Journal of Statistical Software

October 2020, Volume 95, Code Snippet 2. *doi: 10.18637/jss.v095.c02*

ANOVA_robust: A SAS Macro for Parametric Tests of Mean Differences in One-Factor ANOVA Models

Thanh V. Pham
University of South Florida

Jeffrey D. Kromrey
University of South Florida

Yi-Hsin Chen
University of South Florida

Eun Sook Kim
University of South Florida

Diep T. Nguyen
University of South Florida

Yan Wang
University of Massachusetts Lowell

Abstract

Testing the equality of several independent group means is a common statistical practice in the social sciences. The traditional analysis of variance (ANOVA) is one of the most popular methods. However, the ANOVA F test is sensitive to violations of the homogeneity of variance assumption. Many alternative tests have been developed in response to this problem of the F test. These tests include some modifications of the ANOVA F test and others based on the structured means modeling technique. This paper provides a SAS macro for testing the equality of group means using thirteen methods including the regular ANOVA F test. In addition, this paper summarizes the results of a simulation study that compares the performance of these tests in terms of their Type I error rate under different conditions, especially under violations of the homogeneity of variance assumption.

Keywords: analysis of variance, homogeneity of variance assumption, simulation study, homoscedasticity, heteroscedasticity, SAS macro.

1. Introduction

Testing the equality of several independent means is a common statistical practice in the social sciences and the analysis of variance (ANOVA) F test is often used by applied researchers for testing the equality (Tomarken and Serlin 1986). The traditional ANOVA F test, which uses the ordinary least squares (OLS) method, is based on several assumptions including independence, normality, and homogeneity of variance. However, both normality and

homogeneity assumptions are often violated in applied research in the social sciences (Fan and Hancock 2012). It is also well known that the violation of the assumption of equal variances affects the Type I error rate of the ANOVA F test even when the sample sizes are equal across groups (Rogan and Keselman 1977). In response to this problem, many alternative tests have been suggested. These parametric tests include some modifications of the ANOVA F test while some others apply the structured means modeling (SMM) technique. Among the ANOVA-based tests, the common options for applied researchers are: the Welch test (Welch 1951), Brown-Forsythe test (Brown and Forsythe 1974), James second-order test (James 1951), and Alexander-Govern approximation (Alexander and Govern 1994). Simulation studies have shown that under heterogeneous variance conditions, these tests can control the Type I error rate when data are normal but become liberal when data are non-normal (Fan and Hancock 2012). The other methods suggested later include: weighted least squares (WLS) test (Montgomery and Peck 1992), Wilcox test (Wilcox 1988), and mixed model method (Littell, Milliken, Stroup, Wolfinger, and Schabenberger 2006).

A different approach is to apply the SMM technique that does not require the assumption of variance homogeneity. Being developed from the framework of structural equation modeling (SEM), the SMM technique allows variances to be heterogeneous across groups by freely estimating them. The SMM technique can be combined with various estimation methods such as asymptotic distribution free (ADF) estimation (Browne 1982) or maximum likelihood (ML) estimation to test the mean differences. The SMM-based tests were shown to have better performance than ANOVA-based tests in term of power and Type I error rate (Fan and Hancock 2012). The traditional ANOVA F test and the Welch test can be conducted using SAS (SAS Institute Inc. 2013), but test statistics of many other alternative methods are not provided directly in SAS.

In addition to parametric approaches, several non-parametric methods such as permutation, randomization or bootstrap tests have been developed for one-way ANOVA models. With the availability of advanced computing systems and the increase of big and complex data, permutation-based statistical tests have become popular. Characteristics and applications of several permutation methods from univariate to multivariate in the non-parametric framework have been introduced in several books (Pesarin and Salmaso 2010; Basso, Pesarin, Salmaso, and Solari 2009). While both parametric and non-parametric tests can be used for ANOVA models, they have some differences. For example, the null hypotheses tested are not the same. In addition, the parametric procedures are based on given assumptions of normality and homogeneity of variance but the non-parametric tests do not require these assumptions. Parametric tests may be used when the data is normally distributed or has a lightly tailed distribution because of their superior power over the non-parametric counterparts (Conover 1999). However the parametric methods may not be robust when one or all of these assumptions are violated, especially when sample sizes are not sufficiently large or unbalanced. The non-parametric approaches are then often recommended when the data is severely non-normal or sample size is small. Specifically when the data distribution is heavy-tailed such as the lognormal distribution, exponential distribution, or when there are many outliers, non-parametric tests should be used because these procedures have more power than the parametric counterparts (Conover 1999).

In the current study, we focus on comparing parametric approaches that test the same null hypothesis of equal population means. The purpose of this paper is to present a SAS macro that provides all test statistics of the parametric methods mentioned above to examine the

equality of independent means. It should be noted that the non-parametric tests are not included in this macro. The results of a simulation study that compared the performance of these parametric methods are also presented.

