Published in final edited form as:

Biom J. 2008 August ; 50(4): 571–583. doi:10.1002/bimj.200710437.

Testing for Group Effect in a $2 \times k$ Heteroscedastic ANOVA Model

James F. Troendle*

Biometry and Mathematical Statistics Branch, National Institute of Child Health and Human Development, Bld 6100, Bethesda, MD 20892, USA

Summary

It is natural to want to relax the assumption of homoscedasticity and Gaussian error in ANOVA models. For a two-way ANOVA model with $2 \times k$ cells, one can derive tests of main effect for the factor with two levels (referred to as group) without assuming homoscedasticity or Gaussian error. Empirical likelihood can be used to derive testing procedures. An approximate empirical likelihood ratio test (AELRT) is derived for the test of group main effect. To approximate the distributions of the test statistics under the null hypothesis, simulation from the approximate empirical maximum likelihood estimate (AEMLE) restricted by the null hypothesis is used. The homoscedastic ANOVA F -test and a Box-type approximation to the distribution of the heteroscedastic ANOVA F -test are compared to the AELRT in level and power. The AELRT procedure is shown by simulation to have appropriate type I error control (although possibly conservative) when the distribution of the test statistics are approximated by simulation from the constrained AEMLE. The methodology is motivated and illustrated by an analysis of folate levels in the blood among two alcohol intake groups while accounting for gender.

Keywords

Empirical likelihood; Likelihood ratio test; Power; Simulation

1 Introduction

Consider the general case of a particular variable independently observed under two possible treatment conditions (here called groups) and k possible levels of another factor (here called factor levels). Faced with such a problem it is often of interest to compare the mean levels of the variable between the groups. In the usual analysis of variance (ANOVA) jargon this is called testing for a group main effect. The classical approach to this problem is to assume the variable has a Gaussian distribution with equal variance in each of the $2 \times k$ cells determined by combinations of group and factor. This leads naturally to the classical ANOVA F -test (Scheffe, 1959: Chapter 4, pages 90-146).

If one suspects that the variable has a highly skewed distribution, then it is reasonable to suppose that the mean may not be the best measure of center. In this case the median may be more appropriate measure, and rank-based ANOVA methods follow naturally (Puri and Sen, 1985; Toothaker and Newman, 1994). One may also suspect that a nonparametric hypothesis (Akritas et al., 1997) may better reflect the scientific question. However, nonparametric hypotheses are difficult to interpret and in practice one often still prefers using the mean as a basis for inference. Therefore, this paper shall only consider methods of testing the group mean

* e-mail: jt3t@nih.gov, Phone: +001 301 435 6952, Fax: +001 301 402 2084

Conflict of interests statement The author has declared no conflict of interest.

main effect, without assuming a known distribution or equality of variances. Note that the results also do not assume additivity of effects.

The null hypothesis tested here is one of a group main effect. With weights for the levels of the factor, the null restriction is a weighted average of the group mean differences over the levels of the factor. Keselman et al. (1995) and Luh and Guo (2001) consider a different hypothesis that they call “no row main effect”. Their weights are functions of the cell and marginal frequencies. Thus a different study with the same group and factor levels but different cell frequencies results in a different “no row main effect” hypothesis. This doesn't seem interesting because the hypothesis should not depend on a chance event. Keselman et al. (1995) and Luh and Guo (2001) also consider an unweighted hypothesis which would correspond to equal specified weights, but in many cases one might desire unequal but fixed weights. One example would be race as the factor. If the racial proportions in the population were well known, one would probably want to weight by the known proportions, although a given sample might contain quite different proportions. In our formulation, the hypothesis depends on the specified weights, but not on anything random.

To remove the equality of variance assumption, consider a general ANOVA model which allows for arbitrary variance in each cell. Approximations to the limiting distribution of the heteroscedastic ANOVA F -statistic has been given by Welch (1951), Brown and Forsythe (1974), and Brunner et al. (1997). We include the Box-type approximation of Brunner et al. in our comparisons in Section 5. However, instead of approximating the ANOVA F -statistic we approach the problem by removing any distributional assumption and using empirical likelihood (Owen, 2001: Chapter 2, pages 7-28) in deriving testing procedures. An approximation to the empirical likelihood ratio test (AELRT) from the general heteroscedastic ANOVA model will be presented.

To assess the relative power of the procedures, it is necessary to approximate the distribution of the test statistics under the null hypothesis. Our simulations show that the large sample distributional approximations are inadequate for control of type I error. However, an intuitive approximation is to use the constrained approximate empirical maximum likelihood estimate (AEMLE) to approximate the distribution of the test statistics under the null. Simulation from the constrained AEMLE is used to obtain critical cutoffs for testing.

The paper proceeds as follows. The problem is motivated from analyzing folate levels in blood, described in Section 2. In Section 3 the notation and models are introduced. Section 4 contains descriptions of the testing procedures, including the homoscedastic and heteroscedastic ANOVA F -tests, Brunner et al.'s Box-type approximation, and the AELRT. A comparison of level and power is given by simulation in Section 5. Finally, a discussion of the results for the folate example and general recommendations are given in Section 6.

