Citation:

Hartung, J., Argaç, D. & Makambi, K. (2003). Homogeneity tests in meta-analysis. In R. Schulze, H. Holling & D. Böhning (Eds.), *Meta-analysis: New developments and applications in medical and social sciences* (pp. 3–20). Hogrefe & Huber.

1

Homogeneity Tests in Meta-Analysis

Joachim Hartung
Doğan Argaç
Department of Statistics[†]
University of Dortmund

Kepher Makambi
Department of Mathematics and Statistics
Jomo Kenyatta University of Agriculture and Technology

Summary

For the homogeneity problem in meta-analysis, the performance of seven test statistics is compared under homogeneity and heterogeneity of the underlying population (study, group) variances. These are: the classical ANOVA F test, the Cochran test, the Welch test, the Brown-Forsythe test, the modified Brown-Forsythe test, the approximate ANOVA F test and as a proposal, an adjusted Welch test. At the whole, the Welch test proves to be the best one, but for small sample sizes and many groups, it becomes too liberal. In this case the adjusted Welch test is recommended to correct this anomaly. The other tests prove to have changing advantages dependent on the sizes of the parameters involved.

[†]Project "Meta-Analysis in Biometry and Epidemiology" (SFB 475) of the Deutsche Forschungsgemeinschaft (DFG).

1.1 INTRODUCTION

Meta-analysis of results from different experiments (groups, studies) is a common practice nowadays. In the framework of a one-way ANOVA model, serving generally as supporting edifice for meta-analysis, one may be interested in testing the homogeneity hypothesis. However, when the underlying population variances in different populations (studies, groups) are different, the ANOVA F-statistic attains significance levels which are very different from the nominal level (see for example, De Beuckelaer, 1996). In the rubric of the (generalized) Behrens-Fisher problem, a number of alternatives have been suggested.

Using simulation studies for various constellations of number of populations, sample sizes and within population error variances, we compare the actual attained sizes of the classical ANOVA F test, the Cochran test, the Welch test, the Brown-Forsythe test, the modified Brown-Forsythe test, the approximate ANOVA F test and, by adopting an idea of Böckenhoff and Hartung (1998), an adjusted Welch test, simultaneously.

1.2 MODEL AND TEST STATISTICS

Let y_{ij} be the observation on the jth subject of the ith population/study, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$

$$\begin{aligned} y_{ij} &= \mu_i + e_{ij} \\ &= \mu + a_i + e_{ij}\,; \quad i = 1, \ldots, K,\ j = 1, \ldots, n_i, \end{aligned}$$

where μ is the common mean for all the K populations, a_i is the effect of population i with $\sum_{i=1}^{K} a_i = 0$, and e_{ij}, $i = 1, \ldots, K$, $j = 1, \ldots, n_i$ are error terms which are assumed to be mutually independent and normally distributed with

$$E(e_{ij}) = 0, \quad \mathrm{Var}(e_{ij}) = \sigma_i^2\,;\ i = 1, \ldots, K,\ j = 1, \ldots, n_i$$

That is, $e_{ij} \sim \mathcal{N}(0, \sigma_i^2)$; $i = 1, \ldots, K$, $j = 1, \ldots, n_i$.

Interest is in testing the hypothesis $H_0 : \mu_1 = \cdots = \mu_K = \mu$. To test this hypothesis we will make use of the following test statistics:

a) The ANOVA *F* Test

S_{an}, given by

$$S_{an} = \frac{N-K}{K-1} \cdot \frac{\sum_{i=1}^{K} n_i (\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K} (n_i - 1) s_i^2}, \tag{1.1}$$

with $N = \sum_{i=1}^{K} n_i$, $\bar{y}_{i.} = \sum_{j=1}^{K} y_{ij}/n_i$, $\bar{y}_{..} = \sum_{i=1}^{K} n_i \bar{y}_{i.}/N$.

This test was originally meant to test for equality of population means under variance homogeneity and has an F distribution with $K-1$ and

$N - K$ degrees of freedom.

Test: Reject $H_0 : \mu_1 = \cdots = \mu_K$ at level α if $S_{an} > F_{K-1,N-K;1-\alpha}$.

