Research Article *Science, Engineering and Health Studies 2019, 13(1), 29-37*

Impact of homogeneity of variances violation in single factor components of variance model when sampling from finite population

Teerawat Simmachan

Department of Mathematics and Statistics, Faculty of Science and Technology, Thammasat University, Pathum Thani 10120, Thailand
Corresponding author: teerawat@mathstat.sci.tu.ac.th

Received: June 5, 2018; Revised: October 31, 2018; Accepted: November 20, 2018

ABSTRACT

This study aims to appraise the impact of heteroscedasticity in a single factor component of variance model when random effects are sampled from a finite population. Monte Carlo simulation was conducted to evaluate the performance of the F-statistic in the one-way ANOVA via the type I error rate and power. Results suggest that when the null hypothesis is true, the F-test can generally keep the nominal α of 0.05 even the homogeneity of variances is not satisfied, whereas the empirical type I error rates are far from α = 0.01. Further, the heterogeneity of variances is still a problem in the ANOVA for both terms, i.e., the type I error rate and power even for medium heterogeneity cases. The finite F-test always has greater power than the usual F-test in case in which the heteroscedasticity is presented and the random effects (τ_i) are sampled from a finite population. This suggests that a large value of a type II error rate may arise when the sampling fraction (k/N) exceeds five percent. Under this condition, one should avoid use of the ordinary F-test in a single factor ANOVA.

Keywords: ANOVA; finite population; heteroscedasticity; Monte Carlo simulation; variance components

1. INTRODUCTION

Analysis of variance or ANOVA is the most widely used method in statistical analyses (Moder, 2007; Simmachan et al., 2012). The single factor completely randomized design (CRD) is one of the most basic approaches. In a CRD, the data is typically analyzed using a one-way ANOVA model associated with either a fixed effects model or a random effects model, also known as a variance components model or a components of variance model.

The component of variance model is commonly used in manufacturing quality control. Consider the following example. Suppose a company has 50 machines producing bottles for soft drinks, and the process engineer wants to understand the variation in content between bottles. He randomly samples 10 machines from entire machines and produces 40 bottles from each, giving a total sample of 400 bottles. This is a balanced and completely randomized design for a random effects model with 10 treatments and 400 units. It is normally assumed that the factor levels (treatments) are randomly selected from an infinite population (Bennett and Franklin, 1954; Searle et al., 2006; Montgomery, 2013). In this manuscript, the example is drawn from a finite population. However, if this population is a sufficiently large proportion of the randomly selected factor levels, it can be considered infinite (Searle et al., 2006; Montgomery, 2013).

In most studies using finite populations in variance components models, the authors propose

ISSN Online 2630-0087

formulas for both the expected mean squares and variances of the components within the models. Early studies of this type used balanced data (Bennett and Franklin, 1954; Cornfield and Tukey, 1956; Tukey; 1956). Studies using unbalanced data have applied the variances of variance components for single classification (Tukey, 1957), or 3-way nested classification (Mahamunulu, 1963). Attempts have been made to develop the rule of converting expectations under infinite models into expectations under finite models with balanced and unbalanced data, and nested and crossed classifications (Hartley, 1967; Gaylor and Hartwell, 1969; Searle and Fawcett, 1970). The levels of each factor are assumed to be finite. Importantly, they further assume that the populations of random errors are also finite. This is the critical point addressed in the current paper. The n_i observations within a specific factor level are sampled from an infinite number of possible observations, or infinite population. More precisely, n_i replications are a random sample from a $N(\mu + \tau_i, \sigma^2)$ distribution. In other words, the random errors (ε_{ij}) are independent and identically normally distributed with a zero mean and constant variance σ^2or $\varepsilon_{ij} \sim IND(0, \sigma^2)$, and their populations are non-finite.

To address this, Simmachan (2011) began with the simplest model. He derived the expected mean squares for balanced and unbalanced data of the single factor components of a variance model in which the random effects populations are finite but the random errors (ε_{ij}) are drawn from an infinite population. The proofs were presented in Simmachan et al. (2012), along with simulation results for the one-way model in which the random effects (τ_i) are sampled from a finite population and $\varepsilon_{ij} \sim IND(0, \sigma^2)$. Empirical type I error rates and powers of the F-test were used to evaluate the performance of the F-test. However, only balanced data were used, with homogeneity of variances.

