Decision Science Letters 8 (2019) 309–316

Contents lists available at GrowingScience

Decision Science Letters

homepage: www.GrowingScience.com/dsl

The comparison of nonparametric statistical tests for interaction effects in factorial design

Ampai Thongteeraparp*

Department of Statistics, Faculty of Science, Kasetsart University, Bangkok, Thailand 10900

C H R O N I C L E

Article history:
Received October 9, 2018
Received in revised format: October 18, 2018
Accepted November 16, 2018
Available online November 16, 2018

Keywords:
Factorial design
Rank transformation
Modified mean
Adjusted rank transform test
Winsorized mean
Adjusted median transform

A B S T R A C T

Correct application of the classical factorial F-test depends on normality and homogeneity of variance assumptions. If these assumptions are violated the type I error rate will be inflated and power of the test will be decreased. Therefore nonparametric statistical tests have been proposed to analyze the interaction effects in factorial designs. A simulation was conducted to investigate the effect of non-normality on type I error rate and power of the test of the classical factorial F-test and five nonparametric tests namely rank transformation (FR), Winsorized mean (FW), modifies mean (FM), adjusted rank transform (ART) and adjusted median transform (AMT) using program SAS 9.4 with 1,000 replications. The study used 2×2 factorial design with replications of 3, 4 and 6 making sample sizes of 12, 16, and 24, respectively and 3×3 factorial designs with replication of 3 making a sample size of 27 studied at 0.05 level of significance. As a results, when the normality of assumption is satisfied all six statistical tests have the ability to control type I error in all situations. The ART test cannot control type I error rate for 3×3 factorial design when sample size is 27 when normality assumption is violated. For power of the test, the F-test provided the highest test power when the normality of assumption is met. The ART and AMT tests provided approximately the same test power. The AMT and ART tests can be effectively used to analyse the interaction effect between factors A and B in 2×2 factorial design when the sample size is 12 and 16 or 24 respectively and the normality of assumption is not met. Moreover, the results showed that when sample sizes increased, all six statistical tests tended to increase the power of the test.

1. Introduction

Factorial design is used to study the effect of factors on the characteristics of an interest. It is important to recall that the significant of the main effects and interactions are independent. An interaction is the effect that a combination of two or more factors has on the expected value of the response variable. In terms of the parametric perspective, the problem of testing the main effects and interactions are analyzed with Analysis of variance (ANOVA) model. The valid application of the ANOVA F-test depends on assumptions, namely that the observations are independent, the distributions of error are normal, and the observations have homogeneity of variance. In practice, violations of these assumptions are commonly stated many restudies such as O'Gorman (2001). If these assumptions are not met, then the type I error will deviate from the nominal level and this will decrease the power of the test. Therefore, nonparametric approach should be considered to be alternative methods to classical factorial

* Corresponding author.
E-mail address: fsciamu@ku.ac.th (A. Thongteeraparp)

doi: 10.5267/j.dsl.2018.11.003

F-test. The purpose of this study is to compare the classical factorial F-test and five nonparametric tests namely rank transformation (FR), Winsorized mean (FW), modified mean (FM), adjusted rank transform (ART) and adjusted median transform (AMT) for testing the interaction effects in factorial designs by considering their abilities to control type I error and the power of the tests when the normality assumption is not satisfied.

2. Methodology

2.1 Simulation

A simulation study was conducted to investigate the effect of non-normality on type I error rates and test power of the classical factorial F-test (F), rank transformation (FR), Winsorized mean (FW), modified mean (FM), adjusted rank transform (ART) and adjusted median transform (AMT) for testing 2×2 and 3×3 interaction effects in factorial designs. The model for this study is as follows,

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \quad (1)$$

where, Y_{ijk} is experimental response, μ is general mean, α_i is main effect of factor A, β_j is the main effect of factor B, $(\alpha\beta)_{ij}$ is the interaction effect between factor A and B and ε_{ijk} are random error terms. We generate data using program SAS 9.4 with 1,000 replications under the scope of the research as follows:

