VERSITA

Mathematica Slovaca

DOI: 10.2478/s12175-009-0156-x
Math. Slovaca **59** (2009), No. 6, 679–692

ACCURACY OF p-VALUES OF APPROXIMATE TESTS IN TESTING FOR EQUALITY OF MEANS UNDER UNEQUAL VARIANCES

Júlia Volaufová

Dedicated to 70th birthday of Professor Andrej Pázmán

(*Communicated by Gejza Wimmer*)

ABSTRACT. Seemingly, testing for fixed effects in linear models with variance-covariance components has been solved for decades. However, even in simple situations such as in fixed one-way model with heteroscedastic variances (a multiple means case of the Behrens-Fisher problem) the questions of statistical properties of various approximations of test statistics are still alive. Here we present a brief overview of several approaches suggested in the literature as well as those available in statistical software, accompanied by a simulation study in which the accuracy of p-values is studied. Our interest is limited here to the Welch's test, the Satterthwaite-Fai-Cornelius test, the Kenward-Roger test, the simple ANOVA F-test, and the parametric bootstrap test. We conclude that for small sample sizes, regardless the number of compared means and the heterogeneity of variance, the ANOVA F-test p-value performs the best. For higher sample sizes (at least 5 per group), the parametric bootstrap performs well, and the Kenward-Roger test also performs well.

1. Introduction

In many ways the analysis of variance setting, where we compare the means of several populations, is a classic, well studied statistical problem. However, it can happen that a statistician analyzing the data in such a case finds himself in a puzzling situation. The general assumption in a multiple means comparison setting is that the underlying population variances are equal. If this assumption

2000 Mathematics Subject Classification: Primary 62J10, 62F10.
Keywords: heteroscedastic one-way fixed model, approximate F-test, accuracy of p-value.

is not backed up independently of the data, we have a choice of several approximate procedures for testing the equality of means taking into consideration the potential heterogeneity of variances. It is not uncommon to try several tests and, in situations in which the conclusions are inconsistent, the embarrassment is born - what to conclude and how to back up the conclusion? In other words, which procedure dominates the other? This question is recurring in the literature starting from late 30s of the 20th century. There is no unique answer. Several authors proposed approximate tests or studied the size, and in some cases power, of available tests using Monte Carlo simulations. These include, just to name a few, Brown and Forsythe [1], Weerahandi [11], Lee and Ahn [8], and most recently Krishnamoorthy et al. [7]. This list is far from exhaustive.

Most of the authors who investigated the size (or power) of proposed tests looked for a choice of a dominant test in relation to the true population variances and/or their relationship to sample sizes. The glitch in this approach is the fact that the variances are unknown, hence we cannot assess a certain configuration in relation to sample size which would provide adequate information about which approximate test to use. Krishnamoorthy et al. [7] recently compared Welch's test, James's test, the generalized F-test, and in addition, they proposed a parametric bootstrap test (PB). They compared the size and power of these tests with respect to sample sizes and number of means to be compared. In their conclusion, the PB test seemed to perform reasonably well; in fact they claim: "In terms of controlling the Type I error rate, the overall conclusion is that the PB test is the only procedure that performs satisfactorily, regardless the sample sizes, values of error variances, and the number of means being compared".

In the present paper, we focus on investigating the properties of p-values of some of the approximate tests studied by others. Even as we investigate different variance configurations, our focus is to look for a relationship between methods and the configuration of sample sizes, mainly small sample sizes. We are mostly interested in *accuracy* of p-values. The p-value is said to be *accurate at* α if $P(p \leq \alpha) = \alpha$. We investigate the extent to which this property is met for the PB-test and Welch's test, and we consider the Fai and Cornelius generalization of Satterthwaite's approximation, the Kenward-Roger type approximation, and, for comparison, also the basic ANOVA F-test.

2. Approximate test procedures for testing equality of means

The multiple means model with heteroscedastic variances is usually presented as

$$Y_{ij} = \mu_i + \epsilon_{ij}, \qquad i = j, \ldots, k, \;\; j = 1, \ldots, n_i, \tag{1}$$

where the random variables ϵ_{ij} follow the $N(0, \sigma_i^2)$ distribution and all are mutually independent. Here Y_{ij} denotes the response in the jth observation from the ith population. Sample sizes are $n_1, n_2, \ldots, n_k$. Population means are $\mu_1, \mu_2, \ldots, \mu_k$, and population variances are $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$. Denote by $\bar{Y}$ the k-dimensional vector of sample means with the ith entry $\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}$ and by $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$, $i = 1, 2, \ldots, k$, the sample variances.

The main goal is to test the null hypothesis H_0, which states that all k population means are equal, against the alternative that at least one mean is different from the rest. Denote the k-dimensional vector of means by μ and the vector of ones by $\mathbf{1}$. Let L be a $(k-1) \times k$ full row rank contrast matrix, i.e., the $k-1$ rows of L are linearly independent and $L\mathbf{1} = 0$. The null hypothesis can be formally expressed as $L\mu = 0$, which is equivalent to

$$\mu' L' \left(LCL'\right)^{-1} L\mu = 0 \tag{2}$$

for any $k \times k$ symmetric matrix C such that LCL' is nonsingular. It is easy to verify that when C is nonsingular, the quadratic form in μ in (2) is identical to

$$\mu'(C^{-1} - C^{-1}\mathbf{1}(\mathbf{1}'C^{-1}\mathbf{1})^{-1}\mathbf{1}'C^{-1})\mu. \tag{3}$$

That is, (2) is invariant to the choice of L.

