Statistical Papers 37, 291–305 (1996)

Statistical Papers

A closer examination on some parametric alternatives to the ANOVA F-test

A. De Beuckelaer

Received: February 27, 1995; revised version: June 26, 1995

In experiments, the classical (ANOVA) F-test is often used to test the omnibus null-hypothesis $\mu_1 = \mu_2 \ldots = \mu_i = \ldots = \mu_n$ (all n population means are equal) in a one-way ANOVA design, even when one or more basic assumptions are being violated. In the first part of this article, we will briefly discuss the consequences of the different types of violations of the basic assumptions (dependent measurements, non-normality, heteroscedasticity) on the validity of the F-test. Secondly, we will present a simulation experiment, designed to compare the type I-error and power properties of both the F-test and some of its parametric adaptations: the Brown & Forsythe F^*-test and Welch's V_w-test. It is concluded that the Welch V_w-test offers acceptable control over the type I-error rate in combination with (very) high power in most of the experimental conditions. Therefore, its use is highly recommended when one or more basic assumptions are being violated. In general, the use of the Brown & Forsythe F^*-test cannot be recommended on power considerations unless the design is balanced and the homoscedasticity assumption holds.

key words: robust analysis of variance (ANOVA); one-way ANOVA design; Brown & Forsythe F^*(-test); Welch V_w(-test); Monte Carlo simulation.

1. Introduction

Although it has been shown by many authors that the classical (ANOVA) F-test cannot be regarded to be very robust to violations of most of the basic assumptions, some researchers still ignore these findings.

The *general formula* for the ANOVA F-test (useable in balanced or unbalanced designs) follows an F-distribution with K-1 degrees of freedom for the numerator and N-K degrees of freedom for the denominator. Its test value is expressed as:

$$F = \frac{\frac{\sum_{j=1}^{k} n_j (\overline{X}_j - \overline{X})^2}{K-1}}{\frac{\sum_{i=1}^{n_j} \sum_{j=1}^{k} (X_{ij} - \overline{X}_j)^2}{N-K}}$$

with:

X_{ij} = *the* i^{th} *observation of the* j^{th} *experimental group* (j = 1,2,...,K);
K = *the total number of experimental groups*;
N = *the total number of observations in the experiment*;
n_j = *the number of observations in the* j^{th} *experimental group*;
$\overline{X}_j$ = *the mean level in the* j^{th} *experimental group*;
$\overline{X}$ = *the overall mean* = $\frac{\sum_{j}^{k} n_j \overline{X}_j}{N}$

All too often, the F-test is used to test the omnibus null-hypothesis $\mu_1=\mu_2=...=\mu_n$ (all n population means are equal) in experimental conditions in which the F-test is found to be an invalid statistical test.

In order to determine the validness of a statistical test, it is common practice to investigate:

- the degree to which the statistical test keeps control over the nominal type I-error rate;
- the power of the statistical test (the complement of the type II-error rate).

From elementary probability theory it is known that there is a gain in power when the researcher is willing to accept that the empirical type I-error rate becomes inflated (above the nominal type I-error rate). According to Cochran's "rule of thumb" (1954), a statistical test may be considered robust at a nominal type I-error rate of 5% if its empirical type I-error rate falls within the interval [4%;6%].[1] Between alternative robust statistical tests, a choice can be made on the basis of the power properties of the

[1] More liberal definitions of robustness were not considered in our study. Bradley's (1978) most liberal acceptance region for the empirical type I-error is bounded by the extreme values: 0.50 $\alpha_{nominal}$ and 1.50 $\alpha_{nominal}$.

different tests under consideration [cfr. J. Jaccard et al. (1984)]. In our article, we focus on the classical 95% confidence interval since the 99% confidence interval is considered far too conservative for testing the omnibus null-hypothesis in experimental designs with three or more experimental groups [A. De Beuckelaer (1994), p. 3].