As noted above, factor levels (k) are randomly selected from a finite population of size N. In sampling theory, this can be treated as infinite if the sampling fraction denoted by k/N is small. This fraction is related to a finite population correction (fpc) denoted by $(N-k)/(N-1)$ and is used when a sample of size (k) is drawn at random without replacement from a finite population size (N). In practice, the fpc can be ignored if the sampling fraction does not exceed five percent (Cochran, 1977). It is also assumed to have a value of one for a single factor component of a variance model when the number of factor levels is small relative to the population sizes of random effects.

This raises the first question: how to decide whether N is large enough? In other words, if the fraction k/N exceeds five percent, will the data analysis be affected? Further, the data analysis in this model is based upon the F-test in ANOVA. A correct application of this method has three main requirements: (i) independent samples; (ii) normally distributed populations; and (iii) homogeneity of variances. The last requirement is a serious problem in statistical inference (Moder, 2010). The second question concerns the effect that violation of the homogeneity of variances in the one-way components of variance model has on statistical inference in a finite population.

A literature review suggests that these questions have not yet been addressed. It is important to understand the impact that the sampling fraction has when it cannot be ignored and the heteroscedasticity in a single factor component of the variance model. In this study, the type I error rate and power of the test were calculated to evaluate the performance of the F-test in ANOVA for the one-way components of variance model in cases in which the assumption of equality of variances is not satisfied and the random effects (τ_i) are sampled from a finite population. This was done using Monte Carlo simulation.

2. MATERIALS AND METHODS

2.1 A single factor random effects model

For a response variable, y, the model of interest

can be written as

$$y_{ij} = \mu + \tau_i + \varepsilon_{ij} \quad (1)$$

where $i = 1, 2, \ldots, k$ are levels of a factor, $j = 1, 2, \ldots, n_i$ is the sample size per factor level, μ is an overall mean, τ_i are random effects of the i^{th} factor level, and ε_{ij} are random errors. Both τ_i and ε_{ij} have specified distributions, which are traditionally assumed to be normal. When τ_i are sampled from a finite population, the following model assumptions are changed as follows (Simmachan et al., 2012). The treatment effects (τ_i) are selected from a finite population of size N with zero mean and population variance σ_τ^2, and the covariance between each pair of treatment effects is nonzero, so that $COV(\tau_i, \tau_i) \neq 0$.

2.2 Research scope

1) The sampling fractions (k/N) for all combinations of k and N exceed five percent, except a single case where $k = 3$ and $N = 78$. This is shown in Table 1.

Table 1 Sampling fractions under consideration

k/N	$N = 10$	$N = 21$	$N = 36$	$N = 55$	$N = 78$
$k = 3$	0.300	0.143	0.083	0.055	**0.038**
$k = 4$	0.400	0.190	0.111	0.073	0.051

2) The nominal α -levels are 0.01 and 0.05.

3) The differences between variances are expressed by a non-centrality parameter, ϕ^2, as proposed by Games et al. (1972), defined as follows:

$$\phi^2 = \frac{\sum_{i=1}^{k}\left(\sigma_i^2 - \bar{\sigma}^2\right)^2 / k}{\sigma_1^2}$$

where k denotes the number of randomly selected factor levels, σ_1^2 is the first variance, and $\bar{\sigma}^2$ represents the mean of k variances. The details are given in Table 2.

Table 2 Difference levels of error variances

k	Difference levels	σ_i^2	ϕ^2
3	Low	(0.5:1.0:2.0)	0.778
	Medium	(0.5:1.0:3.0)	2.333
	High	(0.5:1.0:5.0)	8.111
4	Low	(0.5:1.0:1.0:2.0)	0.594
	Medium	(0.5:1.0:2.0:3.0)	1.844
	High	(0.5:1.0:3.0:5.0)	6.344

4) For the balanced design, sample sizes per factor level are 5, 10, 15 and 20. For the unbalanced design where $k = 3$, four sets of sample sizes are used: (3:4:5), (3:6:9), (5:10:15), and (10:15:20). For the unbalanced design where $k = 4$, the sample sizes are (3:4:5:6), (4:5:6:10), (5:8:12:15), and (5:10:15:20).