1. Determine distributions of observations as:
   (i) Normal distribution with mean 0 and variance 1
   (ii) Chi-square distribution with 5 degree of freedom
   (iii) t distribution with 2 degree of freedom

2. Determine replications according to levels of factors as:
   (i) 2×2 factorial designs: replications of 3, 4 and 6, making sample sizes of 12, 16, and 24, respectively.
   (ii) 3×3 factorial designs: a single replication of 3, making a sample size of 27.
   Note: Only balanced design (equal number of replications in each cell) is considered.
3. Determine significance level at 0.05

4. The effect of treatment is fixed to test the hypothesis:
   $H_0 : (\alpha\beta)_{ij} = 0$ (There is no interaction between factors A and B)
   $H_1 : (\alpha\beta)_{ij} \neq 0$ (There is interaction between factors A and B)
   There are 2 cases:
   1) The null hypothesis is true: set each parameter as:
      (i) 2×2 factorial designs
      The effect of treatment A: $\alpha_1 = \alpha_2 = 0$
      The effect of treatment B: $\beta_1 = \beta_2 = 0$
      (ii) 3×3 factorial designs
      The effect of treatment A: $\alpha_1 = \alpha_2 = \alpha_3 = 0$
      The effect of treatment B: $\beta_1 = \beta_2 = \beta_3 = 0$
   2) The null hypothesis is not true: set each parameter as:
      (i) 2×2 factorial designs
      The effect of treatment A: $\alpha_1 = -1, \alpha_2 = 1$
      The effect of treatment B: $\beta_1 = 1, \beta_2 = -1$
      (ii) 3×3 factorial designs
      The effect of treatment A: $\alpha_1 = -1, \alpha_2 = 0.5, \alpha_3 = 0.5$
      The effect of treatment B: $\beta_1 = 2, \beta_2 = -1, \beta_3 = -1$

All five statistics and classical factorial F- statistics were computed. It was determined whether H_0 would be rejected for interaction effect at the significance level of 0.05 and repeat 1000 times in each situation. We calculate the approximations of the probability of type I error and the percentages of the power of the test as follows,

$$\text{Probability of type I error} = \frac{\text{the number of reject } H_0 \text{ , when } H_0 \text{is true}}{1000}, \tag{2}$$

$$\text{Percentage of power of the test} = \frac{\text{the number of reject } H_0 \text{ , when } H_0 \text{is not true.}}{1000} \times 100. \tag{3}$$

To assess the ability to control type I error, Bradley (1978) criterion was applied. According to this criterion, the actual type I error rate of a test has to be in the range of $0.025\text{-}0.075$ when testing at the 0.05 level. In this study, a test would be considered to have the ability to control type I error, if its empirical type I error rate falls within the interval $[0.025, 0.075]$. We consider only statistical tests which have the ability to control type I error, if a statistical test has the highest power of the tests and assume that this statistical test is the most effective.

2.2 Statistical Tests

The statistical tests for interaction effects between two factors in this study are examined next.

2.2.1 Classical factorial F-test (F)

The total corrected sum of squares for two-way factorial F- test can be written as:

$$SS_{Total} = \sum_{i}^{a}\sum_{j}^{b}\sum_{k}^{r}\left(Y_{ijk}\text{-}\overline{Y}_{\dots}\right)^2, \tag{4}$$

where Y_{ijk} denotes the observation measured from replication *k* (number of replications), *i* levels (factor A) and *j* levels (factor *B*). $\overline{Y}_{\dots}$ denotes general mean for two way interactions.