2 Folate by Alcohol Intake and Gender

As part of a nutritional assessment survey of college students in Ireland, the Trinity Student Study obtained comprehensive alcohol intake recall from 2509 students. Based on the detailed drinking history the average amount of alcohol consumed per day was determined for each student and categorized into five levels. Of interest was to compare the blood folate level for students in the highest level to that of those in the second lowest level, while controlling for gender. Folate was measured separately in serum and in red blood cells. The number of students with serum folate assay results in the two alcohol groups was 1099 (186 high alcohol females, 256 high alcohol males, 514 low alcohol females, 143 low alcohol males). The numbers are the same for red cell folate except that there was one additional high alcohol male student.

Analysis was planned to be by using a two-way ANOVA model with main effects for alcohol and gender, testing for a main effect of alcohol using equal weights for gender levels. This test is part of the SAS PROC GLM output, found as estimate and associated p -value for the alcohol term. However, this test is only valid for Gaussian data. Thus, for each of the four cells the Shapiro-Wilk test of normality, standard deviation, skewness, and kurtosis was examined. Since the values were positively skewed, a log-transformation was applied and the same characteristics examined for the transformed data (Tables 1 and 2). The log-transformation greatly reduces the skewness, kurtosis, and non-normality of the data for both serum folate and red cell folate. However, there is still some evidence of non-normality in certain cells. A second log-transformation resulted in even more significant departure from normality and was therefore not used. More complicated transformations could be used, but there is no definitive way to optimally choose such a transformation, nor any assurance that the transformed data are truly Gaussian.

The results of testing for main effect of alcohol by applying a two-way ANOVA model to the logged serum folate and red blood cell folate data appear in Table 3. Both serum and red cell folate appear to be lower for the high alcohol students. These analyses are not completely satisfactory because of the evidence that some non-normality remains in the log-transformed data, and also because of unequal variability in the cells. We will revisit this analysis in Section 6 to apply the test using a heteroscedastic ANOVA model.

3 Models and Notation

Consider the general case of a particular variable independently observed under two possible treatment conditions (here called groups) and k possible levels of another factor (here called factor levels). Suppose the theoretical mean of the variable in the first group under factor level j is β_j while the mean of the variable in the second group under factor level j is $\beta_j + \delta_j$, for $j = 1, \dots, k$. Let π_j be the weight associated with factor level j for $j = 1, \dots, k$. A general hypothesis of interest might be that the groups have the same overall weighted mean, called the group main effect null hypothesis,

$$H_{GE}: \delta_1 \pi_1 + \delta_2 \pi_2 + \dots + \delta_k \pi_k = 0.$$

3.1 Two-Way ANOVA Model

Although the above section considers a very general setup, for most applications and for generation of such data a little more structure is usually assumed. A natural model to consider when interested in the means of such data without assuming homoscedasticity is the following heteroscedastic two-way ANOVA model:

$$Y_{lji} = \beta_j + \delta_j I_{[l=2]} + \gamma_{lj} \varepsilon_{lji}, \quad l=1,2, \quad j=1,\dots,k, \quad i=1,\dots,n_{lj}, \quad (1)$$

where l indicates group, j indicates the level of the factor, and i indicates observation number within the cell. The ε_{lji} are assumed iid with mean zero and unit variance. Let

$$n = \sum_{l=1}^2 \sum_{j=1}^k n_{lj}.$$

4 Testing H_{GE}

This section describes several approaches to testing the group main effect null hypothesis, H_{GE} .

4.1 Homoscedastic ANOVA F-Test

A commonly used approach is to assume a homoscedastic ANOVA model, which in this case would be (1) where $\gamma_{lj} \equiv 1$ and without assuming ε_{lji} has unit variance. The test statistic is (see Scheffe, 1959: Chapter 4, pages 90-146).

$$TS_F = SS_A (n - 2k) / SS_e$$

where

$$SS_A = \sum_l W_l \widehat{A}_l^2 - \left(\sum_l W_l \right)^{-1} \left(\sum_l W_l \widehat{A}_l \right)^2,$$

$$W_l = \frac{1}{\sum_j \pi_j^2 / n_{lj}},$$

$$\widehat{A}_l = \sum_j \pi_j \bar{Y}_{lj},$$

$$SS_e = \sum_l \sum_j \sum_i \left(Y_{lji} - \bar{Y}_{lj} \right)^2,$$

and

$$\bar{Y}_{lj} = \frac{1}{n_{lj}} \sum_i Y_{lji}.$$

The test statistic TS_F is approximately distributed as an $F_{1, n-2k}$ under H_{GE} .

4.2 Heteroscedastic ANOVA F-Test

The null hypothesis H_{GE} is a particular contrast for a one-way design formed from the cells of our two-way design. To use the matrix notation usually given, define

$$\mu = (\beta_1 \ \beta_2 \ \cdots \ \beta_k \ \beta_1 + \delta_1 \ \beta_2 + \delta_2 \ \cdots \ \beta_k + \delta_k) \boldsymbol{\iota},$$

$$\bar{Y} = \left(\bar{Y}_{11} \ \bar{Y}_{12} \ \cdots \ \bar{Y}_{1k} \ \bar{Y}_{21} \ \bar{Y}_{22} \ \cdots \ \bar{Y}_{2k} \right) \boldsymbol{\iota},$$

and

$$C = (-\pi_1 - \pi_2 \cdots - \pi_k \pi_1 \pi_2 \cdots \pi_k).$$

Then H_{GE} can be written as $C\mu = 0$. Let

$$\widehat{\gamma}_{lj}^2 = \frac{1}{n_{lj}} \sum_{i=1}^{n_{lj}} (Y_{lji} - \bar{Y}_{lj})^2, \quad l=1,2, \quad j=1,\dots,k,$$