In the following paragraphs, we will briefly discuss the consequences of the different types of violations on the validity of the F-test. Secondly, we will introduce the most important parametric alternatives to the F-test: the Brown & Forsythe F^*-test and the Welch V_w-test. Finally, we will present and discuss the results of a simulation experiment, designed to compare the type I-error and power properties of both the F-test and these parametric adaptations.

2. Consequences of violating the basic assumptions

2.1 Heterogeneous variances

In a balanced one-way ANOVA design with heterogeneous variances, the validity of the F-test is not seriously affected [S. Olejnik (1987)]. Nevertheless, it has been shown by J.C. Rogan and H.J. Keselman (1977) that the type I-error rate varies as a function of the degree of variance heterogeneity.

Glass et al. (1972) showed that in an unbalanced one-way ANOVA design with heterogeneous variances the F-test can not be considered as very robust to violations of the normality or homoscedasticity assumption. His findings support the following statements:

- given that smaller variances are associated with groups having smaller sample sizes, the empirical type I-error rate becomes deflated (the test becomes too *conservative*);
- given that larger variances are associated with groups having smaller sample sizes, the empirical type I-error becomes inflated (the test becomes too *liberal*) [see also Keselman; Rogan & Feir-Walsch (1977)].

2.2 Non-normality

Although simulation studies have shown that the F-test is generally found to be robust to violations of the normality assumption, it is clear that as the distribution form of the data deviates more from normality, the deviation between the nominal and the empirical

type I-error rate enlarges [S. Olejnik (1987); S.E. Maxwell & H.D. Delaney (1990)].

G.W. Milligan et al. (1987) investigated the influence of non-normality on the empirical type I-error rate in an unbalanced two-factor ANOVA design with homogeneous variances. In general, the F-test was found to be robust at a nominal type I-error rate of 5% (+-1%) for both the uniform and the exponential distribution but not for the lognormal distribution.

2.3 Dependent measurements

In case the measurements are dependent, a (significant) correlation in the mean square of both the numerator and the denominator is introduced. This correlation leads to either a (dramatical) increase or a decrease in the experimental type I-errors [D.A. Kenny & C.M. Judd (1986)]. As a consequence, the F-test can not be considered valid in these circumstances.

2.4 Conclusion

In the previous paragraphs, it has been shown that violations to the basic assumptions enlarge the discrepancy between the actual type I-error rate and the nominal type I-error rate. Therefore, it can be argued that, in a situation in which more than one basic assumption is violated, the use of the F-test becomes very unreliable. Especially violations of the independence assumption, of the homoscedasticity assumption and of the balancedness of the design, have great influence on the (empirical) type I-error rates. In particular, the interactions, such as positive/negative pairings of unequal sample sizes and unequal variances, can result in a very large discrepancy between the nominal and the empirical type I-error rate [Keselman; Rogan & Feir-Walsch (1977)].
Given that the F-test is invalid in most "non-standard" situations, the question remains which (valid) statistical procedure can be used to test the omnibus null-hypothesis. In the following paragraph, we will present some important parametric alternatives to the ANOVA F-test.

3. Some parametric alternatives to the ANOVA F-test

Brown & Forsythe's F^*-test and Welch's V_w-test[2] are certainly the most important parametric alternatives. Both test statistics follow (by approximation) an F-distribution with K-1 degrees of freedom for the numerator and an adjusted number of degrees of freedom for the denominator.

Brown & Forsythe's F^* is calculated as follows: $F^* = \frac{\sum_{j=1}^{k} n_j(\bar{X}_j - \bar{X})^2}{\sum_{j=1}^{k} [1-(\frac{n_j}{N})]\, s_j^2}$

with: K = *the number of (experimental) groups*
$\bar{X}$ = *the overall mean*
$\bar{X}_j$ = *the mean of the j th (experimental) group*
s_j = *the standard deviation of the j th (experimental) group*
N = *the total number of observations*

and f degrees of freedom for the denominator $f = \frac{1}{\sum_{j=1}^{k} \frac{g_j^2}{n_j - 1}}$, where

$$g_j = \frac{[1-(\frac{n_j}{N})]s_j^2}{\sum_{j=1}^{k} [1-\frac{n_j}{N}]s_j^2}.$$

The computation of Welch's V_w is given by the next formula[3]:

---

[2] The Brown & Forsythe F^* and Welch V_w-test procedures are not available in SAS or SPSS but only in the BMDP statistical software package.