The most common approach is to base the test statistic for testing H_0 on (3) using $\bar{Y}$ for μ and the covariance matrix $\operatorname{cov}(\bar{Y})$ for C. In model (1), $\operatorname{cov}(\bar{Y}) = \operatorname{Diag}\left\{\frac{\sigma_i^2}{n_i}\right\}$. If the variances σ_i^2 are all known, under H_0, the random variable

$$T = \bar{Y}'(C^{-1} - C^{-1}\mathbf{1}(\mathbf{1}'C^{-1}\mathbf{1})^{-1}\mathbf{1}'C^{-1})\bar{Y} \left(= \bar{Y}' L' \left(LCL'\right)^{-1} L\bar{Y}\right), \tag{4}$$

follows a chi-square distribution with $k-1$ degrees of freedom, χ^2_{k-1}.

In virtually every setting of this sort, the variances σ_i^2s are unknown. Denote by $q = \operatorname{rank}(L)\big(= k-1\big)$. If in (4) instead of C we use its estimate $\hat{C}$, we get a random variable

$$T^* = \bar{Y}' L' \left(L\hat{C}L'\right)^{-1} L\bar{Y} = \bar{Y}'(\hat{C}^{-1} - \hat{C}^{-1}\mathbf{1}(\mathbf{1}'\hat{C}^{-1}\mathbf{1})^{-1}\mathbf{1}'\hat{C}^{-1})\bar{Y}, \tag{5}$$

which, if $\hat{C} = \operatorname{Diag}\left\{\frac{S_i^2}{n_i}\right\}$, takes the form

$$\sum_{i=1}^{k} \left(n_i/S_i^2\right) \bar{Y}_i^2 - \frac{\left[\sum_{i=1}^{k} \left(n_i/S_i^2\right) \bar{Y}_i\right]^2}{\sum_{i=1}^{k} n_i/S_i^2}. \tag{6}$$

T^* in (5) is used as the basis of several of the tests considered here. However, the distribution of T^* under H_0 is no longer a χ^2 distribution.

Let's digress a little and consider the unknown variances all equal to an unknown variance, say σ^2. In this well known special simple case the matrix C takes the form $C = \sigma^2 \operatorname{Diag}\left\{\frac{1}{n_i}\right\} = \sigma^2 W$. The variance σ^2 is estimated by a pooled variance estimator, $\hat{\sigma}^2 = 1/(N-k)\sum_{i=1}^{k}(n_i-1)S_i^2$, with $N = \sum_{i=1}^{k} n_i$. Then if we use $\hat{C} = \hat{\sigma}^2 W$ in the expression for T^* in (5), we get the ratio

$$F^* = \frac{1}{q}T^* = \frac{(1/q)\bar{Y}'(W^{-1} - W^{-1}\mathbf{1}(\mathbf{1}'W^{-1}\mathbf{1})^{-1}\mathbf{1}'W^{-1})\bar{Y}}{\hat{\sigma}^2}, \tag{7}$$

which is a notoriously known F-statistic, which under H_0 follows the $F(q, N-k)$ distribution.

This provides a heuristic motivation to approximate the distribution of T^*/q by F distribution. However, even when all σ^2s are equal, T^*/q does not follow an F distribution unless all n_i are equal, and the setting is more murky when the population variances are different. The idea now is to mimic the common variance case and possibly find a coefficient, say m and degrees of freedom, say ν, both functions of observed s_i^2s, such that the distribution of $F = mF^*$ would be approximated by $F(q, \nu)$. Next we very briefly describe the approximate test statistics for which we shall investigate the behavior of their corresponding p-values.

2.1. Welch's test

W e l c h in [12] derived the coefficient and the denominator degrees of freedom. He showed that choosing

$$m_W = \left[1 + \frac{2(k-2)}{(k^2-1)}\sum_{i=1}^{k}\frac{1}{n_i - 1}\left(1 - \frac{n_i/s_i^2}{\sum_{i=1}^{k} n_i/s_i^2}\right)^2\right]^{-1}, \tag{8}$$

and

$$\nu_W = \left[\frac{3}{(k^2-1)}\sum_{i=1}^{k}\frac{1}{n_i - 1}\left(1 - \frac{n_i/s_i^2}{\sum_{i=1}^{k} n_i/s_i^2}\right)^2\right]^{-1}, \tag{9}$$

the distribution of $F_W = m_W F^*$ is approximately $F(q, \nu_W)$, and hence the p-value resulting from Welch's approximation is $p_W = P(F_W > f_W)$ where f_W is the observed value of F_W.

2.2. Satterthwaite-Fai-Cornelius test

The next test is referred to as Satterthwaite's approximation. However, it is its extension and generalization that can be used for comparison of k means for $k > 2$. Giesbrecht and Burns [4] derived the approximate degrees of freedom for the case of L with a row rank equal to 1 using Satterthwaite's results (see [9] and [10]). Fai and Cornelius [3] extended the results of Giesbrecht and Burns for rank $r(L) = q$, $q > 1$. Recall that $q = k - 1$ in our special case. The main idea of their approach is as follows. Consider the test statistics $F^* = (1/q)\, T^*$, T^* from (5). Using spectral decomposition, there exist a matrix P of eigenvectors and a diagonal matrix Λ of eigenvalues of $L\hat{C}L'$ such that $L\hat{C}L' = P\Lambda P'$. Then

$$F^* = \frac{1}{q}\bar{Y}'L'P\Lambda^{-1}P'L\bar{Y} = \frac{1}{q}\sum_{t=1}^{q}\frac{1}{\lambda_t}(e_t'P'L\bar{Y})^2 = \frac{1}{q}\sum_{t=1}^{q}\left(\frac{e_t'P'L\bar{Y}}{\sqrt{\lambda_t}}\right)^2, \qquad (10)$$

where e_t is a q-vector with 1 on the tth place and zeros elsewhere. Each term in (10) is approximated by a square of Student's t-distribution with δ_t degrees of freedom. Following Giesbrecht and Burns [4], the degrees of freedom δ_t are obtained from the relationship

$$\delta_t \approx \frac{2\lambda_t^2}{\operatorname{approx}\widehat{\operatorname{Var}}(\lambda_t)}. \qquad (11)$$

