Ann. Inst. Statist. Math.
Vol. 47, No. 4, 731–742 (1995)

GENERALIZED F-TESTS FOR UNBALANCED NESTED DESIGNS UNDER HETEROSCEDASTICITY

MALWANE M. A. ANANDA

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

(Received November 4, 1994; revised April 10, 1995)

Abstract. Two-factor fixed-effect unbalanced nested design model without the assumption of equal error variance is considered. Using the generalized definition of p-values, exact tests under heteroscedasticity are derived for testing "main effects" of both factors. These generalized F-tests can be utilized in significance testing or in fixed level testing under the Neyman-Pearson theory. Two examples are given to illustrate the proposed test and to demonstrate its advantages over the classical F-test. Extensions of the procedure for three-factor nested designs are briefly discussed.

Key words and phrases: Nested design, unbalanced models, heteroscedasticity, generalized p-values.

1. Introduction

In many statistical applications involving comparison of two normal populations and ANOVA including nested designs, it is customary to assume that the underline error terms have equal variances. This assumption is made for mathematical tractability rather than anything else. Although, the classical F-test is robust against a moderate departure from this assumption, when the problem of heteroscedasticity is serious, applying the classical F-test with the assumption of equal variance can lead to misleading conclusions (Krutchkoff (1988, 1989)). Krutchkoff argued that transformations cannot resolve the problem and also showed that in many cases the Kruskal-Wallis test is not an alternative solution compared to the classical F-test, although it is less sensitive to the unequal error variance.

In one-way ANOVA, Krutchkoff (1988) and Weerahandi (1994a, 1994b) provided interesting examples to demonstrate the repercussions of applying the classical F-test under serious heteroscedasticity. In particular, this problem can be very serious when the error variances are negatively correlated with the sample sizes. Using the generalized definition of the p-values (see Tsui and Weerahandi (1989)), Weerahandi (1994a) obtained exact unbiased tests for one-way ANOVA problems under heteroscedasticity.

Ananda and Weerahandi (1994) showed that the equal variance assumption is even more serious in higher-way models than in one way-models. Furthermore, they obtained exact unbiased tests for unbalanced two-way ANOVA problems with unequal variances.

In this paper, the fixed level nested design model under heteroscedasticity is considered. In nested design models, it is very reasonable to expect different variances for different factor levels. For instance, consider the following example. Suppose a pharmaceutical company or a software product manufacturing company has two factories, each located in two completely different environments. The company operates two training schools, one in each factory. Also suppose that the training school in the first factory uses 2 different training methods and the school in the second factory uses 3 different training methods. The company is interested in the effect of school (factor A) and training methods (factor B) in learning. In this two factor fixed level nested design model, it is very likely that the variances on learning achievements for the five different training methods are unequal.

As in one-way and two-way ANOVA problems, when heteroscedasticity is serious, it is likely that the classical F-tests will result in misleading conclusions. Using the generalized definition of the p-values, the classical F-tests are extended and exact unbiased tests are obtained for the two factor nested design. These resulting p-values can also be expressed explicitly. Furthermore, a brief discussion of the extensions for three factor nested designs follows.

Each of the generalized tests reported 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 extreme region of the sample space. The test is unbiased in the sense that the probability of the extreme region increases for any departure from the null hypothesis. It should be emphasized that these assertions are not valid under the Neyman-Pearson fixed level testing. In fact, under the Neyman-Pearson theory, exact tests based on the minimal sufficient statistics do not exist for these type of problems. The generalized F-tests developed in this paper can be utilized in fixed level testing as well. Our limited simulation studies have suggested that rejecting a null hypothesis when the generalized p-value is less than α provides an excellent approximate α level test. According to our simulation studies, the generalized F-test is readily size guaranteed for all values of nuisance parameters. In fact, in view of the results in Robinson (1976) and our simulation studies, it is conjectured that, at least in the balanced case, this test is readily size guaranteed for all values of nuisance parameters. However, the proof of such a result is well beyond the scope of this paper. According to other simulation studies reported in the literature (see, for instance, Thursby (1992), Weerahandi and Johnson (1992), 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.

This generalized p-value approach has also been applied in mixed models (Weerahandi (1991), Zhou and Mathew (1994)) and in regression models (Weerahandi (1987), Koschat and Weerahandi (1992)). For a complete coverage and applications of these generalized p-values the reader is referred to Weerahandi (1994*b*).