and

$$S = n \quad \text{diag} \left\{ \frac{\widehat{\gamma}_{11}^2}{n_{11}}, \frac{\widehat{\gamma}_{12}^2}{n_{12}}, \dots, \frac{\widehat{\gamma}_{1k}^2}{n_{1k}}, \frac{\widehat{\gamma}_{21}^2}{n_{21}}, \frac{\widehat{\gamma}_{22}^2}{n_{22}}, \dots, \frac{\widehat{\gamma}_{2k}^2}{n_{2k}} \right\},$$

where $\text{diag} \{ \cdot, \dots, \cdot \}$ is notation for a diagonal matrix with given elements on the diagonal. The heteroscedastic ANOVA test statistic is then

$$\frac{TS_H = n \left(\bar{Y}' C \bar{Y} \right)^2}{CSC'}$$

Under H_{GE} , the test statistic TS_H is asymptotically chi-squared with 1 degree of freedom.

4.3 Box-Type Approximation Test

Brunner et al. (1997) describe a Box-type approximation for the distribution of

$$\frac{TS_B = n \left(\bar{Y}' M \bar{Y} \right)^2}{\text{tr} \{DS\}},$$

where

$$M = \frac{C'C}{CC'}$$

,

$$D = \text{diag} \{m_{11}, \dots, m_{2k,2k}\},$$

and $\text{tr}(\cdot)$ denotes the trace of a square matrix. The approximation leads to null distribution under H_{GE} of $F(df_1, df_2)$ with degrees of freedom given by

$$df_1 = \frac{[\text{tr} \{DS\}]^2}{\text{tr} \{MSMS\}},$$

$$d f_2 = \frac{[\text{tr} (DS)]^2}{\text{tr} (D^2 S^2 \Lambda)},$$

where $\Lambda = \text{diag} ((n_{11} - 1)^{-1}, \dots, (n_{1k} - 1)^{-1}, (n_{21} - 1)^{-1}, \dots, (n_{2k} - 1)^{-1})$.

4.4 Approximate Empirical Likelihood Ratio Test

This section presents an attempt to form a LRT test for H_{GE} by maximization of the empirical likelihood of the data generated by (1). From (1) one obtains that

$$\varepsilon_{lji} = \frac{Y_{lji} - \beta_j - \delta_j I_{[l=2]}}{\gamma_{lj}}, \quad l=1,2, \quad j=1,\dots,k, \quad i=1,\dots,n_{lj} \quad (2)$$

is iid with mean zero and unit variance. Let p_{lji} for $l=1,2, j=1,\dots,k, i=1,\dots,n_{lj}$ be the jumps in the distribution of ε_{lji} . Then the empirical log-likelihood of the data is

$$L = \sum_{r=1}^2 \sum_{s=1}^k \sum_{h=1}^{n_{rs}} \log p_{rsh} \quad (3)$$

with the constraints

$$\sum_{r=1}^2 \sum_{s=1}^k \sum_{h=1}^{n_{rs}} p_{rsh} = 1 \quad (4)$$

$$\sum_{h=1}^{n_{lj}} \frac{p_{ljh}}{\gamma_{lj}} (Y_{ljh} - \beta_j - \delta_j I_{[l=2]}) = 0, \quad l=1,2, \quad j=1,\dots,k \quad (5)$$

$$\sum_{h=1}^{n_{lj}} \frac{p_{ljh}}{\gamma_{lj}^2} (Y_{ljh} - \beta_j - \delta_j I_{[l=2]})^2 = \sum_{h=1}^{n_{lj}} p_{ljh}, \quad l=1,2, \quad j=1,\dots,k. \quad (6)$$

. The constraint (4) ensures that the $\{p_{lji}\}$ define a valid distribution while (5)-(6) ensure the residual distribution has mean zero and unit variance. The empirical maximum likelihood estimates (EMLE) for $\{p_{lji} : l=1,2, j=1,\dots,k, i=1,\dots,n_{lj}\}$ can be obtained by using Lagrange multipliers to solve the maximization problem, (3)-(6), resulting in

$$\widehat{p}_{lji} = \frac{1}{n}$$

$$\widehat{\beta}_j = \frac{1}{n_{1j}} \sum_{i=1}^{n_{1j}} Y_{1ji}$$

$$\widehat{\delta}_j = \frac{1}{n_{2j}} \sum_{i=1}^{n_{2j}} Y_{2ji} - \frac{1}{n_{1j}} \sum_{i=1}^{n_{1j}} Y_{1ji}$$

$$\widehat{\gamma}_{1j}^2 = \frac{1}{n_{1j}} \sum_{i=1}^{n_{1j}} (Y_{1ji} - \widehat{\beta}_j)^2$$

$$\widehat{\gamma}_{2j}^2 = \frac{1}{n_{2j}} \sum_{i=1}^{n_{2j}} (Y_{2ji} - \widehat{\beta}_j - \widehat{\delta}_j)^2,$$

which agree with the natural estimates for the mean and variance under each combination of l and j .