Using statistical differentials and the fact that S_i^2s are independent and their variances are estimated by $\widehat{\operatorname{Var}}(S_i^2) = \frac{2S_i^4}{n_i - 1}$, the denominator of (11) in our case simplifies to

$$\begin{aligned}\operatorname{approx}\widehat{\operatorname{Var}}(\lambda_t) &= \operatorname{approx}\widehat{\operatorname{Var}}(e_t'P'L\hat{C}L'Pe_t)\\ &\approx \sum_{i=1}^{k}\left(\frac{\partial(e_t'P'L\hat{C}L'Pe_t)}{\partial s_i^2}\right)^2\frac{2S_i^4}{n_i-1} = 2\sum_{i=1}^{k}\frac{(e_t'P'Le_i)^4S_i^4}{n_i^2(n_i-1)}.\end{aligned}$$

If F^* follows approximately F distribution with q and, say, ν_S degrees of freedom, then

$$E(F^*) = \frac{\nu_S}{\nu_S - 2}. \qquad (12)$$

The expression (12) together with

$$E(F^*) = \frac{1}{q}\sum_{t=1}^{q}E\left(\frac{e_t'P'L\bar{Y}}{\sqrt{\lambda_t}}\right)^2 = \frac{1}{q}\sum_{t=1}^{q}\frac{\delta_t}{\delta_t - 2} \qquad (13)$$

imply that

$$\nu_S = \frac{2\sum\limits_{t=1}^{q}\frac{\delta_t}{\delta_t-2}}{\sum\limits_{t=1}^{q}\frac{\delta_t}{\delta_t-2} - q}. \qquad (14)$$

It can happen that $\delta_t \leq 2$. For such cases the corresponding terms are left out of the sum in (13). Hence the Satterthwaite-Fai-Cornelius approximation leads to

$$F_S = F^* \approx F(q, \nu_S), \tag{15}$$

with a p-value $p_S = P(F_S > f_S)$, where f_S is the observed value of F_S and the probability is computed from $F(k-1, \nu_S)$.

2.3. Kenward-Roger test

Using Taylor's approximation in several consecutive steps, K e n w a r d and R o g e r [6] extended the results of H a r v i l l e and J e s k e [5] and derived the coefficient m and denominator degrees of freedom ν in a general setting of a linear model with variance-covariance components and general linear hypothesis about the parameter of the mean. Without going into technical details, here we present only the simplified expressions for the coefficient m_{KR} and degrees of freedom ν_{KR} for the special case of model (1).

We shall need the following expressions.

$$\begin{aligned} a &= 2\sum_{i=1}^{k} \left(e_i'(\hat{C}^{-1} - \hat{C}^{-1}\mathbf{1}(\mathbf{1}'\hat{C}^{-1}\mathbf{1})^{-1}\mathbf{1}'\hat{C}^{-1})e_i \right)^2 \frac{s_i^4}{n_i^2(n_i-1)} \\ &= 2\sum_{i=1}^{k} \frac{1}{n_i - 1} \left(1 - \frac{n_i/s_i^2}{\sum_{i=1}^{k} n_i/s_i^2} \right)^2; \end{aligned}$$

$$c_1 = -\frac{21}{2(3k^2+2k+5)(k-1)}; \quad c_2 = \frac{7(k^2+2)}{2(3k^2+2k+5)(k-1)};$$

$$c_3 = \frac{7(k^2+2k+4)}{2(3k^2+2k+5)(k-1)};$$

$$E^* = \left(1 - \frac{a}{k-1}\right)^{-1}; \qquad V^* = \frac{2}{k-1}\left(\frac{1 + c_1 a}{(1-c_2 a)^2(1-c_3 a)}\right).$$

Then

$$\nu_{KR} = 4 + \frac{k+1}{(k-1)\rho - 1}, \qquad \text{and} \qquad m_{KR} = \frac{\nu_{KR}}{E^*(\nu_{KR}-2)}, \tag{16}$$

where

$$\rho = \frac{V^*}{2E^{*2}}.$$

Then the distribution of $F_{KR} = m_{KR}F^*$ can be approximated by $F(k-1, \nu_{KR})$ and the p-value of the Kenward-Roger test is computed from $F(k-1, \nu_{KR})$ and is equal to $p_{KR} = P(F_{KR} > f_{KR})$, where f_{KR} is the observed value of F_{KR}.

