**Paper 134-2015**

# ANOVA_Robust: A SAS® Macro for Various Robust Approaches to Testing Mean Differences in One-Factor ANOVA Models

Thanh V. Pham, EunSook Kim, Diep T. Nguyen, Yan Wang,

Jeffrey D. Kromrey, Yi-Hsin Chen

University of South Florida

## ABSTRACT

Testing the equality of several independent group means is a common statistical problem in social science. The traditional analysis of variance (ANOVA) is one of the most popular methods. However, the ANOVA *F* test is sensitive to violation of the homogeneity of variance assumption. Several alternative tests have been developed in response to this problem of the F test. These tests include a modification of ANOVA *F* test based on Structured Means Modeling technique. This paper provides a SAS macro for testing the equality of group means using thirteen different methods including regular ANOVA F test. In addition, this paper provides the results of simulation study to compare the performance of these tests in terms of their Type I error rate and statistical power under different conditions, especially, under the violation of homogeneity variance assumption.

**KEYWORDS**: Analysis of Variance, homogeneity of variance assumption, simulation study; homoscedasticity; heteroscedasticity.

## INTRODUCTION

The traditional analysis of variance (ANOVA) *F* test is the most common method to test the equality of several independent group means (Tomarken & Serlin 1986). However, ANOVA is sensitive to violation of the homogeneity of variance assumption even when sample sizes are equal (Rogan & Keselman, 1977). Alternative tests have been suggested in response to this problem with the *F* test. Researchers might choose among the Welch test (Welch, 1951), Brown–Forsythe test (Brown & Forsythe, 1974), James second-order test (James, 1951), and Alexander–Govern approximation (Alexander& Govern, 1994). Simulation studies have shown that these alternatives can control the Type I error rate when data are normally distributed and population variances are homogeneous. However, these tests become liberal when data are non-normal and heterogeneous (Fan &Hancock, 2012).

A different approach that does not require the assumption of the homogeneity of variance applies the technique called Structured Means Modeling (SMM). The SMM technique is developed from structural equation modeling (SEM) that allows group variances to be heterogeneous by freely estimating them. Moreover, various estimation methods robust to the violation of normality such as the Asymptotic Distribution Free (ADF) estimation (Browne, 1982) are available in addition to the maximum likelihood (ML) estimation in SEM (Fan & Hancock, 2012). Fan and Hancock (2012) showed that the SMM based tests performed better than ANOVA based tests in term of power and Type I error rate.

SAS can help users to conduct some of the robust ANOVA tests, but it does not provide test statistics of the other alternative methods. The purpose of this paper is to present a SAS macro that provides all test statistics of methods mentioned above to test the equality of independent means. The results of a simulation study to compare the performance of these methods are also presented.

## STATISTICAL METHODS FOR TESTING THE MEAN DIFFERENCES

### ANOVA F TEST

Analysis of Variance (ANOVA) is a common method to test the equality of several independent group means. The statistic *F* is defined by the following equation:

$$\mathrm{F} = \frac{\sum_{\mathrm{i}} \mathrm{n_i}(\mathrm{x_{i.}} - \mathrm{x_{..}})^2/(g-1)}{\sum_{\mathrm{i}}(N_I - 1)\mathrm{s_i}^2/(N-g)}.$$

where

$$\mathrm{N} = \sum_{\mathrm{i}} \mathrm{n_i}$$

$$\mathrm{x_{i.}} = \sum_{\mathrm{i}} \mathrm{x_{ij}}/\mathrm{n_i}$$

$$\mathrm{x}_{..} = \sum_{\mathrm{i}} \mathrm{n_i x_{i.}}/\mathrm{N}$$

$$\mathrm{s_i}^2 = \sum_{\mathrm{i}} \left(\mathrm{x_{ij}} - \mathrm{x_{i.}}\right)^2/(\mathrm{n_i} - 1)$$

And the *F* statistic follows the *F* distribution with $g-1$ and $N-g$ degree of freedom.

#### ALEXANDER AND GOVERN TEST

Alexander and Govern's test defines a weight ($\mathrm{w_i}$) for each group by $w_i = \frac{1/S_i{}^2}{\sum_1^k 1/S_i{}^2}$ where $S_i$ is standard error of group *i*.The variance-weighted estimate of the common mean ($Y^+$) is calculated by$Y^+ = \sum_1^k w_i \bar{Y}_i$. For each of k groups, *t*is defined as$t_i = \frac{\bar{Y}_i - Y^+}{S_i}$ . Statistic $t_i$ is distributed as Student's *t* with $v_i$(= $n_i$-1) degree of freedom. Normalizing transformation of $t_i$ to get $z_i$ by formula:

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

where $\mathrm{a} = \mathrm{v_i} - .5; \mathrm{b} = 48\mathrm{a}^2; \mathrm{c} = [\mathrm{a}\ln(1 + \mathrm{t_i}^2/\mathrm{v_i})]^{1/2}$. $z_i$is used to calculate A statistic by the following equation:

$$A = \sum_1^k z_i{}^2.$$

*A* is distributed as Chi-square with (*k*-1) degree of freedom and can be used to test the equality of independent group means.

#### BROWN-FORSYTHE

Brown and Forsythe testis a modification of ANOVA *F* test. The statistic is called *F**. The formula is:

$$\mathrm{F}^* = \frac{\sum_{\mathrm{i}} \mathrm{n_i}(\mathrm{x_{i.}} - \mathrm{x_{..}})^2}{\sum_{\mathrm{i}}(1 - \mathrm{n_i}/\mathrm{N})\mathrm{s_i}^2}.$$

*F** has an *F*-distribution with *g*-1 and *f* degrees of freedom where *f* is defined by the Satterthwaite approximation:

$$\frac{1}{\mathrm{f}} = \sum_{\mathrm{i}} \mathrm{C_i}^2/(\mathrm{n_i} - 1)$$

and

$$\mathrm{c_i} = (1 - \mathrm{n_i}/\mathrm{N})\mathrm{s_i}^2 \Big/ \left[\sum_{\mathrm{i}} (1 - \mathrm{n_i}/\mathrm{N})\mathrm{s_i}^2\right].$$

#### JAMES' SECOND ORDER TEST

The test statistic for James' test is $Q = \sum_{j=1}^{J} w_j \left(\bar{X}_j - X_w\right)^2$

where $w_j = \frac{n_j}{S_j^2}$ and $X_w = \sum_{j=1}^{J} w_j \bar{X}_j \Big/ \sum_{j=1}^{J} w_j$ .

