Virginia Commonwealth University
VCU Scholars Compass

Theses and Dissertations

Graduate School

2015

Comparing Welch's ANOVA, a Kruskal-Wallis test and traditional ANOVA in case of Heterogeneity of Variance

Hangcheng Liu

Follow this and additional works at: https://scholarscompass.vcu.edu/etd

Part of the Biostatistics Commons

© The Author

Downloaded from
https://scholarscompass.vcu.edu/etd/3985

This Thesis is brought to you for free and open access by the Graduate School at VCU Scholars Compass. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of VCU Scholars Compass. For more information, please contact libcompass@vcu.edu.

COMPARING WELCH ANOVA, A KRUSKAL-WALLIS TEST, AND TRADITIONAL ANOVA IN CASE OF HETEROGENEITY OF VARIANCE

By

Hangcheng Liu

A Thesis

Submitted to the Faculty of
Virginia Commonwealth University
in Partial Fulfillment of the Requirements
for the Master of Science Degrees
in Biostatistics
in the Department of Biostatistics

Richmond, Virginia

JULY 2015

Name: Hangcheng Liu

Date of Degree: July 1, 2015

Institution: Virginia Commonwealth University

Major Field: Biostatistics

Major Professor: Dr. Wen Wan

Title of Study: COMPARING WELCH ANOVA, A KRUSKAL-WALLIS TEST, AND TRADITIONAL ANOVA IN CASE OF HETEROGENEITY OF VARIANCE

Pages in Study: 46

Candidate for Master of Science Degrees

BACKGROUND: Analysis of variance (ANOVA) is a robust test against the normality assumption, but it may be inappropriate when the assumption of homogeneity of variance has been violated. Welch ANOVA and the Kruskal-Wallis test (a non-parametric method) can be applicable for this case. In this study we compare the three methods in empirical type I error rate and power, when heterogeneity of variance occurs and find out which method is the most suitable with which cases including balanced/unbalanced, small/large sample size, and/or with normal/non-normal distributions.

METHODS: Data for three-group comparison are generated via Monte Carlo simulations with the cases of homogeneity/heterogeneity of variance, balanced/unbalanced, normal/non-normal distributions, equal/unequal means, in various sample sizes. The three methods, ANOVA, Welch and Kruskal-Wallis, are used to compare three-group means in a global test (The null hypothesis H_0: all three means are the same vs the alternative hypothesis Ha: at least two means are different) in each simulated dataset in each scenario. When true means are all equal, a type I error rate is calculated to determine the performances of the three methods. When at least two true means are unequal, power of detecting the differences among the three groups is calculated instead.

RESULTS: When simulated data are normally/mild log-normally distributed in balanced/unbalanced designs with homogeneity of variance, the traditional ANOVA controls the nominal type I error the best in the most cases and both Welch and Kruskal-Wallis are competitive to ANOVA, especially in a large sample size. When data are heterogeneous, normal, and balanced, the Welch method controls the nominal type I error the best in all the cases and gain the most power in the most cases, while ANOVA and Kruskal-Wallis exceed the nominal type I error and have the inflation problem in many cases. When data are heterogeneous, normal, and unbalanced, the Welch method controls the nominal type I error the best in all the cases and gain the most power in the most cases, while ANOVA and Kruskal-Wallis are unstable: quite conservative in the cases of the large-variance group having a large sample size and quite inflated in the cases of the large-variance group having a small sample size. When data are heterogeneous, log-normal, and balanced, the Kruskal-Wallis fails to control the nominal type I error rate and Welch and ANOVA have inflation problems in many cases.

CONCLUSIONS: In terms of type I error rate analysis: Traditional ANOVA is the best when data are homogeneous, normal, and balanced/unbalanced; Welch method performs the best when data are heterogeneous, normal, and balanced/unbalanced; Kruskal-Wallis is most unstable when the data are non-normal, while both ANOVA and Welch have type I error inflation problems. In terms of power analysis, three methods may have different powers depending on the settings of simulations.

Key words: ANOVA, Welch, Kruskal-Wallis, three-group comparison, Monte Carlo simulations, homogeneous/heterogeneous, balanced/unbalanced, type I error, power

ACKNOWLEDGEMENTS

I would like to say a big thank you to everyone who helped me complete this thesis. In deed it’s been a long journey but thanks for sticking with me and by me.

Special thanks to Dr. Wen Wan, my thesis advisor, summer project director and thesis committee chair, for your guidance over these two years. Thanks for the suggestions, encouragement and positive feedback.

I would like to thank my thesis committee members, Dr. Donna McClish and Dr. Qin Wang, who took the time to review my work and gave me lots of great feedback.

Finally, I would like to thank members of my family who have been helpful in diverse ways. I could not have done this without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

CHAPTER I

INTRODUCTION

1.1. Introduction

Analysis of variance (ANOVA) is one of the most frequently used methods in statistics (Moder 2007) and it allows us to compare more than two group means in a continuous response variable. An ANOVA model assumes that:

1. The probability distribution of responses in each group is normal.
2. Each probability distribution has the same variance.
3. Samples are independent.

Note that with the normality of the probability distributions and the constant variability, the probability distributions differ only with respect to their means. ANOVA is a very powerful test as long as all prerequisites are met. It is not necessary, nor is it usually possible, that an ANOVA model fit the data perfectly. ANOVA models are reasonably robust against certain types of departures from the model, such as the data not being exactly normally distributed (Kutner et al, 2005).

However, inhomogeneity of variances with data normally distributed may lead to an increased type I error rate and it can be found in a lot of practical trials (Moder 2007). Kutner et al (2005) summarize that a standard remedial measure is to use the weighted least squares method if data are normally distributed. If data are not normally distributed, appropriate data transformation for normality is a standard remedy. When transformations are not successful in stabilizing the group variances and bringing the distribution of the data close to normal, a nonparametric method may be used to compare group means.

Moder (2010) compares the multiple methods in case of heterogeneity of variance with normally distributed data: traditional ANOVA, Welch ANOVA, weighted ANOVA, Kruskal-Wallis test, permutation test using F-statistic as implemented in R-package "coin", permutation test based on Kruskal-Wallis statistic, and a special kind of Hotelling's T^2 method (Moder, 2007; Hotteling, 1931). His simulation results show that traditional ANOVA, permutation tests, and weighted ANOVA cannot control the nominal type I error rate in many cases. Kruskal-Wallis test does not exceed nominal α in some cases but is very conservative. Welch ANOVA may be useful for a small number of groups (3 or 5), but cannot control the nominal α well when the number of groups is 10 or higher. The Hotelling's T^2 test controls α well in all situations of balanced designs, but it is not applicable on unbalanced data (Moder 2010).

