Small sample properties of tests on homogeneity in one–way Anova and Meta–analysis

Joachim Hartung, Doğan Argaç, Kepher H. Makambi
Department of Statistics, University of Dortmund
D–44221 Dortmund, Germany

In the present Monte Carlo study, the empirical Type I error properties and power of several statistics for testing the homogeneity hypothesis in a one–way classification are examined in the case of small sample sizes. We compared these tests under several scenarios: normal populations under heterogeneous variances, nonnormal populations under homogeneous variances, nonnormal populations under heterogeneous variances, balanced and unbalanced sample sizes, and increasing number of populations. Overall, none of the tests considered is uniformly dominating the others. Under normality and variance heterogeneity, the Brown–Forsythe and the Welch test perform well over a wide range of parameter configurations, the modified Brown–Forsythe test by Mehrotra keeps generally the level, but other tests may also perform well, depending on the constellation of the parameters under study. The Welch test becomes liberal when the sample sizes are small and the number of populations is large. We propose a modified version of Welch's test that keeps the nominal level in these cases. With the understanding that methods are unacceptable if they have Type I error rates that are too high, only the testing procedure

---

To appear in:

S t a t i s t i c a l   P a p e r s

associated with the modified Brown–Forsythe test can be recommended both for normal and nonnormal data. Under normality, the modified Welch test can also be recommended.

Key words: meta–analysis, balanced and unbalanced sample sizes, homogeneous and heterogeneous variances, nonnormality

1 Introduction

The problem of testing the homogeneity of several means in a one–way layout of analysis of variance is one of the oldest problems in statistics. This situation arises in many practical settings. For example, a manufacturing company may wish to test whether several machines on production lines produce items of the same quality, and if the items produced are expensive, one can take only a sample of small size for comparison. As a second example, consider the meta–analysis of a series of independent experiments, which address the same question of interest. Here, the goal is to summarize the information provided by the different experiments, see Whitehead and Whitehead (1991), Normand (1999), and Hartung and Knapp (2000). In meta–analysis, it is now common practice to combine the information from the different sources via a one–way model of analysis of variance. A question of interest is to test whether all the experiments share a common effect. This hypothesis is called the homogeneity hypothesis of meta–analysis. According to Hardy and Thompson (1998), the question of homogeneity is an important part of any meta–analysis. The hypothesis of homogeneity in meta–analysis corresponds to testing the homogeneity of the means in the one–way model of analysis of variance.

Under the classical assumptions (normality of the errors, homogeneity of the error variances, and independence of the errors), the Anova F–test is known to be the optimal test, Lehmann (1986). However, when one or more of these basic assumptions are violated, the F–test becomes overly conservative or liberal, depending on the manner and the degree to which these assumptions are violated. In his book, Scheffé (1959) examines the effects of violating these assumptions. He concludes that the effect of violating the normality assumption is slight, at least asymptotically. Replacing the assumption of independence by a serial correlation, he finds that the effect of serial correlation can be disastrous and the F–test is no longer valid, see also Lehmann (1986). The presence of heterogeneous variances can also have a serious effect on the validity of the F–test, especially when the sample sizes are unbalanced. The Type I error rate becomes vastly inflated, when smaller variances are associated with larger group sample sizes, and conversely, when the larger group sample sizes are associated with the larger variances, the empirical rejection rates fall below the nominal level, see e. g. De Beuckelaer (1996). In practice, the assumption of homogeneous error variances is rarely justified, and in fields like meta–analysis, the variances have to be assumed to be heteroscedastic.

In the literature, several alternative tests have been proposed to

account for the heterogeneity of the error variances. In the present paper, we compare the performance of these tests and of the F–test by way of simulation with respect to their attained significance levels and power in the case of small sample sizes under several scenarios: normal data with homogeneous and heterogeneous variances, data from a skew distribution with homogeneous variances, data from a skew distribution under heterogeneity, balanced and unbalanced sample sizes, and increasing number of populations.
The present simulation study differs, at least, in two ways from previously conducted simulation studies, e. g. Mehrotra (1997), De Beuckelaer (1996), and Keselman and Wilcox (1999): we consider tests which have been ignored by other authors and we also investigate the effect of increasing the number of groups and the combined effect of nonnormal data and heterogeneous variances on the validity of the tests. Another goal is to bring all these tests to the attention of researchers working in the area of meta–analysis, since in meta–analysis the only test used for testing the homogeneity hypothesis is Cochran's test. Furthermore, we propose a modified version of the well known Welch test, which is too liberal when the sample sizes are small and the number of groups is large. In these cases the modified Welch test attains levels close to the nominal level.

2 Tests in the one–way Anova model

Let y_{ij} be the observation on the j–th subject of the i–th population, $i = 1, \ldots, K$ and $j = 1, \ldots, n_i$, $K \geq 2$ and $n_i \geq 2$,

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

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

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

That is, $e_{ij} \sim N(0, \sigma_i^2)\,;\ i = 1, \ldots, K,\ j = 1, \ldots, n_i$.
We consider the homogeneity hypothesis $H_0 : \mu_1 = \cdots = \mu_K$ and use parametric procedures to test this hypothesis. We have excluded nonparametric tests such as the Kruskal–Wallis test or the inverse normal scores test, since it is known that these tests are not robust with respect to variance heterogeneity, cf. Lehmann (1975). Also, and more importantly, the null hypothesis for the rank–transformed data may be no longer the same as for the original scale data, as pointed out by Fligner (1981) and Noether (1981) in comments on Conover and Iman (1981). Keselman and Wilcox (1999) propose to replace the hypothesis of homogeneity of the means by the homogeneity hypothesis of the trimmed means, but this concerns a different hypothesis.

We will make use of the following test statistics:

(i) **ANOVA F–test** The F–test, F, is given by

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

with $N = \sum_{i=1}^{K} n_i$, $\bar{y}_{i.} = \sum_{j=1}^{K} y_{ij}/n_i$, $\bar{y}_{..} = \sum_{i=1}^{K} n_i \bar{y}_{i.}/N$, and $s_i^2 = \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2/(n_i - 1)$. 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, denoted by $F_{K-1,N-K}$.

The null hypothesis H_0 is rejected at level α if F exceeds the corresponding $(1-\alpha)$–quantile, i. e. if $F > F_{K-1,N-K;1-\alpha}$. The ANOVA F–test is not robust with respect to heterogeneity in the error variances, see e. g. Brown and Forsythe (1974).

(ii) **Cochran test** The statistic

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

where $w_i = n_i/s_i^2$, $h_i = w_i / \sum_{k=1}^{K} w_k$, was proposed by Cochran (1937), and then modified by James (1951) and Welch (1951).

Cochran's test is the standard test for testing homogeneity in meta–analysis. Under H_0, the Cochran statistic is distributed approximately as a χ^2-variable with $K-1$ degrees of freedom. Reject H_0 at level α if $C > \chi^2_{K-1;1-\alpha}$. James (1951) based his approximation also on the χ^2–distribution, but his test is inferior to Welch's test given below, see Brown and Forsythe (1974), hence we do not consider James's test further.