2.4. Parametric bootstrap

The parametric bootstrap method, as described in detail in Krishnamoorthy et al. [7], involves resampling from a distribution whose parameters are the sample means and sample variances. The reference value for the test is the observed value f^* of F^*. Generating independent $Z_i \sim N(0,1)$ and $\chi^2_{n_i-1}$ random variables, we set $\bar{Y}_{Bi} = \sqrt{\frac{s_i^2}{n_i}} Z_i$ and $S^2_{Bi} = \frac{s_i^2}{n_i-1}\chi^2_{n_i-1}$, $i = 1, 2, \ldots, k$. Plugging into (5) the $\bar{Y}_{Bi}$s and S^2_{Bi}s we get

$$F_{PB} = \frac{1}{q}\left[\sum_{i=1}^{k}\left(n_i/S^2_{Bi}\right)\bar{Y}^2_{Bi} - \frac{\left[\sum\limits_{i=1}^{k}\left(n_i/S^2_{Bi}\right)\bar{Y}_{Bi}\right]^2}{\sum\limits_{i=1}^{k} n_i/S^2_{Bi}}\right]. \tag{17}$$

The p-value of parametric bootstrap test is the proportion of resampled cases in which the observed value of F_{PB}, f_{PB}, exceeds f^*.

2.5. Implementation in SAS

Welch's, Satterthwaite-Fai-Cornelius, and the Kenward-Roger tests are all implemented in more or less modified versions in SAS. Welch's test is available in Proc GLM as an option in the MEANS statement. The Satterthwaite-Fai-Cornelius and Kenward-Roger tests are available in Proc Mixed as options for the choice of denominator degrees of freedom in the MODEL statement.

As stated above, the test statistic F^*, if based on $L'\bar{Y}$, is invariant with respect to a choice of the contrast matrix L. However, the Satterthwaite-Fai-Cornelius approximate degrees of freedom ν_S depend on the choice of L, hence we suggest caution when using this option. The degrees of freedom ν_S are obtained from (14) and the approximate F-distribution from (15) with the following exceptions. If the calculated $\nu_S > N - k$ then ν_S is set to $N - k$, i.e., the p-value is obtained from $F(q, N-k)$. If $\nu_S \leq 0$ then ν_S is set to 1, hence the p-value is obtained using $F(q, 1)$.

The implementation of K e n w a r d - R o g e r is more complex since both the coefficient m_{KR} and the denominator degrees of freedom ν_{KR} get modified depending on how their calculated values turn out. If ν_{KR} from (16) is less than or equal to 1, then ν_{KR} is set to 1 and also m_{KR} is set to 1. Hence the Kenward-Roger p-value p_{KR} is calculated from $p_{KR} = P(F^* > f^*)$ using $F(q, 1)$. If $1 < \nu_{KR} < 2$, then m_{KR} is set to 1 and ν_{KR} is kept as calculated, hence $p_{KR} = P(F^* > f^*)$ using $F(q, \nu_{KR})$. Only if $\nu_{KR} \geq 2$, then $F_{KR} = m_{KR}F^*$, with m_{KR} given by (16) and $p_{KR} = P(F_{KR} > f_{KR})$ using $F(q, \nu_{KR})$.

3. Simulation study

In order to investigate the behavior of p-values, a simulation study was done with all combinations of configurations listed below. Our focus is to study the accuracy of p-values, and hence all simulations were carried out under H_0. Since the statistic T^*, the basis of all the tests, is invariant under H_0 with respect to shifts in the mean, all simulations were done with $\mu_1 = \mu_k = \cdots = \mu_k = 0$. All the calculations were done in SAS IML (Interactive Matrix Language). To be able to investigate the tests, we applied all the restrictions presented in Section 2.5. For each configuration, 10000 simulations were generated and for each the p-values of all the above described tests recorded. The number of means k was set to 3 and 5. For both, equal variances were considered with $\sigma^2 = 1$. For $k = 3$, the unequal variances were given by a vector $(1, 3, 5)$ and for $k = 5$, by $(1, 3, 5, 1, 3)$, and $(1, 3, 5, 6, 7)$. For a balanced case and $k = 3$ the sample sizes were given as $(5, 5, 5)$, $(10, 10, 10)$; and for unbalanced $(2, 3, 3)$, and $(5, 7, 15)$. For $k = 5$, the sample sizes were $(2, 2, 2, 2, 2)$, $(5, 5, 5, 5, 5)$, $(10, 10, 10, 10, 10)$, $(2, 2, 3, 3, 5)$, $(4, 4, 6, 6, 10)$, and $(3, 5, 7, 10, 10)$. For parametric bootstrap, for each of the 10000 simulations, 5000 samples were generated and the proportion of those exceeding the observed f^* was recorded as the p-value.

The figures below illustrate some of the investigated configurations. Since we investigate the accuracy of the p-value, i.e., whether $P(p \leq \alpha) = \alpha$ and the dependence of accuracy on different configurations, the best way to capture it is by plotting the observed p-values on the vertical axis versus the empirical cumulative distribution function (CDF) of the p-value on the horizontal axis. In each figure, the p-value and its the empirical CDF is plotted, obtained from the 10000 simulations. Each panel corresponds to a particular test. Since the range of interest for the size of the test is mainly bounded by 10%, the graphs show only the range from 0 to 0.1. If the p-value is accurate, we expect to see the values right on top of the identity line (the dashed line in all figures). The curve for a conservative p-value is on top or above the identity line. Cases when the p-values are below the identity line indicate anti-conservative (liberal) p-values, overstating the statistical significance of observed differences among the sample means.

In Table 1, for all configurations and all considered tests, the estimates of Type I probabilities are presented for a nominal level of 5%.