Sum of squares for two-way factorial design are calculated as follow,

$$SS_{Cell} = r\left(\sum_{i}^{a}\sum_{j}^{b}\left(\overline{Y}_{ij.}\text{-}\overline{Y}_{\dots}\right)^2\right), \tag{5}$$

$$SS_{Error} = SS_{Total}\text{-}SS_{Cell}, \tag{6}$$

$$SS_{A} = rb\left(\sum_{i}\left(\overline{Y}_{i..}\text{-}\overline{Y}_{\dots}\right)^2\right), \tag{7}$$

$$SS_{B} = ra\left(\sum_{j}\left(\overline{Y}_{.j.}\text{-}\overline{Y}_{\dots}\right)^2\right), \tag{8}$$

$$SS_{AB} = SS_{Cell}\text{-}\left(SS_{A}+SS_{B}\right), \tag{9}$$

where SS_{Total} denotes the total sum of squares, SS_{AB} is the sum of squares for interaction of factor A and B, SS_{Cell} gives the sum of squares for cells or sub-groups, SS_A represents the sum of squares for factor A, SS_B provides the sum of squares for factor B and SS_{Error} is considered for the error sum of squares.

F statistic is computed as $F = \frac{MS_{AB}}{MS_{Error}}$. (10)

where $MS_{AB} = \frac{SS_{AB}}{DF_{AB}}$ denotes the mean square for interaction and $MS_{Error} = \frac{SS_{Error}}{DF_{Error}}$ denotes the mean square for error. The F- test statistic distributed as F-distribution with DF_{AB}= (a-1)(b-1) which is the degree of freedom for interaction and DF_{Error} = ab(r-1) which is the degree of freedom for error term, (Montgomery, 1997).

2.2.2 Rank transformation test (FR)

The rank transformation has been introduced by Conover and Iman (1976). This procedure is just the usual parametric procedure applied to rank of the data. Conover and Iman (1981) stated that the rank transformation procedure is robust and powerful in two way factor with a test for interaction when replication effect are present. From the study of Olejnik and Algina (1985), rank transformation has been recommended as an alternative to factorial F-test, especially when normality assumption is not met. The steps of FR are: (i) rank all observations (Y_{ijk}) by assigning one to the smallest and *n* to the largest. If ties are present, the average rank is assigned to all tied observations. Then, we replace each observation by its rank, (ii) classical factorial F-test on the ranks is used. Therefore, the corrected total sum of squares can be written as:

$$SS_{Total} = \sum_i \sum_j \sum_k \left(Y_{ijk} - \bar{Y}_{R...} \right)^2 , \quad (11)$$

where $\bar{Y}_{R...}$ denote general rank mean.

Computations of the sum of squares for main effects, interaction effect and error for the rank transformation procedure are the same as the classical factorial F-test. In this case, the rank transformation procedure test statistics are computed as follows,

$$FR = \frac{RMS_{AB}}{RMS_{Error}}, \quad (12)$$

where RMS_{AB} denotes the mean square for interaction computed based on ranked observations and RMS_{Error} is the mean square error computed based on ranked observations, respectively.

2.2.3 Winsorized mean test (FW)

Winsorized mean procedure has been studied by Wilcox (1996). It is a robust estimator of the population mean when there are outliers in the sample. The Winsorized mean is computed after the *k* smallest observations are replaced by the (k+1)st smallest observations, and the *k* largest observations are replaced by the (k+1)st largest observations. The steps of Winsorized mean approach are: (i) rank all observations in each treatment combination. (ii) replace the smallest observation in each treatment combination (position: r = 1) by the second smallest (position: r = 2) and replace the largest observation (position: r = r) by the second largest (position: r = r-1). For example, treatment combination a1b1 has 15, 17, 18, 19, 20, the result is 17, 17, 18, 19, 19. (iii) sums of squares are computed using general Winsorized mean by replacing the general arithmetic mean, (iv) the classical factorial F- test is applied on the general Winsorized mean. Therefore, the corrected total sum of squares can be written as follows,

$$SS_{Total} = \sum_i \sum_j \sum_k \left(Y_{ijk} - \overline{Y}_{W...} \right)^2, \quad (13)$$

where $\overline{Y}_{W...}$ denotes general Winsorized mean. Computation of the sum of squares for the main effects, interaction effect and error for the Winsorized mean procedure are the same as for the classical factorial F-test. Thus, test statistics for the Winsorized mean are computed as follows,

$$FW = \frac{WMS_{AB}}{WMS_{Error}}, \quad (14)$$

where WMS_{AB} is the mean square for interaction computed based on Winsorized mean and WMS_{Error} is the mean square error computed based on Winsorized mean.