(iii) **Welch test** The Welch test is given by

$$W = \frac{\sum_{i=1}^{K} w_i(\bar{y}_{i.} - \sum_{j=1}^{K} h_j \bar{y}_{j.})^2}{(K-1) + 2 \cdot (K-2) \cdot (K+1)^{-1} \cdot \sum_{i=1}^{K}(n_i - 1)^{-1}(1-h_i)^2},$$

and Welch (1951) proposed to approximate its distribution using an F–variable. Under H_0, the statistic W has an approximate F distribution with $K-1$ and ν_W degrees of freedom, where

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

The hypothesis H_0 is rejected at level α if $W > F_{K-1,\nu_W;1-\alpha}$.

(iv) **Brown–Forsythe test** This test is also known as the modified F–test and is given by

$$B = \frac{\sum_{i=1}^{K} n_i \ (\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}.$$

Brown and Forsythe use a Satterthwaite approximation to derive the null distribution of the statistic B. When H_0 is true, B is distributed approximately as an F variable with $K-1$ and ν degrees of freedom where

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

We reject H_0 at level α if $B > F_{K-1,\nu;1-\alpha}$.

(v) **modified Brown–Forsythe test** Mehrotra (1997) developed the following test

$$B^* = \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.} - \bar{y}_{..})^2}{\sum_{i=1}^{K}(1 - n_i/N)s_i^2}$$

in an attempt to correct a "flaw" in the original Brown–Forsythe test. The "flaw" in the Brown–Forsythe testing procedure, identified by Mehrotra (1997), is in the specification of the numerator degrees of freedom. Specifically, Brown–Forsythe used $K-1$ numerator degrees of freedom whereas Mehrotra (1997) used a Box (1954) approximation to obtain the numerator degrees of freedom, ν_1, where

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

and ν is given in (iv) above.

Under H_0, B^* is distributed approximately as an F variable with ν_1 and ν degrees of freedom. The null hypothesis H_0 is rejected at level α if $B^* > F_{\nu_1,\nu;1-\alpha}$.

(vi) **approximate ANOVA F–test** Asiribo and Gurland (1990) based their test on

$$F^* = \frac{N-K}{K-1}\cdot\frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{\sum_{i=1}^{K}(n_i-1)s_i^2}.$$

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 F–test.

Under H_0, the test statistic $F^*/\hat{c}$ is distributed approximately as an F-variable with ν_1 and ν_2 degrees of freedom where ν_1 is as given in (v) above,

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

$$\nu_2 = \frac{\left[\sum_{i=1}^{K}(n_i - 1)s_i^2\right]^2}{\sum_{i=1}^{K}(n_i - 1)s_i^4},$$

and H_0 is rejected at level α if $F^* > \hat{c} \cdot F_{\nu_1,\nu_2;1-\alpha}$. We notice that the numerator degrees of freedom for F^* and B^* are equal. Also, $F^*/\hat{c}$ equals B^*, see appendix. The difference between the two test procedures is in the denominator degrees of freedom in the unbalanced case. For balanced sample sizes, the denominator degrees of freedom of the statistics F^* and B^* are the same, and the test statistics F, B, F^* and B^* coincide, see appendix. However, the associated testing procedures are still different, because they use different reference distributions.

(vii) **adjusted Welch test** For small samples in the groups, the Welch test becomes too liberal especially with increasing number of groups. The Welch test uses weights $w_i = n_i/s_i^2$. Böckenhoff and Hartung (1998) have examined these weights, and making use of their results, it follows 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 also Patel et al. (1976). Therefore, an unbiased estimator of n_i/σ_i^2 is $n_i/(c_i s_i^2)$. Let $\varphi_i = (n_i + \delta_1)/(n_i + \delta_2)$, where δ_1 and δ_2 are real numbers chosen such that $1 \leq \varphi_i \leq c_i$; 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 $\delta_2 = -3$). 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 $1 < \varphi_i < c_i$. 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, W^*, is given by

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

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

$$\nu_W^* = \frac{K^2 - 1}{3 \cdot \sum_{i=1}^K (n_i - 1)^{-1}(1-h_i^*)^2},$$

and we reject H_0 at α level if $W^* > F_{K-1,\nu_W^*;1-\alpha}$. When the sample sizes are large, W^* approaches the Welch test, i.e. $(n_i + \delta_1)/(n_i + \delta_2) \xrightarrow{n_i \to \infty} 1$. With small sample sizes, our statistic will help correct the liberality witnessed in the Welch test.

3 Monte Carlo results

We examined the performance of the above tests by way of simulation (10 000 runs for each constellation). Using different constellations of the sample sizes and the error variances, we obtained the simulated actual significance levels for $K = 3$, $K = 6$, and $K = 9$ groups and the power for $K = 3$ and $K = 6$ groups. We started the simulation experiment with $K = 3$ groups. We considered balanced and unbalanced sample sizes with homogeneous and heterogeneous variances. In the case of unbalancedness, we paired the smallest sample size with the smallest variance, and also the smallest sample size with the largest variance. To investigate the effect of the number of groups, K, on the level of the tests, we replicated the experiment for $K = 3$ groups two and three times to give the simulation experiment for $K = 6$ and $K = 9$ groups, respectively. Under normality, the F–test serves as a benchmark in the simulation experiments if the variances are homogeneous. From our experience with the simulations, the choice of the control parameter as $\varphi_i = (n_i+2)/(n_i+1)$ in the adjusted Welch test, W^*, gives reliable results for small sample sizes and a large number of populations.

Since the tests considered have different empirical levels, a fair comparison of their power is not directly possible. A fair comparison must adjust for the latter. This was done using simulated critical values to ensure that all the tests attain the same level (5%). For the sake of completeness, we also give the power of the tests at the nominal unadjusted 5% significance level. Two different configurations of mean differences were used when assessing the power of the tests. In the first pattern, the mean of the first group was set to $\mu_1 = 2$ with the remaining groups having equal means set to zero, $\mu_2 = \mu_3 = 0$. In the second configuration, the means were equally spaced, $\mu_1 = -1$, $\mu_2 = 0$, $\mu_3 = 1$.

The effect of violating the normality assumption was examined using data from a skew distribution. We used the following approach: each observation in the i–th group was generated from a $\chi^2_{\nu_i} - \nu_i$, that is a centered χ^2–distribution, and since this is a location shifted χ^2–distribution, the shape of the distribution is not affected. Using data from this location shifted χ^2–distribution, we ensure that the means of all the groups are the same, as it is needed under the null hypothesis of homogeneity. Choosing the degrees of freedom in all the groups to be equal, $\nu_i = \nu$, we consider the case of homogeneous variances. Now to investigate the dual effect of nonnormality and variance heterogeneity, the variances of the different groups can be chosen to be heteroscedastic by using different degrees of freedom, ν_i, for each group. The degrees of freedom were chosen in such a way that the variance in each group is the same as in the corresponding case of normal data, hence the only difference between the simulations with the normal distribution and the skew distribution is the departure from the normality assumption. In the simulations we have taken the variances $\sigma_i^2 = 2, 4, 6, 10$ which in the nonnormal case considered here lead to shifted $\chi^2_{\nu_i}$–distributions with ($\nu_i = \sigma_i^2/2$) 1,2,3, and 5 degrees of freedom, having the skewness $(2\sqrt{2}/\sqrt{\nu_i})$ $2.8, 2.0, 1.6, 1.3$ and excess $(12/\nu_i)$ $12, 6, 4$, and 2.4, respectively.