Next try to find a maximizer of (3) constrained by H_{GE} as well as (4)-(6). Let λ , Λ_{lj} , Γ_{lj} , and θ be the Lagrange multipliers for (4), (5), (6), and H_{GE} respectively. Then the partial derivatives of the modified log-likelihood are

$$\frac{\partial L}{\partial p_{lji}} = \frac{1}{p_{lji}} - \lambda - \frac{\Lambda_{lj}}{\gamma_{lj}} (Y_{lji} - \beta_j - \delta_j I_{l=2}) - \frac{\Gamma_{lj}}{\gamma_{lj}^2} (Y_{lji} - \beta_j - \delta_j I_{l=2})^2 + \Gamma_{lj} = 0 \quad l=1,2, \quad j=1,\dots,k, \quad i=1,\dots,n_{lj}, \quad (7)$$

$$\frac{\partial L}{\partial \gamma_{lj}} = \frac{\Lambda_{lj}}{\gamma_{lj}^2} \sum_{h=1}^{n_{lj}} p_{ljh} (Y_{ljh} - \beta_j - \delta_j I_{l=2}) + \frac{2\Gamma_{lj}}{\gamma_{lj}^3} \sum_{h=1}^{n_{lj}} p_{ljh} (Y_{ljh} - \beta_j - \delta_j I_{l=2})^2 = 0 \quad l=1,2, \quad j=1,\dots,k, \quad (8)$$

$$\frac{\partial L}{\partial \beta_j} = \frac{\Lambda_{1j}}{\gamma_{1j}} \sum_{h=1}^{n_{1j}} p_{1jh} + \frac{\Lambda_{2j}}{\gamma_{2j}} \sum_{h=1}^{n_{2j}} p_{2jh} + \frac{2\Gamma_{1j}}{\gamma_{1j}^2} \sum_{h=1}^{n_{1j}} p_{1jh} (Y_{1jh} - \beta_j) + \frac{2\Gamma_{2j}}{\gamma_{2j}^2} \sum_{h=1}^{n_{2j}} p_{2jh} (Y_{2jh} - \beta_j - \delta_j) = 0 \quad j=1,\dots,k, \quad (9)$$

$$\frac{\partial L}{\partial \delta_j} = \frac{\Lambda_{2j}}{\gamma_{2j}} \sum_{h=1}^{n_{2j}} p_{2jh} + \frac{2\Gamma_{2j}}{\gamma_{2j}^2} \sum_{h=1}^{n_{2j}} p_{2jh} (Y_{2jh} - \beta_j - \delta_j) - \theta \pi_j = 0, \quad j=1,\dots,k. \quad (10)$$

Solving for a solution to (7)-(10) constrained by (4)-(6) and H_{GE} involves too many parameters to realistically get a numerical exact solution. However, an approximate solution can be obtained quite easily.

Suppose

$$\delta^0 = \sum_{j=1}^k \pi_j \widehat{\delta}_j \neq 0,$$

for otherwise the maximizing $\{\delta_j\}$ would simply be $\{\widehat{\delta}_j\}$. It is intuitive that the maximizing $\{\delta_j\}$ deviate from $\{\widehat{\delta}_j\}$ approximately as they would if the data had a normal distribution. Therefore, consider the problem of finding the MLE of the means of k normal distributions with equal variances constrained by a linear combination of the means equalling zero. In that case the MLE for each mean is exactly the sample mean of that variable minus a factor times the linear combination of sample means, where the factor is proportional to one over the sample size of the variable. In our case one therefore expects

$$\tilde{\delta}_j = \widehat{\delta}_j - \frac{\delta^0}{n_j n^*}, \quad (11)$$

to be approximately the maximizing values, where $n^* = \sum \pi_j / n_j$. Now from (5) one has that for each j , β_j is the mean value for the first group with factor j while δ_j is the mean value for the second group with factor j minus the mean value for the first group with factor j . With these interpretations of $\{\beta_j\}$ and $\{\delta_j\}$, and given that $\{\delta_j\}$ must deviate from $\{\widehat{\delta}_j\}$, it is again intuitive from the normal case that $\{\beta_j\}$ should deviate from $\{\widehat{\beta}_j\}$ by a fraction of the deviation for $\{\delta_j\}$ where the fraction is determined by the relative sample size in the groups for factor j . leading to

$$\tilde{\beta}_j = \widehat{\beta}_j + \frac{n_{2j}}{(n_{1j} + n_{2j})} \frac{\delta^0}{n_j n^*}. \quad (12)$$

With these intuitive approximations, the problem simplifies considerably. From (8), and using (5) and (6), one obtains

$$\tilde{\Gamma}_{lj} = 0.$$

Multiplying Eq. (7) by p_{lji} and summing over i yields

$$n_{lj} = \lambda \sum_{i=1}^{n_{lj}} p_{lji}, \quad l=1,2, \quad j=1,\dots,k$$

which implies that $\sum p_{lji}$ is a constant times n_{lj} . The only solution can be $\tilde{\lambda} = n$ which gives

$$\sum_{i=1}^{n_{lj}} p_{lji} = n_{lj} / n, \quad l=1,2, \quad j=1,\dots,k$$

Now multiplying Eq. (7) by $p_{lji} \left(Y_{lji} - \tilde{\beta}_j - \tilde{\delta}_j I_{l=2} \right)$ and summing over i yields

$$\tilde{\Lambda}_{lj} = \sum_{i=1}^{n_{lj}} \frac{(Y_{lji} - \tilde{\beta}_j - \tilde{\delta}_j I_{l=2})}{\gamma_{lj}},$$

which leads (7) to become

$$\tilde{p}_{lji} = \left[n + \frac{n_{lj} (\bar{Y}_{lj} - \tilde{\beta}_j - \tilde{\delta}_j I_{l=2}) (Y_{lji} - \tilde{\beta}_j - \tilde{\delta}_j I_{l=2})}{\gamma_{lj}^2} \right]^{-1}, \quad l=1,2, \quad j=1,\dots,k, \quad i=1,\dots,n_{lj} \quad (13)$$

where $\bar{Y}_{lj} = \frac{1}{n_{lj}} \sum Y_{lji}$.