2. Statistical methods for testing the mean differences

It is assumed that a test for the mean differences is applied to the data of J groups where J is the total number of groups. Let x_{ij} be the ith observed score in group j and n_j is the number of observations in the jth group. The following notations will be used in describing these tests. N represents the total sample size which is a sum of all group sizes.

$$N = \sum_j n_j$$

$\overline{X}_{\mathrm{j}}$ represents the mean score of group j.

$$\overline{X}_j = \sum_i \frac{x_{ij}}{n_j}$$

$\overline{X}$ represents the grand mean of the entire sample.

$$\overline{X} = \sum_j \frac{n_j \overline{X}_j}{N}$$

${S_j}^2$ and S_j are the variance and standard deviation of group j, respectively.

$${S_j}^2 = \frac{\sum_i (x_{ji} - \overline{X}_j)^2}{n_j - 1}$$

2.1. ANOVA F test (also called OLS)

Researchers often use the analysis of variance (ANOVA) F test to examine the equality of several independent group means. The statistic F is determined by the following equation:

$$F = \frac{\sum_j n_j (\overline{X}_j - \overline{X})^2 / (J-1)}{\sum_j (n_j - 1) {S_j}^2 / (N-J)}.$$

2.2. Alexander and Govern (AG) test

In the Alexander and Govern approximate test (Alexander and Govern 1994), a weight (w_j) for each group is calculated as ($w_j = \frac{1/{S_j}^2}{\sum_j 1/{S_j}^2}$). The variance-weighted estimate of the common mean X^+ is calculated by: ($X^+ = \sum_j w_j \overline{X}_j$). The t statistic for each of J groups is determined as: $t_j = \frac{\overline{X}_j - X^+}{S_j}$ and follows Student's t distribution with $v_j (= n_j - 1)$ degrees of freedom. z_j is achieved by a normalizing transformation of t_j:

$$z_j = c + \frac{(c^3 + 3c)}{b} - \frac{(4c^7 + 33c^5 + 240c^3 + 855c)}{(110b^2 + 8bc^4 + 1000b)},$$

where $a = v_j - .5; b = 48a^2; c = [a \ln(1 + \frac{{t_j}^2}{v_j})]^{1/2}$. z_j is used to calculate the A statistic by:

$$A = \sum_1^J {z_j}^2.$$

A follows a χ^2 distribution with $(J-1)$ degrees of freedom.

2.3. Brown-Forsythe (BF) test

The Brown and Forsythe (Browne 1982) test is a modification of the ANOVA F test. F^* is the test statistic defined as:

$$F^* = \frac{\sum_j n_j(\overline{X}_j - \overline{X})^2}{\sum_j (1 - n_j/N){S_j}^2}$$

F^* has an F-distribution with $(J-1)$ and f degrees of freedom where f is calculated by the Satterthwaite approximation:

$$\frac{1}{f} = \sum_j \frac{{c_j}^2}{(n_j - 1)}$$

and

$$c_j = \frac{(1 - n_j/N)\,{S_j}^2}{\sum_j (1 - n_j/N)\,{S_j}^2}.$$

2.4. James' second order (James) test

The James' test uses the Q statistic which is determined by:

$$Q = \sum_j w_j(\overline{X}_j - X_w)^2,$$

where $w_j = n_j/{S_j}^2$ and $X_w = \sum_j w_j \overline{X}_j / \sum_j w_j$.
The obtained value of Q is compared to a carefully adjusted critical value of χ^2 with $(J-1)$ degrees of freedom (James 1951).

2.5. Welch test (Welch)

The Welch test (Welch 1951) is a modified version of the F test that compares mean differences among multiple groups. This test relies on only independent population and normal distribution assumptions and relaxes the equal population variances requirement. The test statistic is determined as:

$$F' = \frac{\sum_j w_j \frac{(\overline{X}_j - \overline{X}')^2}{J-1}}{1 + \frac{2(J-2)}{J^2-1} \sum_j \left[\left(1 - \frac{w_j}{u}\right)^2 (n_j - 1) \right]},$$

where $w_j = n_j / {S_j}^2; u = \sum_j w_j; \overline{X}' = \sum_j (w_j \overline{X}_j / u)$. The distribution of F' can be estimated by the F distribution, using $v_b = J - 1$, and $\frac{1}{v_w} = (\frac{3}{J^2-1}) \sum_j [\frac{\left(1 - \frac{w_j}{u}\right)^2}{n_j - 1}]$.

2.6. Wilcox test (Wilcox)

Wilcox (1988) introduced an approximate method which is contrasted with the James's second order, to deal with unequal variances. The author made an improvement of this test (Wilcox 1989) and its modification that will be used in this section includes the following setting:

$$D_j = \frac{n_j}{{S_j}^2},$$

$$W_s = \sum_j D_j,$$

$$\tilde{Y} = \sum_j \frac{D_j \tilde{Y}_j}{W_s},$$

where $\tilde{Y}_j = \frac{X_{n_j j}}{n_j} + \frac{\sum_{i=1}^{n_j - 1} \left(1 - \frac{1}{n_j}\right) X_{ij}}{n_j + 1}$. When $H_m = \sum_j D_j (\tilde{Y}_j - \tilde{Y})^2$ surpasses the $(1 - \alpha)$ quantile of a χ^2 distribution with $(J - 1)$ degrees of freedom, the null hypothesis is rejected. Hsiung, Olejnik, and Huberty (1994) demonstrated poor Type I error control of the Wilcox test if the population grand mean differs from zero. To correct this problem, the macro transforms the data by grand mean centering prior to calculation of the Wilcox test.