[3] In the original article of B.L. Welch (1951), the symbols used in the formula of the V_w-test differ from the symbols used here (for example $w_i = n_i/s_i^2$, see p. 334). Because the symbols that we use are consistent with the symbols used in the formula of both the F-test and the F^*-test, we prefer this version over the original one.

$$V_w = \frac{\sum_{j=1}^{k} \frac{n_j}{s_j^2}\left[\bar{X}_j - \frac{\sum_{j=1}^{k} \frac{n_j}{s_j^2}\bar{X}_j}{\sum_{j=1}^{k} \frac{n_j}{s_j^2}}\right]^2}{k-1}}{\left(1+\frac{2}{3}(k-2)\Lambda\right)}$$

with $\Lambda = \frac{3\sum_{j=1}^{k} \frac{\left[1-\frac{\frac{n_j}{s_j^2}}{\sum_{j=1}^{k} \frac{n_j}{s_j^2}}\right]^2}{n_j-1}}{k^2-1}$ ($1/\Lambda$ = the degrees of freedom for the denominator).

For a comprehensive description of both parametric alternatives, we refer to Maxwell & Delaney (1990).

4. Comparison of different ANOVA-alternatives

4.1 Introduction

In order to test the type I-error and the power properties of these parametric alternatives, we based ourselves on a simulation experiment of A.J. Tomarken & R.C. Serlin (1986). In their study, four different commonly recommended ANOVA alternatives for testing mean differences under variance heterogeneity (the ANOVA F-test; Brown & Forsythe's F^*; Welch's V_w; Kruskal & Wallis-test and inverse normal scores) are compared with respect to their type I-error and power properties. Their study served as an example for our simulation experiment.

4.2 Design and sampling procedure of our simulation experiment

Analogously to Tomarken & Serlin's simulation study we did experiments in a three- and four- group design and focused on three types of mean patterns. For both the three- and four-group designs, four cases showed the behaviour of the procedures when

homoscedasticity holds (A conditions) using either a balanced or an unbalanced[4] design. The remaining cases revealed the effects of heteroscedasticity in equal sample-size conditions (B conditions), direct-pairing conditions (C conditions) and inverse-pairing conditions (D conditions)[5]. In exhibit 1, the different mean patterns, the conditions concerning sample size and variance structure (A;B;C;D) and the three-group or four-group experimental conditions are shown. In every experimental condition, the observations in each experimental group were drawn from a normal distribution with given mean and variance.

Exhibit 1: the 3 types of mean patterns used in the 3- and 4- group simulation study of A.J. Tomarken & R.C. Serlin (1986)

The **ES-pattern** (equally spaced means-pattern):

$\mu_1 < \mu_2 < \mu_3$ ($<\mu_4$) with: $\mu_i = \mu_1 + (i-1) \times$ i=1,2,3,(4) [for all experimental conditions];

note:

- adjacent means differ by the same amount;
- in all unbalanced designs decreasing sample sizes are associated with progressively smaller means

The EX-pattern (extreme mean in one experimental group):

[balanced design & homogeneous variances];

note:

- in balanced designs the extreme group has the largest variance;
- in unbalanced designs both direct-pairing and inverse-pairing conditions are investigated;

The **EX1-pattern** (extreme mean in the first experimental group):

$\mu_1 < [\mu_2 = \mu_3 (=\mu_4)]$ [unbalanced design & homogeneous variances or all heterogeneous variance-cases];

note:

- the first (extreme) group has the largest sample size;

The **EX3- or EX4-pattern** (extreme mean in the last experimental group):

$\mu_1 = \mu_2 < \mu_3$ [three group case] or $\mu_1 = \mu_2 < \mu_3 < \mu_4$ [four group case] [unbalanced design & homogeneous variances or all heterogeneous variance-cases];

---

[4] in a *balanced* design the number of observations/replications in each (experimental) group are the same while in an *unbalanced* design they are different.