2. Generalized F-test for two-factor nested designs

Consider a two-factor nested design model with factors A and B; the factor A with I factor levels and the factor B nested within A having $J(1), J(2), \dots, J(I)$ levels respectively yielding a total $J = \sum J(i)$ levels of factor B. Then one can consider the true cell mean of the (i,j) level of factor B, say μ_{ij}, as the sum of a general mean μ, an effect α_i of the i-th level of A, and an effect δ_{ij} of the (i,j) level of factor B,

$$\mu_{ij} = \theta + \alpha_i + \delta_{ij}.$$

Suppose a random sample of size n_{ij} is available from (i,j)-th level of B, $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J(i)$ giving a total sample size $N = \sum_{i=1}^{I} \sum_{j=1}^{J(i)} n_{ij}$. Let X_{ijk}, $i = 1, 2, \dots, I$; $j = 1, 2, \dots, J(i)$; $k = 1, 2, \dots, n_{ij}$ represent these random variables and x_{ijk} represent their observed (sample) values. Sample mean and the sample variance of the (i,j)-th treatment are denoted by $\bar{X}_{ij}$ and S_{ij}^2, $i = 1, \dots, I$; $j = 1, \dots, J(i)$ respectively; that is,

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

Their observed sample values are denoted by $\bar{x}_{ij}$ and s_{ij}^2, $i = 1, \dots, I$; $j = 1, \dots, J(i)$ respectively. Now consider the statistical model with unequal variances:

$$\begin{aligned} &X_{ijk} = \theta + \alpha_i + \delta_{ij} + \epsilon_{ijk}, \\ &\epsilon_{ijk} \sim N(0, \sigma_{ij}^2), \qquad i = 1, 2, \dots, I;\ j = 1, 2, \dots, J(i);\ k = 1, 2, \dots, n_{ij}. \end{aligned} \tag{2.1}$$

In order to have θ, α_i, and δ_{ij} uniquely defined, let us choose the constraints

$$\sum_{i=1}^{I} v_i \alpha_i = 0, \quad \sum_{j=1}^{J(i)} w_{ij} \delta_{ij} = 0 \tag{2.2}$$

where v_i and w_{ij} are nonnegative weights such that $\sum_{i=1}^{I} v_i > 0$ and $\sum_{j=1}^{J(i)} w_{ij} > 0$ for each i.

Consider testing following hypotheses

$$H_{0\delta} : \delta_{ij} = 0, \qquad i = 1, \dots, I,\ j = 1, \dots, J(i) \tag{2.3}$$

$$H_{0\alpha} : \alpha_1 = \alpha_2 = \cdots = \alpha_I = 0 \tag{2.4}$$

against their natural alternative hypotheses.

First, let us consider testing the hypothesis $H_{0\delta}$. Testing this hypothesis can be considered as testing "main effects" of factor B which is equivalent to testing whether the true cell means μ_{ij} depend only on i. In the unbalanced case when variances are equal, it is well known that this hypothesis can be tested using the p-value based on the usual F-statistic (see, for instance, Arnold (1981), pp. 100–101):

$$p = 1 - H_{(J-I),(N-J)}\left[\frac{(N-J)\sum_{i=1}^{I}\sum_{j=1}^{J(i)} n_{ij}(\bar{x}_{ij}^2 - \bar{x}_{i.})^2}{(J-I)\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{n_{ij}}(x_{ijk} - \bar{x}_{ij})^2}\right] \tag{2.5}$$

where $H_{(J-I),(N-J)}$ is the cumulative distribution function of the F-distribution with $(J-I)$ and $(N-J)$ degrees of freedom and $\bar{x}_{i.} = \sum_j n_{ij}\bar{x}_{ij}/\sum n_{ij}$.

2.1 *Test for $H_{0\delta}$ with out the equal variance assumption*

When variances are unequal, the hypothesis $H_{0\delta}$ in (2.3) can be tested on the basis of the p-value

$$p = 1 - E\left\{G_{J-I}\left[\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\frac{\bar{x}_{ij}^2 R_{ij}}{s_{ij}^2} - \sum_{i=1}^{I}\left(\sum_{j=1}^{J(i)}\frac{\bar{x}_{ij}R_{ij}}{s_{ij}^2}\right)^2\left(\sum_{j=1}^{J(i)}\frac{R_{ij}}{s_{ij}^2}\right)^{-1}\right]\right\} \tag{2.6}$$

where G_{J-I} is the cdf of the chi-squared distribution with $(J-I)$ degrees of freedom and the expectation is taken with respect to $R_{ij} \sim \chi^2_{n_{ij}-1}$, $i = 1,\ldots,I$, $j = 1,\ldots,J(i)$ independent chi-square random variables. This p-value can also be written as