4. Conclusions

We studied the behavior of p-values of five different types of tests available for comparing k means. We studied extremely small sample sizes. Figure 1 illustrates that for $k = 3$, if the sample sizes are small (not an unusual setting mainly in basic science applications), independent of whether the variances

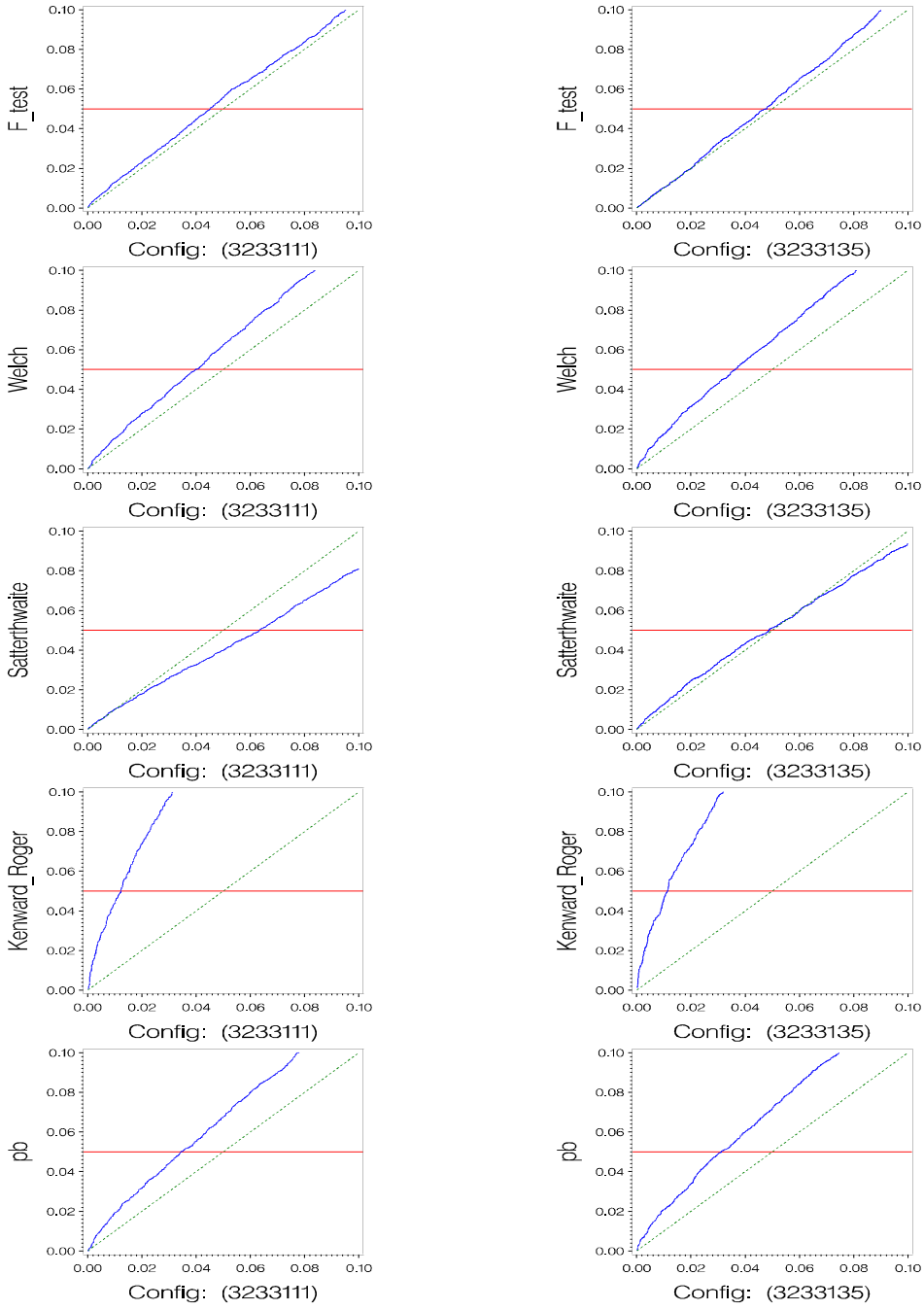

FIGURE 1. p-values for $n = (2, 3, 3)$. Left panels for $\sigma^2 = (1, 1, 1)$ and right for $\sigma^2 = (1, 3, 5)$