[5] *Direct-pairing* means that larger groups are associated with relatively large variances while *inverse-pairing* means that larger groups are associated with relatively small variances.

note:

- the extreme group always has the smallest sample size;
- the extreme group has the extreme mean in the third group in a 3-group design or in the fourth group in a 4-group design;

The **2M-pattern** (two equal means midway between two extreme means):

$\mu_1 < \mu_2 = \mu_3 < \mu_4$ [all 4-group cases]

note:

- two groups with equal means are halfway between two extreme groups;
- in all unbalanced designs, the extreme groups are always those with the largest and smallest sample sizes;
- in all balanced, heterogeneous variance cases, the extreme groups are simply the groups with the largest and smallest variances.

Because Tomarken's study provided empirical evidence to state that the parametric adaptations to the ANOVA F-test are the most powerful statistical procedures, we did not replicate simulations on any of the nonparametric procedures (the Kruskal-Wallis and inverse normal scores tests). Moreover, it has been shown that the Kruskall-Wallis test is not robust to violations of the assumption of identical population distributions.[6] Our main purpose was to investigate the behaviour of the (most powerful) parametric procedures and (eventually)[7] the behaviour of the ANOVA F-test in the various experimental conditions.

One critical point concerning Tomarken's simulation study is that conclusions about superiority in power of a statistical test procedure were essentially based on only one mean-setting in each experimental condition[8] (1 point[9] of the power function). In order to become more reliable results, we replicated their simulation experiment in such a way that three intersective

---

[6] The Kruskall-Wallis test assumes identical population distributions. As a consequence, homogeneity of variance is implicitly assumed. Keselman, Rogan and Feir-Walsch (1977) as well as Tomarken and Serlin (1986) have shown that the type I-error rate could be as large as twice the nominal level when large samples are paired with small variances [cf. Maxwell & Delaney (1990), p. 708].

[7] The behaviour of the ANOVA F-test will be evaluated only in experimental conditions in which this test offers an acceptable control over the experimental type I-error.

[8] Each experimental condition is determined by the number of groups, the mean pattern, the number of observations in each experimental group, and the variance structure of the experimental groups.

[9] Their mean-settings were chosen at an estimated ANOVA-(F-test)power of .70 for a given mean pattern. For more technical details concerning the estimation of the power of the ANOVA F-test (based on Cohen's ANOVA effect size index), the reader is encouraged to consult the original article of A.J. Tomarken and R.C. Serlin (1986).

points[10] on the power curve were chosen in each experimental condition.

Some additional differences between the two simulation experiments are certainly worth mentioning:

- instead of 1000, a total of 2500 test-files were generated for every mean-setting of a specific experimental condition;
- to generate normal random variables, the KR-Algorithm[11] was used instead of the more traditional Box-Muller method, as the former was found to be a faster and more accurate sampling procedure[12] [13] than the latter. The simulation software was written in TURBO PASCAL 6.

4.3 Results

In table 1 it is shown that in the homogeneous variance-cases (A-conditions) both the ANOVA F-test and its parametric adaptations offer an acceptable control over the type I-error. As a consequence, all statistical test procedures in our simulation study could be considered valid in these experimental conditions. Table 1 shows however that, in general, the parametric alternatives to the ANOVA F-test keep the experimental type I- error somewhat closer to its nominal value. The ANOVA F-test tends to be more conservative.

---

[10] Our mean-settings were chosen at an estimated ANOVA-(F-test)-power of .30, .50 and .70 for a given mean pattern. It turned out that in various experimental conditions, the discrepancy between the estimated ANOVA- F-test power and the real power could be large (up to 0.2) depending on the unbalancedness, the variance heterogeneity and the number of subjects in each experimental group.