$$\begin{aligned} p = 1 - E\Bigg\{H_{(J-I),(N-J)}\Bigg[\frac{(N-J)}{(J-I)}\Bigg(&\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\frac{\bar{x}_{ij}^2 Y_{ij}}{s_{ij}^2} \\ &-\sum_{i=1}^{I}\left(\sum_{j=1}^{J(i)}\frac{\bar{x}_{ij}Y_{ij}}{s_{ij}^2}\right)^2\left(\sum_{j=1}^{J(i)}\frac{Y_{ij}}{s_{ij}^2}\right)^{-1}\Bigg)\Bigg]\Bigg\} \end{aligned} \tag{2.7}$$

where $H_{(J-I),(N-J)}$ is the cdf of the F-distribution with $(J-I)$ and $(N-J)$ degrees of freedom and the expectation is taken with respect to Y_{ij}, $i = 1,\ldots,I$, $j = 1,\ldots,J(i)$ random variables which are defined in terms of the independent beta B_{ij} random variables as follows:

$$\begin{cases} Y_{11} = B_{11}B_{12}\cdots B_{1J(1)}B_{21}B_{22}\cdots B_{2J(2)}\cdots B_{I1}B_{I2}\cdots B_{I,J(I)-1} \\ Y_{12} = (1-B_{11})B_{12}\cdots B_{1J(1)}B_{21}B_{22}\cdots B_{2J(2)}\cdots B_{I1}B_{I2}\cdots B_{I,J(I)-1} \\ Y_{13} = (1-B_{12})B_{13}\cdots B_{1J(1)}B_{21}B_{22}\cdots B_{2J(2)}\cdots B_{I1}B_{I2}\cdots B_{I,J(I)-1} \\ \quad\cdots \\ Y_{1J(1)} = (1-B_{1,J(1)-1})B_{1J(1)}B_{21}B_{22}\cdots B_{2J(2)}\cdots B_{I1}B_{I2}\cdots B_{I,J(I)-1} \\ Y_{21} = (1-B_{11})B_{21}B_{22}\cdots B_{2J(2)}\cdots B_{I1}B_{I2}\cdots B_{I,J(I)-1} \\ \quad\cdots \\ \quad\cdots \\ Y_{I,J(I)-1} = (1-B_{I,J(I)-2})B_{I,J(I)-1} \\ Y_{IJ(I)} = (1-B_{I,J(I)-1}) \end{cases} \tag{2.8}$$

where independent B_{ij} beta random variables are defined as:

$$(2.9)\quad \begin{cases} B_{ij} \sim \text{Beta}\left(\sum_{l=1}^{i-1}\sum_{m=1}^{J(l)}(n_{lm}-1)/2 + \sum_{m=1}^{j}(n_{im}-1)/2, (n_{i,j+1}-1)/2\right) \\ \qquad\qquad \text{if} \quad i=1,2,\ldots,I; \ j=1,2,\ldots,J(i)-1 \\ B_{ij} \sim \text{Beta}\left(\sum_{l=1}^{i}\sum_{m=1}^{J(l)}(n_{lm}-1)/2, (n_{i+1,1}-1)/2\right) \\ \qquad\qquad \text{if} \quad i=1,2,\ldots,I-1, \ \text{and} \ j=J(i). \end{cases}$$

The p-value serves to measure the evidence in favor of $H_{0\delta}$. Moreover, it is an exact probability of a well defined extreme region of the sample space and it increases for any departure from the null hypothesis. Practitioners who prefer to take the Neyman-Pearson approach and perform tests at a nominal level α can also find an excellent approximate test by rejecting the null hypothesis when the generalized p-value is less than α.

The derivation of this test is based on the F-test when σ_{ij} values are known and the extended definition of the p-values given in Tsui and Weerahandi (1989), and therefore this test is referred to as the generalized F-test for testing $H_{0\delta}$. The formal derivation of this test is given in Section 4.

This p-value can easily be computed using Monte Carlo simulations or numerical integrations. While it is easy to use Monte Carlo simulations on the first chi-squared representation numerical integration work better for the later representation. Also for a problem with large number of factor levels, the Monte Carlo simulation method is computationally more efficient. However, the later representation is necessary to show the unbiasedness property of this test.

In Monte Carlo simulation, evaluating the expected value in (2.7) is done by simulating a large number of data sets from the beta distributions defined in (2.9). Each of these data sets consist of $J-1$ random numbers, exactly one random number from each of the beta random variables defined in (2.9). After transforming these random numbers to Y_{ij} random numbers using the transformation in (2.8), the expression appearing in the expected value must be evaluated for each data set and then the expected value can be estimated by using their sample mean. The accuracy of this approximation can be assessed by the Monte Carlo variance of the estimate σ_h^2/L, where L is the number of simulations used and σ_h is the sample (simulated) standard deviation of H values.