Moder (2010) recommends to use a more stringent significance level, e.g., set $\alpha = 0.01$ to keep Type I error rate below 0.05 (Keppel, 1992), but this approach is not appropriate because the power is low when heteroscedasticity is not too extreme.

No investigations have been done to evaluate which methods gain the most power in case of heterogeneity of variance, and Moder (2010) did not consider the situation with non-normal distribution. In this study, we will compare the traditional ANOVA, Welch ANOVA, and Kruskal-Wallis in terms of the type I error rate and power, in multiple situations including homogeneity/ heterogeneity (strong & mild), balanced/unbalanced sample sizes, and normal/non-normal distributions.

1.2. Organization of Thesis

Reminder of the thesis is organized as follows. Chapter II gives a brief introduction to the three methods and describes the details of our simulations. Chapter III summarizes the comparison results. Finally, Chapter IV makes conclusions of the study and provides recommendations for the future research. In addition, the appendix presents the simulation data derivative process, the compare process and the SAS programming.

CHAPTER II

METHODS

2.1. Statistical Methods

The three methods were will be examined in case of three-group heterogeneity of variances: the traditional F-test in Analysis of Variance, the Welch's F-test, the Kruskal-Wallis test.

2.1.1. F-test in Analysis of Variance

An F-test for equality of group means was used through an ANOVA model (Kutner et al 2005) in this study, which can be stated as follows:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

where:

Y_{ij} is the value of response variable in the *j*th observation for the *i*th group

μ_i is the mean for the *i*th group

ε_{ij} are independent $N(0, \sigma^2)$

$i = 1,\dots,r$; $j = 1,\dots, n_i$

In order to analyze the differences among the group means (μ_i) the total variability of the Y_{ij} observations (*SSTO*) is calculated and partitioned into two components: treatment sum of squares (*SSTR*) and error sum of squares (*SSE*).

The sample mean for the *i*th group is denoted by $\bar{Y}_{i.}$: $\bar{Y}_{i.} = \frac{\sum_{j=1}^{n_i} Y_{ij}}{n_i}$

The sample grant mean of all observations is denoted by $Y_{..}$: $Y_{..} = \sum_{i=1}^{r} \sum_{j=1}^{n_i} Y_{ij}$

Thus *SSTO* can be calculated as:

$$SSTO = \sum_i \sum_j (Y_{ij} - \bar{Y}_{..})^2$$

And *SSTO* is partitioned into *SSTR* and *SSE*: $SSTO = SSTR + SSE$

where:

$$SSTR = \sum_i n_i(\bar{Y}_{i.} - \bar{Y}_{..})^2$$

$$SSE = \sum_i \sum_j (Y_{ij} - \bar{Y}_{i.})^2$$

Corresponding to the decomposition of the total sum of squares, we can also obtain a breakdown of the associated degrees of freedom: *SSTO* has $n_t - 1$ degrees of freedom associated with it; SSTR has $r - 1$ degrees of freedom associated with it; SSE has $n_t - r$ degrees of freedom associated with it.

The alternative conclusions considered in F-test are:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$$

$$H_a: \text{not all } \mu_i \text{ are equal}$$

And the test statistic to be used between the alternatives is:

$$F^* = \frac{MSTR}{MSE} = \frac{\frac{SSTR}{\sigma^2}/(r-1)}{\frac{SSE}{\sigma^2}/(n_t - r)} \sim F_{r-1, n_t - r}$$

$F^*$is distributed as $F_{(r-1, n_t-r)}$ when H_0 holds and that large values of F^*lead to conclusion H_a, the appropriate decision rule to control the level of significance at α is:

If $F^* \leq F_{(1-\alpha;\, r-1, n_t-r)}$, conclude H_0

If $F^* > F_{(1-\alpha;\, r-1, n_t-r)}$, conclude H_a

where α is the Nominal Type I error and $F_{(1-\alpha;\, r-1, n_t-r)}$ is the $(1-\alpha)100$ percentile of the appropriate *F* distribution.

2.1.2. Welch's F-test

Welch's F-test (Field 2009) is designed to test the equality of group means when we have more than two groups to compare, especially in the cases which didn't meet the homogeneity of variance assumption and sample sizes are small. However, the assumptions of normality and independency are remained.

The main idea of Welch's F-test is using a weight w_i to reduce the effect of heterogeneity. Here w_i is based on the sample size (n_i) and the observed variance (s_i^2) for the ith group:

$$w_i = \frac{n_i}{s_i^2}$$

Then in Welch's F-test, the adjusted grand mean ($\bar{Y}_{welch}$) can be calculated based on a weighted mean for each group:

$$\bar{Y}_{welch} = \frac{\sum_{i=1}^{r} w_i \bar{Y}_{i.}}{\sum_{i=1}^{r} w_i}$$ where $\bar{Y}_{i.}$ is the sample mean for *i*th group; i=1,…,r

Treatment sum of squares ($SSTR_{welch}$) and treatment mean squares ($MSTR_{welch}$) are:

$$SSTR_{welch} = \sum_{i=1}^{r} w_i (\bar{Y}_{i.} - \bar{Y}_{welch})^2$$

$$MSTR_{welch} = \frac{SSTR_{welch}}{r-1}$$

Then we need to calculate a term called lambda (Λ), which is based again on weights:

$$\Lambda = \frac{3\sum_{i=1}^{r} \frac{\left(1 - \frac{w_i}{\sum_{i=1}^{r} w_i}\right)^2}{n_i - 1}}{r^2 - 1}$$

The test statistic to be used between the alternatives is:

$$F_{welch}^{*} = \frac{SSTR_{welch}/(r-1)}{1 + \frac{2\Lambda(r-2)}{3}} \sim F_{r-1,1/\Lambda}$$

The alternatives conclusions considered in Welch's F-test are same as the traditional F-test:

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_r$$

$$H_a: \text{not all } \mu_i \text{ are equal}$$

The decision rule to control the level of significance at α is:

$$\text{If } F^* \leq F_{(1-\alpha;\, r-1, 1/\Lambda)}, \text{ conclude } H_0$$

$$\text{If } F^* > F_{(1-\alpha;\, r-1, 1/\Lambda)}, \text{ conclude } H_a$$

Although Welch's F-test is an adaptation of the F-test and supposed to be more reliable when the assumption of homogeneity of variance was not met, the disadvantage is it has fewer degrees of freedom than the F-test ($1/\Lambda \leq n_t - r$). Thus Welch's F-test is less powerful than the F-test. And since the weight factor is highly related to the sample sizes and the observed variances ($w_i = \frac{n_i}{s_i^2}$), when the number of observations is small, the results of the Welch's F-test may be quite unstable.