2.7. Weighted least squares (WLS) test

Montgomery and Peck (1992) developed a method in which each observation is weighted by the inverse of its variance. A weight for each observation can be obtained by computing the reciprocal of the group variance as follows:

$$w_j = \frac{1}{{S_j}^2}.$$

Generalized least squares (GLS) is used to minimize the weighted sum squares (WSS), which is the sum of the weighted variation of x_{ij} from the grand mean $\overline{X}_j$.

$$WSS = \sum_{j=1}^{J} \sum_{i=1}^{n_j} w_j \left(x_{ij} - \overline{X}_j\right)^2$$

.

2.8. SMM with maximum likelihood estimation (SMM with ML)

Applying the SMM approach to the between-subjects testing of mean equality for a measured variable, the indicator x can be expressed as $x = v_k + \delta$, where v_k is a $p \times 1$ vector of intercept

values and δ is a $p \times 1$ vector of normal errors. The null hypothesis is tested by constraining population means to be equivalent while allowing variances of δ to be heterogeneous.

Estimation within SMM can be handled by using maximum likelihood. If F_{ML} is the ML fit function, the test statistics T_{ML} is calculated as $T_{ML} = (N-1)F_{ML}$, with degrees of freedom equal to $Jp(p+3)/2 - q$, where p is the number of observed variables and q is the number of parameters estimated across all groups.

2.9. SMM with asymptotic distribution free estimation (SMM with ADF)

Browne (1982) proposed using asymptotic distribution free estimation (ADF) for the covariance structure when variables are continuous but not multivariate normally distributed. Muthén (1989) expanded the ADF method by including both mean and covariance structures. The ADF fit function, using the GLS-type fit function, is defined as

$$F_{\text{ADF}} = \sum_j (s_j - \sigma_j)^\top W_j^{-1} (s_j - \sigma_j),$$

where for the jth group, s_j is the vector consisting of p elements of the observed means (s_1) and $p(p+1)/2$ elements of the variance covariance matrix (s_2), σ_j is the model implied counterpart of s_j, and W represents the ADF weight matrix as an estimator of the asymptotic covariance matrix of s. The model parameters are estimated by minimizing the ADF fit function. When this fit function is multiplied by $2N$ (where N is the total sample size), it follows the χ^2 distribution with $(J-1)$ degrees of freedom.

2.10. SMM with Bartlett's correction to the ML test statistics (Bartlett)

Bartlett (1950) suggested a correction to the ML test statistic using the context of explanatory factor analysis with m latent constructs, p observed variables, and small sample sizes. The test statistics T_{BC} is defined as

$$T_{\text{BC}} = \left(N - \frac{p}{3} - \frac{2m}{3} - 11/6\right) F_{\text{ML}},$$

with degrees of freedom $df = Kp* - q$ and $p* = p(p+3)/2$; where N is the total sample size, p is the number of observed variables, and q is the number of parameters estimated across all groups. Applying to the between-subject testing of mean equality, the SMM model is simplified to no latent factor and one observed variable.

2.11. Yuan and Bentler (YB1 and YB2) tests

Yuan and Bentler (1997, 1999) suggested test statistics T_{YB1} and T_{YB2} as the corrections of T_{ADF} for small sample sizes. The statistics T_{YB1} is defined as

$$T_{\text{YB1}} = \frac{T_{\text{ADF}}}{1 + \frac{T_{\text{ADF}}}{N}},$$

where $T_{\text{ADF}} = (N-1)F_{\text{ADF}}$, which follows a central χ^2 distribution with the same df as T_{ADF}.

Based on the logic of the transformation applied to Hotelling's T^2 statistic in multivariate analysis of variance (MANOVA), the second correction, T_{YB2}, of T_{ADF} follows the F distribution.

$$T_{\text{YB2}} = \frac{N - (Jp^* - q)}{(N-1)(Jp^* - q)} T_{\text{ADF}},$$

with numerator and denominator *dfs* of $(Jp^* - q)$ and $(N - Jp^* + q)$, respectively. In the specific case of SMM, the numerator and denominator *dfs* for T_{YB2} reduce to $(J-1)$ and $(N - J + 1)$, respectively.

2.12. Mixed models with heterogeneous variances (Mixed)

Analysis of data from ANOVA designs with heterogeneous variances may be conducted by fitting a mixed model with unequal residual variances (Littell *et al.* 2006). For a one factor ANOVA design, the model may be written as

$$y_{ij} = \mu + \alpha_j + \varepsilon_{ij},$$

where $\varepsilon_{ij} \sim N\left(0, \delta_j^2\right)$. That is, a separate variance is estimated for each group in the ANOVA design. Such a model may be fit using ML or REML estimation. `PROC MIXED` provides a straightforward approach for fitting such a model. This heterogeneous variance solution is obtained with the `GROUP =` option on the `REPEATED` statement (even though a repeated-measures design is not used). That is,

```
REPEATED / GROUP = IV;
```

where `IV` is the name of the independent variable. For such analyses, the Satterthwaite degrees of freedom estimate should be used (Satterthwaite 1946). This is obtained using the `DDFM = SATTERTHWAITE` option on the `MODEL` statement in `PROC MIXED`.