The obtained value of *Q* is compared to a carefully adjusted critical value of $\chi^2$ with (*j* – 1) degrees of freedom (James, 1951).

#### WELCH TEST

Welch (1947) proposed a modification of the *F* test that compares the mean differences of multiple populations. It assumes that populations are independent and normally distributed, but does not require equal population variances. The test statistic can be defined as below:

$$\mathrm{F}' = \frac{\sum_{\mathrm{j}} \mathrm{w_j}\left[\frac{(\bar{\mathrm{X}}_{\mathrm{j}} - \bar{\mathrm{X}}')^2}{\mathrm{J}-1}\right]}{1 + \left[\frac{2(\mathrm{J}-2)}{\mathrm{J}^2-1}\right]\sum_{\mathrm{j}}\left[\left(1 - \frac{\mathrm{w_j}}{\mathrm{u}}\right)^2 (\mathrm{n_j} - 1)\right]}$$

Where $\bar{X}_j$ = the mean of group *j*; J= the number of groups; $w_j=n_j/s_j^2$; u= $\sum_j w_j$;$\bar{X}'=\sum_j \frac{w_j\bar{X}_j}{u}$. The distribution of *F'* can be approximated by the *F* distribution, using $v_B = J - 1$, and $\frac{1}{v_w} = (\frac{3}{j^2-1})\sum_j[\frac{\left(1-\frac{w_j}{u}\right)^2}{n_j-1}]$.

### WILCOX TEST

The Wilcox method was contrasted with James (1951) method. The modification of Wilcox's procedure that will be considered in the remainder of this section consists of setting

$$D_j = n_j/s_j^2$$
$$W_s = \sum D_j$$
$$\tilde{Y} = \sum D_j\tilde{Y}_j/W_s$$

where $\tilde{Y}_j = X_{n_j j}/n_j + \sum_{i=1}^{n_j-1}\left(1-\frac{1}{n_j}\right)X_{ij}/(n_j+1)$and rejecting null hypothesis when $H_m = \sum D_j(\tilde{Y}_j - \tilde{Y})^2$exceeds the 1 –αquantile of a chi-square distribution with (*j* – 1) degrees of freedom.

The Wilcox test has been shown to result in poor Type I error control if the population grand mean differs from zero (Hsiung, Olejnik, & Huberty, 1994). In the macro, the test was modified by grand mean centering the sample values of the dependent variable.

### WEIGHTED LEAST SQUARES

This method weights each observation by the inverse of its variance (Montgomery & Peck, 1992). A weight for each observation can be obtained by computing the reciprocal of the group variance as follows:

$$w_j = \frac{1}{s_j^2},$$

wherew and $s^2$ are the weight and sample variance for group *j*. Generalized least squares is used to minimize

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j} w_j(y_{ij} - \bar{y}_j).$$

The procedures of the weighted least squares used for testing the equality of the group means when HOV is violated are presented as follows: (a) compute the variance within each group; (b) compute the weight for each group, which is the inverse of the group variance; (c) merge the weight data into the original data and then run a weighted ANOVA.

### SMM APPROACH WITH MAXIMUM LIKELIHOOD (ML) ESTIMATION

When the SMM approach is applied to the between-subject testing of measured variable mean equality, indicator x can be expressed as $x = v_k + \delta$ where $v_k$ is a px1 vector of intercept values, δ is a px1 vector of normal errors.The null hypothesis of is thus $H_0: v_1 = v_2 = \cdots = v_k$which istested by constraining population means to be equivalent while still allowing forvariances of δ to be heterogeneous.Estimation within SMM can be handled by using maximum likelihood. The $F_{ML}$ is the ML fit function. The test statistics $T_{ML}$ is a function of $F_{ML}$ as $T_{ML}$ = (N-1) $F_{ML}$, with df equal to Kp(p + 3)/2 -q, where p is the number of observed variables and q is the number of parameters estimated across all groups.

### SMM APPROACH WITH ASYMPTOTIC DISTRIBUTION FREE (ADF) ESTIMATION

When the variables are continuous but not multivariate normally distributed, Browne (1982, 1984) proposed asymptotic distribution free estimation (ADF) for the covariance structure and Muthen (1989) expanded ADF including both mean and covariance structures. Using GLS-type fit function (generalized least squares), ADF fit function is defined as

$$F_{ADF} = 1/2\sum_{g=1}^{G}(s_g - \sigma_g)'W_g^{-1}(s_g - \sigma_g)$$

where for each group G,$s_g$ is the combined vector consisting of p elements of the observed means ($s_1$)and p(p+1)/2 elements of the variance covariance matrix ($s_2$), $\sigma_g$is the model implied counterpart of $s_g$, and W represents the ADF weight matrix as an estimator of the asymptotic covariance matrix of s. Under multivariate normality, the off-diagonal element of W (i.e., $W_{21}$ or $W_{12}$ which is a consistent estimator of the asymptotic covariance between$s_1$and $s_2$) equals zero whereas non-zero $W_{21}$should be computed under non-normality (see Muthen, 1989 for details). Then, the model parameters are estimated by minimizing the ADF fit function. This is also known as weighted least squares (WLS) with continuous variables in the software program Mplus. When this fit function is multiplied by 2n where n is the total sample size, it follows the chi-square distribution with (G -1) degrees of freedom.