TABLE I: Sample Designs for K=3 and K=6.

Pattern		K=3			K=6					
	i	1	2	3	1	2	3	4	5	6
A	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	2	6	10	2	6	10	2	6	10
B	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	2	6	10	2	6	10	2	6	10
C	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	2	6	10	2	6	10	2	6	10
	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	10	6	2	10	6	2	10	6	2
D	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	2	6	10	2	6	10	2	6	10
	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	10	6	2	10	6	2	10	6	2

TABLE II: Sample Design for K=9.

Pattern		K=9								
	i	1	2	3	4	5	6	7	8	9
A	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	5	5	5	5	5	5	5	5	5
	σ_i^2	2	6	10	2	6	10	2	6	10
B	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	10	10	10	10	10	10	10	10	10
	σ_i^2	2	6	10	2	6	10	2	6	10
C	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	2	6	10	2	6	10	2	6	10
	n_i	5	10	15	5	10	15	5	10	15
	σ_i^2	10	6	2	10	6	2	10	6	2
D	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	4	4	4	4	4	4	4	4	4
	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	2	6	10	2	6	10	2	6	10
	n_i	10	20	30	10	20	30	10	20	30
	σ_i^2	10	6	2	10	6	2	10	6	2

TABLE III: Actual Simulated Significance Levels, normal distribution (nominal level 5%) for K=3.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	5.0	4.8	3.3	4.1	3.8	3.8	12.2
	6.0	5.0	3.6	4.6	4.3	4.3	12.8
B	5.1	4.9	3.9	4.9	4.6	4.6	8.4
	5.7	5.1	4.1	5.3	4.7	4.7	8.3
C	5.0	5.3	4.2	5.1	4.8	5.4	10.2
	2.5	4.6	3.6	5.4	4.6	4.6	8.6
	13.4	5.9	4.7	5.8	5.5	6.9	11.6
D	5.2	5.3	4.5	5.1	4.9	5.3	7.7
	2.3	4.9	4.3	5.5	4.6	4.6	6.5
	13.4	5.2	4.2	5.5	5.1	5.8	7.9

TABLE IV: Actual Simulated Significance Levels, normal distribution (nominal level 5%) for K=6.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	5.2	6.2	4.1	4.1	3.3	3.3	22.1
	6.4	6.0	4.4	4.8	3.6	3.6	22.6
B	5.1	5.1	3.7	4.8	4.2	4.2	11.4
	6.1	5.1	3.7	5.5	4.2	4.2	11.6
C	5.0	6.3	4.7	4.7	4.0	4.5	15.5
	2.7	5.7	4.1	5.9	4.4	4.4	13.5
	15.6	6.3	4.7	5.7	4.8	5.7	16.8
D	5.5	5.7	4.8	5.2	4.7	4.9	9.7
	2.3	4.7	3.8	6.1	4.5	4.4	8.2
	15.2	5.2	4.3	5.8	4.6	5.1	9.7

TABLE V: Actual Simulated Significance Levels, normal distribution (nominal level 5%) for K=9.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	5.3	7.3	4.7	4.3	3.2	3.2	28.6
	6.5	7.9	5.5	4.9	3.3	3.3	29.2
B	5.1	6.2	4.3	4.9	4.0	4.0	14.8
	6.6	5.9	4.2	5.9	4.3	4.3	14.5
C	5.3	7.0	4.9	4.9	4.1	4.5	19.3
	2.4	6.3	4.4	6.5	4.4	4.3	17.9
	18.4	7.7	5.5	5.6	4.3	5.2	20.6
D	4.9	5.5	4.1	4.8	4.3	4.4	10.7
	2.2	5.0	3.9	6.2	4.3	4.3	9.7
	18.1	5.5	4.3	5.9	4.5	5.0	10.8

TABLE VI: Actual Simulated Significance Levels, nonnormal distribution (nominal level 5%) for K=3.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	4.4	3.3	2.3	2.9	2.5	2.5	12.3
	6.1	5.9	4.5	4.2	3.8	3.8	15.5
B	4.6	4.7	3.7	3.9	3.2	3.2	8.6
	5.5	5.7	4.7	4.8	4.2	4.2	9.6
C	4.5	7.1	5.9	4.1	3.2	3.4	12.1
	2.5	4.1	3.0	4.2	3.2	3.3	8.5
	13.6	8.4	7.1	6.6	6.1	7.0	15.0
D	4.5	6.7	5.8	4.4	3.6	3.8	8.9
	2.5	5.3	4.5	5.8	4.6	4.6	7.1
	12.8	6.9	6.1	5.7	5.2	5.7	9.6

TABLE VII: Actual Simulated Significance Levels, nonnormal distribution (nominal level 5%) for K=6.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	4.5	5.9	3.7	2.6	1.8	1.8	24.0
	6.4	8.3	6.1	4.1	2.8	2.8	25.2
B	4.4	7.0	5.4	3.6	2.3	2.3	14.7
	6.2	7.0	5.5	5.1	3.3	3.3	14.6
C	4.9	10.6	8.5	3.7	2.3	2.7	21.2
	2.7	5.6	4.1	5.8	3.6	3.7	14.6
	15.6	10.9	8.6	5.2	3.8	4.4	21.6
D	4.7	9.0	7.7	4.3	2.9	3.1	13.7
	2.6	7.0	5.7	6.4	4.2	4.2	11.0
	15.0	8.3	7.1	5.5	3.9	4.3	12.9

TABLE VIII: Actual Simulated Significance Levels, nonnormal distribution (nominal level 5%) for K=9.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	4.5	9.5	5.9	2.4	1.4	1.4	34.3
	6.8	11.1	8.0	4.0	2.4	2.4	34.1
B	4.7	9.6	7.2	3.8	2.0	2.0	20.4
	6.5	9.3	7.1	5.4	3.2	3.2	19.1
C	5.3	13.6	10.6	4.1	2.5	2.6	27.3
	2.4	7.5	4.9	5.9	3.1	3.1	21.4
	18.0	13.6	10.4	4.9	3.0	3.7	28.0
D	5.1	11.0	9.4	4.4	2.8	3.0	17.1
	2.2	8.5	6.9	6.0	3.7	3.7	14.4
	17.7	10.1	8.2	6.1	4.1	4.5	16.2