5) To compute the power of sampling from a finite population, two different finite populations of τ_i values are determined and, without loss of generality, the means of the finite populations are assumed to be zero:

(a) Discrete uniform distribution

N values of τ_i are created from -(N-1)/2 to -(N-1)/2, e.g., $\tau_i \in$ {-4.5, -3.5,…, 3.5, 4.5} for N = 10.

(b) Discrete non-uniform distribution

The frequencies of the τ_i values are shown in Table 3 for each population size N, e.g., for $N = 10$, $\tau_i \in$ {1-C, 2-C, 2-C, 3-C, 3-C, 3-C, 4-C, 4-C, 4-C, 4-C}, where $C = 3$ makes the mean of the distribution zero.

6) R statistical software is used for all simulations, with the number of runs set at 10,000 for each scenario.

Table 3 Frequencies of τ_i values for discrete non-uniform distribution

τ_i	$N = 10$	$N = 21$	$N = 36$	$N = 55$	$N = 78$
1-*C*	1	1	1	1	1
2-*C*	2	2	2	2	2
3-*C*	3	3	3	3	3
4-*C*	4	4	4	4	4
5-*C*		5	5	5	5
6-*C*		6	6	6	6
7-*C*			7	7	7
8-*C*			8	8	8
9-*C*				9	9
10-*C*				10	10
11-*C*					11
12-*C*					12
C	3	13/3	17/3	7	25/3

2.3 Simulation procedure

For analysis of type I error rates, data are generated under the null hypothesis, i.e., $\sigma_\tau^2 = 0$. This implies that all τ_i values are zero. Note that, without loss of generality, μ is assumed to be zero in model (1), and the model then reduces to $y_{ij} = \varepsilon_{ij}$, where $\varepsilon_{ij} \sim N(0, \sigma_i^2)$. The $F_0 = MSTr/MSE$ statistic is calculated for y_{ij} and it is noted whether the null hypothesis is rejected at $\alpha = 0.01$ and 0.05. Next, the empirical type I error rates are estimated, as the proportion of F-tests per 10,000 that reject the null hypothesis. Finally, the empirical type I error rates are compared with Bradley's (1978) criterion. If the rate falls within the interval [0.5α, 1.5α], the F-statistic is deemed to be robust with respect to type I error control.

For power analysis, the empirical power of the F-statistic is computed under two conditions. First, the power of the finite F-test is calculated using random effects (τ_i) sampled without replacement from two finite populations (the discrete uniform and discrete non-uniform distributions). Second, the power of the usual F-test is computed by $\tau_i \sim N(0, \sigma_\tau^2)$ sampled from infinite populations where σ_τ^2 is computed from the two distributions. The goal here is to compare the power with and without the fpc (the sampling fraction can or cannot be ignored). The finite F-test represents the case in which the fpc is used in data analysis, whereas the usual F-test covers the case in which the fpc is not used. Since the focus of this study is the heteroscedasticity case, the power is not computed based upon a non-centrality parameter for the F distribution defined by $\lambda = \sum_{i=1}^{k} n_i \tau_i^2 / \sigma^2$. This method is different from the approach of Simmachan (2011). The procedure is as follows: create two finite populations of τ_i values defined in the scope, then compute the population variance (σ_τ^2), take a simple random sample without replacement of k different τ_i values from the population of N values and generate n_i responses for each of the k treatments. Note again that, without loss of generality, μ is assumed to be zero. Data are generated under the alternative hypothesis, i.e., $\sigma_\tau^2 > 0$, τ_i values in the previous step are used, and the model (1) is reduced to be $y_{ij} = \tau_i + \varepsilon_{ij}$ where $\varepsilon_{ij} \sim N(0, \sigma_i^2)$. The F_0 for y_{ij} is calculated and it is noted whether the null hypothesis is rejected at $\alpha = 0.01$ and 0.05. These steps are repeated 10,000 times. Next, the empirical powers, or the proportion of the 10,000 F-tests that reject the null hypothesis, are estimated. In this step, the simulated power of the finite F-statistic is obtained. To find the power of the usual F-statistic, similar steps are used, but with two changes. In the first step, $\tau_i \sim N(0, \sigma_\tau^2)$ are assumed where σ_τ^2 is calculated from the two finite populations (discrete uniform and discrete non-uniform). In the second step, k different τ_i values are sampled from infinite populations. Finally, the empirical powers of the finite F and usual F-tests are compared. A larger power represents a superior performance.

