $N-IJ$ degrees of freedom. As in standard two-way ANOVA with equal variances, when H_{0AB} is true, $\tilde{S}_I$ has a independent chi-squared distribution with $IJ-1$ degrees of freedom and therefore, under H_{0AB},

$$\frac{\tilde{S}_I \Big/ (IJ-1)}{\tilde{S}_E \Big/ (N-IJ)} \sim F_{IJ-1,N-IJ}. \tag{22}$$

If the null hypothesis is not true this random variable tends to take large values. This suggest that this random variable can be employed to obtain an unbiased test. Define

$$B_{ij} = \frac{\sum_{p=1}^{i-1}\sum_{q=1}^{J} n_{pq}S_{pq}^2/\sigma_{pq}^2 + \sum_{q=1}^{j} n_{iq}S_{iq}^2/\sigma_{iq}^2}{\sum_{p=1}^{i-1}\sum_{q=1}^{J} n_{pq}S_{pq}^2/\sigma_{pq}^2 + \sum_{q=1}^{j+1} n_{iq}S_{iq}^2/\sigma_{iq}^2}$$
$$\text{if} \quad i=1,\ldots,I; \quad j=1,\ldots,J-1$$
$$B_{iJ} = \frac{\sum_{p=1}^{i}\sum_{q=1}^{J} n_{pq}S_{pq}^2/\sigma_{pq}^2}{\sum_{p=1}^{i}\sum_{q=1}^{J} n_{pq}S_{pq}^2/\sigma_{pq}^2 + n_{i+1,1}S_{i+1,1}^2/\sigma_{i+1,1}^2}$$
$$\text{if} \quad i=1,\ldots,I-1.$$

Then it can be shown that the densities of these B_{ij} random variables are beta random variables defined in (11). Furthermore, these B_{ij} random variables and $\tilde{S}_E$ are all independent random variables. Also note that $n_{ij}S_{ij}^2/\sigma_{ij}^2 = \tilde{S}_E Y_{ij}$ for all $i=1,\ldots,I; j=1,\ldots,J$; where Y_{ij}'s are products of beta random variables defined in (12). Now define a potential test variable as

$$\begin{aligned} T(\mathbf{X};\mathbf{x},\zeta) &= \frac{\tilde{S}_I\left(\sigma_{11}^2,\ldots,\sigma_{IJ}^2\right)}{\tilde{s}_I\left(s_{11}^2\sigma_{11}^2/S_{11}^2,\ldots,s_{IJ}^2\sigma_{IJ}^2/S_{IJ}^2\right)} \\ &= \frac{\tilde{S}_I\left(\sigma_{11}^2,\ldots,\sigma_{IJ}^2\right)}{\tilde{s}_I\left(n_{11}s_{11}^2/(\tilde{S}_E Y_{11}),\ldots,n_{IJ}s_{IJ}^2/(\tilde{S}_E Y_{IJ})\right)} \\ &= \frac{\tilde{S}_I\left(\sigma_{11}^2,\ldots,\sigma_{IJ}^2\right)}{\tilde{S}_E\tilde{s}_I\left(n_{11}s_{11}^2/Y_{11},\ldots,n_{IJ}s_{IJ}^2/Y_{IJ}\right)}. \end{aligned} \tag{23}$$

The observed value of $T(\mathbf{X})$ is $t(\mathbf{x}) = 1$. From (22) it is clear that, under the null hypothesis H_{0AB}, the distribution of $\tilde{S}_I/\tilde{S}_E$ does not depend on any

nuisance parameters. Since the Y_{ij} terms are product of beta random variables, the distribution of $\tilde{s}_I\left(n_{11}s_{11}^2/Y_{11},\ldots,n_{IJ}s_{IJ}^2/Y_{IJ}\right)$ also does not depend on any nuisance parameters. Therefore, under the null hypothesis, the distribution of $T(\mathbf{X})$ does not depend on any nuisance parameters. Furthermore, if H_{0AB} is not true, then $\tilde{S}_I$ has a noncentral chi-squared distribution and consequently T tends to take larger values for deviations from H_{0AB}. Hence, T is a test variable that can be employed to test the null hypothesis H_{0AB} and the p-value is given by

$$\begin{aligned} p &= \Pr\Big(T(\mathbf{X}) \geq t(\mathbf{x})\Big) = \Pr\Big[\tilde{S}_I\Big(\sigma_{11}^2,\ldots,\sigma_{IJ}^2\Big) \geq \tilde{s}_I\Big(\frac{s_{11}^2\sigma_{11}^2}{S_{11}^2},\ldots,\frac{s_{IJ}^2\sigma_{IJ}^2}{S_{IJ}^2}\Big)\Big] \\ &= \Pr\Big[\tilde{S}_I \geq \tilde{S}_E\tilde{s}_I\Big(n_{11}s_{11}^2/Y_{11},\ldots,n_{IJ}s_{IJ}^2/Y_{IJ}\Big)\Big] \\ &= \Pr\Big[\frac{\tilde{S}_I\Big/(I-1)(J-1)}{\tilde{S}_E\Big/(N-IJ)} \geq \frac{N-IJ}{(I-1)(J-1)}\tilde{s}_I\Big(\frac{n_{11}s_{11}^2}{Y_{11}},\ldots,\frac{n_{IJ}s_{IJ}^2}{Y_{IJ}}\Big)\Big] \\ &= 1 - E\Big\{H_{(I-1)(J-1),(N-IJ)}\Big(\frac{N-IJ}{(I-1)(J-1)}\tilde{s}_I\Big[\frac{n_{11}s_{11}^2}{Y_{11}},\ldots,\frac{n_{IJ}s_{IJ}^2}{Y_{IJ}}\Big]\Big)\Big\} \end{aligned}$$