[11] cf. A.J. Kinderman & J.G. Ramage (1976).

[12] cf. A.J. Kinderman & J.G. Ramage (1976) and A.M. Law & D. Kelton (1991).

[13] Many thanks to Dr. J. Annaert (University of Antwerp-RUCA) who delivered us a PASCAL-routine that can be used to generate normal random variables following the KR-methodology.

Table 1: Empirical type I-error probabilities at a nominal type I-error rate of .05

Variances -->	HOMOGENEOUS			HETEROGENEOUS											
	6 / 6 / 6			12 / 4 / 1			6 / 2 / 1			1 / 4 / 12			1 / 2 / 6		
Number of observations	F	F^*	V_w	F	F^*	V_w	F	F^*	V_w	F	F^*	V_w	F	F^*	V_w
20 / 20 / 20	0.050 (A1)	0.054 0.054	0.054 0.054	*0.064 (B1)*	0.059 0.059	*0.038 0.038*	*0.062 (B5)*	0.059 0.059	0.048 0.049						
12 / 12 / 12	0.043 (A2)	0.046 0.048	0.048 0.049	*0.072 (B2)*	*0.063 0.064*	0.057 0.058	*0.061 (B6)*	0.056 0.058	0.045 0.046						
30 / 20 / 10	0.043 (A3)	0.046 0.046	0.048 0.048	*0.020 (C1)*	*0.060 0.061*	0.047 0.048	*0.022 (C5)*	*0.062 0.062*	0.050 0.050	*0.157 (D1)*	0.058 0.060	0.047 0.049	*0.143 (D5)*	0.055 0.056	0.042 0.043
18 / 12 / 6	0.052 (A4)	0.058 0.060	0.058 0.060	*0.020 (C2)*	0.059 0.060	0.049 0.046	*0.022 (C6)*	0.057 0.057	0.053 0.054	*0.165 (D2)*	*0.062 0.065*	0.050 0.052	*0.141 (D6)*	0.056 0.059	0.050 0.052
Variances -->	HOMOGENEOUS			HETEROGENEOUS											
	6 / 6 / 6 / 6			12 / 6 / 4 / 1			6 / 3 / 2 / 1			1 / 4 / 6 / 12			1 / 2 / 3 / 6		
Number of observations	F	F^*	V_w	F	F^*	V_w	F	F^*	V_w	F	F^*	V_w	F	F^*	V_w
20/20/20/20	0.046 (A5)	0.049 0.049	0.048 0.048	*0.065 (B3)*	*0.062 0.062*	0.044 0.044	*0.062 (B7)*	0.060 0.060	0.053 0.054						
12/12/12/12	0.042 (A6)	0.047 0.047	0.050 0.051	*0.065 (B4)*	0.057 0.058	0.049 0.051	*0.069 (B8)*	*0.066 0.068*	0.054 0.056						
30/24/16/10	0.050 (A7)	0.053 0.053	0.050 0.051	*0.025 (C3)*	*0.063 0.063*	0.049 0.050	*0.029 (C7)*	*0.062 0.062*	0.052 0.052	*0.140 (D3)*	0.056 0.056	0.049 0.050	*0.128 (D7)*	0.058 0.059	0.056 0.056
18/14/10/ 6	0.045 (A8)	0.050 0.051	0.053 0.055	*0.029 (C4)*	*0.074 0.074*	0.049 0.050	*0.032 (C8)*	*0.070 0.070*	0.051 0.053	*0.142 (D4)*	*0.062 0.060*	0.052 0.055	*0.133 (D8)*	0.048 0.050	0.045 0.046

note:
- because the adjusted degrees of freedom for the denominator are real numbers that are situated between two integer values, **two** different **empirical type I-error rates** are calculated for the parametric alternatives to the ANOVA F-test. One empirical type I-error is calculated for the lower integer value and one for the higher integer value;
- empirical type I-errors printed in italics exceed the critical bounds indicated by Crochran's "rule of thumb" (4% and 6% at a nominal type I-error of 5%);
- the naming of the different experimental conditions is given between brackets in the column of the (ANOVA) F-test;
- to measure the type I-error probabilities all population means were set to zero;
- for every experimental condition 2500 test-files were generated.