3. RESULTS

As can be seen from Table 4, the type I error rate analysis suggests that the F-test can generally

control the nominal α of 0.05. Additionally, the F-statistic can preserve the nominal α of 0.05 for up to 79.17 percent of the 48 total scenarios. In the case of α = 0.01, the empirical type I error rates (Table 5) are far from 0.01 even in medium heterogeneity cases. The F-statistic can preserve the nominal α of 0.01 only in 47.92 percent of the scenarios. However, the F-test performed better on balanced data than on unbalanced data. Figures 1 and 2 present the data from Table 4 and 5 in graphical form. Here, the vertical axis shows the empirical type I error rates and the horizontal axis the 48 scenarios.

The results of power analysis are visualized in Figure 3 and 4. Given the very large number of results for all combinations of distributions, α, n_i, k, N, and σ_i^2, only selected results are presented. Figure 3 shows the empirical power of the two F-tests with balanced and unbalanced data, a discrete uniform distribution, α = 0.05, k = 3, and high heterogeneity of variances. The upper and lower rows show the power for balanced and unbalanced data, respectively. In balanced cases, the horizontal axis represents the sample size at each factor level. In the unbalanced cases, it represents the total observations. Subfigures (a)-(f) show the empirical power at N values of 10, 21, and 55. As can be seen, the empirical power of the finite F-test is larger than that of the usual F-test in both balanced and balanced cases.