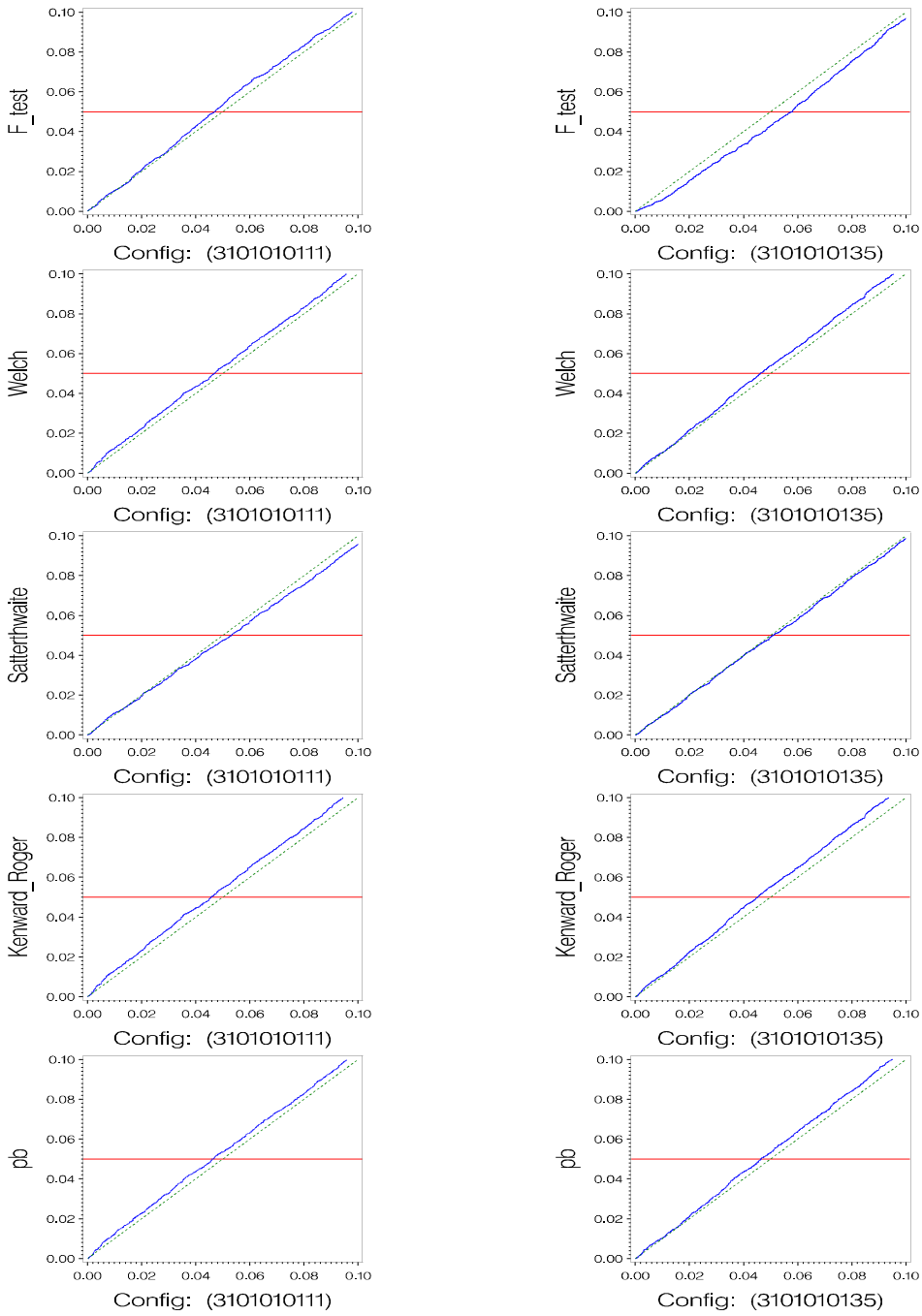

FIGURE 2. p-values for $n = (10, 10, 10)$. Left panels for $\sigma^2 = (1, 1, 1)$ and right for $\sigma^2 = (1, 3, 5)$