Table 2: Relative[14] superiority in power of the Brown & Forsythe F^*-test and the Welch V_w-test in HOMOGENEOUS VARIANCE cases[1 2]

3 GROUPS		F	F^*	V_w	Undecided	Total
Balanced Design						
[A1;A2]	ES-pattern	0	6	0	0	6
	EX-pattern	1	4	0	1 x F or F^*	6
Unbalanced Design						
[A3;A4]	ES-pattern	3	3	0	0	6
	EX1-pattern	5	1	0	0	6
	EX3-pattern	2	4	0	0	6
4 GROUPS		F	F^*	V_w		
Balanced Design						
[A5;A6]	ES-pattern	0	4	1	1 x F or F^* or V_w	6
	EX-pattern	0	5	0	1 x F or F^*	6
	2M-pattern	1	4	0	1 x F or F^* or V_w	6
Unbalanced Design						
[A7;A8]	ES-pattern	4	1	0	1 x F or F^*	6
	EX1-pattern	3	2	0	1 x F or V_w	6
	EX4-pattern	4	0	0	1 x F or F^* 1 x F or F^* or V_w	6
	2M-pattern	3	1	0	2 x F or F^*	6
TOTAL NUMBER OF RUNS						72

note:

- for every experimental condition 2500 test-files were generated at 3 intersective points of the power curve (3 different mean-settings with an estimated F-test power of .30; .50 and .70);
- the Welch V_w-test was found to be the least powerful test-procedure in almost every experimental condition;
- the numbers in the columns indicate the number of times that the F-test, the Brown & Forsythe F^*-test or the Welch V_w-test was found to be the most powerful statistical test.

[14] Tables with the actual power values of the different tests under consideration can be obtained from the author.

Table 3: Relative[15] superiority in power of the Brown & Forsythe F^*-test and the Welch V_w-test in HETEROGENEOUS VARIANCE cases

3 GROUPS		F^*	V_w	Undecided (F^* or V_w)	Total
Balanced Design					
[B1;B2;B5;B6]	ES-pattern	2	**10**	0	12
	EX1-pattern°	**12**	0	0	12
	EX3-pattern°°	0	**12**	0	12
Unbalanced Design					
Direct-pairing	ES-pattern	3	**7**	2	12
[C1;C2;C5;C6]	EX1-pattern°	**12**	0	0	12
	EX3-pattern°°	3	**6**	3	12
Inverse-pairing	ES-pattern	0	**12**	0	12
[D1;D2;D5;D6]	EX1-pattern°°	4	**8**	0	12
	EX3-pattern°	**12**	0	0	12
4 GROUPS		F^*	V_w		
Balanced Design					
[B3;B4;B7;B8]	ES-pattern	0	**11**	1	12
	EX1-pattern°	**12**	0	0	12
	EX4-pattern°°	0	**10**	2	12
	2M-pattern	0	**12**	0	12
Unbalanced Design					
Direct-pairing	ES-pattern	0	**10**	2	12
[C3;C4;C7;C8]	EX1-pattern°	**11**	0	1	12
	EX4-pattern°°	0	**7**	5	12
	2M-pattern	1	**8**	3	12
Inverse-pairing	ES-pattern	0	**12**	0	12
[D3;D4;D7;D8]	EX1-pattern°°	4	**8**	0	12
	EX4-pattern°	**12**	0	0	12
	2M-pattern	0	**12**	0	12
TOTAL NUMBER OF RUNS					252

e: ° the extreme means are associated with the groups having larger variances;
°° the extreme means are associated with the groups having smaller variances;
- for every experimental condition 2500 test-files were generated at 3 intersective points of the power curve (3 different mean-settings with an estimated F-test power of .30; .50 and .70);
- the power of the F-test is not evaluated since it is shown that the F-test is not able to offer a good control of the type I-error rate in HETEROGENEOUS VARIANCE cases;
- the numbers in the columns indicate the number of times that the Brown & Forsythe F^*-test or the Welch V_w-test was found to be the most powerful statistical test.