3. Description of the SAS macro

The SAS macro `ANOVA_robust` is written in base SAS and SAS/STAT. There are three input arguments that are required by the macro.

- `DATA` specifying the name of the SAS data set containing the data to be analyzed;
- `Y` identifying the name of the dependent variable;
- `GROUP` indicating the name of the independent variable.

Default values are provided for each argument. Observations with missing values for either the independent or dependent variable are deleted from the analyses.

4. Example

In this section, an example will demonstrate the use of the `ANOVA_robust` macro with a particular data set. There are two SAS codes that can be downloaded in this project. One is

Tests of Mean Differences

Independent Variable:	Group
N of Groups:	5
Dependent Variable:	Y
Total N of Observations:	33

Group	N	Mean	Variance
1	6	3.0000	4.4000
2	7	10.4286	14.9524
3	7	12.7143	14.5714
4	5	19.2000	32.2000
5	8	30.1250	38.1250

Test	Value	p value	DF
ANOVA-Typed			
Alexander-Govern Test	39.1575	p < .001	4
Brown-Forsythe Test	35.5206	p < .001	4, 19.52
James' Second Order Test	166.4407	p < .01	4
Mixed Model	41.6102	p < .001	4, 9.18
Regular ANOVA F	34.7226	p < .001	4, 28
Welch Test	36.0493	p < .001	4, 12.97
Wilcox Test	100.9498	p < .001	4
Weighted Least Squares Test	41.6102	p < .001	4, 28
Structured Means Modeling			
SMM with ML estimator	35.6174	p < .001	4
SMM with ADF estimator	142.1882	p < .001	4
Barlett Correction Test	34.3189	p < .001	4
Yuan and Bentler Test 1	26.7838	p < .001	4
Yuan and Bentler Test 2	32.2145	p < .001	4, 29

Figure 1: The exemplary output of the SAS ANOVA_robust macro.

the SAS macro (named anovarobmacro.sas) and the other is an execution SAS code (named code.sas) to execute the SAS macro. Below is the content of the execution SAS code. The data set one contains 33 observations on an independent variable (group) and a dependent variable (y). The user can modify the code to obtain the data. The macro is called after the data extraction step. In this example, the SAS macro is placed in the local folder. The dependent and independent variables are defined as y and group, respectively, in the data extraction step.

```
data one;
input group y @@;
datalines;
1 5 1 1 1 2 1 6 1 1 1 3 2 13 2 13 2 6 2 11 2 4 2 14 2 12 3 12 3 16
3 9 3 18 3 7 3 14 3 13 4 17 4 13 4 16 4 23 4 27 5 22 5 30 5 27 5 32 5 32
5 43 5 29 5 26
;
run;
%include "anovarobmacro.sas";
%ANOVA_robust(data = one, y = y, group = group);
run;
```

The output sample of the macro is demonstrated in Figure 1. General information of the data set is presented in the first output section while the second output section shows the results

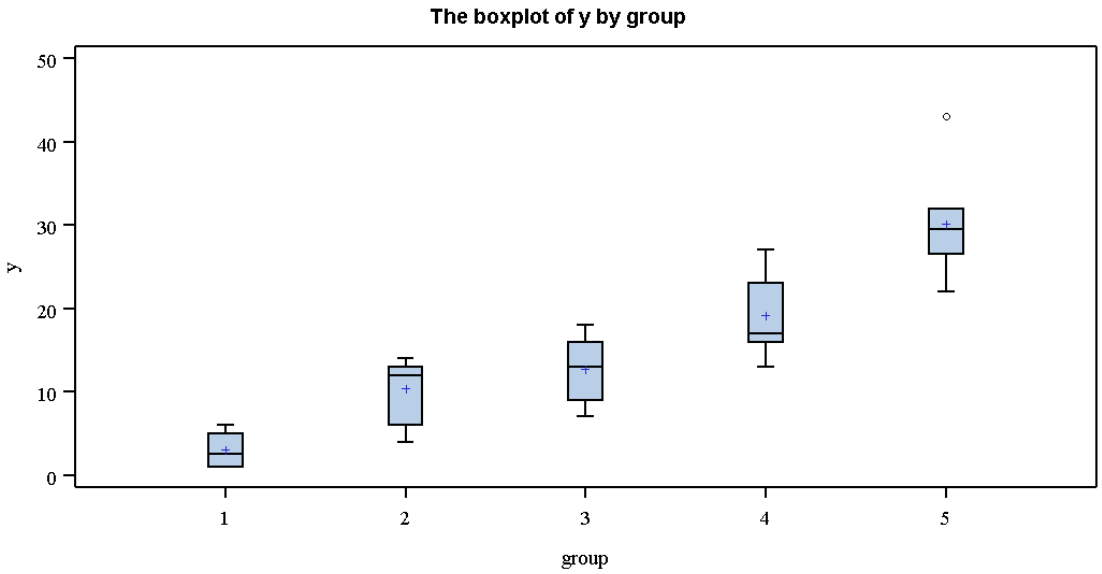

Figure 2: The boxplot output of the SAS `ANOVA_robust` macro for sample distributions across groups.