#### SMM WITH BARLETT'S CORRECTION TO THE ML TEST STATISTICS

Within the context of EFA with m latent constructs and small sample sizes, Bartlett (1951) suggested a correction to the ML test statistic, which translates to:

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

With degree of freedom: df = Kp* - q; N= total sample size; p= number of observed variables, m= group's observed mean vector; q = number of parameters estimated across all groups.
$T_{BC}$ should more closely follow a w2 distribution with (Kp* – q) df than the usual $T_{ML}$ statistic. This adjusted statistic is equivalent to applying a multiplicative correction to $T_{ML}$ (or to any test statistic) of the form:

$$c = 1 – [(2p+4m+5)/6\ (N-1)]$$

#### YUAN AND BENTLER

Yuan and Bentler(1997, 1999), yielding test statistics $T_{YB1}$ and $T_{YB2}$ that make corrections to $T_{ADF}$ for small sample sizes. Specifically,

$$T_{\mathrm{YBI}} = \frac{\mathrm{T_{ADF}}}{1 + \frac{\mathrm{T_{ADF}}}{\mathrm{N}}}$$

Where $T_{ADF}$=(N-1) /$F_{ADF}$ ,which follows a central χ2 distribution with the same model df as $T_{ADF}$ (when $H_0$is true).

Their second modification to ADF appeals to the F distribution, transforming $T_{ADF}$ based upon the logic of the transformation applied to Hotelling'$sT^2$ statistic in multivariate analysis of variance (MANOVA). Observing that $T^2$ is a quadratic form, similar in structure to the ADF fit function, they proposed to rescale $T_{ADF}$ to an F-distributed statistic,

$$T_{YB2}= \left[\mathrm{N} - (\mathrm{K_p} * - \mathrm{q})\right] / \left[(\mathrm{N} - 1)(\mathrm{K_p} * - \mathrm{q})\right] \mathrm{T_{ADF}}$$

with numerator and denominator df of Kp* – q and N – (Kp* – q), respectively.
Inthe specific case of RMM, the numerator and denominator df for TYB2 reduce toK - 1 and N – K +1, respectively.

#### PROC MIXED

PROC MIXED provides an elegant test for mean differences while adjusting for unequal variances. 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. This is obtained using the DDFM = SATTERTHWAITE option on the MODEL statement in PROC MIXED.