Notice that the problem has now been reduced to finding $\{\gamma_{lj}\}$ such that (4), (6), and $0 \leq \tilde{p}_{lji} \leq 1$ hold. In general it will not be possible to find such an exact solution. However, notice that if one relaxes the restriction (6), then the problem is reduced to finding for each group and factor a γ_{lj} such that

$$\sum_{i=1}^{n_{lj}} \frac{\gamma_{lj}^2}{n\gamma_{lj}^2 + n_{lj}c_{lji}} = 1, \quad (14)$$

where $c_{lji} = (\bar{Y}_{lj} - \tilde{\beta}_j - \tilde{\delta}_j I_{l=2})(Y_{lji} - \tilde{\beta}_j - \tilde{\delta}_j I_{l=2})$. This is a series of univariate numerical problems which have many solutions (the left-hand side of (14) is a ratio of polynomials in γ_{lj}). An appealing strategy is to use a univariate root finding method to find solutions to (14) from a grid of initial γ_{lj} , and then to take the solution that comes closest to satisfying (6).

Notice that $\tilde{p}_{lji} \geq 0$ implies a lower bound for $\tilde{\gamma}_{lj}^2$ of $\max \{-n_{lj}c_{lji}/n : i = 1, \dots, n_{lj}\}$. I have implemented a method which, for each combination of l and j , uses the FORTRAN routine ZREAL to find solutions to (14) for each initial γ_{lj} in a grid from the lower bound to 10 with grid size 0.1. The solution with minimum log of the ratio of right and left hand sides of (6) is taken as $\tilde{\gamma}_{lj}$. Refer to the resulting estimates marked with $\sim$ as the approximate EMLE constrained by H_{GE} .

The approximate ELRT test statistic is then

$$TS_{ELRT} = \sum_{l=1}^2 \sum_{j=1}^k \sum_{i=1}^{n_{lj}} \left(\log \hat{p}_{lji} - \log \tilde{p}_{lji} \right). \quad (15)$$

5 Simulations

The level and power of the testing procedures were compared by simulation of data from model (1) with $k = 2$. Four error distributions were considered: Gaussian, uniform, exponential, and Log-normal. each distribution was scaled and shifted to give zero mean and unit variance. Skewness and kurtosis for the scaled distributions are: Gaussian (0, 0), uniform (0, -1.2), exponential (2, 6), log-normal (2.8, 17.2). Each simulation had proportion of factor level one

of p_1 . Regardless of factor level, the group level was also one with proportion p_1 . Thus the simulations with $p_1 = 0.5$ had proportion 0.25 in each cell, while the simulations with $p_1 = 0.65$ had smallest cell with proportion 0.1225. Set $\pi_j = p_j$ for $j = 1, 2$. Each simulation under a null configuration consisted of 100 000 replications while each under a non-null configuration consisted of 10 000 replications.

Since the simulations are based on fairly small sample sizes and the asymptotic null distributions are not attainable, one might also consider approximating the null distribution by simulation from the AEMLE constrained by the null hypothesis. A similar approach was found to work quite well by Troendle (2002) and Fokianos and Troendle (2007) for testing the nonparametric Behrens-Fisher hypothesis in the two-sample problem. In this case, simulation from the constrained AEMLE means drawing deviates from

$$Y_{lji}^* = \tilde{\beta}_j + \tilde{\delta}_j I_{\{l=2\}} + \tilde{\gamma}_{lj} \tilde{\varepsilon}_{lji}, \quad l=1, 2, \quad j=1, \dots, k, \quad i=1, \dots, n_{lj},$$

where $\tilde{\varepsilon}_{lji}$ are iid from the discrete distribution determined by jumps of size p_{rsh} at residual $(Y_{rsh} - \tilde{\beta}_s - \tilde{\delta}_s I_{\{r=2\}}) / \tilde{\gamma}_{rs}$ for $r = 1, 2, s = 1, \dots, k, h = 1, \dots, n_{rs}$. To implement a test in this manner, m new data sets are generated by Monte Carlo simulation. Let T be the observed value of the test statistic for a particular test. The test statistic is then re-calculated for each of the m simulated data sets, denoting the b -th such value T^{*b} . The p -value used to determine acceptance/rejection is then calculated as in a random permutation test

$$p\text{-value} = \frac{\#\{b : T^{*b} \geq T\} + 1}{m+1}.$$

Each replication under the null hypothesis used $m = 19$ simulated data sets to approximate the null hypothesis, while each replication under an alternative hypothesis used $m = 199$ simulated data sets.