Now let us consider testing the hypothesis $H_{0\alpha}$ given in (2.4). Unlike the previous hypothesis, this hypothesis depends on the weights chosen for w_{ij}, so that these weights must be chosen prior to choosing the sampling scheme. In the case of equal variances, it is well known that this hypothesis can be tested using the usual F-test and, in particular with the weights $w_{ij}=n_{ij}$ this F-statistic can be written in a closed-form formula (cf. Arnold (1981), pp. 101–102) which yields the p-value

$$(2.10)\qquad p = 1 - H_{(I-1),(N-J)}\left[\frac{(N-J)\sum_{i=1}^{I} n_{i.}(\bar{x}_{i.}^2 - \bar{x}_{..})^2}{(I-1)\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{n_{ij}}(x_{ijk}-\bar{x}_{ij})^2}\right].$$

In the case of general weights w_{ij} and proportional weights $w_{ij} = n_{ij}/\sigma_{ij}^2$ in particular, the generalized F-tests with out the assumption of equal variances are derived and the resulting p-values are expressed explicitly.

2.2 *Generalized F-test for $H_{0\alpha}$ with general weights w_{ij}*

Consider the testing problem of $H_{0\alpha}$ with the constraint given in (2.1) with a given set of weights w_{ij}. Let us define the generalized sum of squares $\tilde{S}_\alpha$,

$$\tilde{S}_\alpha(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{I,J(I)}^2) = \sum_{i=1}^{I}\sum_{j=1}^{J(i)} \frac{n_{ij}}{\sigma_{ij}^2}(\bar{X}_{ij} - \hat{\theta} - \hat{\delta}_{ij})^2 \tag{2.11}$$

where

$$\hat{\theta} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J(i)}[w_{ij}(\sum_{k=1}^{J(i)} w_{ik}\bar{X}_{ik})/(\sum_{k=1}^{J(i)} w_{ik}^2\sigma_{ik}^2/n_{ik})]}{\sum_{i=1}^{I}\sum_{j=1}^{J(i)}[w_{ij}(\sum_{k=1}^{J(i)} w_{ik})/(\sum_{k=1}^{J(i)} w_{ik}^2\sigma_{ik}^2/n_{ik})]} \tag{2.12}$$

and

$$\hat{\delta}_{ij} = \bar{X}_{ij} - \hat{\theta} - \frac{w_{ij}\sigma_{ij}^2(\sum_{k=1}^{J(i)} w_{ik}\bar{X}_{ik} - \hat{\theta}\sum_{k=1}^{J(i)} w_{ik})}{n_{ij}\sum_{k=1}^{J(i)}(w_{ik}^2\sigma_{ik}^2/n_{ik})} \quad \text{for all} \quad i = 1, 2, \ldots, I;\ j = 1, 2, \ldots, J(i). \tag{2.13}$$

When variances are unequal, the hypothesis $H_{0\alpha}$ can be tested using the p-value

$$p = 1 - E\left\{G_{I-1}\left[\tilde{s}_\alpha\left(\frac{n_{11}s_{11}^2}{R_{11}}, \frac{n_{12}s_{12}^2}{R_{12}}, \ldots, \frac{n_{I,J(I)}s_{I,J(I)}^2}{R_{I,J(I)}}\right)\right]\right\} \tag{2.14}$$

where G_{I-1} is the cdf of the chi-squared distribution with $(I-1)$ degrees of freedom and expectation is taken with respect to independent $R_{ij} \sim \chi^2_{n_{ij}i-1}$ variables. With respect to Y_{ij} variables, this can be written as

$$p = 1 - E\left\{H_{(I-1),(N-J)}\left[\frac{(N-J)}{(I-1)}\tilde{s}_\alpha\left(\frac{n_{11}s_{11}^2}{Y_{11}}, \frac{n_{12}s_{12}^2}{Y_{12}}, \ldots, \frac{n_{I,J(I)}s_{I,J(I)}^2}{Y_{I,J(I)}}\right)\right]\right\}. \tag{2.15}$$

Here H is the cdf of F-distribution with degrees of freedom $(I-1)$ and $(N-J)$. The sketch of proof is also given in Section 4.