The ANOVA test has the weakness of not being robust with respect to heterogeneity in the intra-population error variances (Brown & Forsythe, 1974).

b) The Welch Test

$$S_{we} = \frac{\sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2}{\left((K-1) + 2 \cdot \frac{K-2}{K+1} \cdot \sum_{i=1}^{K} \frac{1}{n_i - 1}(1-h_i)^2\right)}, \tag{1.2}$$

where $w_i = n_i / s_i^2, \quad h_i = w_i / \sum_{k=1}^{K} w_k$, was an extension of testing the equality of two means to more than two means (see Welch, 1951) in the presence of variance heterogeneity within populations.

Under H_0, the statistic S_{we} has an approximate F distribution with $K-1$ and ν_g degrees of freedom, where

$$\nu_g = \frac{(K^2-1)/3}{\sum_{i=1}^{K} \frac{1}{n_i - 1}(1-h_i)^2}.$$

Test: Reject H_0 at level α if $S_{we} > F_{K-1,\nu_g;1-\alpha}$.

c) Cochran's Test

$$S_{ch} = \sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2, \tag{1.3}$$

was proposed by Cochran (1937) and then modified by Welch. We take it into our comparisons in order to get better comprehension and insight of the behavior of both statistics.

Under H_0, the Cochran statistic is distributed approximately as a χ^2-variable with $K-1$ degrees of freedom.

Test: Reject H_0 at level α if $S_{ch} > \chi^2_{K-1;1-\alpha}$.

d) Brown-Forsythe (B-F) Test

This one is also known as the modified F test and is given by

$$S_{b-f} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}. \tag{1.4}$$

When H_0 is true, S_{b-f} is distributed approximately as an F variable with $K-1$ and ν degrees of freedom where

$$\nu = \frac{\left(\sum_{i=1}^{K}(1 - n_i/N)s_i^2\right)^2}{\sum_{i=1}^{K}(1 - n_i/N)^2 s_i^4/(n_i - 1)}. \tag{1.5}$$

Test: Reject H_0 at level α if $S_{b-f} > F_{K-1,\nu;1-\alpha}$.

Using a simulation study Brown and Forsythe (1974) demonstrated that their statistic is robust under inequality of variances. If the population variances are homogeneous, the B-F test is closer to ANOVA than Welch.

e) Mehrotra (Modified Brown-Forsythe) Test

$$S_{b-f(m)} = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}, \tag{1.6}$$

was proposed by (Mehrotra, 1997) in an attempt to correct a "flaw" in the B-F test.

Under H_0, $S_{b-f(m)}$ is distributed approximately as an F variable with ν_1 and ν degrees of freedom where

$$\nu_1 = \frac{\left(\sum_{i=1}^{K}(1 - n_i/N)s_i^2\right)^2}{\sum_{i=1}^{K} s_i^4 + \left(\sum_{i=1}^{K} n_i s_i^2/N\right)^2 - 2 \cdot \sum_{i=1}^{K} n_i s_i^4/N} \tag{1.7}$$

and ν is given in Equation 1.5 above.

Test: Reject H_0 at level α if $S_{b-f(m)} > F_{\nu_1,\nu;1-\alpha}$.

The flaw mentioned above is in the estimation of the numerator degrees of freedom by $K-1$ instead of ν_1.

f) The Approximate ANOVA *F* Test

$$S_{aF} = \frac{N-K}{K-1} \cdot \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(n_i-1)s_i^2}, \tag{1.8}$$

by Asiribo and Gurland (1990). This test gives an approximate solution to the problem of testing equality of means of normal populations in case of heteroscedasticity by making use of the classical ANOVA test.

Under H_0, the statistic S_{aF} is distributed approximately as an F-variable with ν_1 and ν_2 degrees of freedom where ν_1 is as given in Equation 1.7 above and

$$\nu_2 = \frac{\left(\sum_{i=1}^{K}(n_i-1)s_i^2\right)^2}{\sum_{i=1}^{K}(n_i-1)s_i^4}. \tag{1.9}$$

Test: Reject H_0 at level α if $S_{aF} > \hat{c} \cdot F_{\nu_1,\nu_2;1-\alpha}$, where

$$\hat{c} = \frac{N-K}{N(K-1)} \frac{\sum_{i=1}^{K}(N-n_i)s_i^2}{\sum_{i=1}^{K}(n_i-1)s_i^2}. \tag{1.10}$$