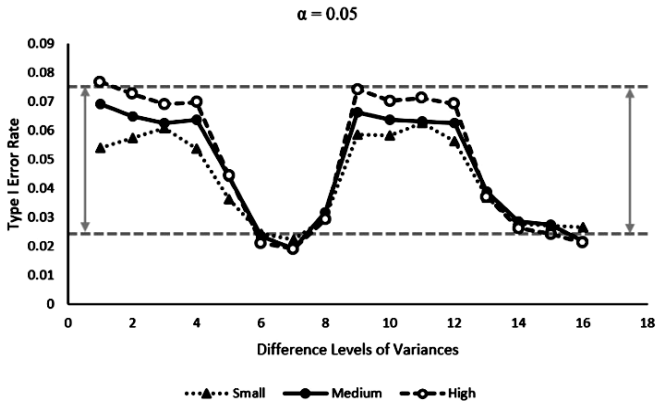

Figure 1 Empirical type I error rate at α = 0.05

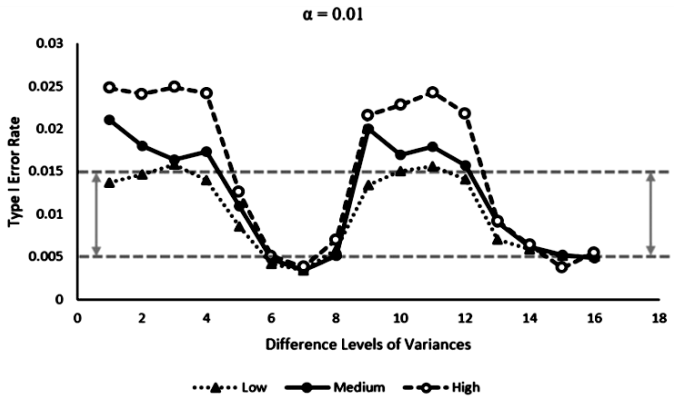

Figure 2 Empirical type I error rate at α = 0.01

Table 4 Empirical type I error rate for nominal α = 0.05

k	n_i	Difference levels of variances		
		Small	Medium	High
3	(5:5:5)	0.0540*	0.0692*	0.0768
	(10:10:10)	0.0575*	0.0648*	0.0727*
	(15:15:15)	0.0609*	0.0626*	0.0691*
	(20:20:20)	0.0538*	0.0637*	0.0698*
	(3:4:5)	0.0363*	0.0440*	0.0444*
	(3:6:9)	0.0241	0.0234	0.0210
	(5:10:15)	0.0221	0.0190	0.0190
	(10:15:20)	0.0296*	0.0316*	0.0292*
4	(5:5:5:5)	0.0587*	0.0662*	0.0742*
	(10:10:10:10)	0.0583*	0.0637*	0.0702*
	(15:15:15:15)	0.0627*	0.0632*	0.0712*
	(20:20:20:20)	0.0563*	0.0626*	0.0692*
	(3:4:5:6)	0.0367*	0.0388*	0.0371*
	(4:5:6:10)	0.0278*	0.0285*	0.0262*
	(5:8:12:15)	0.0271*	0.0275*	0.0241
	(5:10:15:20)	0.0266*	0.0217	0.0211

Note: An F-test that controls the type I error is indicated by ''.*

Table 5 Empirical type I error rate for nominal α = 0.01

k	n_i	Difference levels of variances		
		Small	Medium	High
3	(5:5:5)	0.0137*	0.0211	0.0248
	(10:10:10)	0.0147*	0.0180	0.0241
	(15:15:15)	0.0158	0.0164	0.0249
	(20:20:20)	0.0140*	0.0173	0.0242
	(3:4:5)	0.0086*	0.0109*	0.0126*
	(3:6:9)	0.0042	0.0047	0.0051*
	(5:10:15)	0.0034	0.0034	0.0038
	(10:15:20)	0.0059*	0.0051*	0.0070*
4	(5:5:5:5)	0.0134*	0.0200	0.0216
	(10:10:10:10)	0.0150*	0.0170	0.0228
	(15:15:15:15)	0.0156	0.0179	0.0243
	(20:20:20:20)	0.0141*	0.0157	0.0218
	(3:4:5:6)	0.0070*	0.0090*	0.0092*
	(4:5:6:10)	0.0059*	0.0062*	0.0064*
	(5:8:12:15)	0.0051*	0.0052*	0.0037
	(5:10:15:20)	0.0050*	0.0048	0.0055*

Note: An F-test that controls the type I error is indicated by ''.*

However, both tests have higher power when applied to balanced data rather than unbalanced data. Further, as the sample size per factor level and experimental units increase, the power also increases. In addition, as the population size of random effects inflates, the difference between the two powers becomes small. In the cases where $N = 55$, the two powers are approximately the same. For $N = 78$, the two statistics had identical power (data not shown). Importantly, the sampling fraction affects the data analysis when using the F-test. This is because the sampling fraction becomes large, if N is not sufficiently large, so that the finite F-test and the usual F-test perform differently. Figure 4 shows the power for $k = 3$ and $k = 4$, balanced data, discrete uniform distribution, $N = 10$, and $\alpha = 0.05$. The first and second rows show powers for $k = 3$ and $k = 4$, respectively. Subfigures (a)-(f) present the powers in cases of low, medium, and high heteroscedasticity, respectively. Note that, as the sample size per factor level increases, both simulated powers of the F-statistics also increase. Correspondingly, the empirical powers for $k = 4$ are larger than those for $k = 3$, because the total units are larger. Moreover, the power of the two F-tests converges as the number of factor levels increases. Heteroscedasticity affects hypothesis testing because the empirical power of the two F-statistics decreases as the difference levels of error variances increase. However, this effect seems to be better in case with balanced data and a large total unit size. Interestingly, the power of the finite F-test dominates that of the usual F-test in all scenarios.