and this complete the proof of (10).

Similarly, the proof of (17) can be obtain by considering the test variable

$$T(\mathbf{X}) = \tilde{S}_A(\sigma_{11}^2,\ldots,\sigma_{IJ}^2)/\tilde{s}_A(s_{11}^2\sigma_{11}^2/S_{11}^2,\ldots,s_{IJ}^2\sigma_{IJ}^2/S_{IJ}^2)$$

with the help of the F-distribution on which (16) is based.

Acknowledgement

The authors thank the reviewers for their constructive comments and suggestions which led to a substantial improvement of this article.

References

Arnold, S. F. (1981). *The Theory of Linear Models and Multivariate Analysis.* John Wiley, New York.

Barnard, G. A. (1984). Comparing the means of two independent samples. *Appl. Statist.* **33**, 266-271.

Fujikoshi, Y. (1993). Two-Way ANOVA models with unbalanced data. *Discrete Math.* **116**, 315-334.

Griffiths, W. and Judge, G. (1992). Testing and estimating location vectors when the error covariance matrix is unknown. *J. Econom.* **54**, 121-138.

Koschat, M. A. and Weerahandi, S. (1992). Chow-type tests under heteroscedasticity. *J. Business & Econom. Statist.* **10**, 221-228.

Krutchkoff, R. G. (1988). One-way fixed effects analysis of variance when the error variances may be unequal. *J. Statist. Comput. Simul.* **30**, 259-271.

Krutchkoff, R. G. (1989). Two-way fixed effects analysis of variance when the error variances may be unequal. *J. Statist. Comput. Simul.* **32**, 177-183.

Lindman, H. R. (1992). *Analysis of Variance in Experimental Design.* Springer-Verlag, New York.

Meng, X.-L.(1994). Posterior predictive p-values. *Ann. Statist.* **22**, 1142-1160.

Rice, W. R. and Gains, S. D. (1989). One-way analysis of variance with unequal variances. *Proc. Natl. Acad. Sci.* **86**, 8183-8184.

Robinson, G. K. (1976). Properties of student's t and of the Behrens-Fisher solution to the two means problem. *Ann. Statist.* **4**, 963-971.

Scheffe, H. (1959). *The Analysis of Variance.* John Wiley, New York.

Searle, S. R. (1971). *Linear Models.* John Wiley, New York.

Seber, G. A. F. (1977). *Linear Regression Analysis.* John Wiley, New York.

Thursby, J. G. (1992). A comparison of several exact and approximate tests for structural shift under heteroscedasticity. *J. Econom.* **53**, 363-386.

Tsui, K. and Weerahandi, S. (1989). Generalized p-values in significance testing of hypotheses in the presence of nuisance parameters. *J. Amer. Statist. Assoc.* **84**, 602-607.

Weerahandi, S. (1987). Testing regression equality with unequal variances. *Econometrica* **55**, 1211-1215.

Weerahandi, S. (1991). Testing variance components in mixed models with generalized p values. *J. Amer. Statist. Assoc.* **86**, 151-153.

Weerahandi, S. (1995a). ANOVA under unequal error variances. *Biometrics* **51**, 589-599.

Weerahandi, S. (1995b). *Exact Statistical Methods for Data Analysis.* Springer-Verlag, New York.

Weerahandi, S. and Johnson, R. A. (1992). Testing reliability in a stress-strength model when X and Y are normally distributed. *Technometrics* **34**, 83-91.

Weerahandi, S. and Tsui, K. W. (1996). Comment on "Posterior predictive assessment of model fitness via realized discrepancies" by Gelman, Meng, and Stern. *Statist. Sinica* **6**, 792-796.

Welch, B. L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika* **38**, 330-336.

Zhou, L. and Mathew, T. (1994). Some tests for variance components using generalized p-values. *Technometrics* **36**, 394-402.

IMSL MATH AND STAT/LIBRARY (1989), FORTRAN subroutines for mathematical and statistical applications, IMSL Inc.

Department of Mathematical Sciences, University of Nevada, Las Vegas, NV 89154, U.S.A.

E-mail: anada@nevada.edu

Bell Communications Research, Morristown, NJ 07960, U.S.A.

E-mail: swrs@bellcore.com

(Received May 1995; accepted August 1996)