TABLE IX: Simulated Power for K=3,

adjusted and unadjusted (*cursive*),

at $\mu_1 = 2, \mu_2 = 0, \mu_3 = 0$,

normal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	27.1	25.1	25.1	27.1	27.1	27.1	25.2
	28.8	*24.8*	*19.7*	*25.6*	*24.6*	*24.6*	*44.6*
	17.4	24.0	24.0	17.4	17.4	17.4	24.1
	20.7	*25.0*	*19.8*	*16.7*	*15.5*	*15.5*	*45.8*
B	57.8	55.4	55.4	57.8	57.8	57.8	55.4
	57.8	*55.1*	*50.8*	*57.0*	*56.0*	*56.0*	*65.2*
	38.5	56.5	56.5	38.5	38.5	38.5	56.5
	42.8	*57.0*	*52.3*	*40.2*	*37.5*	*37.5*	*67.4*
C	37.7	33.7	33.1	36.1	36.1	36.1	33.9
	39.0	*35.4*	*30.6*	*35.7*	*34.5*	*37.7*	*48.3*
	28.0	47.5	46.4	27.4	27.4	27.4	47.5
	14.9	*46.9*	*41.5*	*28.8*	*25.2*	*25.0*	*58.6*
	23.2	18.1	17.8	21.1	21.1	21.1	18.3
	38.9	*18.9*	*15.7*	*21.2*	*20.2*	*23.7*	*30.6*
D	70.2	66.6	65.9	69.5	69.5	69.5	66.6
	71.8	*68.0*	*64.4*	*70.0*	*69.4*	*70.7*	*74.6*
	57.7	81.9	81.5	57.9	57.9	57.9	81.9
	41.2	*82.3*	*80.0*	*62.4*	*58.3*	*58.0*	*85.6*
	43.2	34.8	34.2	41.1	41.1	41.1	34.8
	61.5	*35.5*	*32.0*	*42.8*	*41.2*	*43.7*	*43.8*

TABLE X: Simulated Power for K=6,
adjusted and unadjusted (*cursive*),
at $\mu_1 = \mu_4 = 2, \mu_2 = \mu_5 = 0, \mu_3 = \mu_6 = 0$,
normal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	37.0	29.2	29.2	37.0	37.0	37.0	29.0
	39.6	*34.1*	*26.0*	*35.2*	*31.0*	*31.0*	*65.6*
	21.6	30.0	30.0	21.6	21.6	21.6	30.0
	26.4	*34.2*	*26.5*	*20.3*	*16.6*	*16.6*	*67.2*
B	77.0	72.2	72.2	77.0	77.0	77.0	72.1
	77.2	*72.3*	*67.5*	*76.5*	*74.7*	*74.7*	*85.3*
	52.4	73.6	73.6	52.4	52.4	52.4	73.6
	58.6	*74.5*	*69.5*	*56.0*	*49.1*	*49.1*	*86.5*
C	54.1	46.2	45.4	53.4	53.4	53.4	46.5
	53.5	*49.3*	*42.0*	*50.0*	*46.7*	*49.5*	*68.8*
	32.8	59.7	59.3	32.5	32.5	32.5	59.9
	19.8	*62.2*	*54.5*	*37.3*	*30.6*	*30.2*	*79.8*
	30.7	20.9	20.2	27.9	27.9	27.9	21.0
	53.6	*25.3*	*20.6*	*29.0*	*25.2*	*29.2*	*44.3*
D	88.4	84.2	83.7	87.8	87.8	87.8	84.4
	89.0	*85.4*	*82.3*	*88.2*	*87.3*	*87.8*	*91.1*
	73.4	95.2	95.1	74.1	74.1	74.1	95.2
	58.4	*95.4*	*94.2*	*78.7*	*72.3*	*71.9*	*97.4*
	59.6	47.5	47.0	57.6	57.6	57.6	47.6
	79.4	*49.4*	*44.5*	*61.2*	*56.2*	*58.3*	*61.6*

TABLE XI: Simulated Power for K=9,

adjusted and unadjusted (*cursive*),

at $\mu_1 = \mu_4 = \mu_7 = 2, \mu_2 = \mu_5 = \mu_8 = 0, \mu_3 = \mu_6 = \mu_9 = 0$,

normal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	47.3	33.1	33.1	47.3	47.3	47.3	33.1
	48.4	*42.8*	*32.9*	*43.8*	*38.3*	*38.3*	*78.0*
	25.7	33.8	33.8	25.7	25.7	25.7	33.8
	32.9	*45.1*	*35.0*	*26.4*	*20.3*	*20.3*	*80.3*
B	88.6	83.3	83.3	88.6	88.6	88.6	83.3
	88.7	*84.4*	*79.5*	*88.1*	*86.3*	*86.3*	*93.9*
	65.6	85.6	85.6	65.6	65.6	65.6	85.6
	71.9	*86.4*	*81.6*	*69.4*	*60.9*	*60.9*	*94.9*
C	66.7	51.7	50.9	65.2	65.2	65.2	52.0
	65.3	*60.0*	*51.5*	*61.3*	*57.3*	*60.1*	*81.2*
	42.0	71.1	70.6	42.9	42.9	42.9	71.2
	25.3	*74.8*	*67.4*	*46.5*	*37.5*	*37.1*	*90.5*
	38.0	24.2	23.8	34.6	34.6	34.6	24.3
	64.7	*32.7*	*26.2*	*36.0*	*30.5*	*34.4*	*56.1*
D	96.1	93.3	93.0	95.8	95.8	95.8	93.4
	95.9	*93.9*	*92.0*	*95.4*	*94.9*	*95.1*	*97.1*
	84.9	98.8	98.7	85.3	85.3	85.3	98.8
	71.8	*98.8*	*98.4*	*89.7*	*84.7*	*84.6*	*99.5*
	72.0	56.9	55.7	70.4	70.4	70.4	56.9
	89.2	*61.6*	*55.8*	*73.8*	*68.1*	*70.2*	*74.0*

TABLE XII: Simulated Power for K=3, adjusted and unadjusted (*cursive*), at $\mu_1 = -1, \mu_2 = 0, \mu_3 = 1$, normal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	21.8	19.4	19.4	21.8	21.8	21.8	19.5
	21.1	*17.8*	*13.8*	*18.7*	*17.7*	*17.7*	*35.8*
	14.4	15.5	15.5	14.4	14.4	14.4	15.6
	17.1	*16.5*	*12.3*	*14.1*	*13.0*	*13.0*	*33.1*
B	46.2	45.0	45.0	46.2	46.2	46.2	44.9
	45.4	*43.0*	*38.6*	*44.3*	*43.3*	*43.3*	*53.1*
	29.2	35.3	35.3	29.1	29.1	29.1	35.2
	32.0	*34.5*	*31.0*	*30.2*	*28.1*	*28.1*	*45.3*
C	38.4	33.7	33.7	36.7	36.7	36.7	33.9
	38.9	*34.6*	*29.9*	*36.0*	*34.8*	*38.3*	*49.2*
	26.9	32.8	32.4	26.6	26.6	26.6	32.9
	17.2	*34.0*	*29.7*	*29.1*	*25.9*	*25.5*	*44.9*
	22.5	22.8	23.1	20.3	20.3	20.3	23.2
	40.5	*24.4*	*21.0*	*21.5*	*20.5*	*24.3*	*39.6*
D	70.6	67.8	67.8	70.2	70.2	70.2	67.8
	71.1	*68.2*	*65.0*	*69.5*	*68.7*	*70.1*	*74.5*
	53.4	65.2	65.0	53.8	53.8	53.8	65.1
	39.8	*65.4*	*62.7*	*57.1*	*53.2*	*52.9*	*70.3*
	45.6	51.2	51.3	43.3	43.3	43.3	51.2
	65.8	*51.4*	*48.3*	*44.3*	*42.0*	*45.1*	*60.7*