2.1.3. Kruskal-Wallis test

The Kruskal–Wallis test (Kruskal; Wallis 1952) is a non-parametric method for testing whether samples originate from the same distribution. Since it is a non-parametric test, it does not assume that the response variable is normally distributed. And like most non-parametric tests, it perform the test on ranked data. It is commonly used when we don't have a large sample size and can't clearly demonstrate that our data are normally distributed.

To do this test, first we rank all data (Y_{ij}) from 1 to n_t ignoring group membership, assign any tied values the average of the ranks they would have received had they not been tied. Denote R_{ij} as the rank of observation j from group i, considered as a new response variable.

The alternatives conclusions considered in Kruskal–Wallis test are:

H_0: all mean ranks of the groups are equal

H_a: not all mean ranks of the groups are equal

The test statistic to be used between the alternatives is:

$$K = (n_t - 1)\frac{\sum_{i=1}^{r} n_i(\bar{R}_{i.} - \bar{R})^2}{\sum_{i=1}^{r}\sum_{j=1}^{n_i}(R_{ij} - \bar{R})^2} \sim F_{r-1,n_t-r}$$

Where:

r is the total number of groups

n_t is the total number of observations across all groups

n_i is the number of observations in group i

R_{ij} is the rank (among all observations) of observation j from group i

$\bar{R}_{i.} = \frac{\sum_{j=1}^{n_i} R_{ij}}{n_i}$ is the average rank in ith group

$\bar{R} = \frac{1}{2}(n_t + 1)$ is the average rank of all observations

The decision rule to control the level of significance at α is:

If $K \leq F_{(1-\alpha;\, r-1,n_t-r)}$, conclude H_0

If $K > F_{(1-\alpha;\, r-1,n_t-r)}$, conclude H_a

The advantage to using the Kruskal–Wallis test in heterogeneity of variance cases is that the ranked data can clearly meet the normality assumption of one-way analysis of variance (McDonaldc et al, 2014). However, you lose information when you substitute ranks for the original values, which can make this a somewhat less powerful test than a one-way ANOVA. Moreover, some researchers hold the opinion that even though ANOVA need the normality assumption, it is not very sensitive to deviations from normality (McDonaldc et al, 2014) and in some non-normal distributions, the performance of ANOVA is better than Kruskal–Wallis test.

2.2. Simulation methods

Monte Carlo simulation studies are performed using SAS 9.4. 10,000 datasets are generated for each simulation scenario including balanced/unbalanced sample sizes, equal/unequal means, homogeneity/ heterogeneity of variance, normal/non-normal data sets. Each statistical method is assessed to compare three group means in a simulated dataset. We focus on three groups because it is quite common and can be found in a lot of trials.

2.2.1. Balanced/Unbalanced

A balanced design means all groups have the same sample size, while an unbalanced design represents unequal sample size among groups. In our balanced simulation studies, the sample sizes are 5, 10, 20, or 40 per group, respectively. For example $n_1 = n_2 = n_3 = 10$ or $n_1 = n_2 = n_3 = 20$. In our unbalanced studies, the total sample size is 60 but sample size in each group is different. For example $n_1 = 10, n_2 = 20,$ and $n_3 = 30$ or $n_1 = 40, n_2 = 10,$ and $n_3 = 10$.

2.2.2. Equal/Unequal Means

In our simulated data sets, group means are varied between 0 and 2. And for equal means studies, we set the group means equal (e.g., $\mu_1 = 0, \mu_2 = 0, \mu_3 = 0$), and in unequal means studies, at least two group means are unequal (e.g., $\mu_1 = 0, \mu_2 = 0, \mu_3 = 2$). Equal-means studies are used to compare the three statistical methods in terms of empirical type I error rate and unequal-means studies are used to compare power of detecting mean differences among groups.

2.2.3. Homogeneity/ Heterogeneity of Variance

We generated homogeneous data (where group variances are equal) as well as heterogeneous data (where at least two group variances are unequal). The values of standard deviation are varied between 1 and 4. For example in a homogeneity design we set the equal standard deviation for each group: $\sigma_1 = \sigma_2 = \sigma_3 = 1$; in a heterogeneity design, at least two standard deviations are not equal: for example $\sigma_1 = 1, \sigma_2 = 1, \sigma_3 = 4$.

2.2.4. Normal/Non-normal

In this simulation log-normally distributed data are generated and considered as non-normal data. We set parameters a and b to generated the log-normal data Y as:

$$Y = e^{a+b*Z} \quad \text{where: } Z \sim N(0,1)$$

Thus the mean value of log-normal data Y is: $E(Y) = e^{a+\frac{b^2}{2}}$

And its variance is: $Var(Y) = \left(e^{b^2} - 1\right) * e^{2a+b^2}$

In this study, for empirical type I error rate comparison in log-normal cases, and we set the mean of log-normal data $E(Y) = 1$, the scale parameter b varies between 0.25, 0.5 and 0.75. Thus the standard deviations ($\sigma(Y) = \sqrt{Var(Y)}$) varies between 0.254, 0.533 and 0.869.

2.3. Comparison methods

In each simulation scenario, a test is considered to be significant when a p-value is less than the nominal α=0.05. The number of significant tests will be counted in 10,000 simulated datasets in a scenario and the rejection rate will be calculated.

The comparison procedures will be divided into two parts:

a. In the situation that the null hypothesis (H_0: $\mu_1 = \mu_2 = \mu_3$) is true, the rejection rate of the null hypothesis will be considered as the empirical type I error rate for each test method. The method that have the closest empirical type I error rate to the nominal α=0.05is considered as the best among these three methods.

b. In the situation that the alternative hypothesis (Ha: not all μ_i are equal) was true, the rejection rate of the null hypothesis will be considered as the power for each test method. The method that have largest power is considered as the best among these three methods.