## THE ANOVA_Robust MACRO

The macro ANOVA_Robust is written in Base SAS and SAS/STAT. Only three arguments are required for the macro: the name of the SAS data set containing the data to be analyzed (data), the name of the dependent variable (y), and the name of the independent variable (group).Default values are provided for each argument. Observations with missing values for either the independent or dependent variable are deleted from the analyses.

```
%MACRO ANOVA_robust (DATA = _Last_, y = y, group = group);
ODS listing close;
DATA wmDATA;SET &DATA;
KEEP &y &group;
IF CMISS(of _all_) then delete;
*compute the VARiance, mean, sample size for each group;
PROC SORT DATA =wmDATA; BY &group;
```

```
PROC MEANS DATA =wmDATA NOPRINT;BY &group;VAR &y;
OUTPUT OUT=groupi MEAN=grp_mean VAR=grp_VAR N=grp_n SUM=grp_sum idgroup(last
OUT(y)=last_score);
*compute the VARiance, mean, sample size for Big sample;
PROC MEANS DATA=wmDATA NOPRINT;
VAR &y;OUTPUT OUT=Big_sample MEAN=B_mean VAR=B_VAR N=Big_N;
DATA two;     SET groupi; id = 1; KEEP grp_mean grp_VAR grp_n id;
DATA grp_stat; SET two;
DATA three;   SET big_sample;
       id = 1;KEEP B_mean B_VAR Big_n id;
CALL symput('Sample_size', Big_N);
DATA four;    MERGE two three;    BY id;
       ci_c = (1-grp_n/big_n)*grp_VAR;
PROC MEANS DATA = four NOPRINT;
       VAR ci_c;     OUTPUT OUT=fourplus SUM = ci_cSum;
DATA five;    SET fourplus;
       id = 1;       g = _freq_;   KEEP id ci_cSum g;
CALL SYMPUT('N_group', g);
DATA six;     MERGE four five;    BY id;
       ci = ci_c/ci_cSum;  fi = ci*ci/(grp_n-1);
       Big_F1i = grp_n*(grp_mean-B_mean)*(grp_mean-B_mean);
PROC MEANS DATA = six NOPRINT;
       VAR fi big_f1i;     OUTPUT OUT=seven SUM = fi_sum Big_F1;
* Alexander Govern (1994) Method;*compute the weight for each group;
DATA weights;SET groupi;
 SE2 = grp_VAR/grp_n; * Square of Equation (1);
cell_wt_num = 1/SE2;
PROC MEANS NOPRINT DATA = weights;
VAR cell_wt_num; OUTPUT OUT = sumry SUM = sum_SE2;
DATA wt2;
IF _n_ = 1 THEN SET sumry; RETAIN sum_SE2;
SET weights;
w_i = cell_wt_num / sum_SE2;* Equation (2);
wtd_mean = w_i * grp_mean;
PROC MEANS NOPRINT DATA = wt2;VAR wtd_mean;
OUTPUT OUT = sumry2 SUM = Y_plus; * Equation (3);
DATA ts;
IF _n_ = 1 THEN SET sumry2;RETAIN Y_plus;
SET wt2;
t_i = (grp_mean - Y_plus) / SQRT(SE2); * Equation (4);
* Elements of Equation (8);
 a = grp_n - 1.5; b = 48*a**2; c = SQRT(a*log(1 + t_i**2/(grp_n - 1)));
z_i = c + (c**3 + 3*c)/b - (4*c**7 + 33*c**5 + 240*c**3 + 855*c)/(10*b**2 + 8*b*c**4 +
1000*b); * Equation (8);
z_i_squared = z_i**2;
PROC MEANS NOPRINT DATA = ts; VAR z_i_squared;
OUTPUT OUT = AG SUM = A N = n_groups; * Equation (9);
DATA Alex_Gov (KEEP = Labl dof obt_value P_value);SET AG;
Length labl $40 dof $ 15;
prob = 1 - PROBCHI(A,n_groups - 1);
Labl= "Alexander-Govern Test";
dof = TRIM(LEFT(n_groups - 1));
obt_value = INPUT(ROUND(A,.0001), $15.);
p_value = INPUT(ROUND(prob,.0001), $15.);
IF Prob<0.001 then p_value = "p < .001";
* Brown and Forsythe (BF) test
DATA eight;   SET seven;    ID = 1;
DATA final_BF (KEEP = LablObt_valuep_valuedof);
       Length labl $40 dof $ 15;  MERGE five eight;   BY ID;
       Big_f2 = ci_cSum;   d1 = g-1;     d2 = 1/fi_sum;      value = Big_f1/Big_f2;
       Obt_value = INPUT(ROUND(value,.0001), $15.);
       p = 1 - PROBF(Obt_value,d1,d2);  p_value = INPUT(ROUND(p,.0001), $15.);
       Labl = "Brown-Forsythe Test                          ";
```

```
       DOF = TRIM(LEFT(d1)) || ', ' || TRIM(LEFT(ROUND(d2,.01)));
       IF p <0.001 then p_value = "p < .001";
* James' second-order test;
DATA cells;SET Groupi;dj = grp_n/grp_VAR; djYbar = dj*grp_mean;
PROC MEANS NOPRINT DATA = cells; VAR djdjYbar;
OUTPUT OUT = sums SUM = D Sum_djYbar n = ngroups;
DATA back;
IF _n_ = 1 THEN SET sums;
RETAIN D Sum_djYbar ngroups;
SET cells;
Ystar = Sum_djYbar/D;
Uj = dj*(grp_mean - Ystar)**2;
vj = grp_n - 1;
 r10 = ((dj/D)**0)*(1/vj**1);
 r11 = ((dj/D)**1)*(1/vj**1);
 r12 = ((dj/D)**2)*(1/vj**1);
 r20 = ((dj/D)**0)*(1/vj**2);
 r21 = ((dj/D)**1)*(1/vj**2);
 r22 = ((dj/D)**2)*(1/vj**2);
 r23 = ((dj/D)**3)*(1/vj**2);
wt_VAR = (1 - dj/D)**2/vj;
PROC MEANS NOPRINT DATA = back;
VAR Uj r10 r11 r12 r20 r21 r22 r23 wt_VAR; ID ngroups;
OUTPUT OUT = obt_James SUM = U r10 r11 r12 r20 r21 r22 r23 wt_VAR;
DATA obt_James; SET obt_James;
* Long computation of critical value for James'' test, alpha = .10:
c10 = CINV(.90, ngroups - 1);
x2 = (c10**1) / (ngroups - 1);
x4 = (c10**2) / ((ngroups - 1)*(ngroups + 1));
x6 = (c10**3) / ((ngroups - 1)*(ngroups + 1)*(ngroups + 3));
x8 = (c10**4) / ((ngroups - 1)*(ngroups + 1)*(ngroups + 3)*
      (ngroups + 5));
 h10 = c10 + (1/2)*( 3*x4 + x2) * wt_VAR +
   (((1/16) * (3*x4 + x2)**2 * (1 - (ngroups - 3)/c10) * wt_VAR**2)
 +   (1/2) * (3*x4 + x2) * ((8*r23 - 10*r22 + 4*r21 - 6*r12**2 +
8*r12*r11 - 4*r11**2) +   (2*r23 - 4*r22 + 2*r21 - 2*r12**2 + 4*r12*r11 - 2*r11**2) *
(x2 -1 ) + (1/4) *   (-1*r12**2 + 4*r12*r11 - 2*r12*r10 - 4*r11**2 + 4*r11*r10 -
r10**2) *   (3*x4 - 2*x2 - 1)) +   (r23 - 3*r22 + 3*r21 - r20)*(5*x6 + 2*x4 + x2) +
   (3/16) * (r12**2 - 4*r23 + 6*r22 - 4*r21+r20)*(35*x8 + 15*x6 + 9*x4 + 5*x2) +
   (1/16) * (-2*r22 + 4*r21 - r20 + 2*r12*r10 - 4*r11*r10 + r10**2) *
   (9*x8 - 3*x6 - 5*x4 - x2) +
   (1/4)*(-1*r22 + r11**2) * (27*x8 + 3*x6 + x4 + x2) +
   (1/4) * (r23 - r12*r11) * (45*x8 + 9*x6 + 7*x4 + 3*x2));
* Long computation of critical value for James'' test, alpha = .05;
c05 = CINV(.95, ngroups - 1);
x2 = (c05**1) / (ngroups - 1);
x4 = (c05**2) / ((ngroups - 1)*(ngroups + 1));
x6 = (c05**3) / ((ngroups - 1)*(ngroups + 1)*(ngroups + 3));
x8 = (c05**4) / ((ngroups - 1)*(ngroups + 1)*(ngroups + 3)*
      (ngroups + 5));