2.2.4 Modified mean test (FM)

Mendeş and Yiğit (2013) presented the procedure of the modified mean. This procedure is computed by dividing the rank data set into two groups as Set 1 and Set 2. Then the arithmetic means of both groups are calculated as $\overline{Y}_{Set1}$ and $\overline{Y}_{Set2}$, respectively. We replace $\overline{Y}_{Set1}$ with the smallest number and replace $\overline{Y}_{Set2}$ with the largest number. The modified mean test is obtained as follows: (i) rank all observations in each treatment combination, (ii) calculate the smallest adjusted average (EK_{ij}) and calculate the largest adjusted average (EB_{ij}), where EK_{ij} denotes the average of observations which are lower than $\overline{Y}_{ij}$ and EB_{ij} denotes the average of observations which are greater than $\overline{Y}_{ij}$ (iii) in each treatment combination, replace the smallest observation by EK_{ij} and the largest observation by EB_{ij}. Afterwards, the mean of modified data set are calculated. Computations of the sum of squares for main effects, interaction effect and error for the modified mean, the procedure are the same as the classical factorial F-test. Therefore, the corrected total sum of squares can be written as follows,

$$SS_{Total} = \sum_i \sum_j \sum_k \left(Y_{ijk} - \overline{Y}_{M...} \right)^2, \quad (15)$$

where $\overline{Y}_{M...}$ denote general modified mean.
Test statistics for the modified mean are computed as below:

$$FM = \frac{MMS_{AB}}{MMS_{Error}}. \quad (16)$$

where MMS_{AB} denotes the mean square for interaction computed based on the modified mean observations and MMS_{Error} denotes the mean square error computed based on modified mean.

2.2.5 Adjusted rank transform test (ART)

ART is based on the rank transformation introduced by Conover and Iman (1981). Wobbrock et al. (2011) presented the aligned rank transform for nonparametric factorial data. The method consists aligning the observation before assigning the rank and analyses the adjusted data with classical F-test. The main idea of ART is to remove the unwanted effects from the response variable in order to study one effect at a time. Kelley and Sawilowsky (1997) found good results for the adjusted rank transform test and indicated that the test aligned by means had superior power when compared with the classical F-test if the distribution is heavy tailed or skewed. The procedure of adjusted rank transform test are: (i) subtract the average of all observations in level *i* from factor A $\left(\overline{Y}_{i..} \right)$ and the average of all

observations in level *j* from factor B $\left(\overline{Y}_{.j.}\right)$. Thus, the adjusted value is $Y_{ijk} - \overline{Y}_{i..} - \overline{Y}_{.j.}$ (ii) rank all adjusted values, if ties are present, the average rank is assigned to all tied observations, then, replace observations by rank of observations. (iii) using the rank of observation compute the sum of squares for main effects, interaction effect and error for the adjusted rank transform test in the same process as that for the classical factorial F-test.

2.2.6 Adjusted median transform test (AMT)

AMT is also based on the rank transformation introduced by Conover and Iman (1981). The procedure of AMT is developed from the idea of the ART using the median instead of mean by following the suggestion of Sawilowsky (1990) who recommended for using alignments other than the mean for further study of the aligned rank transform test for interaction. The procedures of adjusted median transform test are: (i) subtract the median of all observations in level *i* from factor A $\left(\tilde{Y}_{i..}\right)$ and the median of all observations in level j from factor B $\left(\tilde{Y}_{.j.}\right)$. Thus, the adjusted value is $Y_{ijk} - \tilde{Y}_{i..} - \tilde{Y}_{.j.}$ (ii) rank all adjusted values, if ties are present, the average rank is assigned to all tied observations, then, replace observations by rank of observations. (iii) using the rank of observation compute the sum of squares for main effects, interaction effect and error for the adjusted median transform test in the same process as that for the classical factorial F-test.