of 13 tests for group-mean equality. The general information of the data set includes the names of the independent and dependent variables, the number of groups as well as the total sample size. The first section also provides statistical information for each group including group sizes, group means, and the variances of each group. The second section presents the obtained value, associated p value, and degrees of freedom for each `ANOVA_robust` test. These tests are classified into two groups: the group of tests using an ANOVA-type approach and the group of tests using the SMM technique. The macro also provides the boxplot of sample distributions across groups as shown in Figure 2.

5. A simulation study

A simulation study was conducted to investigate the performance of the testing methods in terms of Type I error control. Six simulation factors were included: (1) number of groups, (2) group size, (3) group size pattern, (4) variance pattern, (5) maximum group variance ratio, and (6) population distribution. In addition, the performance of each method was examined at three nominal alpha levels: 0.01, 0.05, and 0.10. The combination of these factors created a total of 2,736 conditions. Five thousand samples were generated for each simulated condition. The Type I error rates were evaluated as the simulation outcome.

Three tests using the SMM technique, including ADF, YB1, and YB2 tests, did not provide solutions with some small sample size conditions (a sample with three or fewer observations in a group). They were treated as missing data in analyzing the simulation outcomes. The outcomes of homogeneous and heterogeneous conditions were examined separately. The η^2 analysis for effect size was conducted to explore the significant impact of the research design factors on the variability in the estimated Type I error rates.

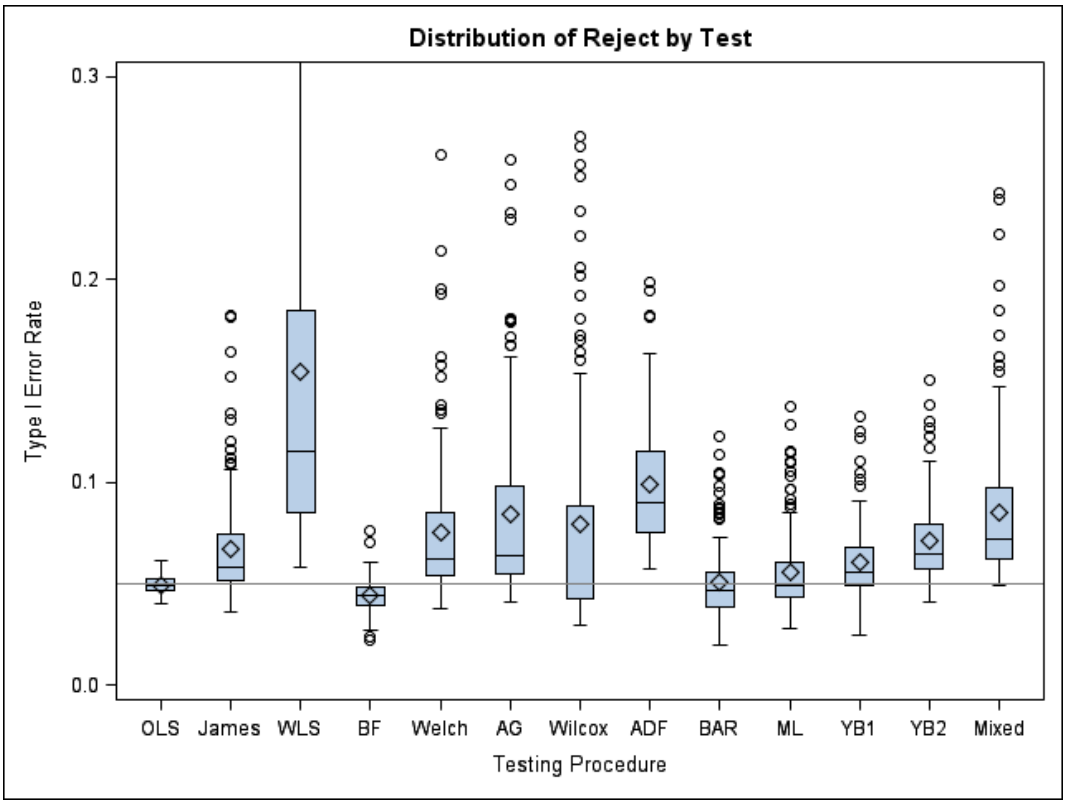

Figure 3: Rejection rate distributions at the 0.05 significance level of homogeneous conditions.

5.1. Type I error rates estimates with homogeneous conditions

Figures 3 presents boxplots of the rejection rate distributions at the 0.05 significance level of homogeneous conditions. Under the condition of equal variance, the OLS method (i.e., the ANOVA F test) showed the best performance. Among the other approaches, BF, Bartlett, and SMM with ML controlled Type I error adequately.