[15] cf. footnote 14.

Tomarken & Serlin concluded that in the homogeneous variance-case the F-test was consistently the most powerful test-procedure.

With respect to conditions in which the homoscedasticity assumptions holds, the replications in our simulation study could not support an absolute superiority in power of the F-test. Especially in (3- and 4- group) balanced designs, the Brown & Forsythe's F^*-test procedure was found to be more powerful than the F-test (see table 2). Although we must admit that the difference in power is very small (between 0.00 and 0.157) in our simulation experiment. Given that the F^*-test keeps better control over the empirical type I-error rate and proofs to be somewhat more powerful than the F-test, we would recommend it for general use in balanced designs when the homoscedasticity assumption is met.

As it is shown (cf. table 1) that the ANOVA F-test is invalid in heterogeneous variance-cases, a choice has to be made between the parametric alternatives.

For the heterogeneous variance-cases Tomarken & Serlin state that: "On the basis of superior control of type I-errors and greater power, the Welch's V_w-test proved to be the procedure of choice when means were equally spaced, when extreme means were paired with small variances and when two identical means were situated midway between two extreme means. This recommendation applies whether sample sizes are equal or directly or inversely paired with variances." [A.J. Tomarken & R.C. Serlin (1986), p. 90]. Brown & Forsythe's F^* was found to be the optimal test-procedure only if extreme means were paired with large variances (see also table 3).

As far as type I-error and power considerations in heterogeneous variance-cases are concerned, our findings are identical to those contained in Tomarken & Serlin's article. Nevertheless, we are convinced that, even when extreme means were paired with large variances, the use of Brown & Forsythe's F^*-test-procedure must not be recommended in heterogeneous variance-cases because of its inability to control the type I-error rate in an acceptable way in most of the experimental conditions [16] (see table 1). Therefore, we cannot agree with M.B. Brown & A.B. Forsythe (1974) who stated earlier that: "The choice between W and F^* depends upon whether extreme means are thought to have extreme variances (W) or not (F^*)." [M.B. Brown & A.B. Forsythe (1974), p. 131].

---

[16] In particular we consider the experimental conditions: B2, B3, B8, C1, C3, C4, C5, C7, C8, D2, D4 in which extreme means were paired with large variances. In these experimental conditions the Brown & Forsythe F^*-test could not guarantee an acceptable control of type I-errors.

Researchers who are willing to accept a more liberal definition of robustness might consider the Brown & Forsythe F^*-test as a valid statistical test (cf. table 1). However, it is absolutely necessary to realize that additional violations of the independence and normality assumption, which were not considered in our simulation experiment, could turn the Brown & Forsythe F^*-test into a completely invalid statistical test.

4.4 Conclusion

Summarizing the above findings, we would recommend the use of Welch's V_w-test in case the homoscedasticity assumption does not hold. Only when there is enough empirical evidence to believe that variances are identical, would we recommend the use of the ANOVA F-test in unbalanced designs and the Brown & Forsythe F^*-test in balanced designs. However, when one is not sure whether the homoscedasticity assumption does or does not hold, the use of the Welch V_w-test must be recommended. These recommendations apply when the normality and the independence assumtion is met. Clinch & Keselman (1982) found that the Welch V_w-test becomes somewhat liberal for skewed distributions while the Brown & Forsythe F^*-test provided acceptable control over the type I-error rate, following Bradley's liberal definition of robustness. Further research is needed to clarify the impact of these additional violations of assumptions on the robustness of these parametric alternatives.

Acknowledgement

I would like to thank the referee for helpful comments on a previous version of this paper.