3. Research Results

3.1 The ability to control type I error

Table 1 shows the empirical type I error rates of the classical factorial F-test and five nonparametric tests namely rank transformation (FR), Winsorized mean (FW), modifies mean (FM), adjusted rank transform (ART) and adjusted median transform (AMT) where two-way factorial designs are used for significant level 0.05. The results show that for 2×2 factorial design all five statistical tests and classical factorial F-test have the ability to control type I error for all distribution. Thus all six statistical tests are robust to the normal assumption condition. The results for 3×3 factorial design show that when the normal assumption is violated, ART does not have the ability to control the type I error rate. However, all six statistical tests still have the ability to control type I error rate for the t distribution that is all six statistical tests still robust when the distribution is symmetry or not much deviate from the normal. Furthermore, the increase in the number of replication has positively affected keeping type I error rates at nominal level. When the level of factors A and B increased ART test tended to decrease the ability to control type I error.

3.2 Power of the test

To consider the power of the test, the results in Table 2 show that for 2×2 factorial design the classical F-test and FW test provided approximately the same test power while ART test and AMT test provided approximately the same test power. The classical F-test provided the highest test power for all number of replications when the normality assumption holds. While the distributions are Chi-square and t distribution, AMT test provided the highest test power when the sample size is 12 and ART test provided the highest test power when the sample size is 16 or 24. For 3×3 factorial design classical F-test and FW test provided approximately the same test power. F-test and FW test have the highest test power when the normality assumption is satisfied. While the distribution are chi-square and t distribution, ART test provided the highest test power. Moreover, the result show that when sample sizes increased, all six statistical tests tended to increase the power of the test.

Table 1
The empirical Type I error rate for the six statistical tests

n	a×b	Number of replications	Distribution of error	Statistical test					
				F	FR	FW	FM	ART	AMT
12	2×2	3	Normal	0.065	0.065	0.054	0.049	0.064	0.053
			Chi-square	0.040	0.054	0.040	0.036	0.049	0.053
			t	0.044	0.050	0.044	0.042	0.052	0.053
16	2×2	4	Normal	0.050	0.050	0.049	0.046	0.055	0.049
			Chi-square	0.043	0.051	0.042	0.037	0.063	0.057
			t	0.048	0.057	0.047	0.042	0.052	0.048
24	2×2	6	Normal	0.045	0.054	0.045	0.044	0.045	0.050
			Chi-square	0.046	0.052	0.045	0.044	0.056	0.055
			t	0.046	0.050	0.046	0.044	0.055	0.053
27	3×3	3	Normal	0.053	0.060	0.052	0.047	0.056	0.060
			Chi-square	0.057	0.064	0.056	0.053	0.085*	0.072
			t	0.052	0.064	0.052	0.048	0.064	0.070

Note: *means the statistical test cannot control type I error.

Table 2
Power of the test for the six statistical tests

n	a×b	Number of replications	Distribution of error	Statistical test					
				F	FR	FW	FM	ART	AMT
12	2×2	3	Normal	**0.867**	0.404	0.867	0.851	0.859	0.860
			Chi-square	0.405	0.271	0.402	0.361	0.429	**0.452**
			t	0.724	0.370	0.721	0.701	0.709	**0.734**
16	2×2	4	Normal	**0.952**	0.567	0.951	0.948	0.950	0.947
			Chi-square	0.522	0.392	0.521	0.481	**0.584**	0.582
			t	0.835	0.545	0.835	0.817	**0.854**	0.824
24	2×2	6	Normal	**0.999**	0.861	**0.999**	**0.999**	0.996	0.995
			Chi-square	0.668	0.592	0.664	0.632	**0.786**	0.765
			t	0.950	0.805	0.950	0.944	**0.965**	**0.965**
27	3×3	3	Normal	**0.982**	0.453	**0.982**	0.977	0.971	0.719
			Chi-square	0.460	0.281	0.458	0.417	**0.540**	0.452
			t	0.823	0.411	0.822	0.804	**0.847**	0.844

Note: bold number means the statistical test has the the highest test power.