The results for $\alpha = 0.01$ (not shown) follow the same pattern as those $\alpha = 0.05$, but the simulated power for $\alpha = 0.05$ is larger than that for $\alpha = 0.01$. The results are very similar for the discrete non-uniform distribution and discrete uniform distribution (not shown). Importantly, the power of the finite F remains better than that of the usual F.

4. DISCUSSION AND CONCLUSIONS

We now return to the two questions that motivate this study. The first concerns hypothesis testing under conditions in which the sampling fraction exceeds five percent. We find that this has a real effect on the power of the F-test in ANOVA, in that the finite F always dominates the usual F although the error variances are heterogeneous. This effect is known to be a problem in balanced designs with homogeneity of variance (Simmachan, 2011). Precisely, if N is insufficiently large relative to a large k or k/N, the type II error rate will also have a large value. This important result suggests that a large value of a type II error rate may arise when a traditional F-test is used with the one-way components of variance model, the random effects are sampled from a finite population, and the population size of the random effects is relatively small. This suggests avoiding the use of the standard F-test when

the sampling fraction exceeds five percent.

The second question is whether the ANOVA is affected by heterogeneity of variances in a model with a finite population. As with a fixed effects model and an infinite population, heteroscedasticity is known to be a problem in ANOVA even in cases of the medium heterogeneity (Liu, 2015; Moder, 2007; Moder, 2010). This inflates the type I error rate as the significance level decreases, even when the sample sizes per factor level are equal. Similarly, homoscedasticity violation increases the type II error rate as the nominal α-level is decreased, despite claims in the literature (and in textbooks) that the F-test is robust. The power also reduces a significantly under homogeneity of variances (Hecke, 2010). This suggests that Hotelling's T^2 test (Moder, 2007) or the Welch test on rank (Cribbie et al., 2007) should be preferred when heteroscedasticity is present.

The results reported in this study suggest directions for future work. One would be to consider situations in which $k > 4$. Violation of normality should also be considered, as this often appears in practical applications. The values of τ_i may have other distributions, both discrete and continuous. The important question is whether the finite F-test always has greater power than the usual F-test. If this is so, a proof of the expected mean squares is required. If it is not so, then a counter example must be presented. An extension to other designs including blocking, factorial experiments and nested designs should be conducted.