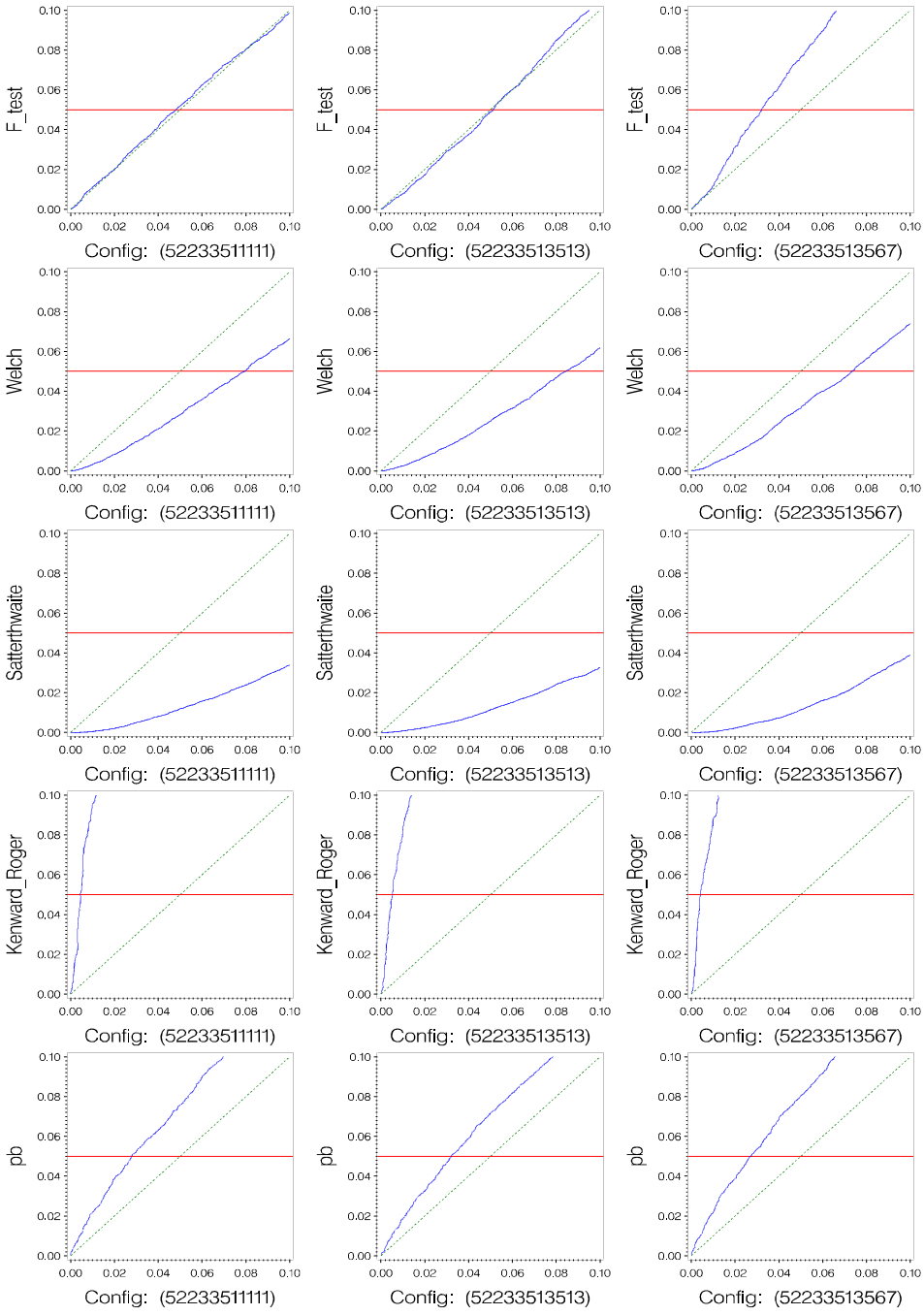

FIGURE 3. p-values for $n = (2, 2, 3, 3, 5)$. Left panels for $\sigma^2 = (1, 1, 1, 1, 1)$, center for $\sigma^2 = (1, 3, 5, 1, 3)$, and right for $\sigma^2 = (1, 3, 5, 6, 7)$

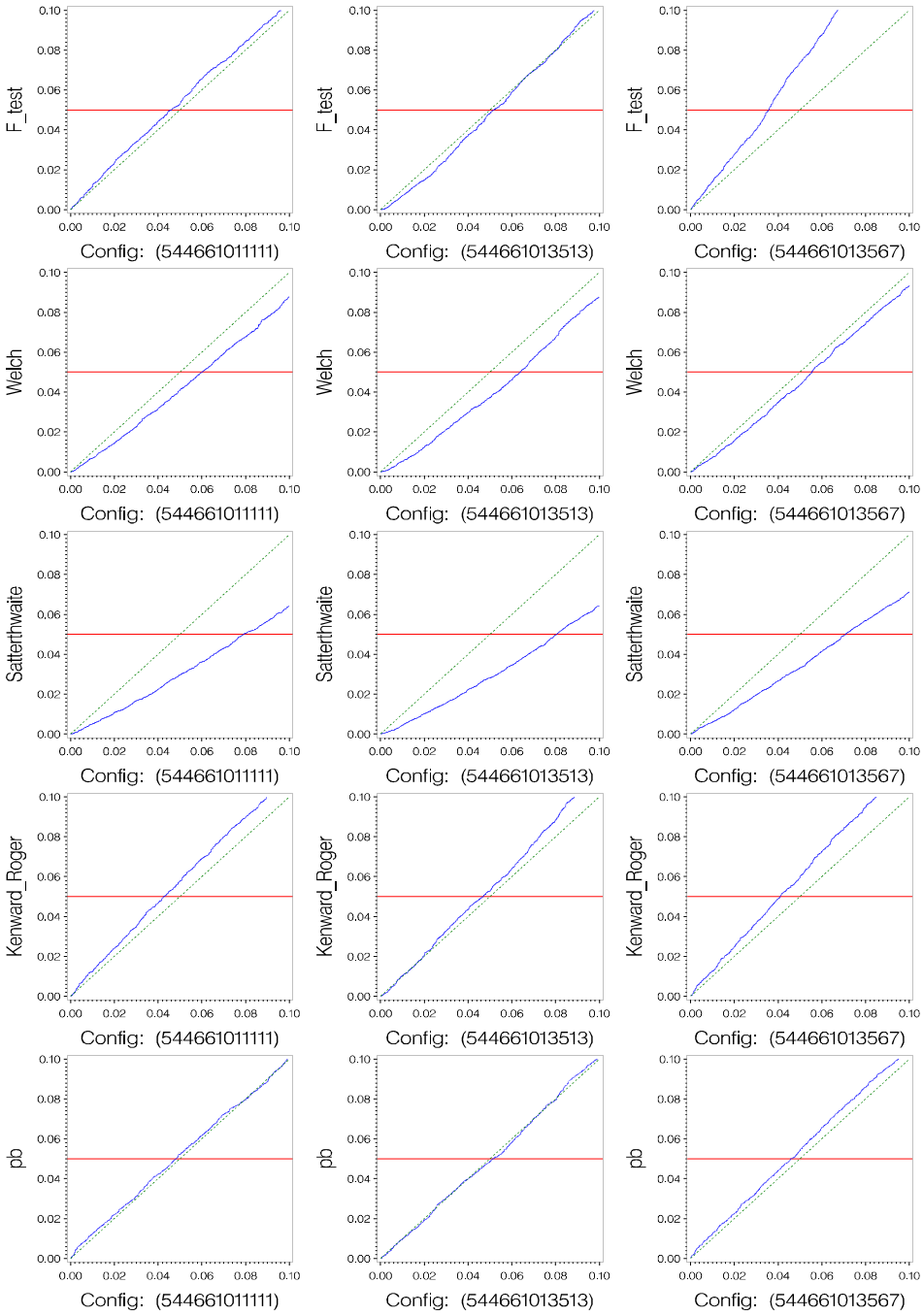

FIGURE 4. p-values for $n = (4, 4, 6, 6, 10)$. Left panels for $\sigma^2 = (1, 1, 1, 1, 1)$, center for $\sigma^2 = (1, 3, 5, 1, 3)$, and right for $\sigma^2 = (1, 3, 5, 6, 7)$

TABLE 1. Actual nominal levels of tests for different configurations at 5% level of significance