We notice that the numerator degrees of freedom for S_{aF} and $S_{b-f(m)}$ are equal. Further, for $n_i = n, i = 1, \ldots, K$, that is, for balanced samples, the test statistic and the degrees of freedom for both the numerator and denominator of these two statistics are also equal. That is, for balanced designs

$$S_{aF} = S_{b-f(m)} = \frac{nK}{K-1} \cdot \frac{\sum_{i=1}^{K}(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K} s_i^2},$$

and

$$\nu = \nu_2 = (n-1) \cdot \frac{\left(\sum_{i=1}^{K} s_i^2\right)^2}{\sum_{i=1}^{K} s_i^4}.$$

g) The Adjusted Welch Test

The Welch Test uses weights $w_i = n_i/s_i^2$. We know that

$$E(w_i) = E\left(\frac{n_i}{s_i^2}\right) = c_i \cdot \frac{n_i}{\sigma_i^2},$$

where $c_i = (n_i - 1)/(n_i - 3)$, see Patel, Kapadia, and Owen (1976, pages 39-40). Therefore, an unbiased estimator of n_i/σ_i^2 is $n_i/c_i s_i^2$.

Now, let $\varphi_i = (n_i + \delta_1)/(n_i + \delta_2)$, where δ_1 and δ_2 are arbitrary real numbers; and then define the general weights by $w_i^* = n_i/\varphi_i s_i^2$. That is, for the Welch test, $w_i = w_i^*$ with $\varphi_i = 1$ ($\delta_1 = 0$, and $\delta_2 = 0$) and if we take the unbiased weights, $w_i = n_i/c_i s_i^2$, then $\varphi_i = c_i$, ($\delta_1 = -1$ and $\delta_2 = -3$).

For small samples in the groups, the Welch test becomes too liberal especially with increasing number of groups. Also, in our experience, using the unbiased weights in the Welch test makes the test too conservative. A reasonable compromise in this situation is to choose φ_i such that $1 \leq \varphi_i \leq c_i$.

This defines a new class of Welch type test statistics whose properties can be adjusted accordingly by choosing the control parameter, φ_i, appropriately. Our proposed test, which we shall henceforth call the *adjusted Welch test*, uses the weights $w_i^* = n_i/\varphi_i s_i^2$ in the Welch test, where $1 \leq \varphi_i \leq c_i$. That is the adjusted Welch test, S_{aw}, is given by:

$$S_{aw} = \frac{\sum_{i=1}^{K} w_i^* (\bar{y}_{i.} - \sum_{j=1}^{K} h_j^* \bar{y}_{j.})^2}{\left((K-1) + 2 \cdot \frac{K-2}{K+1} \cdot \sum_{i=1}^{K} \frac{1}{n_i - 1}(1 - h_i^*)^2\right)}, \tag{1.11}$$

where $h_i^* = w_i^* / \sum_{i=1}^{K} w_i^*,\ i = 1, \ldots, K$.
Under H_0, the adjusted Welch statistic, S_{aw}, is distributed approximately as an F-variable with $K - 1$ and ν_g^* degrees of freedom, with

$$\nu_g^* = \frac{(K^2 - 1)/3}{\sum_{i=1}^{K} \frac{1}{n_i - 1}(1 - h_i^*)^2}.$$

Test: Reject H_0 at α level if $S_{aw} > F_{K-1,\nu_g^*;1-\alpha}$.

When the sample sizes are large, S_{aw} approaches the Welch test, that is, $(n_i + \delta_1)/(n_i + \delta_2) \overset{n_i \to \infty}{\longrightarrow} 1$. With small sample sizes, our statistic will help correct the liberality witnessed in the Welch test.

To assess the relative performance of these test statistics in terms of the actual levels of significance attained, we will consider levels between 4% and 6% to be satisfactory, that is, following Cochran's rule of thumb (cf. Cochran, 1954).

1.3 SIMULATION STUDY AND DISCUSSION

In order to see the effect of balancedness and unbalancedness, as well as variance homogeneity and heterogeneity, a simulation study was conducted with sampling experiments determined by the number of studies, sample sizes and the variances in each study. In the first sampling experiment the following patterns and combinations of the number of studies, sample sizes and variances were considered (cf. Tables 1.1, 1.2, 1.3, and 1.4): Balanced samples and homogeneous variances, unbalanced samples combined with homogeneous variances. The next experiment investigated the effect of variance heterogeneity on the empirical Type I error rates. We matched balanced and unbalanced sample sizes with heterogeneous variances. In the unbalanced sample size cases, large sample sizes were separately paired with small and large variances. To investigate the effect of a large number of studies, we started with $K = 3$ studies and made independent replications to give $K = 6, 2 \times (.)$, $K = 9$, $3 \times (.)$, and $K = 18, 6 \times (.)$. We will use the term small sample to refer to $n_i = 5$, and moderate for $n_i = 10$, $15, i = 1, \ldots, K$. However, if any of the sample sizes, n_i, is greater or equal to 20, then the constellation will be taken to be of large samples.