CHAPTER III

RESULTS

3.1. Results for Normal Balanced Heterogeneity Data

3.1.1. Type I Error Analysis for Normal Balanced Heterogeneity Data

Table 1 and Figure 1 show the empirical type I error rates of the F-test (by a traditional ANOVA), Welch's F and Kruskal-Wallis test for a normal balanced homogeneous variance situation. Data are normally distributed: $Y \sim N(\mu_i, {\sigma_i}^2)$. Equal sample sizes are 5, 10, 20, and 40. Equal means are $\mu_1 = \mu_2 = \mu_3 = 0$. Standard deviations $(\sigma_1 : \sigma_2 : \sigma_3)$ was given by $1:1:1$. The nominal α level is 0.05. (The results of Table 1 are visually presented in Figure 1 and thus Table 1 is moved to the end of the References. The same arrangement is applied to the other tables.) Blue lines represent the results of F-test, red lines for Welch's F-test and green lines for Kruskal-Wallis test.

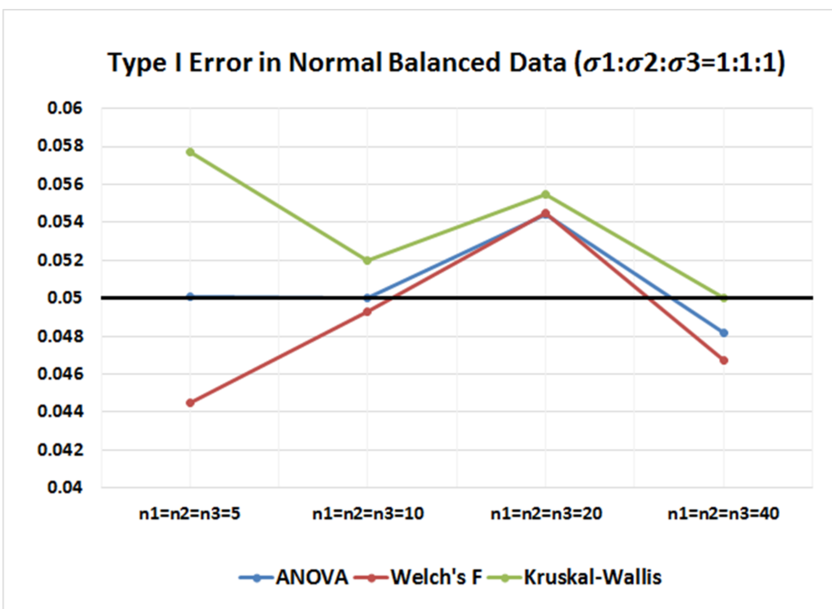

Figure 1.

Empirical type I error rate of equal sample sizes 5, 10, 20 and 40, Equal means $\mu_1 = \mu_2 = \mu_3 = 0$, ratio of $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1$, nominal α of 0.05.

From Table 1 and Figure 1, we can see that in this normal balanced homogeneity case ($\mu_1 = \mu_2 = \mu_3 = 0, \sigma_1 : \sigma_2 : \sigma_3 = 1:1:1$), when sample sizes are small, the calculated type I error rates of the three methods are relatively different: the ANOVA F-test has the closest type I error rate to the nominal α of 0.05. When the sample size gets larger, the three methods have similar type I error rates and all control the nominal α value well.

Similar to the homogeneity case in Figure 1, Table 2 and Figure 2 show the results in the empirical type I error rates of three methods in nine different heterogeneity scenarios, plus the homogeneity scenario shown in Figure 1. The group means are all the same ($\mu_1 = \mu_2 = \mu_3 = 0$), and the standard deviation are varied between 1 and 4 ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1 \sim 1:4:4$). The simulation settings (both sample sizes and standard deviation values) for each scenario is presented at the bottom of the figure.

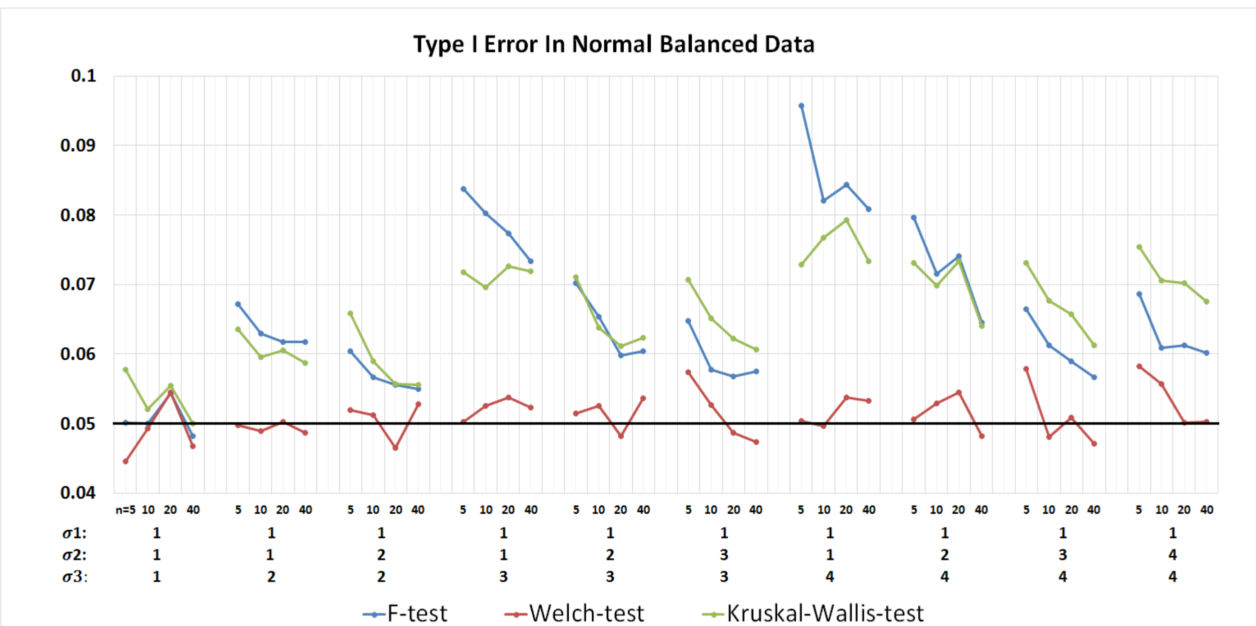

Figure 2.
Empirical type I error rate of equal sample sizes 5, 10, 20 and 40,

Equal means $\mu_1 = \mu_2 = \mu_3 = 0$, ratios of $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1 \sim 1:4:4$, nominal α of 0.05.

In Figure 2, unlike the homogeneity set ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1$), the performances of the three statistic methods are various. In every heterogeneity simulation set, the empirical type I error rate of Welch's F-test are all around 0.05 (red lines in the figure are all close the black horizontal line) and performs the best in normal balanced heterogeneity data. However, the type I error rates of the F-test and Kruskal-Wallis test are all greater than 0.05 and exceed the nominal type I error by up to 4.57%. As the sample size gets larger, the empirical error rates are closer to 0.05.