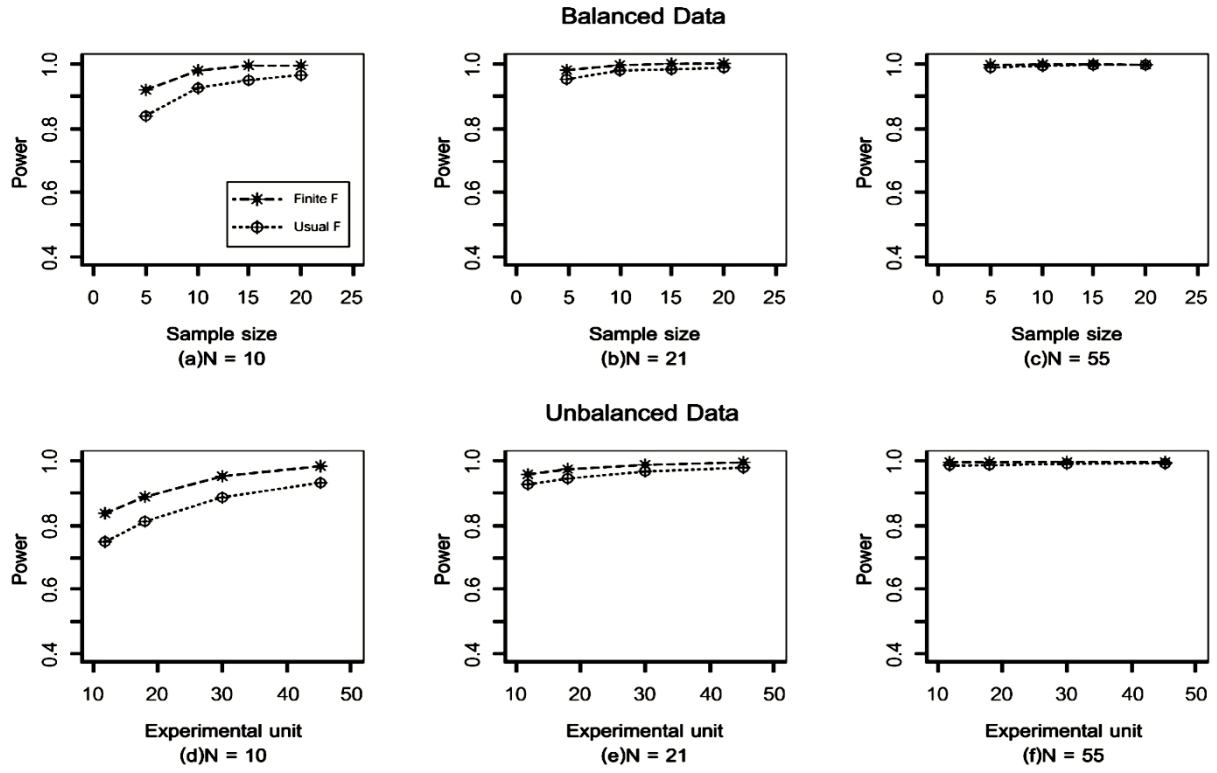

Figure 3 Empirical powers for balanced and unbalanced data with discrete uniform, $k = 3$ and $\alpha = 0.05$