Table 1.1 reports the actual significance levels for $K = 3$, Table 1.2 for $K = 6$, Table 1.3 for $K = 9$ and Table 1.4 for $K = 18$. For the adjusted Welch test, S_{aw}, we have taken $\varphi_i = (n_i + 2)/(n_i + 1), i = 1, \ldots, K$. From these Tables, we make the following observations in order of the various tests presented in Section 1.2 above:

a) The ANOVA *F* Test

In the case when the number of populations, $K = 3$:

i. for balanced samples sizes and homoscedastic cases, the test, as expected, keeps the nominal level;

ii. for balanced and heterogeneous variance cases, the test keeps control of the significance level. This trend is maintained with increasing sample sizes;

iii. for unbalanced and homoscedastic cases, the test keeps the nominal level;

iv. for the unbalanced and heterogeneous cases, if small samples are matched with small variances, the test tends to be conservative. However, when small sample sizes are paired with large variances, the test becomes liberal. This pattern remains largely unchanged even if the sample sizes are increased.

For $K = 6$, $K = 9$ and $K = 18$, the observations made in i. to iv. above still hold; except for balanced designs and heterogeneous variances where the test becomes more liberal with increasing number of populations.

Table 1.1 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 3$

Sample Sizes	Variances	$\hat{\alpha}\%$						
(n_1, n_2, n_3)	$(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.0	4.8	12.2	4.1	3.8	3.8	3.3
	(1,3,5)	6.0	5.0	13.5	4.6	4.2	4.2	3.6
(10,10,10)	(4,4,4)	5.1	4.9	8.4	4.9	4.6	4.6	3.9
	(1,3,5)	5.7	4.7	8.2	5.1	4.5	4.5	3.9
(20,20,20)	(4,4,4)	5.1	4.9	6.5	5.0	4.9	4.9	4.2
	(1,3,5)	5.6	4.8	6.4	5.4	4.7	4.7	4.2
(40,40,40)	(4,4,4)	4.9	4.9	5.6	4.8	4.8	4.8	4.5
	(1,3,5)	5.9	5.2	5.8	5.8	5.0	5.0	4.8
(5,10,15)	(4,4,4)	5.0	5.3	10.2	5.1	4.8	5.4	4.2
	(1,3,5)	2.4	4.9	8.9	5.6	4.7	4.5	3.8
	(5,3,1)	12.3	5.4	11.5	5.3	5.0	6.2	4.4
(10,20,30)	(4,4,4)	5.2	5.3	7.7	5.1	4.9	5.3	4.5
	(1,3,5)	2.2	4.9	6.5	5.5	4.6	4.5	4.2
	(5,3,1)	12.9	5.5	8.1	5.6	5.2	5.9	4.5
(20,40,60)	(4,4,4)	4.8	4.9	5.9	4.9	4.7	4.9	4.4
	(1,3,5)	2.1	5.1	5.8	5.7	4.7	4.6	4.5
	(5,3,1)	12.5	4.9	6.4	5.5	5.0	5.4	4.4

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.2 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 6$

Sample Sizes	Variances	$\hat{\alpha}\%$						
$2\times$ (n_1, n_2, n_3)	$2\times$ $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.2	6.2	22.1	4.1	3.3	3.3	4.1
	(1,3,5)	6.6	6.1	22.4	4.8	3.7	3.7	4.3
(10,10,10)	(4,4,4)	5.1	5.1	11.4	4.8	4.2	4.2	3.7
	(1,3,5)	6.3	5.2	12.0	5.6	4.3	4.3	3.7
(20,20,20)	(4,4,4)	4.8	4.7	7.7	4.7	4.3	4.3	3.8
	(1,3,5)	6.0	4.8	7.7	5.7	4.4	4.4	4.0
(40,40,40)	(4,4,4)	4.7	4.6	6.0	4.6	4.4	4.4	4.2
	(1,3,5)	6.8	5.4	6.9	6.6	5.0	5.0	4.9
(5,10,15)	(4,4,4)	5.0	6.3	15.5	4.7	4.0	4.5	4.7
	(1,3,5)	2.4	5.5	13.1	5.9	4.3	4.2	3.8
	(5,3,1)	16.3	6.7	16.7	5.7	4.6	5.5	5.0
(10,20,30)	(4,4,4)	5.5	5.7	9.7	5.2	4.7	4.9	4.8
	(1,3,5)	2.3	5.2	8.3	6.5	4.8	4.7	4.2
	(5,3,1)	16.3	5.7	10.2	6.3	4.8	5.5	4.7
(20,40,60)	(4,4,4)	5.2	5.3	7.2	5.2	4.8	5.0	4.6
	(1,3,5)	2.6	5.5	7.1	6.7	5.1	5.0	4.7
	(5,3,1)	15.3	4.8	6.7	6.3	4.9	5.2	4.1