Now consider a special set of weights $w_{ij} = n_{ij}/\sigma_{ij}^2$. In equal variance set up, these weights reduce to $w_{ij} = n_{ij}$ which is an attractive set of weights due to the simple formula of the F-statistic (see Arnold (1981), p. 102). Also, in the balanced case, using the weights $w_{ij} = n_{ij}(= n)$ is same as using the weights $w_{ij} = 1$, which implies that every nested factor is equally important. With these set of weights

$w_{ij} = n_{ij}/\sigma_{ij}^2$, the generalized sum of squares $\tilde{S}_\alpha$ given in (2.11) can be written more compactly as

$$\tilde{S}_\alpha^* = \sum_{i=1}^{I}\left(\sum_{j=1}^{J(i)} n_{ij}\bar{X}_{ij}/\sigma_{ij}^2\right)^2\left(\sum_{j=1}^{J(i)} n_{ij}/\sigma_{ij}^2\right)^{-1} - \left(\sum_{i=1}^{I}\sum_{j=1}^{J(i)} n_{ij}\bar{X}_{ij}/\sigma_{ij}^2\right)^2\left(\sum_{i=1}^{I}\sum_{j=1}^{J(i)} n_{ij}/\sigma_{ij}^2\right)^{-1} \tag{2.16}$$

and therefore the p-value for testing $H_{0\alpha}$ can be computed conveniently using the formula

$$p = 1 - E\left\{G\left[\sum_{i=1}^{I}\left(\left(\sum_{j=1}^{J(i)} \bar{x}_{ij}R_{ij}/s_{ij}^2\right)^2\left(\sum_{j=1}^{J(i)} R_{ij}/s_{ij}^2\right)^{-1}\right) - \left(\sum_{i=1}^{I}\sum_{j=1}^{J(i)} \bar{x}_{ij}R_{ij}/s_{ij}^2\right)^2\left(\sum_{i=1}^{I}\sum_{j=1}^{J(i)} R_{ij}/s_{ij}^2\right)^{-1}\right]\right\}. \tag{2.17}$$

One can use the solution given in (2.12) and (2.13) to obtain an explicit solution for the unbalanced case of the conventional F-test (with the assumption of equal variance) with general weights w_{ij}. The explicit solution for the conventional F-test can be written as

$$p = 1 - H_{(I-1),(N-J)}\left[\frac{(N-J)}{(I-1)s_E}\sum_{i=1}^{I}\sum_{j=1}^{J(i)} n_{ij}(\bar{x}_{ij} - \hat{\hat{\theta}} - \hat{\hat{\delta}}_{ij})^2\right] \tag{2.18}$$

where $\hat{\hat{\theta}}$, $\hat{\hat{\delta}}_{ij}$ and, s_E are identical to $\hat{\theta}$, $\hat{\delta}$ and, $\tilde{s}_E$ (defined in equations (2.12), (2.13) and (4.2), respectively) except that the σ_{ij} values in all expressions equal 1.

3. Examples

The objective of this section is to demonstrate the usefulness of the proposed exact tests over the classical F-tests when heteroscedasticity is serious. Let us consider a problem with a moderate heteroscedasticity.

Example 1. Consider the fixed effect nested design data set given in Table 1, with two factors A and B; A with 2 levels A_1, A_2 and; B nested within A having 2 (say B_1, B_2) and 3 (say B_3, B_4, B_5) levels respectively.

This is a typical data set with a moderate heteroscedasticity. This data set was generated by the following two-factor nested design model with following population parameter configurations: $\theta = 50.0$, $\alpha_1 = 1.0$, $\alpha_2 = -1.0$, $\delta_{ij} = 0$ for all ij. Then their cell means are $\mu_{11} = 51.0$, $\mu_{12} = 51.0$, $\mu_{21} = 49.0$, $\mu_{22} = 49.0$, $\mu_{23} = 49.0$. Cell standard deviations (population) were taken as $\sigma_{11} = 1.0$, $\sigma_{12} = 2.2$, $\sigma_{21} = 2.9$, $\sigma_{22} = 2.5$, $\sigma_{23} = 1.0$.

Table 1. Sample means and sample standard deviations.

		Sample Size	Mean	Standard Deviation
A_1	B_1 :	$n_{11} = 10$	$\bar{x}_{11} = 51.13$	$s_{11} = 1.29$
	B_2 :	$n_{12} = 7$	$\bar{x}_{12} = 49.15$	$s_{12} = 2.49$
A_2	B_3 :	$n_{21} = 6$	$\bar{x}_{21} = 50.01$	$s_{21} = 2.58$
	B_4 :	$n_{22} = 9$	$\bar{x}_{22} = 49.26$	$s_{22} = 1.19$
	B_5 :	$n_{23} = 8$	$\bar{x}_{23} = 48.99$	$s_{23} = 0.99$

Table 2. p-values with and without the assumption of equal variance.