Tables 4 and 5 show the simulated type I error rates under certain null hypothesis configurations ($\delta_1 = \delta_2 = 0$) and desired type I error rate of 5%. Values of the variance parameters $\{\gamma_{lj}\}$ are given in matrix form as γ . Bold values indicate cases where the type I error rate exceeds two simulation standard errors over the desired 5%. Notice that the asymptotic null distributions are not reliable for type I error control when the variances are unequal. Furthermore, using simulation from the AEMLE to approximate the distribution under the null hypothesis improves somewhat the worst cases but does not yield strict type I error control. However, in all cases investigated, the AELRT using simulation from the AEMLE has type I error less than the desired upper bound of 5%.

Tables 6 and 7 show the simulated power under certain alternative configurations. To generate data under an alternative, allow $\delta_1 \neq 0$ as shown in the table. Since many of the tests presented in Tables 6 and 7 have poor type I error control, we focus on comparing only the TS_B and AELRT tests using simulation from the AEMLE to approximate the null distribution. Notice that while the TS_B procedure does have higher power than the AELRT, the power advantage seems to diminish in the cases with corresponding type I error for TS_B closer to the desired 5%. From this it appears that the AELRT is a robust choice for testing H_{GE} when not knowing anything about the error distribution or variances.

6 Discussion

Consider again the motivating analysis, described in Section 2, of folate levels in the blood for two alcohol intake groups (high and low) while accounting for gender. Recall that using an ordinary ANOVA model on the logged serum folate values gave a p -value for testing H_{GE} of 0.015 and for red cell folate 0.052. Using the AELRT with equal weights and using 99 999 simulated datasets to approximate the null distribution gives p -values 0.024 (simulation standard error approximately 0.0005) and 0.062 (simulation standard error approximately 0.0008) for serum and red cell folate respectively. The Box-type approximation method of Brunner et al. (1997) gives p -values that appear rather large ($p = 0.041$ for serum folate and $p = 0.127$ for red cell folate). When simulation from the AEMLE is used to approximate the null distribution of the heteroscedastic test statistic, the p -values are very similar to those of the AELRT ($p = 0.021$ for serum folate and $p = 0.063$ for red cell folate). The methods based on a heteroscedastic model are less enthusiastic about the significance of the alcohol effect. Since the method based on the ordinary ANOVA model requires the transformed data to be exactly Gaussian with equal variance to be valid, the AELRT from the heteroscedastic model is recommended.

The assumption of the classical ANOVA model is that the cells have equal variance and follow a Gaussian distribution. This assumption is almost always questionable, and yet there are no methods that give valid tests for a group main effect in a $2 \times k$ heteroscedastic model for small sample sizes. It has been shown here how to obtain an approximation to the empirical maximum likelihood estimate, constrained by the group main effect null hypothesis. Moreover, it has been shown how to create an approximate empirical likelihood ratio test for the group main effect null hypothesis. To decide critical cutoffs, one can approximate the distribution of the AELRT under the null hypothesis by simulation of data from the constrained AEMLE. The resulting test was shown to have appropriate level control (although conservative for small sample sizes), across a variety of error distributions and variance heterogeneity. This test should be used whenever the researcher is uncertain about variance homogeneity or the error distribution.

Acknowledgements

This research was supported by the Intramural Research Program of the NIH, NICHD. However, the ideas and statements are the authors and do not represent those of the NIH or the US Government.

References

- Akritis MG, Arnold SF, Brunner E. Nonparametric Hypotheses and Rank Statistics for Unbalanced Factorial Designs. *Journal of the American Statistical Association* 1997;92:258–265.
- Brown MB, Forsythe AB. The Small Sample Behaviour of Some Statistics Which Test the Equality of Several Means. *Technometrics* 1974;16:129–132.
- Brunner E, Dette H, Munk A. Box-Type Approximations in Nonparametric Factorial Designs. *Journal of the American Statistical Association* 1997;92:1494–1502.
- Fokianos K, Troendle JF. Inference for the Relative Treatment Effect With the Density Ratio Model. *Statistical Modeling* 2007;7:155–173.
- Keselman HJ, Carriere KC, Lix LM. Robust and Powerful Nonorthogonal Analyses. *Psychometrika* 1995;60:395–418.
- Luh W-M, Guo J-H. Using Johnson's Transformation and Robust Estimators With Heteroscedastic Test Statistics: An Examination of the Effects of Non-Normality and Heterogeneity in the Non-Orthogonal Two-Way ANOVA Design. *Journal of Mathematical and Statistical Psychology* 2001;54:79–94.
- Owen, AB. Empirical Likelihood. Chapman & Hall/CRC; Boca Raton: 2001.
- Puri, ML.; Sen, PK. Nonparametric Methods in General Linear Models. John Wiley & Sons; New York: 1985.

- Scheffe, H. The Analysis of Variance. John Wiley & Sons; New York: 1959.
- Toothaker LE, Newman D. Nonparametric Competitors to the Two-Way ANOVA. Journal of Educational and Behavioral Statistics 1994;19:237–273.
- Troendle JF. A Likelihood Ratio Test for the Nonparametric Behrens-Fisher Problem. Biometrical Journal 2002;7:813–824.
- Welch BL. On the Comparison of Several Mean Values. Biometrika 1951;38:330–336.

Table 1
Measures of normality for cells of two-way ANOVA for serum folate.