TABLE XIII: Simulated Power for K=6, adjusted and unadjusted (*cursive*), at $\mu_1 = \mu_4 = -1, \mu_2 = \mu_5 = 0, \mu_3 = \mu_6 = 1$, normal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	28.3	22.0	22.0	28.3	28.3	28.3	21.80
	28.9	*25.6*	*19.1*	*25.4*	*22.3*	*22.3*	*56.0*
	17.1	17.7	17.7	17.1	17.1	17.1	17.6
	21.6	*21.6*	*16.1*	*17.5*	*14.5*	*14.5*	*50.5*
B	63.0	58.0	58.0	63.0	63.0	63.0	57.9
	62.2	*57.5*	*51.5*	*61.3*	*58.9*	*58.9*	*73.0*
	37.6	44.8	44.8	37.6	37.6	37.6	44.8
	43.1	*46.5*	*41.1*	*40.8*	*35.2*	*35.2*	*63.7*
C	53.7	44.5	44.7	52.5	52.5	52.5	44.9
	53.4	*47.3*	*40.5*	*49.5*	*46.5*	*49.2*	*68.7*
	33.0	41.6	41.9	32.8	32.8	32.8	41.7
	20.7	*44.1*	*37.2*	*36.1*	*29.6*	*29.4*	*63.8*
	29.7	27.7	27.5	27.2	27.2	27.2	27.8
	56.2	*34.1*	*28.2*	*28.5*	*24.4*	*28.9*	*57.7*
D	88.1	85.0	85.1	87.4	87.4	87.4	85.2
	89.0	*86.0*	*83.2*	*88.0*	*87.0*	*87.6*	*91.5*
	67.2	82.6	82.5	68.0	68.0	68.0	82.6
	54.7	*82.7*	*79.8*	*72.8*	*67.1*	*66.7*	*88.5*
	62.1	67.7	68.0	60.2	60.2	60.2	67.9
	83.8	*69.4*	*65.6*	*63.6*	*57.7*	*60.4*	*79.5*

TABLE XIV: Simulated Power for K=9,
adjusted and unadjusted (*cursive*),
at $\mu_1 = \mu_4 = \mu_7 = -1, \mu_2 = \mu_5 = \mu_8 = 0, \mu_3 = \mu_6 = \mu_9 = 1$,
normal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	36.6	25.3	25.3	36.6	36.6	36.6	25.3
	35.6	*32.8*	*24.5*	*31.4*	*26.9*	*26.9*	*69.1*
	19.8	19.4	19.4	19.8	19.8	19.8	19.3
	25.5	*28.2*	*20.7*	*20.3*	*15.4*	*15.4*	*63.6*
B	74.8	67.7	67.7	74.8	74.8	74.8	67.7
	75.6	*70.9*	*64.1*	*74.9*	*72.1*	*72.1*	*85.4*
	47.5	56.0	56.0	47.5	47.5	47.5	56.0
	53.1	*56.9*	*50.0*	*51.0*	*43.2*	*43.2*	*77.1*
C	67.0	49.7	49.8	65.0	65.0	65.0	49.9
	65.9	*59.6*	*51.2*	*62.2*	*58.3*	*60.9*	*81.4*
	41.0	50.2	50.6	41.9	41.9	41.9	50.3
	26.3	*54.3*	*46.6*	*45.4*	*37.8*	*37.4*	*76.8*
	36.9	32.6	33.1	33.2	33.2	33.2	32.8
	67.2	*43.4*	*35.9*	*36.1*	*29.7*	*34.2*	*69.2*
D	96.0	93.6	93.5	95.7	95.7	95.7	93.6
	96.3	*94.4*	*92.7*	*95.9*	*95.4*	*95.7*	*97.5*
	78.6	91.3	91.3	79.0	79.0	79.0	91.4
	67.2	*92.5*	*90.5*	*84.3*	*79.3*	*79.1*	*96.0*
	76.5	80.0	79.8	74.6	74.6	74.6	80.0
	92.9	*81.9*	*78.2*	*76.8*	*70.6*	*72.6*	*89.8*

TABLE XV: Simulated Power for K=3,

adjusted and unadjusted (*cursive*),

at $\mu_1 = 2, \mu_2 = 0, \mu_3 = 0$,

nonnormal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	39.6	41.5	41.5	39.6	39.6	39.6	41.1
	35.1	*35.2*	*28.7*	*29.3*	*27.5*	*27.5*	*58.8*
	24.9	36.7	36.7	24.9	24.9	24.9	36.6
	29.5	*42.0*	*36.3*	*24.1*	*22.7*	*22.7*	*61.4*
B	65.3	68.9	68.9	65.3	65.3	65.3	68.7
	61.8	*68.0*	*63.6*	*60.2*	*58.3*	*58.3*	*76.8*
	43.5	62.8	62.8	43.5	43.5	43.5	62.8
	48.1	*66.8*	*63.3*	*45.4*	*42.4*	*42.4*	*75.3*
C	43.4	37.2	35.4	41.3	41.3	41.3	37.4
	42.6	*41.4*	*34.4*	*35.4*	*33.0*	*37.6*	*59.9*
	32.0	60.5	59.6	32.9	32.9	32.9	60.6
	21.8	*59.7*	*54.3*	*32.8*	*29.2*	*29.8*	*71.0*
	22.2	8.7	8.3	14.9	14.9	14.9	8.7
	38.9	*14.2*	*11.1*	*14.6*	*13.4*	*17.5*	*27.3*
D	72.5	73.7	72.8	74.3	74.3	74.3	74.1
	72.5	*80.7*	*77.2*	*75.3*	*73.2*	*74.3*	*86.7*
	58.7	90.0	89.7	61.8	61.8	61.8	90.1
	43.4	*89.3*	*87.5*	*63.7*	*58.5*	*58.2*	*91.7*
	42.6	23.3	22.9	37.3	37.3	37.3	23.3
	61.0	*31.5*	*28.0*	*39.3*	*36.3*	*39.8*	*41.8*

TABLE XVI: Simulated Power for K=6,
adjusted and unadjusted (*cursive*),
at $\mu_1 = \mu_4 = 2, \mu_2 = \mu_5 = 0, \mu_3 = \mu_6 = 0$,
nonnormal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	46.2	50.3	50.3	46.2	46.2	46.2	49.8
	43.9	*54.5*	*45.3*	*35.6*	*29.2*	*29.2*	*82.3*
	28.8	47.6	47.6	28.8	28.8	28.8	47.6
	33.6	*57.8*	*50.5*	*26.2*	*21.2*	*21.2*	*81.4*
B	80.1	81.3	81.3	80.1	80.1	80.1	81.1
	78.6	*86.8*	*83.3*	*76.6*	*71.5*	*78.6*	*93.9*
	54.5	78.3	78.3	54.5	54.5	54.5	78.3
	61.3	*83.7*	*80.6*	*57.9*	*50.0*	*50.0*	*91.3*
C	55.6	41.8	40.6	56.6	56.6	56.6	42.5
	54.1	*61.9*	*53.3*	*48.6*	*41.4*	*45.1*	*82.5*
	34.7	76.0	75.5	35.9	35.9	35.9	75.6
	24.7	*77.7*	*72.5*	*38.7*	*30.4*	*30.6*	*89.7*
	30.7	8.8	8.5	24.9	24.9	24.9	8.8
	52.1	*21.2*	*15.8*	*22.9*	*18.2*	*22.2*	*42.0*
D	88.4	89.0	88.4	91.5	91.5	91.5	89.1
	87.9	*95.4*	*93.5*	*90.5*	*87.9*	*88.3*	*97.8*
	70.9	97.6	97.6	73.2	73.2	73.2	97.6
	58.3	*98.2*	*97.7*	*78.1*	*69.9*	*69.7*	*99.2*
	59.3	33.5	32.5	58.9	58.9	58.9	33.5
	79.1	*47.9*	*42.1*	*61.2*	*53.9*	*56.6*	*61.7*