h05 = c05 + (1/2)*( 3*x4 + x2) * wt_VAR +
   (((1/16) * (3*x4 + x2)**2 * (1 - (ngroups - 3)/c05) * wt_VAR**2)
 +
   (1/2) * (3*x4 + x2) * ((8*r23 - 10*r22 + 4*r21 - 6*r12**2 +
8*r12*r11 - 4*r11**2) +
   (2*r23 - 4*r22 + 2*r21 - 2*r12**2 + 4*r12*r11 - 2*r11**2) *
(x2 -1 ) + (1/4) *
   (-1*r12**2 + 4*r12*r11 - 2*r12*r10 - 4*r11**2 + 4*r11*r10 -
r10**2) *
   (3*x4 - 2*x2 - 1)) +
   (r23 - 3*r22 + 3*r21 - r20)*(5*x6 + 2*x4 + x2) +
   (3/16) * (r12**2 - 4*r23 + 6*r22 - 4*r21+r20)*(35*x8 + 15*x6
 + 9*x4 + 5*x2) +
```

```
   (1/16) * (-2*r22 + 4*r21 - r20 + 2*r12*r10 - 4*r11*r10 +
r10**2) *
   (9*x8 - 3*x6 - 5*x4 - x2) +
   (1/4)*(-1*r22 + r11**2) * (27*x8 + 3*x6 + x4 + x2) +
   (1/4) * (r23 - r12*r11) * (45*x8 + 9*x6 + 7*x4 + 3*x2));

* Long computation of critical value for James'' test, alpha = .01:
c01 = CINV(.99, ngroups - 1);
x2 = (c01**1) / (ngroups - 1);
x4 = (c01**2) / ((ngroups - 1)*(ngroups + 1));
x6 = (c01**3) / ((ngroups - 1)*(ngroups + 1)*(ngroups + 3));
x8 = (c01**4) / ((ngroups - 1)*(ngroups + 1)*(ngroups + 3)*
      (ngroups + 5));
h01 = c01 + (1/2)*( 3*x4 + x2) * wt_VAR +
   (((1/16) * (3*x4 + x2)**2 * (1 - (ngroups - 3)/c01) * wt_VAR**2)
 +   (1/2) * (3*x4 + x2) * ((8*r23 - 10*r22 + 4*r21 - 6*r12**2 +
8*r12*r11 - 4*r11**2) +
   (2*r23 - 4*r22 + 2*r21 - 2*r12**2 + 4*r12*r11 - 2*r11**2) *
(x2 -1 ) + (1/4) *
   (-1*r12**2 + 4*r12*r11 - 2*r12*r10 - 4*r11**2 + 4*r11*r10 -
r10**2) *   (3*x4 - 2*x2 - 1)) +
   (r23 - 3*r22 + 3*r21 - r20)*(5*x6 + 2*x4 + x2) +
   (3/16) * (r12**2 - 4*r23 + 6*r22 - 4*r21+r20)*(35*x8 + 15*x6
 + 9*x4 + 5*x2) +
   (1/16) * (-2*r22 + 4*r21 - r20 + 2*r12*r10 - 4*r11*r10 +
r10**2) *
   (9*x8 - 3*x6 - 5*x4 - x2) +
   (1/4)*(-1*r22 + r11**2) * (27*x8 + 3*x6 + x4 + x2) +
   (1/4) * (r23 - r12*r11) * (45*x8 + 9*x6 + 7*x4 + 3*x2));
IF U > h01 THEN result = 'p < .01';
ELSE IF U > h05 THEN result = 'p < .05';
ELSE IF U > h10 THEN result = 'p < .10';
ELSE result = 'p > .10';
dof = TRIM(LEFT(ngroups - 1));
*Welch Test: compute the weight for each group;
DATA Groupi; SET Groupi; grp_wgt = 1/grp_VAR;
*MERGE the weights into the original DATA;
DATA two; MERGE wmDATA Groupi; BY &group;
*conduct the weighted ANOVA;
PROC GLM DATA=two; WEIGHT grp_wgt; CLASS &group;
Model &y = &group /ss3; ODS OUTPUT overallANOVA = WLS_out;
TITLE 'Weighted Least Squared Analysis with GLM';
PROC GLM DATA = wmDATA; CLASS &group;
MODEL &y = &group; MEANS &group / welch;
ODS OUTPUT overallANOVA = ANOVA_out Welch = Welch_out;
TITLE 'Regular ANOVA and Welch Test via GLM';
PROC MIXED DATA = wmDATA; CLASS &group;
MODEL &y = &group / ddfm = satterthwaite;
REPEATED / group = &group; ODS OUTPUT Tests3 = Mixed1_out;
TITLE 'Heterogeneous VARiance Model with PROC Mixed';
*Wilcox Test: Grand mean center the outcome variable;
PROC MEANS DATA=wmDATAnoprint; VAR &y; OUTPUT OUT =g_mean MEAN=grand_mean;
DATA Wilcox_d; SET wmDATA;
IF _n_ = 1 THEN SET g_mean; RETAIN grand_mean;y_c = &y - grand_mean;
PROC SORT DATA=Wilcox_d; BY &group;
PROC MEANS DATA=Wilcox_d NOPRINT; BY &group; VAR y_c;
OUTPUT OUT=groupi_w MEAN=grp_mean VAR=grp_VAR N=grp_n SUM=grp_sum idgroup(last
out(y_c)=last_score);
*Define Dj, Y_tilda_j, and their product;
DATA two_w; SET groupi_w;   D=grp_n/grp_VAR; grp_sum1 = grp_sum - last_score;
Ytilda_j = last_score/grp_n + grp_sum1*(1-1/grp_n)/(grp_n+1);DYtilda_j = D*Ytilda_j;
PROC MEANS DATA=two_wnoprint; VAR D DYtilda_j; OUTPUT OUT=groupD SUM=W SumDYtilda;
DATA three_w; SET groupD; Ytilda = SumDYtilda/W;
```

```
DATA four_w; MERGE two_w three_w; BY _TYPE_; DROP W SumDYtilda;
Hm_j = D*(Ytilda_j - Ytilda)**2;
PROC MEANS DATA=four_wnoprint; VAR Hm_j; OUTPUT OUT=five_w SUM=Hm N=ngroups;
DATA Wilcox (keep = Labldofobt_valueP_value); SET five_w;
 Length labl $40 dof $ 15; p = 1 - probchi(Hm,ngroups-1); Labl = "Wilcox Test";
 obt_value = INPUT(ROUND(Hm,.0001), $15.);  p_value = INPUT(ROUND (p,.0001), $15.);
 dof = TRIM(LEFT(ngroups-1)); IF P <0.001 THEN p_value = "p < .001";
* SMM Tests;
%LET macstate = ' '; *initialize a macro variable that will become a SAS statement;
%DO grp = 1 %TO &N_group; * Parse original DATA SET BY value of group;
       DATA g&grp; SET wmDATA;
              IF &group = &grp;
* The following statement builds the macro variable with the SAS statements for proc
calis;
              %LET macstate = &macstate||'group '||TRIM(LEFT(&grp))||' / '||" DATA =
g&grp;"; * Note double quote to let macro value enter;
%END;
*transfer the macro variable into a regular SAS variable,then into macstate2;
DATA jk; myVAR = &macstate;
CALL symput('macstate2', myVAR);
PROC CALIS COVPATTERN=saturated MEANPATTERN=eqmeanvec OUTFIT=outfit METHOD= ADF;
VAR Y;
&macstate2
FITINDEX NoIndexType On(only)=[chisqdfprobchirmseaaiccaicsbc];
TITLE 'SMM with ADF: Testing Equal MEANS Only';
DATA ADF_fit; SET outfit;
IF FitIndex ~= 'Number of Observations' and FitIndex ~= 'Chi-Square'
And FitIndex ~= 'Chi-Square DF' and FitIndex ~= 'Pr> Chi-Square' then delete;
Keep FitIndex FitValue;
PROC TRANSPOSE DATA=adf_fit OUT=adf_fit; VAR FitValue;
DATA YB1 (KEEP = Labl dof obt_value P_value);
 SET ADF_fit;
 Length labl $40 dof $ 15;
   N = COL1; T_ADF = COL2; d1 = COL3; value = T_ADF/(1 + (T_ADF/N));
   p = 1 - PROBCHI(value,d1); dof = TRIM(LEFT(d1));
obt_value = INPUT(ROUND(value,.0001), $15.); p_value = INPUT(ROUND(p,.0001), $15.);
Labl = "Yuan and Bentler Test 1";
IF P <0.001 then p_value = "p < .001";
* Yuan and Bentler (1999) test statistic 2;
DATA YB2 (KEEP = Labl dof obt_value P_value); SET ADF_fit;
Length labl $40 dof $ 15;
   N = COL1; T_ADF = COL2; d1 = COL3; d2 = N - d1;
   value = ((N - d1)/((N - 1)*d1))*T_ADF; p = 1 - PROBF(value,d1,d2);