Cell	Sample Size	p-value	Shapiro-Wilk		Skewness	Kurtosis
			S.D.			
Serum Folate						
High Alcohol Females	186	<0.0001	7.79		1.55	2.90
High Alcohol Males	256	<0.0001	6.43		1.02	0.97
Low Alcohol Females	514	<0.0001	8.21		1.05	1.16
Low Alcohol Males	143	<0.0001	7.46		1.02	0.54
Log (Serum Folate)						
High Alcohol Females	186	0.35	0.55		-0.00	-0.23
High Alcohol Males	256	0.43	0.48		-0.23	-0.03
Low Alcohol Females	514	0.002	0.57		-0.31	-0.25
Low Alcohol Males	143	0.55	0.51		-0.05	-0.42

Troendle

Table 2
Measures of normality for cells of two-way ANOVA for red cell folate.

Cell	Sample Size	p-value	Shapiro-Wilk S.D.	Skewness	Kurtosis
Red Cell Folate					
High Alcohol Females	186	<0.0001	174.40	1.06	1.46
High Alcohol Males	257	0.0002	159.72	0.68	0.87
Low Alcohol Females	514	<0.0001	190.35	0.93	1.28
Low Alcohol Males	143	<0.0001	191.83	1.14	2.51
Log (Red Cell Folate)					
High Alcohol Females	186	0.71	0.40	-0.15	0.10
High Alcohol Males	257	0.05	0.35	-0.42	0.39
Low Alcohol Females	514	0.09	0.42	-0.26	0.08
Low Alcohol Males	143	0.28	0.38	-0.13	0.34

Table 3

Results of tests for main effect of alcohol.

	Estimate	<i>t</i> -value	<i>p</i> -value
Log (Serum Folate) Alcohol = High	-0.09	-2.45	0.015
Log (Red Cell Folate) Alcohol = High	-0.05	-1.94	0.052

Table 4

Type I error (%) of the tests for null configurations. N , U , E , and L respectively denotes a Gaussian, Uniform, Exponential, and Log-Normal distribution.

γ	n	p_1	ε	Asymptotic Null			Simulated AEMLE		
				TS_F	TS_B	$AELRT$	TS_F	TS_B	$AELRT$
$\begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{pmatrix}$	50	0.5	N	5.04	4.81	4.89	4.89	4.89	4.17
			U	5.01	4.86	4.93	4.93	4.93	4.33
			E	4.77	4.44	4.64	4.64	4.66	3.71
	100	0.65	L	4.53	4.24	4.39	4.40	4.40	3.66
			N	4.90	4.99	5.07	4.94	4.94	4.16
			U	4.92	4.96	5.19	4.97	4.97	4.49
	100	0.35	E	4.85	5.10	5.12	4.99	4.99	3.38
			L	4.79	4.89	4.98	4.86	4.86	3.38
			N	4.96	4.90	5.17	4.97	4.97	4.17
	100	0.5	U	4.95	4.96	5.31	4.93	4.93	4.41
			E	4.85	5.11	5.19	5.04	5.04	3.20
			L	4.71	4.98	5.08	5.00	5.00	3.16
$\begin{pmatrix} 1.0 & 1.0 \\ 1.5 & 2.0 \end{pmatrix}$	100	0.5	N	4.97	4.92	4.91	4.91	4.91	4.26
			U	5.03	4.92	4.98	4.98	4.98	4.69
			E	5.00	4.75	4.78	4.78	4.78	3.35
	50	0.5	L	4.67	4.65	4.66	4.66	4.66	3.19
			N	4.71	4.79	4.91	5.00	5.00	3.32
			U	4.82	4.85	4.93	5.00	5.00	3.78
	100	0.65	E	5.38	5.30	5.24	5.31	5.31	3.42
			L	5.44	5.36	5.33	5.41	5.41	3.65
			N	8.72	4.83	5.17	4.86	4.86	4.35
	100	0.35	U	8.78	4.93	5.24	4.94	4.94	4.70
			E	8.97	6.02	6.52	5.72	5.72	3.94
			L	8.95	6.18	6.38	5.86	5.86	4.13
	100	0.35	N	2.01	4.98	4.78	4.87	4.87	3.84
			U	1.97	5.03	4.95	4.93	4.93	4.10

γ	n	p_1	ε	Asymptotic Null		Simulated AEMLE		
				TS_F	TS_B	TS_F	TS_B	AELRT
	100	0.5	E	2.48	4.89	4.93	4.88	2.69
			L	2.70	4.76	4.82	4.71	2.64
			N	5.13	4.96	4.95	4.95	4.19
			U	5.15	4.99	4.97	4.97	4.54
			E	5.48	5.35	5.25	5.25	3.64
			L	5.46	5.43	5.34	5.34	3.59

Table 5
Type I error (%) of the tests for null configurations. N , U , E , and L respectively denotes a Gaussian, Uniform, Exponential, and Log-Normal distribution.