Table 3
Summary of results for the six statistical tests

n	**a×b**	**Number of replications**	**Distribution of error**	**Statistical test**					
				F	**FR**	**FW**	**FM**	**ART**	**MED**
12	2×2	3	Normal	(1)	**	**	(1)	**	**
			Chi-square	**	**	**	**	**	(1)
			t	**	**	**	**	**	(1)
16	2×2	4	Normal	(1)	**	**	**	**	**
			Chi-square	**	**	**	**	(1)	**
			t	**	**	**	**	(1)	**
24	2×2	6	Normal	(1)	**	(1)	(1)	**	**
			Chi-square	**	**	**	**	(1)	**
			t	**	**	**	**	(1)	(1)
27	3×3	3	Normal	(1)	**	**	(1)	**	**
			Chi-square	**	**	**	**	(1)	**
			t	**	**	**	**	-	(1)

Note: - means the statistical test does not have the ability to control type I error.
** means the statistical test has the ability to control type I error.
(1) means the statistical test has the ability to control type I error and has the highest power.

4. Conclusion and Discussion

O'Gorman (2001) presented that some nonparametric tests could be used in place of classical F-test when normality assumption is not satisfied. However the performance of these nonparametric tests may differ based on the experiment condition such as distribution, number of factors, number of replications, etc. In general the parametric factorial F-test would recommend if the normality assumption is not violated because it provides the greatest power and would hold the type I error rate at nominal level. In this study, the results have shown that the classical F-test had the ability to control type I error rate and had the highest test power when the normality assumption was satisfied. However, one can conclude that the shape of the distribution did not affect the ability to control type I error much but the level of factors A and B and the number of replications did. As the level of factors A and B or the number of replications increased, ART test tended to decrease the ability to control type I error. To consider the power of the test, the F-test provided the highest test power when normality assumption was satisfied, if the assumption of normality is suspicious AMT test and ART test are recommended. The ART test is an alternative nonparametric statistical test for testing the interaction effect between factors A and B in 2×2 factorial designs when the sample size is 16 or 24 and the distribution of error is Chi-square. The AMT test is recommended for testing the interaction of 3×3 factorial designs when the sample size is 27. Sample size affected the power of the test; when the sample size increased, all six statistical tests tended to increase the power of the test.

References

Bradley, J. V. (1978). Robustness. *British Journal of Mathematical and Statistical Psychology*, *31*, 144–152.

Conover, W., & Iman, R. L. (1976). On some alternative procedures using ranks for the analysis of experimental designs. *Communications in Statistics-Theory and Methods*, *5*(14), 1349-1368.

Conover, W. J., & Iman, R. L. (1981). Rank transformations as a bridge between parametric and nonparametric statistics. *The American Statistician*, *35*(3), 124-129.

Kelley, D. L. and Sawilowsky, S.S. (1997). Nonparametric alternatives to the F-statistics in analysis of variance. *Journal of Computer Simulations, 58*, 343–359.

O'Gorman, T. W. (2001). A comparison of the F-test, Friedman's test, and several aligned rank tests for the analysis of randomized complete blocks. *Journal of Agricultural, Biological, and Environmental Statistics*, *6*(3), 367-378.

Montgomery, D. C. (1997). *Design and analysis of experiments*. 4th edn, John Wiley & Sons, Inc., New York, USA.

Olejnik, S. F., & Algina, J. (1985). A review of nonparametric alternatives to analysis of covariance. *Evaluation Review*, *9*(1), 51-83.

Sawilowsky, S. S. (1990). Nonparametric tests of interaction in experimental design. *Review of Educational Research*, *60*(1), 91-126.

Wilcox, R. R. (1996). A note on testing hypotheses about trimmed means. *Biometrical Journal*, *38*(2), 173-180.

Mendeş, M., & Yiğit, S. (2013). Type I error and test power of different tests for testing interaction effects in factorial experiments. *Statistica Neerlandica*, *67*(1), 1-26.

Wobbrock, J. O., Findlater, L., Gergle, D., & Higgins, J. J. (2011, May). The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 143-146). ACM.