obt_value = INPUT(ROUND(value,.0001), $15.); p_value = INPUT(ROUND(p,.0001), $15.);
labl = "Yuan and Bentler Test 2";
DoF = TRIM(LEFT(d1)) || ', ' || TRIM(LEFT(ROUND(d2,.01)));
IF P <0.001 THEN p_value = "p < .001";
DATA finalSMM_ADF (KEEP = Labl dof obt_value P_value); SET adf_fit;
       Length labl $40 dof $ 15; Labl = 'ADF';
       DoF = TRIM(LEFT(COL3)); obt_value = INPUT(ROUND(COL2,.0001), $15.);
       p_value = INPUT(ROUND (col4,.0001), $15.);
       IF col4 <0.001 THEN p_value = "p < .001";
PROC CALIS COVPATTERN=saturated MEANPATTERN=eqmeanvec OUTFIT=MLoutfit; VAR Y;
&macstate2
Fitindex NoIndexType On(only)=[chisq df probchi rmsea aic caic sbc];
TITLE 'ML';
DATA ML_fit; SET MLoutfit;
IF FitIndex ~= 'Number of Observations' AND FitIndex ~= 'Chi-Square'
AND FitIndex ~= 'Chi-Square DF' AND FitIndex ~= 'Pr> Chi-Square' THEN DELETE;
KEEP FitIndex FitValue;
PROC TRANSPOSE DATA=ML_fit OUT=ML_fit; VAR FitValue;
DATA Bartlett (KEEP = Labl dof obt_value P_value); SET ML_fit;
Length labl $40 dof $ 15; N = COL1; T_ML = COL2; d1 = COL3;
```

```
   T_BC = (T_ML/(N-1))*(N-1/3-0/3-11/6);p = 1 - PROBCHI(T_bc,d1); value = T_BC;
obt_value = INPUT(ROUND(value,.0001), $15.); p_value = INPUT(ROUND(p,.0001), $15.);
Labl = "Barlett Correction"; DoF = TRIM(LEFT(d1));
IF p <0.001 THEN p_value = "p < .001";
DATA finalSMM_ML (KEEP = Labl dof obt_value P_value);SET ML_fit;
 Length labl $40 dof $ 15; Labl = 'ML';DoF = TRIM(LEFT(COL3));
 obt_value = INPUT(ROUND(COL2,.0001), $15.);p_value = INPUT(ROUND(col4,.0001), $15.);
       IF col4 <0.001 THEN p_value = "p < .001";
DATA welch1; SET Welch_out;
IF source NE "Error";
 Length labl $40 dof $ 15; df1 = df; obt_value = INPUT(ROUND(Fvalue,.0001), $15.);
 p_value = INPUT(ROUND(ProbF,.0001), $15.); Labl = "Welch Test"; MERGEID = 1;
       IF ProbF<0.001 THEN p_value = "p < .001";
DATA welch2 (KEEP = MERGEID df2); SET Welch_out;
IF source ="Error";
       df2 = ROUND(df,.01); MERGEID = 1;
DATA finalWelch (KEEP = Labl dof obt_value P_value); MERGE welch1 welch2; BY MERGEID;
       Dof = TRIM(LEFT(df1)) || ', ' || TRIM(LEFT(df2));

DATA finalJames (KEEP= Labl obt_value P_value dof); SET obt_James;
 Length labl $40 dof $ 15; Labl= "James' Second Order Test";
 obt_value = INPUT(ROUND(U,.0001), $15.); p_value = result;
DATA WLS1 (KEEP = Labl dof obt_value P_value MERGEID df1); SET WLS_out;
 IF source ="Model"; MERGEID = 1;Length labl $40 dof $ 15;  df1 = TRIM(LEFT(df));
 obt_value = INPUT(ROUND(Fvalue,.0001), $15.);
 p_value = INPUT(ROUND(ProbF,.0001), $15.); Labl = "Weighted Least Squares Test";
       IF ProbF<0.001 then p_value = "p < .001";
DATA WLS2 (KEEP = MERGEID df2); SET WLS_out; IF source ="Error";
       df2 = ROUND(df,.01); MERGEID = 1;
DATA finalWLS (KEEP = Labl dof obt_value P_value); MERGE WLS1 WLS2; BY MERGEID;
       dof=TRIM(LEFT(df1)) || ', ' || TRIM(LEFT(df2));
DATA ANOVA1 (KEEP = Labl dof obt_value P_value MERGEID df1); SET ANOVA_out;
IF source ="Model"; MERGEID = 1;  Length labl $40 dof $ 15; df1 = TRIM(LEFT(df));
 obt_value = INPUT(ROUND(Fvalue,.0001), $15.);
 p_value = INPUT(ROUND(ProbF,.0001), $15.);
       Labl = "Regular ANOVA F";  IF ProbF<0.001 THEN p_value = "p < .001";
DATA ANOVA2 (KEEP = MERGEID df2); SET ANOVA_out;
IF source ="Error"; df2 = ROUNF(df,.01); MERGEID = 1;
DATA finalANOVA (KEEP = Labl dof obt_value P_value); MERGE ANOVA1 ANOVA2;
       BY MERGEID; dof=TRIM(LEFT(df1)) || ', ' || TRIM(LEFT(df2));
DATA finalmixed (KEEP = Labl dof obt_value P_value); SET mixed1_out;
       Length labl $40 dof $ 15; d1 = numdf; d2 = dendf;
       dof=TRIM(left(d1)) || ', ' || TRIM(LEFT(ROUND(d2,.01)));
       obt_value = INPUT(ROUND(Fvalue,.0001), $15.);
       p_value = INPUT(ROUND (ProbF,.0001), $15.);
       Labl = "PROC MIXED"; IF ProbF<0.001 THEN p_value = "p < .001";
DATA SMM (KEEP = Labl dof obt_value P_value);
 Length labl $40 obt_value p_value dof $15.;
labl = 'Structural MEANS Modeling '; Obt_value = '  '; p_value = '   ';dof = '   ';
DATA grpstats; SET groupi;
*assign group label, n, mean, variance to same variable name as in finalprint;
labl = INPUT(&group, $40.); Obt_value = INPUT(grp_n, $15.);
 grp_mn2 = PUT(grp_mean, F15.4); grp_vr2 = PUT(grp_VAR, F15.4);
p_value = input(grp_mn2, $15.); dof = INPUT(grp_vr2, $15.);
*put grp_VARgrp_mean grp_mn2;* Start printing output;
DATA spanner1; Length labl $40 obt_value p_value dof $15.;
labl = 'Group'; Obt_value = 'N'; p_value = 'Mean'; dof = 'VARiance';
DATA spanner2; Length labl $40 obt_value p_value dof $15.;
labl = 'Test'; Obt_value = 'Value'; p_value = 'p value'; dof = 'DF';
DATA spanner_blank; Length labl $40 obt_value p_value dof $15.;
labl = '  '; Obt_value = '  '; p_value  = '   '; dof = '   ';
DATA finalprint;
```

```
SET Alex_Gov final_bf finalJames finalmixed finalANOVA finalWelch Wilcox finalWLS
spanner_blank SMM spanner_blank finalSMM_ML finalSMM_ADF Bartlett YB1 YB2;
ODS listing;
OPTIONS PAGENO=1 NODATE;
DATA _null_;
SET spanner1 spanner_blank grpstats spanner_blank spanner_blank spanner2 spanner_blank
finalprint;
varname = SYMGET('group'); n_names = SYMGET('n_group'); dv_name = SYMGET('y');
n_name = SYMGET('Sample_size'); G_name = STRIP(VARname); ng_name = STRIP(n_names);
d_name = STRIP(dv_name); nn_name = STRIP(n_name); FILE PRINT NOTITLES HEADER=PAGETOP;
PUT @3 labl @46 obt_value $15. -R @61 p_value $15. - R @76 dof $15. -R;
return;
page top:
PUT @3"Tests of Mean DIFferences" //
             @3"Independent VARiable:"        @30G_name$32. -R /
             @3"N of Groups:"                 @30ng_name$32. -R /
             @3"Dependent VARiable:"          @30d_name$32. -R /
             @3"Total N of Observations:"     @30nn_name$32. -R ///;
RUN;
%MEND ANOVA_robust;
```