TABLE XVII: Simulated Power for K=9,

adjusted and unadjusted (*cursive*),

at $\mu_1 = \mu_4 = \mu_7 = 2, \mu_2 = \mu_5 = \mu_8 = 0, \mu_3 = \mu_6 = \mu_9 = 0$,

nonnormal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	54.6	56.3	56.3	54.6	54.6	54.6	55.9
	51.9	*69.1*	*59.3*	*42.5*	*33.8*	*33.8*	*92.2*
	33.0	54.1	54.1	33.0	33.0	33.0	54.2
	38.7	*70.3*	*62.9*	*29.7*	*22.0*	*22.0*	*90.6*
B	89.2	90.6	90.6	89.2	89.2	89.2	90.6
	87.9	*94.7*	*92.5*	*86.3*	*81.0*	*81.0*	*98.5*
	66.2	88.2	88.2	66.2	66.2	66.2	88.1
	71.5	*92.9*	*90.5*	*68.0*	*57.8*	*57.8*	*97.2*
C	64.7	49.0	47.4	66.8	66.8	66.8	49.6
	65.1	*75.7*	*66.2*	*60.9*	*50.7*	*54.4*	*92.9*
	43.4	86.0	85.8	44.9	44.9	44.9	85.8
	28.2	*88.2*	*84.1*	*47.3*	*34.9*	*35.0*	*96.4*
	37.5	9.2	8.9	32.8	32.8	32.8	9.2
	63.1	*28.8*	*21.8*	*30.4*	*22.3*	*26.8*	*55.1*
D	95.2	95.4	95.2	97.0	97.0	97.0	95.4
	95.3	*99.1*	*98.6*	*96.8*	*95.2*	*95.5*	*99.7*
	82.8	99.5	99.4	85.0	85.0	85.0	99.5
	70.4	*99.7*	*99.6*	*88.2*	*81.3*	*81.1*	*99.9*
	72.1	40.5	39.5	72.7	72.7	72.7	40.5
	89.1	*60.4*	*53.6*	*74.6*	*66.5*	*68.7*	*75.7*

TABLE XVIII: Simulated Power for K=3,
adjusted and unadjusted (*cursive*),
at $\mu_1 = -1, \mu_2 = 0, \mu_3 = 1$,
nonnormal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	31.9	31.7	31.7	31.9	31.9	31.9	31.4
	28.2	*26.5*	*21.7*	*22.5*	*20.7*	*20.7*	*46.7*
	10.9	9.0	9.0	10.9	10.9	10.9	9.1
	14.0	*11.4*	*8.6*	*9.5*	*8.2*	*8.2*	*30.8*
B	53.6	51.1	51.1	53.6	53.6	53.6	50.9
	51.5	*51.8*	*48.0*	*49.8*	*48.0*	*48.0*	*60.7*
	25.7	29.8	29.8	25.7	25.7	25.7	30.0
	30.4	*33.6*	*29.3*	*27.0*	*23.3*	*23.3*	*45.6*
C	47.3	47.0	46.8	48.7	48.7	48.7	47.2
	45.5	*50.4*	*46.7*	*45.4*	*43.6*	*45.7*	*62.0*
	25.8	46.8	46.9	30.7	30.7	30.7	46.6
	14.0	*45.4*	*41.7*	*30.2*	*25.2*	*24.9*	*55.2*
	29.0	29.1	29.0	27.5	27.5	27.5	29.2
	47.9	*39.3*	*35.7*	*30.9*	*29.7*	*32.8*	*53.9*
D	75.1	69.6	69.5	70.9	70.9	70.9	69.9
	74.0	*73.8*	*71.7*	*69.9*	*68.9*	*70.3*	*79.0*
	56.3	71.1	70.9	58.1	58.1	58.1	71.2
	39.2	*69.6*	*67.9*	*59.4*	*54.8*	*54.7*	*73.6*
	49.6	54.1	54.2	46.5	46.5	46.5	54.1
	66.6	*59.2*	*56.8*	*48.0*	*46.6*	*48.6*	*66.4*

TABLE XIX: Simulated Power for K=6,
adjusted and unadjusted (*cursive*),
at $\mu_1 = \mu_4 = -1, \mu_2 = \mu_5 = 0, \mu_3 = \mu_6 = 1$,
nonnormal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	36.1	38.2	38.2	36.1	36.1	36.1	37.8
	33.2	*40.7*	*33.5*	*25.6*	*20.6*	*20.6*	*69.5*
	15.3	12.0	12.0	15.3	15.3	15.3	12.2
	19.3	*20.7*	*14.4*	*12.1*	*8.7*	*8.7*	*54.8*
B	68.0	62.7	62.7	68.0	68.0	68.0	62.5
	65.0	*69.2*	*64.7*	*62.4*	*57.0*	*57.0*	*81.3*
	36.6	42.2	42.2	36.6	36.6	36.6	42.2
	42.3	*51.8*	*45.8*	*38.3*	*29.0*	*29.0*	*68.4*
C	59.1	55.4	55.7	58.9	58.9	58.9	55.9
	59.0	*68.9*	*64.2*	*54.4*	*49.2*	*51.6*	*82.9*
	30.3	62.5	62.7	33.9	33.9	33.9	62.2
	19.8	*64.8*	*60.3*	*37.2*	*28.0*	*27.8*	*78.3*
	36.2	36.4	36.5	35.6	35.6	35.6	36.3
	60.2	*55.1*	*49.9*	*34.4*	*29.9*	*33.3*	*72.0*
D	90.2	84.9	84.8	86.6	86.6	86.6	84.9
	89.2	*90.5*	*89.0*	*84.7*	*81.9*	*83.0*	*93.7*
	69.2	85.7	85.6	68.9	68.9	68.9	85.7
	55.3	*87.0*	*85.3*	*73.1*	*66.3*	*66.1*	*91.1*
	64.3	70.1	70.2	62.0	62.0	62.0	70.1
	82.9	*78.8*	*76.1*	*63.4*	*57.4*	*59.6*	*85.5*

TABLE XX: Simulated Power for K=9,

adjusted and unadjusted (*cursive*),

at $\mu_1 = \mu_4 = \mu_7 = -1, \mu_2 = \mu_5 = \mu_8 = 0, \mu_3 = \mu_6 = \mu_9 = 1$,

nonnormal distribution.