Hypothesis	Generalized p-value	Classical p-value
$H_{0\delta}$	0.334	0.144
$H_{0\alpha}$	0.009	0.117

We analyze this data using both methods, the proposed generalized p-values and the classical p-values (with the assumption of equal variance). Recall, while the p-value for testing $\delta_{ij} = 0$ is independent of the chosen weights, the p-value for testing $\alpha_i = 0$ is dependent on them, and therefore to test the later hypothesis weights proportional to the sample sizes $w_{ij} = n_{ij}/\sigma_{ij}^2$ were used in this analysis. Results are given in Table 2.

It is remarkable that, while the generalized p-value obtained for $H_{0\delta}$ is much larger than the classical p-value, the p-value for $H_{0\alpha}$ is smaller than the classical p-value. According to the actual population parameters for which the data were generated, this is exactly what one would expect from a better test. The classical F-test has failed to detect significant differences between α's despite the fact that the data actually provides sufficient information to do so. Lack of the power of a test at this magnitude is unacceptable in practical applications.

Example 2. This is a simulation study to compare classical p-values and generalized p-values. The simulation is based on 5000 iterations from population parameter configurations and sample sizes based on the previous example, i.e., $\theta = 50.0$, $\alpha_1 = 1.0$, $\alpha_2 = -1.0$, $\delta_{ij} = 0$ for all ij, $\sigma_{11} = 1.0$, $\sigma_{12} = 2.2$, $\sigma_{21} = 2.9$, $\sigma_{22} = 2.5$, $\sigma_{23} = 1.0$, $n_{11} = 10$, $n_{12} = 7$, $n_{21} = 6$, $n_{22} = 9$, $n_{23} = 8$.

According to the population parameters on which the simulation is based, larger p-values are preferable for testing $H_{0\delta}$ and smaller p-values are preferable for testing $H_{0\alpha}$. These results clearly show that for both hypotheses, even under moderate heteroscedasticity, the tests based on generalized p-values are more efficient in detecting the true state than the classical p-values.

According to these findings, in situations where the equal variance assumption is not reasonable, it is very important to carry out these generalized tests to avoid misleading conclusions. We also recommend these procedures even if there is no

Table 3. Simulation results based on 5000 iterations.

Hypothesis	Average generalized p-values	Average classical p-values	Percentage of trials with generalized p-values less than classical p-values
$H_{0\delta}$	0.49	0.45	39%
$H_{0\alpha}$	0.004	0.03	87%

strong evidence to support the equal variance assumption. In any case, if one can detect a significant difference using the generalized p-values and cannot detect it using the classical p-values, the former conclusion must be used than the latter, since the former conclusion is based on weaker assumptions.

4. Derivation of tests

Tsui and Weerahandi (1989) extended the definition of the p-value and extreme regions by means of a test variable (a function defined on the sample space), so that one can get exact solutions for problems such as the Behrens-Fisher problem, where the solution otherwise does not exist. To test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ based on an observable random vector $\boldsymbol{X}$, they defined a test variable-$T(X; x, \zeta)$ with the following two properties: 1. The distribution function of $T(\boldsymbol{X}; \boldsymbol{x}, \zeta_0)$ and $t_{obs} = T(\boldsymbol{x}; \boldsymbol{x}, \zeta)$ both do not depend on nuisance parameters δ, where $\zeta_0 = (\theta_0, \delta)$; 2. $\Pr(T(\boldsymbol{X}; \boldsymbol{x}, \zeta) \geq t) \geq \Pr(T(\boldsymbol{X}; \boldsymbol{x}, \zeta_0) \geq t)$ for all θ and given any fixed t, $\boldsymbol{x}$ and δ. Here $\zeta = (\theta, \delta)$ is the vector of unknown parameters, δ the vector of nuisance parameters and, x the observed value of X. Then, the generalized p-value is defined as $p = \Pr(T(\boldsymbol{X}; \boldsymbol{x}, \zeta_0) \geq t_{obs})$. 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.

A generalized p-value serve to measure the evidence in favor or against a null hypotheses. It is the exact probability of a well defined extreme region of the sample space. Furthermore, this probability increases for any departure from the null hypothesis. This concept of generalizing p-values is consistent with the way Fisher treated the problem of significance testing rather than Neyman-Pearson treatment of fixed level testing.