$(n_1,\ldots,n_k)-(\sigma_1^2,\ldots,\sigma_k^2)$	F-test	Welch	S-F-C	K-R	PB
233-111	0.045	0.034	0.063	0.012	0.034
233-135	0.048	0.036	0.049	0.011	0.031
555-111	0.051	0.048	0.057	0.042	0.047
555-135	0.060	0.049	0.057	0.042	0.048
5715-111	0.049	0.052	0.062	0.049	0.052
5715-135	0.020	0.048	0.057	0.045	0.048
101010-111	0.047	0.047	0.053	0.046	0.046
101010-135	0.058	0.046	0.051	0.045	0.046
22222-11111	0.048	0.078	0.113	0.011	0.015
22222-13513	0.067	0.084	0.116	0.012	0.018
22222-13567	0.066	0.088	0.122	0.014	0.019
22335-11111	0.048	0.079	0.133	0.004	0.028
22335-13513	0.051	0.084	0.132	0.005	0.032
22335-13567	0.032	0.073	0.120	0.004	0.027
55555-11111	0.049	0.052	0.065	0.033	0.041
55555-13513	0.061	0.055	0.066	0.037	0.043
55555-13567	0.065	0.065	0.076	0.043	0.053
446610-11111	0.046	0.060	0.079	0.042	0.048
446610-13513	0.052	0.064	0.080	0.047	0.051
446610-13567	0.036	0.055	0.071	0.041	0.046
3571010-11111	0.050	0.069	0.080	0.054	0.058
3571010-13513	0.057	0.062	0.080	0.048	0.053
3571010-13567	0.027	0.051	0.064	0.040	0.044
1010101010-11111	0.051	0.051	0.059	0.047	0.049
1010101010-13513	0.061	0.052	0.059	0.047	0.049
1010101010-13567	0.064	0.056	0.063	0.052	0.053

are equal or unequal the standard ANOVA type F-test has a p-value which is reasonably accurate. The second best for such small sample sizes is the generalized Satterthwaite-Fai-Cornelius test. The Kenward-Roger test is extremely conservative, as we can conclude from Figure 1, and we see the same in Table 1; its estimate of Type I probability at 5% is around 1%. If we have a balanced case with moderately large sample sizes, as shown in Figure 2, all tests behave very well. For unequal variances the PB goes hand in hand with Welch's test although again, the Satterthwaite-Fai-Cornelius method is dominant in this setting. The situation changes dramatically if we increase the number of compared means. For $k = 5$ and small sample sizes, $n(= 2, 2, 3, 3, 5)$, Figure 3 illustrates the settings. In all cases the Satterthwaite-Fai-Cornelius and Welch tests are extremely anti-conservative; on the other hand, the Kenward-Roger test is too

conservative in all of these three cases. The ANOVA type F-test seems to behave the best unless the variances are extremely different (see Figure 3, right panels). In the latter case it is questionable which of the two is better, the F-test or the PB test. As soon as the sample sizes increase, we can confirm the good behavior of the PB test, as shown on Figure 4, however the Kenward-Roger test is the second best in such cases. For configurations not shown we observed a very similar pattern. Our recommendation hence is for small sample sizes up to 5 per group even if we are suspicious of heterogeneity of variances, just use the ANOVA type F-test. For higher sample sizes one might choose to go with the PB test, whose simplicity is certainly appealing, or with the Kenward-Roger test, which is readily available or easily programmable.

REFERENCES

[1] BROWN, M. B.—FORSYTHE, A. B.: *The small sample behavior of some statistics which test the equality of several means*, Technometrics **16** (1974), 129–132.
[2] CASELLA, G.—BERGER, R. L.: *Statistical Inference* (2nd ed.), Duxbury, Belmont, CA, 2002.
[3] FAI, A. H. T.—CORNELIUS, P. L.: *Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments*, J. Stat. Comput. Simul. **54** (1996), 363–378.
[4] GIESBRECHT, F. G.—BURNS, J. C.: *Two-stage analysis based on a mixed model: large sample asymptotic theory and small sample simulation results*, Biometrics **41** (1985), 477–486.
[5] HARVILLE, D. A.—JESKE, D. R.: *Mean squared error of estimation or prediction under a general linear model*, J. Amer. Statist. Assoc. **87** (1992), 724–731.
[6] KENWARD, M. G.—ROGER, J. H.: *Small sample inference for fixed effects from restricted maximum likelihood*, Biometrics **53** (1997), 983–997.
[7] KRISHNAMOORTHY, K.—LU, FEI—MATHEW, T.: *A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models*, Comput. Statist. Data Anal. **51** (2007), 5731–5742.
[8] LEE, S.—AHN, C. H.: *Modified ANOVA for unequal variances*, Comm. Statist. Simulation Comput. **32** (2003), 987–1004.
[9] SATTERTHWAITE, F. E.: *Synthesis of variance*, Psychometrika **6** (1941), 309–316.
[10] SATTERTHWAITE, F. E.: *An approximate distribution of estimates of variance components*, Biometrics Bull. **2** (1946), 110–114.
[11] WEERAHANDI, S.: *ANOVA under unequal variances*, Biometrics **51** (1995), 589–599.
[12] WELCH, B. L.: *On the comparison of several mean values: an alternative approach*, Biometrika **38** (1951), 330–336.

Received 22. 2. 2008
Accepted 19. 5. 2009

Biostatistics Program
LSUHSC School of Public Health
1615 Poydras St., Suite 1400
New Orleans, LA 70112
USA
E-mail: jvolau@lsuhsc.edu