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.3 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 9$

Sample Sizes	Variances	$\hat{\alpha}\%$						
$3\times$ (n_1, n_2, n_3)	$3\times$ $(\sigma_1^2, \sigma_2^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	5.3	7.3	28.6	4.3	3.2	3.2	4.7
	(1,3,5)	6.5	7.8	28.7	4.7	3.3	3.3	5.1
(10,10,10)	(4,4,4)	5.1	6.2	14.8	4.9	4.0	4.0	4.3
	(1,3,5)	7.0	6.0	14.5	6.2	4.5	4.5	4.3
(20,20,20)	(4,4,4)	5.2	5.4	9.1	5.1	4.6	4.6	4.4
	(1,3,5)	6.6	5.1	9.1	6.3	4.5	4.5	4.2
(40,40,40)	(4,4,4)	5.0	5.2	7.0	5.0	4.7	4.7	4.5
	(1,3,5)	6.6	4.9	6.9	6.5	4.8	4.8	4.4
(5,10,15)	(4,4,4)	5.3	7.0	19.3	4.9	4.1	4.5	4.9
	(1,3,5)	2.2	6.6	16.9	6.2	4.1	4.0	4.6
	(5,3,1)	18.7	7.6	20.9	5.5	4.1	5.1	5.5
(10,20,30)	(4,4,4)	4.9	5.5	10.7	4.8	4.3	4.4	4.1
	(1,3,5)	2.1	5.2	9.6	6.4	4.6	4.5	4.0
	(5,3,1)	17.6	5.9	10.7	5.8	4.3	4.8	4.6
(20,40,60)	(4,4,4)	5.2	5.5	8.0	5.3	5.1	5.2	4.9
	(1,3,5)	2.3	5.4	7.3	7.0	5.2	5.1	4.0
	(5,3,1)	18.1	5.3	7.6	6.4	4.8	4.9	4.5

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

Table 1.4 Actual Simulated Significance Levels (Nominal Level 5%) for $K = 18$

Sample Sizes	Variances	$\hat{\alpha}$%						
$6\times$ (n_1, n_2, n_3)	$6\times$ $(\sigma_1^2, \sigma_3^2, \sigma_3^2)$	S_{an}	S_{we}	S_{ch}	S_{b-f}	$S_{b-f(m)}$	S_{aF}	S_{aw}
(5,5,5)	(4,4,4)	4.9	11.7	46.3	3.8	2.5	2.5	7.1
	(1,3,5)	7.1	12.4	46.7	4.1	3.2	3.2	7.8
(10,10,10)	(4,4,4)	5.3	7.0	20.9	5.1	4.0	4.0	4.4
	(1,3,5)	7.0	6.8	21.1	6.2	3.8	3.8	4.2
(20,20,20)	(4,4,4)	4.8	5.2	11.3	4.8	4.2	4.2	4.1
	(1,3,5)	7.0	5.2	11.0	6.7	4.7	4.7	4.0
(40,40,40)	(4,4,4)	4.8	5.3	7.9	4.8	4.6	4.6	4.4
	(1,3,5)	7.1	5.3	8.0	6.9	5.0	5.0	4.4
(5,10,15)	(4,4,4)	5.2	9.7	28.6	4.8	3.5	4.0	6.4
	(1,3,5)	1.7	8.2	26.6	6.8	4.5	4.4	5.1
	(5,3,1)	24.9	10.2	30.0	5.7	3.7	4.6	7.1
(10,20,30)	(4,4,4)	5.3	6.2	14.0	5.3	4.5	4.7	4.4
	(1,3,5)	1.5	5.9	13.0	7.0	4.5	4.4	4.2
	(5,3,1)	24.1	6.1	14.4	6.5	4.4	4.8	4.2
(20,40,60)	(4,4,4)	4.8	5.2	8.5	4.8	4.3	4.4	4.0
	(1,3,5)	1.5	5.3	8.5	6.5	4.4	4.4	4.5
	(5,3,1)	23.2	5.0	8.2	6.3	4.3	4.5	4.0