γ	n	p_1	ε	Asymptotic Null		Simulated AEMLE		
				TS_F	TS_B	TS_F	TS_B	AELRT
$\begin{pmatrix} 1.0 & 1.0 \\ 2.0 & 3.0 \end{pmatrix}$	50	0.5	N	4.73	4.82	4.79	4.94	2.05
			U	4.91	4.94	4.70	4.80	2.37
			E	6.20	6.24	5.95	6.06	3.02
			L	6.43	6.64	6.19	6.35	3.34
	100	0.65	N	11.10	4.92	5.23	4.90	4.11
			U	11.13	5.04	5.22	4.97	4.61
			E	11.73	6.50	7.23	6.14	4.63
			L	11.95	6.83	7.24	6.40	4.76
	100	0.35	N	1.34	4.98	4.46	4.85	3.75
			U	1.30	5.02	4.60	4.83	4.16
			E	2.11	5.34	4.82	5.21	2.73
			L	2.29	5.45	4.89	5.28	2.79
	100	0.5	N	5.23	4.95	4.91	4.91	4.09
			U	5.28	5.06	5.02	5.02	4.59
			E	6.10	5.86	5.67	5.67	3.74
			L	6.21	6.14	5.91	5.91	3.91

Table 6

Power (%) of the tests for alternative configurations. N , U , E , and L respectively denotes a Gaussian, Uniform, Exponential, and Log-Normal distribution.

γ	n	ϕ_1	p_1	ε	Asymptotic Null			Simulated AEMLE		
					TS_F	TS_B	TS_F	TS_B	TS_F	AELRT
$\begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{pmatrix}$	50	1.7	0.5	N	83.4	82.6	82.5	82.6	82.6	74.7
				U	83.3	83.3	82.7	82.7	82.7	76.4
				E	83.4	82.7	81.9	81.9	81.9	65.5
				L	84.4	83.1	82.7	82.6	82.6	64.1
	100	0.9	0.65	N	79.0	78.7	78.6	78.3	78.3	74.9
				U	79.1	78.8	79.2	77.8	77.8	75.6
				E	79.8	81.8	82.2	81.1	81.1	73.3
				L	80.6	83.1	83.1	82.1	82.1	72.2
	100	1.7	0.35	N	79.9	79.1	79.2	78.3	78.3	75.0
				U	79.7	79.4	80.3	78.4	78.4	76.0
				E	80.2	77.0	77.2	75.7	75.7	63.0
				L	81.2	77.9	77.2	76.4	76.4	66.2
$\begin{pmatrix} 1.0 & 1.0 \\ 1.5 & 2.0 \end{pmatrix}$	100	1.2	0.5	N	84.9	84.8	84.0	84.0	84.0	82.0
				U	84.8	84.6	83.7	83.7	83.7	83.1
				E	84.5	83.8	83.9	83.9	83.9	77.8
				L	85.3	84.6	84.4	84.4	84.4	77.4
	50	2.3	0.5	N	77.8	78.9	77.4	77.7	77.7	37.4
				U	77.9	79.0	77.6	77.9	77.9	35.4
				E	83.8	84.0	83.2	83.6	83.6	41.2
				L	85.5	84.9	84.8	85.2	85.2	44.8
	100	1.2	0.65	N	78.5	69.1	69.5	68.5	68.5	64.7
				U	78.5	68.5	68.9	68.0	68.0	66.5
				E	82.2	75.4	77.6	72.8	72.8	62.7
				L	83.2	77.1	79.8	75.1	75.1	61.2
	100	2.3	0.35	N	67.5	80.2	78.9	79.3	79.3	71.2
				U	67.1	80.2	79.3	79.3	79.3	73.7

γ	n	δ_1	p_1	ε	Asymptotic Null			Simulated AEMLE		
					TS_F	TS_B		TS_F	TS_B	AELRT
	100	1.8	0.5	E	71.3	81.8		81.9	80.7	54.3
				L	73.4	82.2		82.0	81.5	52.3
				N	87.7	87.3		86.8	86.8	79.5
				U	87.4	87.5		86.7	86.7	81.0
				E	91.0	90.0		90.3	90.3	73.6
				L	92.0	91.2		91.3	91.3	71.8

Table 7
Power (%) of the tests for alternative configurations. N , U , E , and L respectively denotes a Gaussian, Uniform, Exponential, and Log-Normal distribution.

γ	n	ϕ_1	p_1	ε	Asymptotic Null			Simulated AEMLE		
					TS_F	TS_B	$AELRT$	TS_F	TS_B	$AELRT$
$\begin{pmatrix} 1.0 & 1.0 \\ 2.0 & 3.0 \end{pmatrix}$	50	3.2	0.5	N	80.2	81.6	20.1	78.3	78.7	20.1
				U	80.3	81.5	15.8	78.2	78.5	15.8
				E	89.0	89.0	26.1	88.2	88.9	26.1
	100	1.6	0.65	L	90.4	90.1	30.7	89.5	90.0	30.7
				N	80.4	67.0	49.0	67.7	66.5	49.0
				U	80.1	66.5	50.4	66.7	65.3	50.4
	100	3.0	0.35	E	85.9	74.4	48.3	77.3	71.0	48.3
				L	87.3	76.6	49.5	80.1	74.2	49.5
				N	59.8	79.5	57.4	76.3	78.4	57.4
	100	2.2	0.5	U	59.3	79.8	60.2	77.0	78.6	60.2
				E	63.8	83.5	45.0	79.0	82.5	45.0
				L	66.5	84.1	43.5	79.2	83.6	43.5
	100			N	80.2	79.9	45.6	78.8	78.8	45.6
				U	80.0	79.9	47.0	78.4	78.4	47.0
				E	86.2	84.8	43.7	84.6	84.6	43.7
				L	88.0	87.1	43.2	86.3	86.3	43.2