The η^2 analysis revealed several design factors that substantially associated with the variability of Type I error rates including testing method, population shape, group size and group size pattern. The effect of the group size on the Type I error rates was found to vary by testing methods and population shape. The variability of Type I error rates of all methods by population shape and group size is presented in Table 1. As observed in Table 1, the OLS and BF tests adequately controlled Type I error around 0.05 across all conditions under homogeneity of variance. On the contrary, Type I error rates of WLS and SMM with ADF methods were often above 0.07. The Wilcox test showed reasonable Type I error control for all conditions under equal variances except for the small group size conditions. For the SMM methods without ADF (i.e., Bartlett, ML, YB1, and YB2), the Type I error rates were reasonably controlled even with small group size. However, when the distribution was extremely leptokurtic (kurtosis = 25) and the group size was small, the Type I error rates were slightly inflated. The James, Welch, AG, Wilcox, and Mixed methods failed to control for the Type I error rates when sample size was small.

Shape	Group size	OLS	James	WLS	BF	Welch	AG	Wilcox	ADF	BAR	ML	YB1	YB2	Mixed
1 $(0, 0)$	5	0.05	0.07	0.24	0.04	0.09	0.10	0.13	0.12	0.04	0.05	0.04	0.06	0.10
	10	0.05	0.05	0.11	0.05	0.06	0.05	0.04	0.10	0.04	0.04	0.06	0.07	0.07
	20	0.05	0.05	0.08	0.05	0.05	0.05	0.04	0.07	0.04	0.04	0.05	0.06	0.06
2 $(1, 3)$	5	0.05	0.06	0.22	0.04	0.07	0.09	0.11	0.11	0.03	0.04	0.03	0.05	0.08
	10	0.05	0.05	0.11	0.04	0.05	0.05	0.04	0.09	0.03	0.04	0.05	0.06	0.06
	20	0.05	0.05	0.08	0.05	0.05	0.05	0.04	0.07	0.04	0.05	0.05	0.06	0.06
3 $(1.5, 5)$	5	0.05	0.06	0.22	0.04	0.07	0.09	0.11	0.11	0.03	0.04	0.03	0.05	0.08
	10	0.05	0.05	0.11	0.04	0.06	0.06	0.04	0.10	0.04	0.05	0.06	0.07	0.07
	20	0.05	0.06	0.08	0.05	0.06	0.06	0.04	0.08	0.05	0.05	0.06	0.06	0.06
4 $(2, 6)$	5	0.05	0.07	0.24	0.03	0.08	0.12	0.11	0.14	0.06	0.07	0.05	0.07	0.09
	10	0.05	0.08	0.15	0.04	0.09	0.10	0.07	0.14	0.07	0.08	0.09	0.10	0.10
	20	0.05	0.08	0.11	0.04	0.08	0.08	0.06	0.10	0.07	0.07	0.08	0.09	0.09
5 $(0, 25)$	5	0.05	0.14	0.33	0.05	0.16	0.20	0.21	0.17	0.09	0.11	0.07	0.10	0.17
	10	0.05	0.08	0.13	0.05	0.08	0.08	0.07	0.12	0.06	0.06	0.08	0.09	0.09
	20	0.05	0.05	0.08	0.05	0.06	0.05	0.04	0.07	0.05	0.05	0.05	0.06	0.06
6 $(0, -1)$	5	0.05	0.10	0.28	0.05	0.12	0.14	0.17	0.15	0.06	0.07	0.06	0.09	0.14
	10	0.05	0.07	0.12	0.05	0.07	0.07	0.06	0.11	0.05	0.05	0.07	0.08	0.08
	20	0.05	0.06	0.08	0.05	0.06	0.06	0.05	0.08	0.05	0.05	0.06	0.06	0.07

Table 1: The Type I error rates of thirteen robust ANOVA tests by population shape and group size at nominal alpha level of 0.05 for homogeneous conditions. Note: For population shapes, the values within parentheses are skewness and kurtosis, respectively. For example 1 (0, 0) indicates normal distribution with skewness = 0 and kurtosis = 0.