Note. For a definition of S_{an}, S_{we}, S_{ch}, S_{b-f}, $S_{b-f(m)}$, S_{aF}, and S_{aw} see Equations 1.1, 1.2, 1.3, 1.4, 1.6, 1.8, and 1.11.

b) The Welch Test

For $K = 3$:

i. for balanced sample sizes and homoscedastic cases, the test keeps the required nominal level. Increasing the individual sample sizes has no significant effect on the attained levels of significance;

ii. for balanced and heterogeneous variances, the test keeps the nominal level and this is not significantly affected by enlarging the sample sizes;

iii. for unbalanced and homogeneous variance case, the test performs well and there is no significant effect in increasing the sample sizes;

iv. for unbalanced and heterogeneous variance cases, the test attains acceptable significance levels both for small and large sample sizes.

For $K = 6$:

i. for balanced sample sizes and homoscedastic cases, the test keeps the required nominal level but seems to be a bit liberal when the sample sizes are small. Increasing the individual sample sizes has the effect of making the attained levels of significance closer to the nominal level;

ii. for balanced and heterogeneous variances, the test keeps the nominal level but, is a bit more liberal when the sample sizes are small;

iii. for unbalanced and homogeneous variance cases, if one of the sample sizes is small, the test becomes liberal otherwise, it attains acceptable levels of significance;

iv. The test attains acceptable significance levels for unbalanced and heteroscedastic cases, except when one of the sample sizes is small and is paired with a large variance.

For $K = 9$:

i. for balanced samples sizes and homoscedastic cases, the Welch test is liberal for small samples but works quite well for moderate to large sample sizes;

ii. for balanced and heterogeneous variances, the test is liberal for small variance cases but, works well for moderate to large samples;

iii. for unbalanced and homogeneous variance cases, for small sample sizes the test becomes liberal but, in general it attains acceptable levels of significance;

iv. in the case of unbalanced samples and heterogeneous variances, the test is slightly liberal when one of the sample sizes is small. Otherwise, increasing the sample sizes makes the test better.

For $K = 18$:

i. for balanced samples and equal variances in all the groups, the test is liberal for small and moderate samples. For large samples, the test keeps good control of the significance level;

ii. for equal sample sizes in all the groups and heterogeneous variances, the test is liberal for small and moderate samples but, attains acceptable significance levels for large sample sizes;

iii. for unbalanced and homogeneous variances, the test attains acceptable significance levels only when all the sample sizes are above or equal to 20. Otherwise, the test is liberal;

iv. for unbalanced samples and unequal variances in the groups, the test controls the nominal level when all the sample sizes are equal or greater than 20. When one of the samples is of size 10 or less, the test is more liberal when relatively large variances are combined with relatively small sample sizes.

c) Cochran's Test

For $K = 3$:

i. for balanced sample sizes and homogeneous variances, the test is largely liberal. This liberality is much more pronounced for very small samples and reduces drastically for moderate to large samples;

ii. for balanced sample sizes and heterogeneous variances, the test attains levels which are above the acceptable upper limit but, there is some improvement as the sample sizes increase;

iii. for unbalanced sample sizes and homogeneous variances, liberality is clear but reduces with increased sample sizes;

iv. for unbalanced samples and heterogeneous variances, the test is generally liberal but, if small variances are paired with small sample sizes, the attained levels are relatively lower than in case of large variances paired with small sample sizes. Increasing the sample sizes improves the attained significance levels.

For $K = 6$:

i. for homogeneous variances and balanced samples, the liberality of the test persists but it is interesting to note, for example, that when the all the sample sizes are $n_i = 5, i = 1, \ldots, K$, the level attained is 22.1%. This improves appreciably by about a half to 11.4% when the sample sizes are doubled, $n_i = 10, i = 1, \ldots, K$;

ii. for balanced samples and heterogeneous variances, the improvement in the attained levels similar to i) is observed; but the test remains liberal;

iii. for unbalanced samples and homogeneous variances, the test is liberal and improves with increased sample sizes;

iv. for unbalanced samples and heterogeneous variances, the test is liberal and, as expected, attains better levels as the sample sizes increase.