## MACRO EXECUTION

In this example, the data set ONE contains observations on an independent variable (GROUP) and a dependent variable (Y). The macro is called after the data step.

```
DATA one;
INPUT group y;
CARDS;
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;
```

```
%ANOVA_robust(data = one, y = Y, group = Group);
RUN;
```

## OUTPUT EXAMPLE OF MACRO

An OUTPUT sample of the macro is demonstrated in Table 1. The first part of the table shows the DATA information. The second part includes the obtained value and associated *p*-value for each ANOVA_robust test.

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 |
|---|---|---|---|
| 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 |
| PROC MIXED | 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 |

Structural Means Modeling

| | | | |
|---|---|---|---|
| ML | 35.6174 | p < .001 | 4 |
| ADF | 142.1882 | p < .001 | 4 |
| Barlett Correction | 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 |

**Table 1- Macro output**

## SIMULATION STUDY

The simulation study was conducted to compare the methods for Type I error control. Six design factors were simulated: (1) number of groups, (2) cell size, (3) cell size pattern, (4) variance pattern, (5) maximum group variance ratio, and (6) population distribution ($\gamma1$ = 0.00 and $\gamma2$ =0.00, $\gamma1$ = 1.00 and $\gamma2$ = 3.00, $\gamma1$ = 1.50 and $\gamma2$ =5.00, $\gamma1$ = 2.00 and $\gamma2$ = 6.00, $\gamma1$ = 0.00 and $\gamma2$ =25.00, and $\gamma1$ = 0.00 and $\gamma2$ =-1.00, where $\gamma1$ and $\gamma2$ represent skewness and kurtosis, respectively).Non-normal populations were generated by implementing Fleishman's transformation (Fleishman, 1978). Since the space is limited, sample patterns and variance patterns are available upon request. The performance of thirteen testing approaches was examined at different nominal alpha levels: .01, .05, and .10. For each condition, 5,000 samples were generated. Type I error rate control was evaluated as the simulation outcomes. ANOVA analyses with eta-square effect sizes were conducted to explore the significant impacts of the research design factors on the variability in the estimated Type I error.

The simulation results show that ADF, YB1, and YB2 test do not provide solutions with some small sample size conditions (a minimum of four observations per group are required for these approaches). We treated them as missing data in analyzing simulation outcome. Boxplots were first examined to describe the distributions of Type I error rate estimates across all null conditions at each nominal alpha level. Figures 1 and 2 present boxplots of the rejection rate distributions at the .05 significance level of homogeneous null conditions and heterogeneous null conditions, respectively. Regarding the homogeneous conditions, the OLS method shows the best performance. Among the other approaches, BF, Wilcox, Bartlett, and SMM with ML controlled Type I error adequately. However, under the heterogeneous conditions, OLS method shows poor performance as expected. The BF, Bartlett and SMM with ML provide the best overall Type I error control.