Pattern	$\hat{\alpha}\%$						
	F	W	W^*	B	B^*	F^*	C
A	42.7	41.8	41.8	42.7	42.7	42.7	41.5
	40.3	*53.2*	*43.8*	*30.6*	*22.7*	*22.7*	*82.6*
	19.0	13.6	13.6	19.0	19.0	19.0	13.8
	23.2	*30.6*	*21.7*	*15.2*	*9.6*	*9.6*	*70.1*
B	78.3	73.2	73.2	78.3	78.3	78.3	73.1
	75.6	*81.7*	*76.9*	*73.5*	*66.4*	*66.4*	*91.7*
	47.0	53.0	53.0	47.0	47.0	47.0	52.8
	53.4	*66.0*	*59.6*	*49.2*	*37.8*	*37.8*	*82.6*
C	68.3	64.5	64.5	66.0	66.0	66.0	64.8
	68.9	*81.1*	*76.8*	*61.6*	*54.8*	*57.2*	*92.1*
	40.3	74.5	74.7	42.5	42.5	42.5	74.3
	24.0	*78.0*	*73.4*	*44.8*	*33.2*	*33.1*	*89.9*
	43.1	41.8	42.1	41.4	41.4	41.4	41.9
	69.7	*66.5*	*60.5*	*40.0*	*33.5*	*36.9*	*83.5*
D	96.2	92.1	92.1	93.9	93.9	93.9	92.1
	96.0	*96.7*	*95.9*	*92.7*	*90.4*	*91.1*	*98.4*
	81.9	92.7	92.7	81.1	81.1	81.1	92.8
	68.9	*95.2*	*94.3*	*84.2*	*78.5*	*78.4*	*97.4*
	76.8	80.1	80.2	73.3	73.3	73.3	80.1
	91.5	*88.6*	*86.5*	*74.4*	*67.3*	*69.3*	*93.4*

4 Discussion

First, we consider the attained significance levels, and we observe that the F–test is not robust with respect to variance heterogeneity, especially when the sample sizes are unbalanced. When larger sample sizes are paired with the larger variances, the actual Type I error rate falls below the nominal level, and if the smaller sample sizes are paired with the larger variances, the actual Type I error rate becomes inflated. Cochran's test, the standard test in meta–analysis, should definitely not be used, since it is always liberal, confer also the simulation results conducted for the example, Table XVII, where the sample sizes are much larger. Under variance heterogeneity, both the Welch and the Brown–Forsythe tests are suitable alternatives to the F–test in most of the configurations considered in the simulation experiments, but the other tests may also perform well, depending on the configuration of the parameters involved. When the variances are homogeneous, the F–test is the optimal test, see Lehmann (1986), and should be used (the observed deviations from the nominal level are due to simulation), the researcher should not follow in this respect the recommendations of De Beuckelaer (1996) who recommends the Brown–Forsythe test in these cases. With increasing number of groups, from $K = 3$ to $K = 9$, the empirical Type I error rates of the modified Brown–Forsythe and the approximate F–test become smaller, especially for small sample sizes, and the tests may get conservative. The remaining tests attain higher Type I error rates with increasing number of groups in most cases and can become liberal, except the modified Welch test which attains acceptable levels for small sample sizes.

Concerning the power of the tests, we discuss only the behaviour of the test statistics with respect to the adjusted power, confer the corresponding remarks in section 3. Since the test statistics of the Brown–Forsythe, the modified Brown–Forsythe, and the approximate Anova F–test are the same, see appendix, these tests have the same (adjusted) power. There is no practically significant difference between the Cochran, the Welch, and the modified Welch test with respect to power in the simulations considered. As it was to be expected, the F–test is the most powerful test when the variances are homogeneous. In the balanced case the Brown–Forsythe test, the modified Brown–Forsythe test, and the approximate Anova F–test attain the same power as the F–test because the test statistics coincide. If the variances are homogeneous, the Cochran, the Welch, and the modified Welch test are slightly less powerful than the F–test, when the sample sizes are balanced; in case of unbalancedness, the difference between the power of the F–test and the latter tests is larger. Under variance heterogeneity and balanced sample sizes, the Cochran, the Welch, and the modified Welch tests are always more powerful than the other tests regardless of the alternative hypothesis. In case of unbalancedness and variance heterogeneity, the Cochran, the Welch, and the modified Welch tests still continue to be most powerful when the alternative is $\mu_1 = -1$, $\mu_2 = 0$, $\mu_3 = 1$. If the alternative hypothesis considered is $\mu_1 = 2$, $\mu_2 = \mu_3 = 0$, the

Cochran, the Welch, and the modified Welch tests are most powerful when the smallest variance is associated with the smallest sample size. If the smallest variance is paired with the largest sample size, the Brown–Forsythe, the modified Brown–Forsythe, the approximate Anova F–test and the F–test are most powerful.

In the case of the nonnormal distribution and homogeneous variances, the F–test still gives acceptable levels, even for the small sample sizes considered here. If the distribution is nonnormal and the variances are heterogeneous, the F–test behaves the same way as in the case of normal distributions with heterogeneous variances: when the largest sample size is paired with the largest variance, the F–test is conservative, and if the smallest sample size is paired with the largest variance, the F–test is liberal. Hence, the departure from the normality assumption seems to be a minor affair compared to the presence of variance heterogeneity. Among the remaining tests, none gives acceptable levels for all cases considered: the tests can become liberal or conservative or keep the nominal level, depending on the particular configuration of the parameters involved (sample sizes and variances).

Now to summarize the above findings concerning the attained significance levels and the power of the tests, we first want to emphasize that when judging the performance of tests by way of simulation, it is necessary to consider first their Type I error properties. Overall, the Welch test, the Brown–Forsythe test (especially for small sample sizes), and the modified Welch test provide acceptable control of Type I errors when the variances are heterogeneous. But with the understanding that methods are unacceptable if they have Type I error rates that are too high, only the testing procedure associated with the modified Brown–Forsythe test can be recommended both for normal and nonnormal data. Under normality, the modified Welch test can also be recommended.

As a last remark, we want to emphasize that the results from our simulation studies should not be overgeneralized. The conclusions drawn are limited to experimental settings where the sample sizes, the variances, the number of groups, and the distributions do not differ markedly from those considered here, confer especially the following example.

5 Example

To illustrate the application of the various homogeneity tests discussed above, we give an example from clinical trials. The data are taken from Li, Shi and Roth (1994). In eight randomized controlled trials the effectiveness of a new drug, amlodipine, in the treatment of angina was examined. The response variable, the change in work capacity, was compared for patients who received either the drug or placebo. The change in work capacity is the ratio of the exercise time after the patient receives the intervention (drug or placebo) to before the patient receives the intervention. The logarithms of the observed changes are assumed to be approximately normally distributed. We present here the data for the placebo group, and address in particular the homogeneity with respect to the placebo group, since this is also recommended by Chalmers (1991) to investigate the homogeneity question in a meta–analysis.

TABLE XV: Change in work capacity in the treatment of angina, placebo group data, taken from Li, Shi and Roth (1994).