For $K = 9$:

i. for homogeneous variances and balanced samples, the test attains unacceptable significance levels for very small sample sizes, for example, 28.6% for all $n_i = 5$. This scenario dramatically improves to 14.8% when all $n_i = 10$ and 7.0% when $n_i = 40$;

ii. for balanced and heterogeneous variances, at very small samples the levels attained are unacceptable. This improves as the sample sizes increase. However, the test still remains liberal;

iii. for unbalanced samples and homogeneous variances, even though the test is liberal, an improvement of the attained levels with increased samples is clear;

iv. for unbalanced samples and heterogeneous variances, the test is liberal and improves as the samples sizes increase. Note, that when small samples are combined with small variances the attained significance levels are relatively better than when small samples are combined with large variances.

For $K = 18$:

For all cases of balancedness and unbalancedness, and homogeneity and heterogeneity, the test attains unacceptable significance levels. Notice the extreme liberality when the sample sizes are small.

d) Brown-Forsythe (B-F) Test

For $K = 3$:

For balanced samples and homoscedastic cases, the test attains acceptable significance levels which remain significantly unaffected by increases in individual sample sizes. This observation is largely true for balanced sample sizes combined with heterogeneous variances, unbalanced sample sizes combined with homogeneous variances and unbalanced samples combined with heterogeneous variances.

For K=6:

i. for balanced samples and homogeneous variances, the test attains acceptable levels for all sample sizes;

ii. for balanced samples and heterogeneous variances, the test keeps significance levels for small samples but, for large samples, the test becomes liberal;

iii. for unbalanced samples and homogeneous variances, the levels attained by this test are acceptable;

iv. for unbalanced samples and heterogeneous variances, the test attains acceptable levels for small samples. In the other cases the test becomes liberal.

For $K = 9$:

i. for balanced sample sizes and homogeneous variances, the test attains levels which are within the acceptable range;

ii. for balanced samples and heterogeneous variances, the test becomes liberal for large sample sizes;

iii. for unbalanced samples and homoscedastic cases, the test attains good levels;

iv. for unbalanced samples and heterogeneous variances, the test is in general liberal, except for the case when a small sample is paired with a large variance.

For $K = 18$:

i. for equal sizes in the groups and homogeneous variances, the test attains an acceptable significance level except when the sample sizes are small, in which case it tends to be a bit conservative;

ii. for equal sample sizes and heterogeneous variances, the test attains acceptable levels only when the samples are small. Otherwise, the test becomes liberal even for large sample sizes;

iii. for unbalanced samples and equal variances, the test keeps the significance level within acceptable limits. This is true for small to large sample cases;

iv. for unbalanced samples and unequal variances, the test is always liberal,with the liberality being more pronounced when relatively small samples are paired with relatively small variances.

e) The Modified Brown-Forsythe Test

For $K = 3$:

for balanced or unbalanced sample sizes matched with homogeneous or heterogeneous variances, the levels attained are acceptable, except for one case where the test is a bit conservative.

For $K = 6$:

i. for balanced samples and homogeneous variances, the test is a bit conservative for small samples. Otherwise, with moderate to large sample sizes, the test attains acceptable levels. This is also true for balanced samples combined with heterogeneous variances;

ii. for unbalanced samples matched with homogeneous or heterogeneous variances, the test attains acceptable significance levels.

For $K = 9$ and $K = 18$, all the observations for $K = 6$ above are true but, the degree of conservativeness in cases of small samples increases with increasing K.

f) The Approximate ANOVA *F* Test

This test is very similar to the modified Brown-Forsythe test given above. As noted in Section 1.2 above, when the sample sizes in the groups are equal, the results are expected to be the same. Checking Tables 1.1-1.4, we see that the attained levels are all equal for the two tests for balanced designs. Consequently, for $K = 3$, 6, 9 and 18, the observations made above for the modified Brown-Forsythe test also hold for this test (approximate ANOVA F test).