In the mild heterogeneity case ($\sigma_1 : \sigma_2 : \sigma_3 = 1:2:2$), when the sample size is 40 per group, the type I error inflation problem by both F-test and Kruskal-Wallis test is mild (around 0.0580). But when the heterogeneity of variances gets stronger, the inflation problem by the two methods gets much worse. For example, the type error rates by the two methods are inflated to 0.0809 and 0.0734, respectively, in the 8th setting ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:4$, $n_1 : n_2 : n_3 = 40$).

The results of the 4th and 6th settings ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:3$ and $1:3:3$) and results of 7th and 10th settings ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:4$ and $1:4:4$) indicate that F-test has the very strong inflation problem when one group has relatively bigger variance and the other two smaller variances ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:3 \ \ and \ 1:1:4$). However when one group has smaller variance and the other two larger variances ($\sigma_1 : \sigma_2 : \sigma_3 = 1:3:3$ and $1:4:4$), the performance of F-test is much better than in the cases with ($\sigma_1 : \sigma_2 : \sigma_3 = 1:1:3 \ \ and \ 1:1:4$). These results are consistent with the opinion by Underwood (1997), who concluded that analysis of variance has problems with heterogeneity in balanced samples only when the variance of one group is markedly larger than the others.

3.1.2. Power Analysis for Normal Balanced Heterogeneity Data

From the type I error analysis we can see that when there is mild heterogeneity of variances ($\sigma_1: \sigma_2: \sigma_3 = 1:1:2$ and $1:2:2$) there is relatively mild inflation of type I error for the three methods. In these cases, we can't decide which method is the best only depend on the results of type I error analysis. In this situation, a statistical method that has the largest power will be considered as the best method. Therefore we conduct the power analysis in these two cases with mild heterogeneity of variances to compare the power of these three methods when they have relatively similar type I error rates.

In the simulation settings that the alternative hypothesis (Ha: not all μ_i are equal) was true, we calculated the rejection rates of F-test, Welch's F and Kruskal-Wallis test, respectively. The rejection rates for each test were calculated in 10,000 simulated datasets, and these rejection rates are considered as power of these tests. The larger the power is, the better a method performs.

Table 3 and Figure 3 showed the calculated powers of F-test, Welch's F and Kruskal-Wallis test in a normal balanced situation with unequal group means, equal sample sizes are 5, 10, 20, and 40 and$\mu_1: \mu_2: \mu_3 = 0:0:1 \sim 0:2:2$. Standard deviations ($\sigma_1: \sigma_2: \sigma_3$) were given as $\sigma_1: \sigma_2: \sigma_3 = 1:1:2$, nominal α of 0.05.

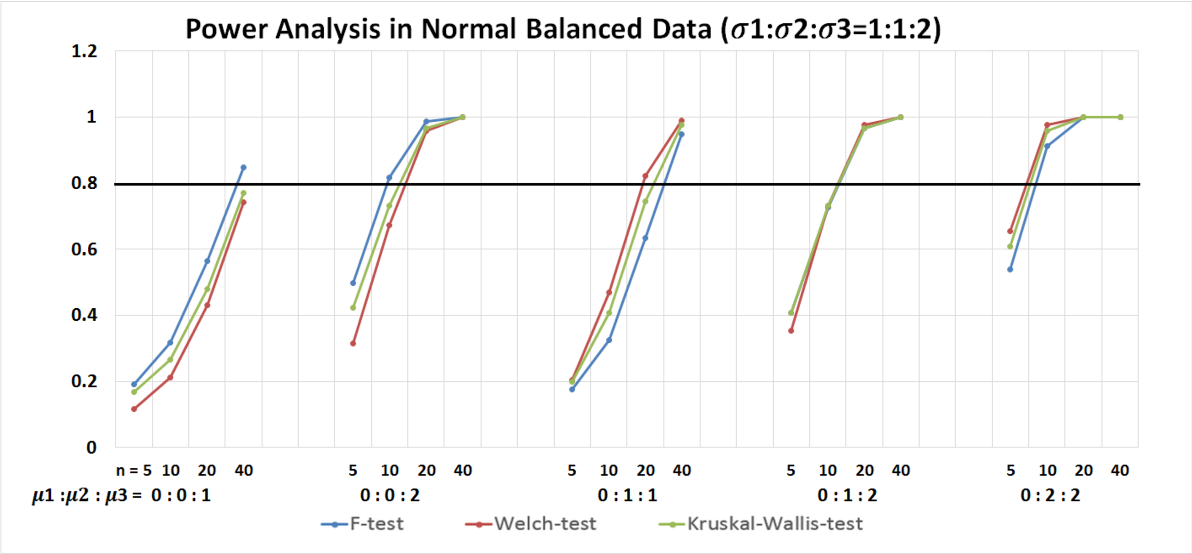

Figure 3. Power of equal sample sizes 5, 10, 20 and 40, unequal means $\mu_1 : \mu_2 : \mu_3 = 0 : 0 : 1 \sim 0 : 2 : 2$, Ratio of $\sigma_1 : \sigma_2 : \sigma_3 = 1 : 1 : 2$.

Table 4 and Figure 4 showed the power of F-test, Welch's F and Kruskal-Wallis test in a normal balanced situation with unequal group means: equal sample sizes were 5, 10, 20, and 40, unequal means are $\mu_1 : \mu_2 : \mu_3 = 0 : 0 : 1 \sim 0 : 2 : 2$. Standard deviations $(\sigma_1 : \sigma_2 : \sigma_3)$ were given as $\sigma_1 : \sigma_2 : \sigma_3 = 1 : 2 : 2$, nominal α of 0.05.