References

BRADLEY, J.V. (1978), Robustness?, *The British Journal of Mathematical and Statistical Psychology*, vol. 31, pp. 144-152.

BROWN, M.B. & FORSYTHE, A.B. (1974), The Small Sample Behavior of Some Test Statistics Which Test the Equality of Several Means, *Technometrics*, vol. 16, number 1, pp. 129-132.

CLINCH, J.J. & KESELMAN, H.J. (1982), Parametric Alternatives to the Analysis of Variance, *Journal of Educational Statistics*, vol. 7, pp. 207-214.

COCHRAN, W.G. (1954), Some Methods for Strengthening the Common X^2-tests, *Biometrics*, vol. 10, pp. 417-451.

DE BEUCKELAER, A. (1994), *Testing the Omnibus Null-hypothesis in "Non-Standard" Situations*, Antwerp: RUCA - STE, 14 p. (Working Paper; 1994:08)

GLASS, G.V., PECKHAM, P.D. & SANDERS, J.R. (1972), Consequences of Failure to Meet Assumptions Underlying the Fixed Effects Analysis of Variance and Covariance, *Review of Educational Research*, vol. 42, pp. 237–287.

JACCARD, J, BECKER, M.A., WOOD, G. (1984), Pairwise Multiple Comparison Procedures: A Review, *Psychological Bulletin*, vol. 96, number 3, pp. 589–596.

KENNY, D.A. & JUDD, C.M. (1986), Consequences of Violating the Independence Assumption in Analysis of Variance, *Psychological Bulletin*, vol. 99, number 3, pp. 422–431.

KESELMAN, H.J., ROGAN, J.C. & FEIR-WALSCH, B.J. (1977), An Evaluation of Some Nonparametric and Parametric Tests for Location Equity, *The British Journal of Mathematical and Statistical Psychology*, vol. 30, pp. 213–221.

KINDERMAN, A.J. & RAMAGE, J.G. (1976), Computer Generation of Normal Random Variates, *Journal of the American Statistical Association*, vol. 71, number 356, pp. 893–896.

LAW, A.M. & KELTON, W.D. (1991), *Simulation Modeling & Analysis*, New York, McGraw-Hill, 759 p.

MAXWELL, S.E. & DELANEY, H.D. (1990), *Designing Experiments and Analyzing Data*, California, Wadsworth, 902 p.

MILLIGAN, G.W., WONG, D.S. & THOMPSON, P.A. (1987), Robustness Properties of Nonorthogonal Analysis of Variance, *Psychological Bulletin*, vol. 101, number 3, pp. 464–470.

OLEJNIK, S.F. (1987), Conditional ANOVA for Mean Differences When Population Variances Are Unknown, *Journal of Experimental Education*, vol. 55, number 3, pp. 141–148.

ROGAN, J.C. & KESELMAN, H.J. (1977), Is the ANOVA F-test Robust to Variance Heterogeneity When Sample Sizes are Equal?: An Investigation via a Coefficient of Variation, *American Educational Research Journal*, vol. 14, number 4, pp. 493–498.

TOMARKEN, A.J. & SERLIN, R.C. (1986), Comparison of ANOVA Alternatives Under Variance Heterogeneity and Specific Noncentrality Structures, *Psychological Bulletin*, vol. 99, number 1, pp. 90–99.

WELCH, B.L. (1951), On the Comparison of Several Mean Values: An Alternative Approach, *Biometrika*, vol. 38, pp. 330–336.

WILCOX, R.R., CHARLIN, V.L. & THOMPSON, K.L. (1986), New Monte Carlo Results on the Robustness of the ANOVA F, W, and F^* Statistics, *Communications in Statistics – Simulation and Computation*, vol. 15, pp. 933–943.

Alain De Beuckelaer
University of Antwerp – RUCA
Faculty of Applied Economics
Computer Science and Operations Management
Middelheimlaan 1
B-2020 Antwerpen
BELGIUM