For unbalanced samples and $K = 3$, 6, 9 and 18, the test attains levels which are very close to the Brown-Forsythe test for all values of K, save for small differences. For example, in one case, $K = 3$, when the sample combination is (5,10,15) and the corresponding variances (5,3,1), the approximate ANOVA F test is liberal (6.2%) whereas the modified Brown-Forsythe attains an acceptable level. Further, even though both tests attain acceptable significance levels, the approximate ANOVA test has relatively higher levels when relatively large variances are paired with relatively small sample sizes and when the variances in the groups are all equal. However, when relatively small variances are combined with small sample sizes, the approximate ANOVA F test attains relatively lower significance levels compared to the modified Brown-Forsythe test.

g) The Adjusted Welch Test

For $K = 3$:

i. for balanced samples and homogeneous variances, the test keeps control of the nominal significance level only when the sample sizes are large otherwise, the test is conservative;

ii. for balanced samples and heterogeneous variances, the test attains acceptable significance level only when the sample sizes are large otherwise, the test is conservative;

iii. for unbalanced samples and homogeneous variances, the adjusted Welch test attains acceptable levels;

iv. for unbalanced samples and heterogeneous variances, the actual levels are within the acceptable limits.

For $K = 6$:

i. when the sample sizes and error variances in the groups are equal, the test does not seem to keep good control of the nominal significance level;

ii. for equal sample sizes and heterogeneous variances, the test attains reliable significance levels only when all the sample sizes are equal to 40;

iii. for unbalanced samples and homogeneous variances, the test attains acceptable significance levels;

iv. for unbalanced samples and heterogeneous variances, the actual levels attained are within the limits (4%, 6%).

For $K = 9$:

For all cases of balancedness, unbalancedness, homoscedasticity, and heteroscedasticity, the actual levels attained are acceptable.

For K=18:

For balanced and small sample sizes the test is liberal. For moderate to large sample sizes the test attains acceptable levels. In the case of unbalancedness, when one sample is small the test is liberal except when a small sample size is paired with a small error variance. In this case the test attains its level. For moderate to large samples, the levels attained are in the acceptable range.

In general, we see that for the Welch test, increasing the number of populations has no significant effect on the levels attained, except that for balanced small samples the test becomes too liberal. In this case the adjusted Welch test is preferred. For the Cochran test, increasing the number of populations has the effect of dramatically inflating the significance levels when the sample sizes are small and sometimes also for moderate samples. For the Brown-Forsythe test, increasing the number of studies does not significantly affect the levels attained by the test. For the modified Brown-Forsythe test, the attained significance levels are all within the acceptable range for $K = 3$, 6, 9, and 18, save sometimes for small samples where the test becomes more conservative with an increasing number of populations. This behavior is also true for the approximate ANOVA F test. The modified Brown-Forsythe test rarely attains significance levels above 5%. This is true regardless of the number of populations.

1.4 CONCLUSION

The modified Brown-Forsythe and the approximate ANOVA F test are relatively least affected by changes in the sample sizes and number of populations except when the number of groups is large and the corresponding sample sizes are small, in which case these tests become too conservative. The Brown-Forsythe test should not be used when the number of populations, K, is large with large sample sizes and heterogeneous variances. We will recommend the test for small individual samples regardless of the number of populations. The Welch test can be recommended in the case of heterogeneous variances, except when the sample sizes are small and the number of studies is large. In this case we recommend a suitable adjustment of the adjusted Welch test. This test (adjusted Welch test) has the advantage that the weights, $w_i^*, i = 1, \ldots, K$, have a control parameter, $\varphi_i, i = 1, \ldots, K$, which can be adjusted accordingly.

In case of homoscedasticity regardless of balanced or unbalanced designs, the ANOVA F test is the optimal test (cf. Lehmann, 1986) and the observed deviations from the nominal level are due to the simulation experiment.

REFERENCES

Asiribo, O., & Gurland, J. (1990). Coping with variance heterogeneity. *Communications in Statististics – Theory and Methods*, *19*, 4029–4048.

Böckenhoff, A., & Hartung, J. (1998). Some corrections of the significance level in meta-analysis. *Biometrical Journal*, *40*, 937–947.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*, 129–132.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Journal of the Royal Statistical Society*, *Supplement 4*, 102–118.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2-tests. *Biometrics*, *10*, 417–451.

De Beuckelaer, A. (1996). A closer examination on some parametric alternatives to the ANOVA *F*-test. *Statistical Papers*, *37*, 291–305.

Lehmann, E. L. (1986). *Testing statistical hypotheses* (2nd ed.). New York: Wiley.

Mehrotra, D. V. (1997). Improving the Brown–Forsythe solution to the generalized Behrens–Fisher problem. *Communications in Statistics – Simulation and Computation*, *26*, 1139–1145.

Patel, J. K., Kapadia, C. P., & Owen, D. B. (1976). *Handbook of statistical distributions.* New York: Marcel Dekker.

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