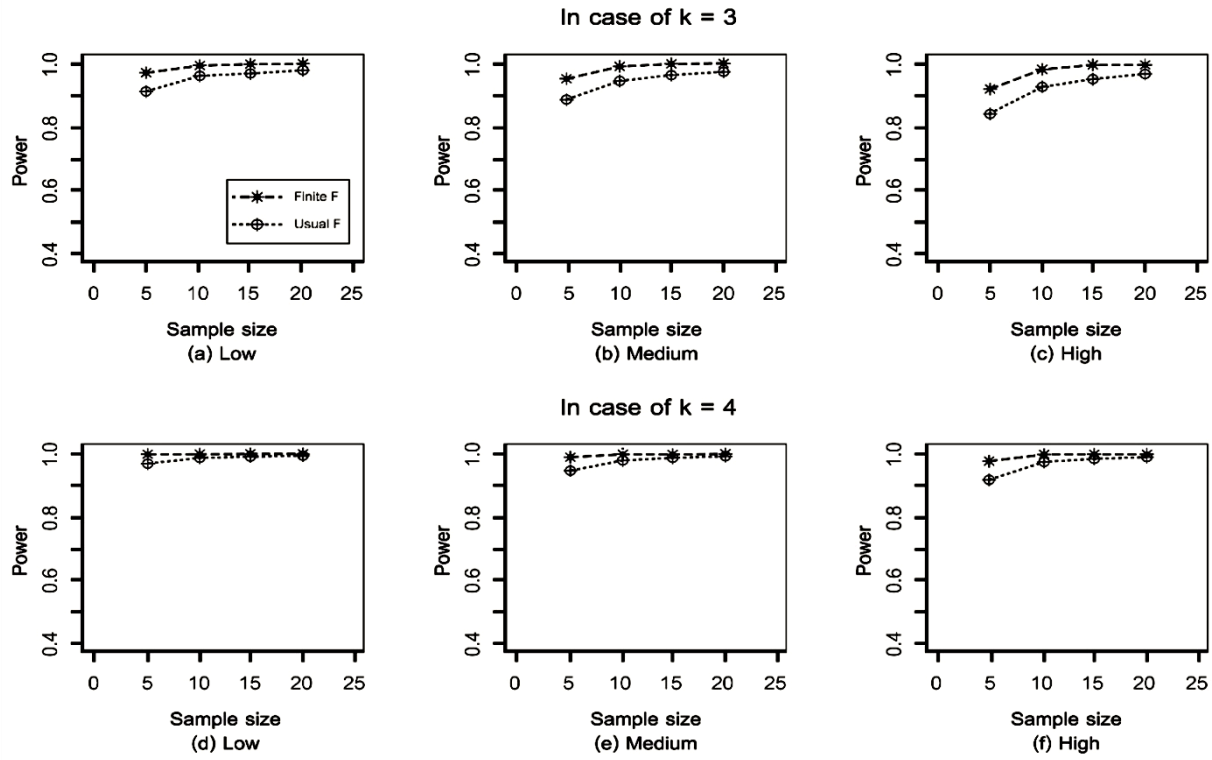

Figure 4 Empirical powers for k = 3 and k = 4 with discrete uniform, balanced data, $N = 10$ and $\alpha = 0.05$

ACKNOWLEDGEMENTS

Author gratefully acknowledges the financial support provided by Thammasat University Research Fund under the TU Research Scholar, Contact No. 1/ 5/ 2559. Special thanks are also extended to the referees for providing valuable suggestions.

REFERENCES

Bennett, C. A. and Flanklin, N. L. (1954). *Statistical Analysis in Chemistry and the Chemical Industry*, John Wiley & Sons, New York, pp. 393.

Bradley, J. V. (1978). Robustness? *The British Journal of Mathematical and Statistical Psychology*, 31(2), 144-152.

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed., John Wiley & Sons, New York: pp. 25.

Cornfield, J. and Tukey, J. W. (1956). Average values of mean squares in factorials. *Annals of Mathematical Statistics*, 27, 907-949.

Cribbie, R. A., Wilcox, R. R., Bewell, C., and Keselman, H. J. (2007). Test for treatment group equality when data are nonnormal and heteroscedastic. *Journal of Modern Applied Statistical Methods*, 6(1), 117-132.

Game, P. A., Winkler, H. B., and Probert, D. A. (1972). Robust tests for homogeneity of variance. *Educational and Psychological Measurement*, 32, 887-909.

Gaylor, D. W. and Hartwell, T. D. (1969). Expected mean squares for nested classifications. *Biometrics*, 25, 427-430.

Hartley, H. O. (1967). Expectations, variances and covariances of ANOVA mean squares by synthesis. *Biometrics*, 23, 105-114.

Hecke, T. V. (2010). Power study of anova versus Kruskal-Wallis test. *Journal of Statistics and Management Systems*, 15(2-3), 241-247.

Liu, H. (2015). *Comparing Welch's ANOVA, a Kruskal-*

Wallis Test and Traditional ANOVA in Case of Heterogeneity of Variance (master's thesis). Virginia Commonwealth University, Richmond, Virginia, United States.

Mahamunulu, D. M. (1963). Sampling variances of the estimates of variance components in the unbalanced 3-way nested classification. *Annals of Mathematical Statistics*, 34, 521-527.

Moder, K. (2007). How to keep the type I error rate in ANOVA if variances are heteroscedastic. *Austrian Journal of Statistics*, 36(3), 179-188.

Moder, K. (2010). Alternatives to F-Test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modelling*, 52(4), 343-353.

Montgomery, D. C. (2013). *Design and Analysis of Experiments*, 8th ed., John Wiley & Sons, New York, pp. 116-118.

Searle, S. R., Casella, G., and McCulloch, C. E. (2006). *Variance Components*, John Wiley & Sons, New Jersey, pp. 16-18.

Searle, S. R. and Fawcett, R. F. (1970). Expected mean squares in variance components models having finite populations. *Biometrics*, 26, 243-254.

Simmachan, T. (2011). *Analytical Method for the Random Effects One-Way ANOVA Model when Sampling from a Finite Population* (master's thesis). Thammasat University, Thailand.

Simmachan, T., Borkowski, J. J., and Budsaba, K. (2012). Expected mean squares for the random effects one-way ANOVA model when sampling from a finite population. *Thailand Statistician*, 10(1), 121-128.

Tukey, J. W. (1956). Variances of variance components: I Balanced designs. *Annals of Mathematical Statistics*, 27, 722-736.

Tukey, J. W. (1957). Variances of variance components: II The unbalanced single classification. *Annals of Mathematical Statistics*, 28, 43-56.