First let us derive the test for $H_{0\delta}$ given in (2.6) and (2.7). Define the standardized error sum of squares $\tilde{S}_E$ and standardized error sum of squares $\tilde{S}_\delta$ as:

$$(4.1) \quad \tilde{S}_\delta(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{I,J(I)}^2) = \sum_{i=1}^{I} \sum_{j=1}^{J(i)} \frac{n_{ij}(\bar{X}_{ij})^2}{\sigma_{ij}^2} - \sum_{i=1}^{I} \left(\left(\sum_{j=1}^{J(i)} \frac{n_{ij}\bar{X}_{ij}}{\sigma_{ij}^2} \right)^2 \left(\sum_{j=1}^{J(i)} \frac{n_{ij}}{\sigma_{ij}^2} \right)^{-1} \right)$$

$$(4.2) \quad \tilde{S}_E(\sigma_{11}^2, \sigma_{12}^2, \ldots, \sigma_{I,J(I)}^2)$$

$$= \sum_{i=1}^{I} \sum_{j=1}^{J(i)} \sum_{k=1}^{n_{ij}} \frac{1}{\sigma_{ij}^2}(X_{ijk} - \bar{X}_{ij})^2 = \sum_{i=1}^{I} \sum_{j=1}^{J(i)} n_{ij} S_{ij}^2/\sigma_{ij}^2.$$

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 $N - J$ degrees of freedom. When $H_{0\delta}$ is true, it can easily be shown that, $\tilde{S}_\delta$ has an independent chi-squared distribution with $J - I$ degrees of freedom and therefore $(\tilde{S}_\delta/(k-1))/(\tilde{S}_E/(N-k))$ has a F-distribution with degrees of freedom $(J - I)$ and $(N - J)$. Now define

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

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

Then it can be shown that the densities of these B_{ij} random variables are beta random variables defined in (2.9). Furthermore, these B_{ij} random variables and $\tilde{S}_E$ are all independent random variables. Also notice that $n_{ij}S_{ij}^2/\sigma_{ij}^2 = \tilde{S}_E Y_{ij}$ for all $i = 1, 2, \ldots, I$; $j = 1, 2, \ldots, J(i)$; where Y_{ij}'s are the product of beta random variables defined in (2.8). Now define a potential test variable as

$$T(\boldsymbol{X}) = \tilde{S}_\delta(\sigma_{11}^2, \ldots, \sigma_{I,J(I)}^2)/\tilde{s}_\delta(s_{11}^2\sigma_{11}^2/S_{11}^2, \ldots, s_{I,J(I)}^2\sigma_{I,J(I)}^2/S_{I,J(I)}^2). \tag{4.3}$$

The observed value of $T(\boldsymbol{X})$ is $t(\boldsymbol{x}) = 1$. Also notice that the expression in the above equation can be written as $\tilde{S}_\delta(\sigma_{11}^2, \ldots, \sigma_{I,J(I)}^2)/(\tilde{S}_E\tilde{s}_\delta(n_{11}s_{11}^2/Y_{11}, \ldots, n_{I,J(I)}s_{I,J(I)}^2/Y_{I,J(I)}))$. Under the null hypothesis, clearly the distribution of $\tilde{S}_\delta/\tilde{S}_E$ does not depend on any nuisance parameters. Since Y_{ij} terms are products of beta random variables, the distribution of $T(\boldsymbol{X})$ (under the null hypothesis) does not depend on any nuisance parameters. If the null hypothesis is not true, then $\tilde{S}_\delta$ has a noncentral chi-squared distribution and consequently T tends to take larger values for deviations from $H_{0\delta}$. Hence, T is a test variable that can be employed to test the null hypothesis $H_{0\delta}$ and the p-value is given by

$$\begin{aligned} p &= \Pr[\tilde{S}_\delta(\sigma_{11}^2, \ldots, \sigma_{I,J(I)}^2) \geq \tilde{s}_\delta(s_{11}^2\sigma_{11}^2/S_{11}^2, \ldots, s_{I,J(I)}^2\sigma_{I,J(I)}^2/S_{I,J(I)}^2)] \\ &= \Pr[\tilde{S}_\delta \geq \tilde{S}_E\tilde{s}_\delta(n_{11}s_{11}^2/Y_{11}, \ldots, n_{I,J(I)}s_{I,J(I)}^2/Y_{I,J(I)})] \end{aligned} \tag{4.4}$$

which can be written in the form of (2.7). The representation (2.6) can be obtained from equation (4.4) by replacing S_{ij} variables in terms of chi-squared R_{ij} variables and rewriting the probability as an expectation with respect to chi-squared $\tilde{S}_\delta$ variable. This completes the proof of (2.6).

Now let us consider the derivation of the test given in (2.14). One can show that the minimum of

$$\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{n_{ij}} \frac{1}{\sigma_{ij}^2}(x_{ijk} - \theta - \delta_{ij})^2$$

subject to the constraints $\sum_j w_{ij}\delta_{ij} = 0$, $i = 1, \ldots, I$ occurs at $\hat{\theta}$ and $\hat{\delta}_{ij}$ given in (2.12) and (2.13) respectively. So it can easily be shown that the generalized sum of squares $\tilde{S}_\alpha$ given in (2.11) has a chi-squared distribution with degrees of freedom $(I - 1)$, and it is independent of the standardized error sum of squares $\tilde{S}_E$ given in (4.2). Now considering the test variable given in (4.3) by replacing $\tilde{S}_\alpha$ and $\tilde{s}_\alpha$ in the place of $\tilde{S}_\delta$ and $\tilde{s}_\delta$ respectively, the rest of the proof follows similarly.