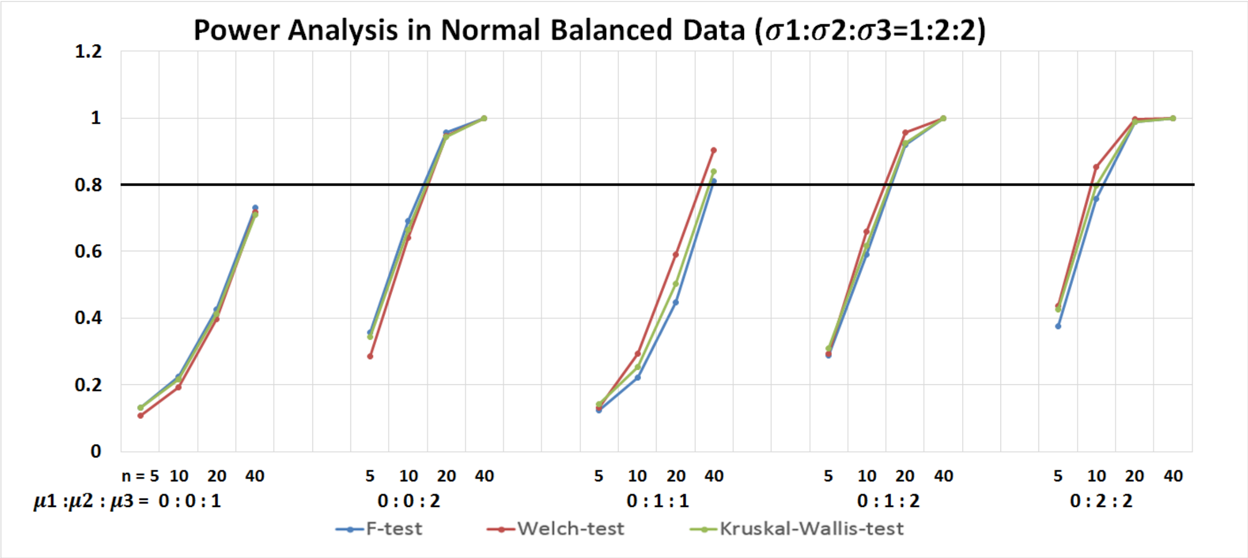

Figure 4.
Power of equal sample sizes 5, 10, 20 and 40, unequal means $\mu_1 : \mu_2 : \mu_3 = 0 : 0 : 1 \sim 0 : 2 : 2$, Ratio of $\sigma_1 : \sigma_2 : \sigma_3 = 1 : 2 : 2$.

Figures 3 and 4 showed us in some settings, Welch's test is the best, but in some other cases it's not. We can try to find the reason in the mathematical analysis.

In ANOVA, the F ratio is calculated by:

$$F^* = \frac{MSTR}{MSE} = \frac{\frac{SSTR}{\sigma^2} / (r-1)}{\frac{SSE}{\sigma^2} / (n_t - r)}$$

In the power analysis above, we set group number $r = 3$, and the sample sizes for each group are the same $n_1=n_2=n_3=n$. That means, if every simulation is followed the distributions we set, the F ratio will be:

$$F^* = \frac{[(\mu_1 - \bar{\mu})^2 + (\mu_2 - \bar{\mu})^2 + (\mu_3 - \bar{\mu})^2]/2}{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)/(3n-3)} \propto \frac{Var(\mu_i)}{(\sigma_1^2 + \sigma_2^2 + \sigma_3^2)} * (n-1)$$

The F ratio will only related to the variance of means ($Var(\mu_i)$), sum of variances ($\sigma_1^2 + \sigma_2^2 + \sigma_3^2$) and sample sizes ($n - 1$). In a particular setting (e.g.$\sigma_1 : \sigma_2 : \sigma_3 = 1:1:2, \mathrm{n1} = \mathrm{n2} = \mathrm{n3} = 5, 10, 20, 40$), the sum of variances and sample sizes are fixed, thus the F ratio is only depend of the variance of means. From the results of the setting $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:2,\ \mu_1 : \mu_2 : \mu_3 = 0:0:1$ ($Var(\mu_i) = 0.33$) and $\mu_1 : \mu_2 : \mu_3 = 0:0:2$($Var(\mu_i) = 1.33$) we can see in the higher variance of means settings, the results of ANOVA are better than the lower variance of means settings.

In Welch's F-test, since the weight factor ($w_i = \frac{n_i}{s_i^2}$) is used, when sample sizes are the same, the big variance group will have smaller weight. That means in the calculation of variance of means, the model in Welch's test should focus more on the means in small variance groups. The results in the setting $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:2,\ \mu_1 : \mu_2 : \mu_3 = 0:0:1$ ($Var(\mu_i) = 0.33$) and $\mu_1 : \mu_2 : \mu_3 = 0:1:1$($Var(\mu_i) = 0.33$) confirmed this assumption. Since the variance of means in these two settings are the same, the results of traditional ANOVA are alike, however, the variance of means in small variance groups are not the same, in the setting $\mu_1 : \mu_2 : \mu_3 = 0 : 1 : 1$Welch's F-test works much better than in the setting $\mu_1 : \mu_2 : \mu_3 = 0 : 0 : 1$.

From Figure 3 and 4, all the results of Kruskal-Wallis test are all between the traditional ANOVA and Welch's F-test. And Figures 3 and 4 also showed us that in normal balanced unequal heterogeneity data, the sample size plays the most important role on the outcome of power. The power of these three tests increase rapidly when the sample size gets larger.

In the power analysis, we know that the performances of these three methods are depend on the setting of group means. We should choose appropriate statistical method in particular situation.

3.2. Results for Non-normal (log-normal) Balanced Heterogeneity Data

Table 5 and Figure 5 showed the empirical type I error rates of F-test, Welch's F and Kruskal-Wallis test for a log-normal balanced situation. Data are lognormal distributed: $Y \sim \ln N(\mu_i, \sigma_i)$. Equal sample sizes are 5, 10, 20, and 40, respectively. Equal means are $\mathrm{E}(\overline{Y_1}) = \mathrm{E}(\overline{Y_2}) = \mathrm{E}(\overline{Y_3}) = 1$. Scale parameters (b_1, b_2, b_3) varies between 0.25, 0.5 and 0.75, presenting mild skewness (skewness is calculated by $(e^{b^2} + 2)\sqrt{e^{b^2} - 1}$)), and the standard deviations ($\sigma(Y) = \sqrt{Var(Y)}$) varies between 0.254, 0.533 and 0.869. The scale parameter b decides the shape of the distribution. When b is small like 0.25 and 0.5, the distribution of log-normal is very similar to normal distribution. In this study, we only consider small b cases because when the b parameter is big, we can easily know the data are not normal distributed and it doesn't make sense to use ANOVA on non-normal data.