Study	Sample size	Mean	Variance
1	48	−0.0027	0.0007
2	26	0.0270	0.1139
3	72	0.0443	0.4972
4	12	0.2277	0.0488
5	34	0.0056	0.0955
6	31	0.0943	0.1734
7	27	−0.0057	0.9891
8	47	−0.0057	0.1291

The samples are highly unbalanced, and since the variance estimators differ considerably, the assumption of variance homogeneity seems not to be justified. We perform now the tests presented above to investigate the heterogeneity question among the studies. The results are summarized in the next table, Table XVI.
The null hypothesis of homogeneity is only rejected by the Cochran test, whereas the remaining tests do not reject the null hypothesis. For illustration, we simulated the empirical levels of the tests using the sample sizes and the variance estimators given in Table XV, assuming that the variance estimators are the true values.

TABLE XVI: Value of the test statistics and corresponding critical values (level $\alpha = 5\%$) for the data of TABLE XV.

Statistic	Value of the statistic	Critical value
F	0.41	2.04
W	2.06	2.13
W^*	1.93	2.13
B	0.44	2.11
B^*	0.44	2.71
$F^*/\hat{c}$	0.44	2.68
C	15.17	14.07

We obtained the following empirical levels: 8.4. (F, ANOVA F–test), 5.1 (W, Welch), 4.4 (W^*, adjusted Welch), 9.7 (B, Brown–Forsythe), 5.2 (B^*, modified Brown–Forsythe), 5.3 ($F^*/\hat{c}$, approximate ANOVA F–test), and 8.0 (C, Cochran). Note that, in general, the behaviour of the tests in a particular example cannot be deduced from Monte Carlo results, although the general impression of the tests obtained in section 3 seems to be confirmed in case of the placebo group data. But of course it should be clear that here we could not simulate with the true, unknown variances underlying the data.

Acknowledgements

The financial support of the Sonderforschungsbereich 475, projects "Anova" and "Meta–Analysis", of the German Research Community (DFG) is gratefully acknowledged.

Further, we would like to thank the referees for their valuable comments on the first version of the paper.

Appendix

It will be shown that the test statistics of the approximate ANOVA F–test, the Brown–Forsythe, and the modified Brown–Forsythe are the same both in balanced and unbalanced samples:

$$\begin{aligned}
F^*/\hat{c} &= \frac{N(K-1)}{N-K} \cdot \frac{\sum_{i=1}^{K}(n_i-1)s_i^2}{\sum_{i=1}^{K}(N-n_i)s_i^2} \cdot \frac{N-K}{K-1} \\
&\quad \cdot \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{\sum_{i=1}^{K}(n_i-1)s_i^2} \\
&= \frac{N\sum_{i=1}^{K} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{\sum_{i=1}^{K}(N-n_i)s_i^2} \\
&= \frac{\sum_{i=1}^{K} n_i(\bar{y}_{i.}-\bar{y}_{..})^2}{\sum_{i=1}^{K}(1-n_i/N)s_i^2} \\
&= B^* = B.
\end{aligned}$$

Next, it will be shown that the denominator degrees of freedom of the approximate ANOVA F–test and of the modified Brown–Forsythe test are the same if the sample sizes are balanced $n_i = n,\ i = 1, \ldots, K$:

$$\begin{aligned}
\nu_2 &= \frac{\left[\sum_{i=1}^{K}(n_i-1)s_i^2\right]^2}{\sum_{i=1}^{K}(n_i-1)s_i^4} \\
&= \frac{\left[\sum_{i=1}^{K}(n-1)s_i^2\right]^2}{\sum_{i=1}^{K}(n-1)s_i^4}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{(n-1)\left[\sum_{i=1}^{K} s_i^2\right]^2}{\sum_{i=1}^{K} s_i^4} \\
&= \nu.
\end{aligned}
$$

Hence, in balanced cases the reference distributions of the approximate ANOVA F–test and of the modified Brown–Forsythe are the same and since the test statistics are also the same, both test procedures coincide in balanced cases.

Also, for balanced sample sizes, $N = \sum_{i=1}^{K} n_i = Kn$ and we have:

$$
\begin{aligned}
\hat{c} &= \frac{N-K}{N(K-1)} \cdot \frac{\sum_{i=1}^{K}(N-n_i)s_i^2}{\sum_{i=1}^{K}(n_i-1)s_i^2} \\
&= \frac{N-K}{N(K-1)} \cdot \frac{\sum_{i=1}^{K}(N-n)s_i^2}{\sum_{i=1}^{K}(n-1)s_i^2} \\
&= \frac{Kn-K}{Kn(K-1)} \cdot \frac{\sum_{i=1}^{K}(N-n)s_i^2}{\sum_{i=1}^{K}(n-1)s_i^2} \\
&= \frac{K(n-1)}{Kn(K-1)} \cdot \frac{(Kn-n)\sum_{i=1}^{K} s_i^2}{(n-1)\sum_{i=1}^{K} s_i^2} \\
&= \frac{(K-1)n}{(K-1)n} = 1.
\end{aligned}
$$

We conclude that in balanced samples the test statistics of the approximate ANOVA F–test, the modified Brown–Forsythe test, the Brown–Forsythe test, and the ANOVA F–test are the same, $F^*/\hat{c} = F^* = B^* = B = F$.

References

Asiribo, O., Gurland, J. (1990). Coping with variance heterogeneity. Commun. Statist.– Theory Meth., 19, 4029–4048.

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

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one–way classification. Annals of Mathematical Statistics, 25, 290–302.

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.

Chalmers, T. C. (1991). Problems induced by meta–analyses. Statistics in Medicine, 10, 971–980.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. J. Roy. Stat. Soc. Supp., 4, 102–118.

Conover, W. J., Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. American Statistician, 35, 124–129.

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

Fligner, M. A. (1981). Comment on rank transformations as a bridge between parametric and nonparametric statistics. American Statistician, 35, 131–133.

Hardy, R. J., Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine, 17, 841–856.

Hartung, J., Knapp, G. (2000). On tests of the overall treatment effect in the meta–analysis with normally distributed responses. Statistics in Medicine, to appear.

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

Keselman, H. J., Wilcox, R. R. (1999). The 'improved' Brown and Forsythe test for mean equality: some things can't be fixed. Commun. Statist.– Simula., 28, 687–698.

Lehmann, E. L. (1975) Nonparametrics. Holden Day, San Francisco.

Lehmann, E. L. (1986) Testing Statistical Hypotheses. 2nd edn., Wiley, New York.

Li, Y., Shi, L., Roth, H. D. (1994). The bias of the commonly-used estimate of variance in meta–analysis. Commun. Statist.– Theory Meth., 23, 1063–1085.

Mehrotra, D. V. (1997). Improving the Brown–Forsythe solution to the generalized Behrens–Fisher problem. Commun. Statist.–Simula., 26, 1139–1145.

Noether, G. E. (1981). Comment on rank transformations as a bridge between parametric and nonparametric statistics. American Statistician, 35, 129–130.

Normand, S. T. (1999). Meta–analysis: Formulating, evaluating, combining, and reporting. Statistics in Medicine, 18, 321–359.

Patel, J. K., Kapadia, C. P., Owen, D. B. (1976) Handbook of Statistical Distributions. Marcel Dekker, New York.

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

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

Whitehead, A., Whitehead, J. (1991). A general parametric approach to the meta-analysis of randomized clinical trials. Statistics in Medicine, 10, 1665–1677.

Joachim Hartung
Doğan Argaç
Kepher H. Makambi
Department of Statistics
University of Dortmund
D–44221 Dortmund
Germany