5. Generalized F-test for three factor nested design

Consider the three-factor nested design model with factors A, B, and C; A with I factor levels; B is nested within A having factor levels $J(1), J(2), \ldots, J(I)$ respectively; C is nested within B having factor levels $K(i,j)$, $i = 1, \ldots, I$; $j = 1, \ldots, J(i)$. Also suppose a random sample of size n_{ijk} is available from each level of C and $\bar{x}_{ijk}$ and s_{ijk}^2 be the sample mean and sample standard deviation. Then the model can be written as

$$\begin{aligned}&X_{ijkm} = \theta + \alpha_i + \beta_{ij} + \delta_{ijk} + \epsilon_{ijkm},\\ &\epsilon_{ijkm} \sim N(0, \sigma_{ijk}^2), \quad i = 1, 2, \ldots, I; \; j = 1, 2, \ldots, J(i);\\ &k = 1, 2, \ldots, K(i,j); \; m = 1, \ldots, n_{ijk}.\end{aligned} \tag{5.1}$$

In order to define the model uniquely, let us impose the restriction

$$\sum_{i=1}^{I} u_i\alpha_i = 0, \quad \sum_{j=1}^{J(i)} v_{ij}\beta_{ij} = 0 \quad \text{for all } i, \qquad \text{and}$$
$$\sum_{k=1}^{K(i,j)} w_{ij}\delta_{ijk} = 0 \quad \text{for all } i, j.$$

Suppose we are interested in testing the hypothesis $H_0 : \delta_{ijk} = 0$. Testing of this hypothesis is independent of the chosen weights and when variances are equal this can be done using the usual F-test (Scheffe (1959)). Under heteroscedasticity, this hypothesis can be tested using the p-value

$$\begin{aligned}p = 1 - E\Bigg\{G\Bigg[&\sum_{i=1}^{I}\sum_{j=1}^{J(i)}\sum_{k=1}^{K(i,j)} \frac{\bar{x}_{ijk}^2 R_{ijk}}{s_{ijk}^2}\\ &- \sum_{i=1}^{I}\sum_{j=1}^{J(i)}\left(\sum_{k=1}^{K(i,j)} \frac{\bar{x}_{ijk}R_{ijk}}{s_{ijk}^2}\right)^2\left(\sum_{k=1}^{K(i,j)} \frac{R_{ijk}}{s_{ijk}^2}\right)^{-1}\Bigg]\Bigg\}\end{aligned} \tag{5.2}$$

where G is the cdf of chi-squared distribution with degrees of freedom $(\sum_{ij} K(i,j) - \sum_i J(i))$ and expectation is taken with respect to independent $R_{ijk} \sim \chi^2_{n_{ijk}-1}$ chi-squared random variables.

Testing of the hypotheses $H_0 : \alpha_i = 0$, for all i and $H_0 : \beta_{ij} = 0$, for all i, j can also be done in a similar manner.

Acknowledgements

I would like to thank Dr. Samaradasa Weerahandi of Bell Communications Research for reading the first draft of this paper and making suggestions to improve the presentation of the results, and also the referees and the associate editor for their valuable comments.

References

Ananda, M. M. A. and Weerahandi, S. (1994). Two-way ANOVA with unequal cell frequencies and unequal variances, Tech. Report, Department of Mathematical Sciences, University of Nevada, Las Vegas (submitted).

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

Koschat, M. A. and Weerahandi, S. (1992). Chow-type tests under heteroscedasticity, *Journal of Business and Economic Statistics*, **10**, 221–227.

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

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

Robinson, G. K. (1976). Properties of student's t and of the Behrens-Fisher solution to the two sample problem, *Ann. Statist.*, **4** (5), 963–971.

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

Thursby, J. G. (1992). A comparison of several exact and approximate tests for structural shift under heteroskedasticity, *J. Econometrics*, **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. (1994*a*). ANOVA under unequal error variances, *Biometrics* (to appear).

Weerahandi, S. (1994*b*). *Exact Statistical Methods for Data Analysis*, Springer, 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** (1), 83–91.

Zhou, L. and Mathew, T. (1994). Some tests for variance components using generalized p-values, *Technometrics*, **36** (4), 394–402.