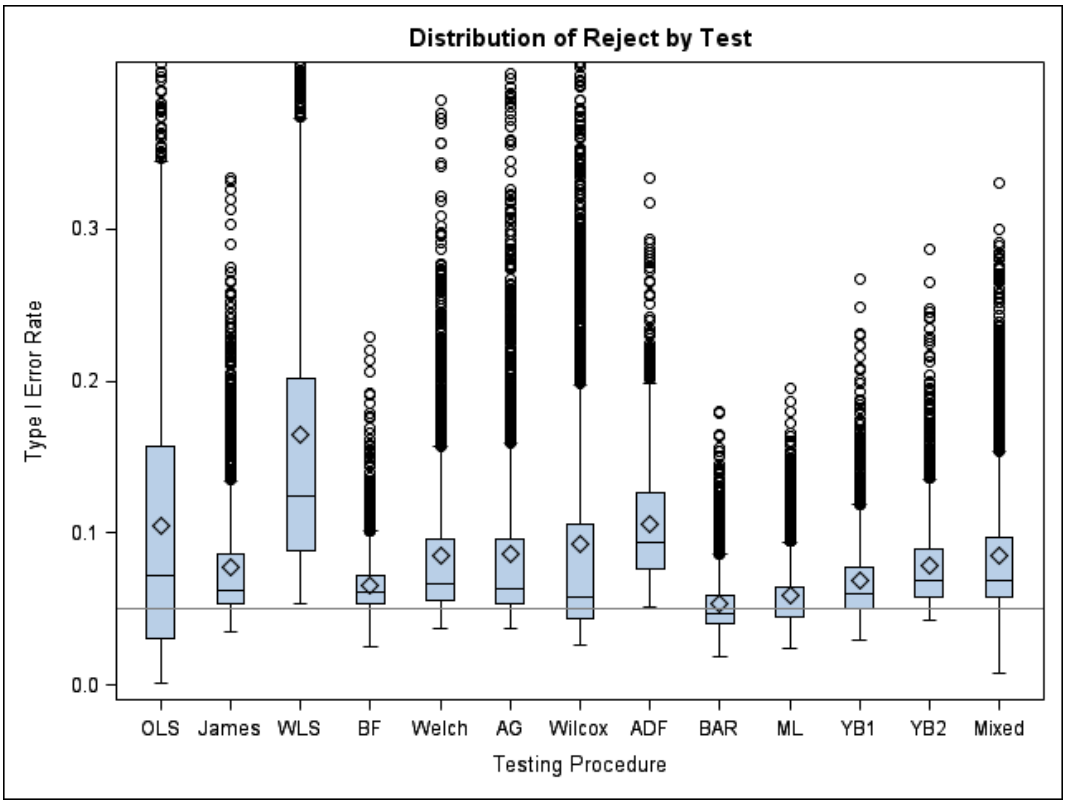

Figure 4: Rejection rate distributions at the 0.05 significance level of heterogeneous conditions.

5.2. Type I error rates estimates with heterogeneous conditions

The η^2 analysis showed that the Type I error rates were affected by the interaction of cell size and testing method, as well as the interaction of variance pattern and testing method. The population shape and variance pattern also had a significant effect on the Type I error rates. Figure 4 presents the distribution of Type I error rates estimates across testing methods. WLS and SMM with ADF had the highest Type I error rates for all simulation conditions of heterogeneous variance as shown in Figure 4. The impact of variance pattern on the method in terms of Type I error control is presented in Table 2. As expected, the ANOVA F test (OLS) showed poor performance in controlling for Type I error under unequal variance conditions. It was conservative when the large group had the large variance and was liberal when the large group were associated with the small variance. The best performing methods were SMM with Bartlett and SMM with ML which controlled Type I error rates around 0.05 across all variance patterns. Following SMM with Bartlett and SMM with ML, the Wilcox and James tests controlled Type I error adequately. The Welch, AG, and the BF methods were the next good performers in terms of controlling for Type I error.

6. Conclusion

ANOVA is a popular method used to compare the means of several groups. While there are many statistical tests for independent group means, there is no one suitable for every research

Variance pattern	OLS	James	WLS	BF	Welch	AG	Wilcox	ADF	BAR	ML	YB1	YB2	Mixed
Extreme	0.05	0.06	0.14	0.07	0.07	0.07	0.07	0.14	0.05	0.05	0.06	0.06	0.07
Split	0.04	0.07	0.14	0.06	0.07	0.07	0.07	0.10	0.05	0.05	0.07	0.08	0.07
Progress	0.03	0.07	0.14	0.05	0.07	0.07	0.07	0.10	0.05	0.05	0.06	0.07	0.08
Extreme Inversely	0.17	0.08	0.17	0.09	0.09	0.09	0.10	0.11	0.05	0.06	0.07	0.08	0.09
Split Inversely	0.20	0.10	0.19	0.06	0.11	0.10	0.12	0.12	0.06	0.07	0.08	0.09	0.10
Progress Inversely	0.14	0.09	0.19	0.05	0.10	0.11	0.12	0.12	0.06	0.07	0.08	0.09	0.10

Table 2: The Type I error rates of thirteen robust ANOVA tests by variance pattern at nominal alpha level of 0.05 for heterogeneous conditions. Note: For variance pattern, Extreme indicates one group has a different variance than the other groups; Split is when half the number of groups have each similar variance ratios; Progressive means that the population variances increased in a progressive way among groups; Progressive Inversely refers to the same variance patterns as in Progressive but in the reverse group order and it is similar to other reverse variance patterns.

situation. Therefore, it is important for applied researchers to have guidelines on selecting an appropriate approach for their research scenario. As noted in the simulation results part, the traditional ANOVA test had the best performance with equal group variances. However, it did not work well when the variances are heterogeneous. Among the other tests, BF, Bartlett, and SMM with ML seem to be robust to the violation of homogeneity assumption. While SAS and other statistical software (e.g., SPSS, IBM Corporation 2017, Stata, StataCorp 2017) do not provide all the robust tests for independent group mean comparison, this macro provides researchers with the ability to easily conduct these tests.