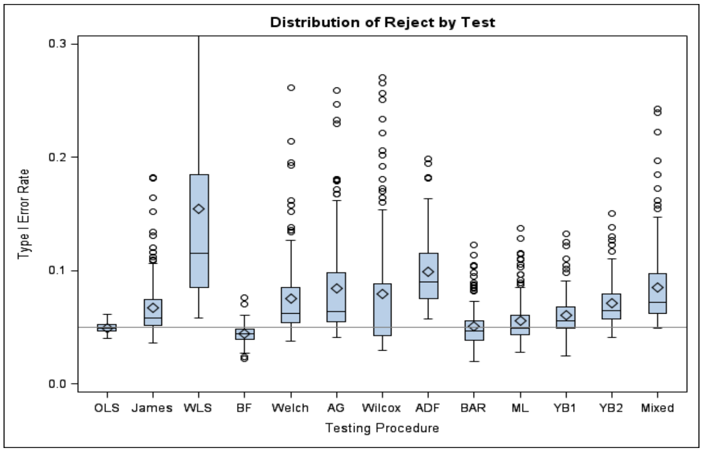


**Figure 1: Rejection rate distributions at the .05 significance level of homogeneous conditions**

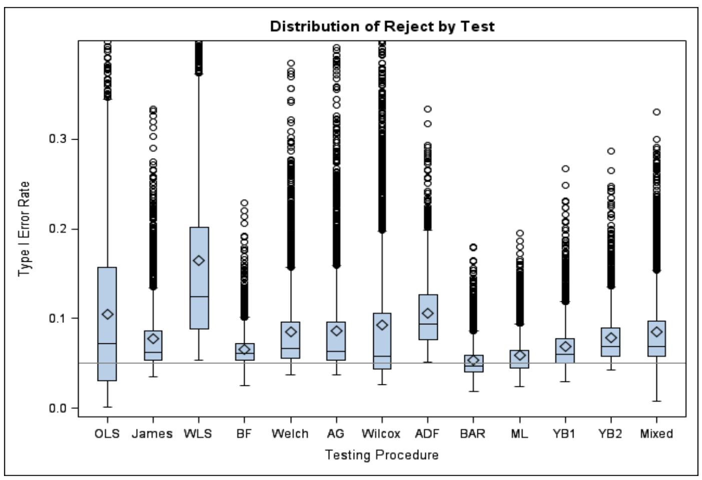


**Figure 2: Rejection rate distributions at the .05 significance level of heterogeneous conditions**

Eta-squared analyses were conducted to see which simulated factors have significant effect on the type I error rate. The testing method, population shape, the cell size and sample size pattern have significant effect but the effect of cell size on the Type I error is varied at different testing methods and population shape. For the heterogeneous conditions, the Type I error rate is affected by the interaction of cell size and testing method. The population shape also have significant effect on the Type I error rate.

## 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 single research 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, traditional ANOVA test has the best performance in homogeneous conditions. However, it is not working well under the heterogeneous conditions. Among the other tests, BF, Wilcox, Bartlett, and SMM with ML seem to be robust to the violation of homogeneity assumption.  While SAS does not provide all the robust tests for independent group mean comparison, this macro will provide the researcher with the ability to easily conduct these tests.

## REFERENCES

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational Statistics, 19*, 91–101.

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Psychology: Statistical Section, 3*, 77–85.

Brown, M. B., & Forsythe, A. B. (1974). The small sample behavior of some statistics which test the equality of several means. *Technometrics*, *16*(1), 129-132.

Browne, M. W. (1982). Covariance structures in DM Hawkins (Ed.), *Topics in applied multivariate analysis* (pp. 72–141).

Fan, W., & Hancock, G. R. (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.

Fleishman, A. I. (1978). A method for simulating non-normal distributions.*Psychometrika*, *43*(4), 521-532.

Hsiung, T.,Olejnik, S. &Huberty, C. J. (1994). Comment on a Wilcox test statistic for comparing means when variances are unequal. *Journal of Educational Statistics*, *19*, 111-118.

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

Lix, L. M., Keselman, J. C., &Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66,* 579–619.

Montgomery, D. C., & Peck, E. A. (1992*). Introduction to Linear Regression Analysis* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

Muthe´n, B. (1989). Multiple-group structural modeling with non-normal continuous variables. *British Journal of Mathematical and Statistical Psychology, 42*, 55–62.

O'brien, R. G. (1979). A general ANOVA method for robust tests of additive models for variances. *Journal of the American Statistical Association*, *74*(368), 877-880.

Rogan J. C. &Keselman H. J. (1977). Is the ANOVA F-Test Robust to Variance Heterogeneity When Sample Sizes areEqual?: An Investigation via a Coefficient of Variation. American Educational Research Journal Fall 1977, Vol 14, No. 4, Pp. 493-498.

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*, *14*(4), 493-498.

Satorra A. (1992). Asymptotic Robust Inferences in the Analysis of Mean and Covariance Structures. *Sociological Methodology*, *22*, 249-278.

Tomarken, A. J., &Serlin, R. C. (1986). Comparison of ANOVA alternatives under variance heterogeneity and specific noncentrality structures. *Psychological Bulletin*, *99*(1), 90.

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

Wilcox, R. R. (1988). A new alternative to the ANOVA F and new results on James's second-order method. *British Journal of Mathematical and Statistical Psychology, 41*, 109–117.

Yuan, K. H.,&Bentler,P.M. (1997).Mean and covariance structure analysis:Theoretical and practical improvements. *Journal of the American Statistical Association, 92*, 767–774.

Yuan, K. H., &Bentler, P. M. (1999). F tests for mean and covariance structure analysis. *Journal of Educational and Behavioral Statistics, 3*, 225–243.

Yuan, K.-H., &Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological methodology, 30*, 165–200.

## CONTACT INFORMATION

Your comments and questions are valued and encouraged.Contact the author at:

Thanh V. Pham

Educational Measurement, Research & Evaluation Program

Department of Educational and Psychological Studies

College of Education, EDU 105

University of South Florida

4202 East Fowler Avenue, EDU 105

Tampa, FL 33620-5650

Phone: (813) 974-6064

Fax: (813) 974-5910

E-mail: tvpham2@mail.usf.edu