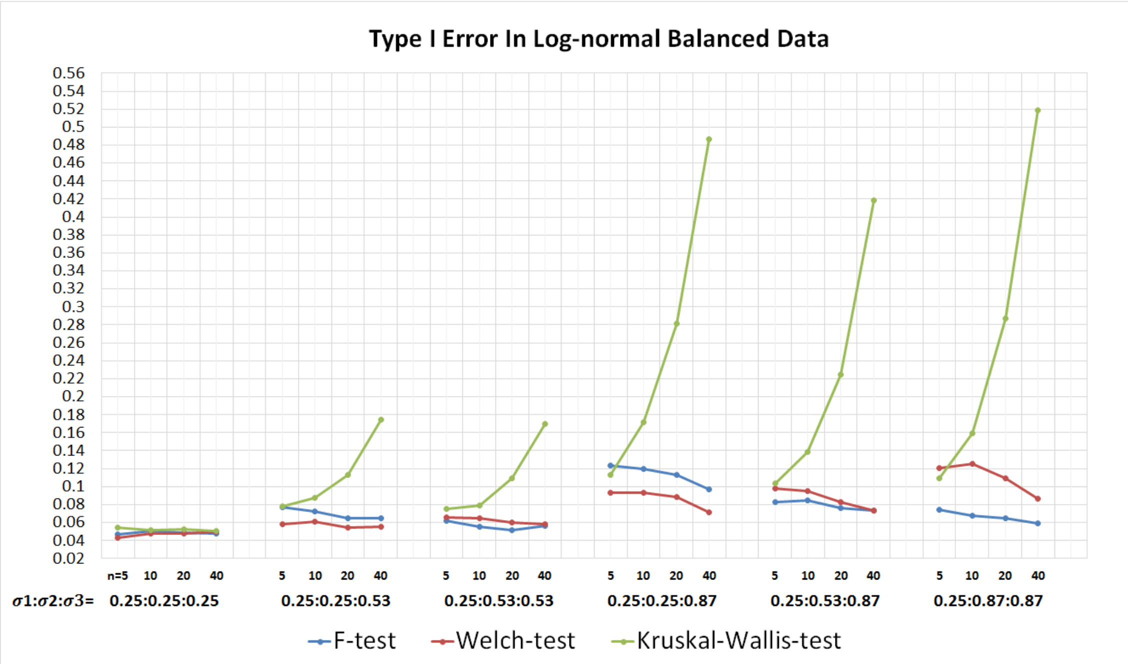

Figure 5 (1).

Empirical type I error rates of log-normal data, equal sample sizes 5, 10, 20 and 40, equal means $\mathrm{E}(\overline{Y_1}) = \mathrm{E}(\overline{Y_2}) = \mathrm{E}(\overline{Y_3}) = 1$, $\sigma(Y_1){:}\,\sigma(Y_2){:}\,\sigma(Y_3) = 0.25{:}\,0.25{:}\,0.25 \sim 0.25{:}\,0.87{:}\,0.87$ Ratios of scale parameter $b_1{:}\,b_2{:}\,b_3 = 0.25{:}\,0.25{:}\,0.25 \sim 0.25{:}\,0.75{:}\,0.75$, nominal α of 0.05.

Figure 5 (2).

Empirical type I error rates of log-normal data, equal sample sizes 5, 10, 20 and 40, equal means $\mathrm{E}(\overline{Y_1}) = \mathrm{E}(\overline{Y_2}) = \mathrm{E}(\overline{Y_3}) = 1$, $\sigma(Y_1){:}\,\sigma(Y_2){:}\,\sigma(Y_3) = 0.25{:}\,0.25{:}\,0.25 \sim 0.25{:}\,0.87{:}\,0.87$ Ratios of scale parameter $b_1{:}\,b_2{:}\,b_3 = 0.25{:}\,0.25{:}\,0.25 \sim 0.25{:}\,0.75{:}\,0.75$, nominal α of 0.05.

In Figure 5(1), the results of all these heterogeneity simulation settings are different with the homogeneity set (the first setting in Figure 5(1), $\sigma(Y_1){:}\,\sigma(Y_2){:}\,\sigma(Y_3) = 0.25{:}\,0.25{:}\,0.25$) . Unlike the homogeneity set, the results of these three statistical methods showed us they are more or less inappropriate in log-normal heterogeneity cases. In all heterogeneity simulation sets, the Kruskal-Wallis test has the type I error inflation problem and this inflation gets larger as the sample size increases.

Figure 5 (2), since Kruskal-Wallis test has been showed very inappropriate in this log-normal heterogeneity case, we exclude the results of Kruskal-Wallis test and only show the results of F-test and Welch's F-test in much smaller scales than in Figure 5 (1). In these heterogeneity cases, the empirical type I error rate of Welch's F-test are all above 0.05 even in mild heterogeneity cases (e.g., in case scale parameter $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.25 : 0.5$ and $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.5 : 0.5$). And when the heterogeneity get more and more severe, the results get worse. Thus Welch's F-test is not suitable for severe log-normal balanced heterogeneity data.

The results of F-test in log-normal settings are similar to those in the normal settings in section 3.1.1., when mild heterogeneity happens (scale parameter $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.5 : 0.5$) its results are still acceptable (The type I error of F-test is 0.0550). From the results of 2^{nd} and 3^{rd} settings (scale parameter $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.25 : 0.5$ and $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.5 : 0.5$) and results of 4^{th} and 6^{th} settings (scale parameter $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.25 : 0.75$ and $\sigma_1 : \sigma_2 : \sigma_3 = 0.25 : 0.75 : 0.75$), we can know F-test presents the inflation problems when there is a big variance group in the data. When there is a small variance group in the data, the performance of F-test is better. In all these 5 heterogeneity settings, the empirical type I error rate of the F-test are all above 0.05.

Thus all these three methods are not suitable for log-normal heterogeneity cases.

3.3. Results for Normal Unbalanced Heterogeneity Data

3.3.1. Type I Error Analysis

Table 6 and Figure 6 showed the empirical type I error rates of F-test, Welch's F and Kruskal-Wallis test for 9 sets of normal unbalanced situation along with a set of normal balanced situation. Data are normally distributed: $Y \sim N(\mu_i, \sigma_i)$.For these three groups in the analysis, the total sample size is 60 and the sample size for each group varies from 10 to 40. Equal means are $\mu_1 = \mu_2 = \mu_3 = 0$. Standard deviations $(\sigma_1 : \sigma_2 : \sigma_3)$ was given between 1 and 4 $(\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1 \sim 1:4:4)$. The nominal α level is 0.05.