References

Alexander RA, Govern DM (1994). "A New and Simpler Approximation for ANOVA under Variance Heterogeneity." *Journal of Educational Statistics*, **19**(2), 91–101. doi:10.3102/10769986019002091.

Bartlett MS (1950). "Tests of Significance in Factor Analysis." *British Journal of Statistical Psychology*, **3**(2), 77–85. doi:10.1111/j.2044-8317.1950.tb00285.x.

Basso D, Pesarin F, Salmaso L, Solari A (2009). *Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R*. 3rd edition. Springer-Verlag, New York.

Brown MB, Forsythe AB (1974). "The Small Sample Behavior of Some Statistics Which Test the Equality of Several Means." *Technometrics*, **16**(1), 129–132. doi:10.1080/00401706.1974.10489158.

Browne MB (1982). *Topics in Applied Multivariate Analysis*. Cambridge University Press, Cambridge.

Conover WJ (1999). *Practical Nonparametric Statistics*. 3rd edition. John Wiley & Sons, New York.

Fan W, Hancock GR (2012). "Robust Means Modeling: An Alternative for Hypothesis Testing of Independent Means under Variance Heterogeneity and Nonnormality." *Journal of Educational and Behavioral Statistics*, **37**(1), 137–156. doi:10.3102/1076998610396897.

Hsiung T, Olejnik S, Huberty CJ (1994). "Comment on a Wilcox Test Statistic for Comparing Means When Variances Are Unequal." *Journal of Educational Statistics*, **19**(2), 111–118. doi:10.3102/10769986019002111.

IBM Corporation (2017). *IBM SPSS Statistics 25*. IBM Corporation, Armonk. URL https://www.ibm.com/analytics/spss-statistics-software/.

James GS (1951). "The Comparison of Several Groups of Observations When the Ratios of the Population Variances Are Unknown." *Biometrika*, **38**(3–4), 324–329. doi:10.1093/biomet/38.3-4.324.

Littell RC, Milliken GA, Stroup WW, Wolfinger RD, Schabenberger O (2006). *SAS for Mixed Models*. 2nd edition. SAS Institute Inc, Cary.

Montgomery DC, Peck EA (1992). *Introduction to Linear Regression Analysis*. John Wiley & Sons, New York.

Muthén B (1989). "Multiple-Group Structural Modelling with Non-Normal Continuous Variables." *British Journal of Mathematical and Statistical Psychology*, **42**(1), 55–62. doi:10.1111/j.2044-8317.1989.tb01114.x.

Pesarin F, Salmaso L (2010). *Permutation Tests for Complex Data: Theory, Applications and Software*. John Wiley & Sons, New York.

Rogan JC, Keselman HJ (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*, **14**(4), 493–498. doi:10.3102/00028312014004493.

SAS Institute Inc (2013). *SAS/STAT Software, Version 13.2*. SAS Institute Inc., Cary. URL https://www.sas.com/.

Satterthwaite FE (1946). "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin*, **2**(6), 110–114. doi:10.2307/3002019.

StataCorp (2017). *STATA Statistical Software: Release 15*. StataCorp LLC, College Station. URL https://www.stata.com/.

Tomarken AJ, Serlin RC (1986). "Comparison of ANOVA Alternatives under Variance Heterogeneity and Specific Noncentrality Structures." *Psychological Bulletin*, **99**(1), 90–99. doi:10.1037/0033-2909.99.1.90.

Welch BL (1951). "On the Comparison of Several Means: An Alternative Approach." *Biometrika*, **38**(3–4), 330–336. doi:10.1093/biomet/38.3-4.330.

Wilcox RR (1988). "A New Alternative to the ANOVA F and New Results on James Second-Order Method." *British Journal of Mathematical and Statistical Psychology*, **41**(1), 109–117. doi:10.1111/j.2044-8317.1988.tb00890.x.

Wilcox RR (1989). "Adjusting for Unequal Variances When Comparing Means in One-Way and Two-Way Fixed Effects ANOVA Models." *Journal of Educational and Behavioral Statistics*, **14**(3), 269–278. doi:10.3102/10769986014003269.

Yuan KH, Bentler PM (1997). "Mean and Covariance Structure Analysis: Theoretical and Practical Improvements." *Journal of the American Statistical Association*, **92**(438), 767–774. doi:10.1080/01621459.1997.10474029.

Yuan KH, Bentler PM (1999). "F Tests for Mean and Covariance Structure Analysis." *Journal of Educational and Behavioral Statistics*, **24**(3), 225–243. doi:10.3102/10769986024003225.

Affiliation:

Thanh V. Pham
Rightpath Research & Innovation Center
Department of Child & Family Studies
College of Behavioral and Community Sciences
University of South Florida
Tampa, Florida 33620, United States of America
E-mail: tvpham2@mail.usf.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
October 2020, Volume 95, Code Snippet 2 *Submitted:* 2016-07-01
doi:10.18637/jss.v095.c02 *Accepted:* 2018-09-30