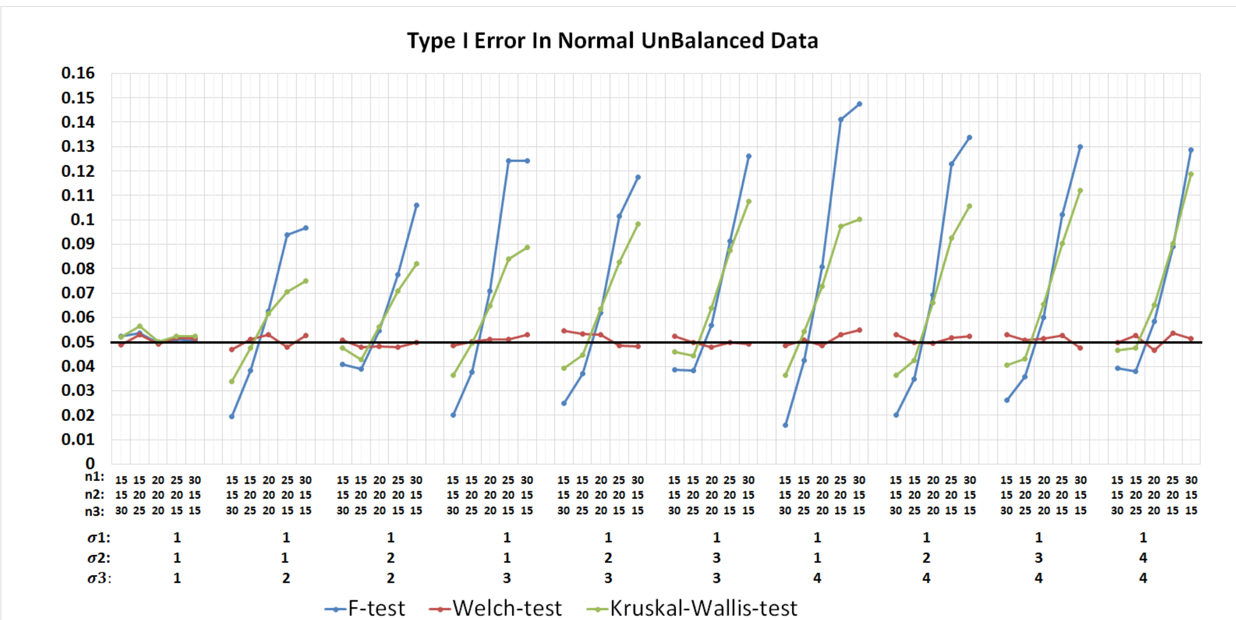

Figure 6
Empirical type I error rate of unequal sample sizes, equal means $\mu_1 = \mu_2 = \mu_3 = 0$,
Ratios of $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1 \sim 1:4:4$, nominal α of 0.05.

Table 6 and Figure 6 reflect the situations that in normal unbalanced heterogeneity data, the empirical type I error rate depends to a high degree on how sample sizes and variances are combined. The result of 7th setting $(\sigma_1 : \sigma_2 : \sigma_3 = 1:1:4, \mathrm{n1:n2:n3} = 30:15:15)$ shows when a small number of observations comes with a high variance, then the type I error rates of all three methods exceed the nominal α level, the F-test has the worst type I error rate (0.1476), the Kruskal-Wallis test has the second worse type I error rate (0.1002), however the result of the Welch's F test doesn't exceed nominalα level too much (0.0547). And compared to the results of 2nd and 7th settings, we can see when this "high variance" get larger, the results of F-test and Kruskal-Wallis test will get worse.

If small sample sizes comes with small standard deviations (e.g. $\sigma_1 : \sigma_2 : \sigma_3 = 1 : 1 : 4, \mathrm{n1} : \mathrm{n2} : \mathrm{n3} = 15 : 15 : 30$), then Welch's F-test works well and the type I error rate is around the nominal α level. However the other two methods are too conservative, their type I error rates are much smaller than the nominal α level (F-test 0.0160, Kruskal-Wallis test 0.0363).

In other situations where the highest variance is not bound to high or low sample sizes, F-test and Kruskal-Wallis test are more or less inappropriate. After comparing these results, we can say among these three methods Welch's F-test is most appropriate in normal unbalanced heterogeneity data.

CHAPTER IV

DISCUSSION

4.1. Conclusion and Implications

After doing these comparisons, now we can go back to the questions that motivated this study. The first question is under what conditions is heterogeneity of variances really a problem in ANOVA. We can state that heterogeneity of variances is always a problem in ANOVA even in the moderate heterogeneity cases, i.e., when one group variance is smaller than the others. And the problem is worse when the heterogeneity get more and more extreme, i.e., when one group variance is very larger than the others. Thus we can know one effect of heterogeneity on ANOVA is the inflation of the type I error rate.

We arrived at several conclusions from the present analyses to answer the second question - which method should be chosen when there is heterogeneity of variance in three-group cases:

(1) F-test in analysis of variance is unsuitable in three-group heterogeneity cases, even though it is a robust test when data are homogeneous, normal and equal/unequal sample sizes.

(2) Welch's test performs the best in three-group heterogeneity cases when data are normal and equal and unequal sample sizes. And this result is consistent with the results from Moder (2010) that Welch's F-test is useful for a small number of group cases.

(3) Kuskal-Wallis test is acceptable in mild heterogeneity cases ($\sigma_1 : \sigma_2 : \sigma_3 = 1 : 1 : 2, \sigma_1 : \sigma_2 : \sigma_3 = 1 : 2 : 2$) when data are normal and equal sample sizes.

To summarize, no one of these three methods can be suitable in every heterogeneity case, but if we have doubts about the homogeneity of variances, Welch's test is recommended for normal and balanced/unbalanced designs. For a specific situation, we should use the method which works best to do the analysis.

4.2. Limitation and Future Work

Although the present study and analyses used many situations of heterogeneity data, we only considered the 3 group situations. In the analysis of non-normal data, we only used log-normal data, thus the conclusion we got from log-normal data may not suitable for all non-normal cases. Only three analysis methods were examined.

The recommendations we have as the future work of this research are listed as follows:

(1) We need to expand to comparing more than 3 groups and try other kinds of non-normal data.

(2) More statistical methods are needed to be examined in this study, like Hotelling's T^2 test.

(3) We can try to reduce the variance by transform the original data to relieve the effect of heterogeneity.

(4) Prior to the analysis of variance, the test of homogeneity of variance (e.g., Barlett's test) should be used to assess the homogeneity of within-group variances.

(5) Real world data may be used in this study to see which way to deal with the heterogeneity is most applicable.

REFERENCE

[1]. Kutner, M.H., Nachtsheim, C.J., Neter, J.,Li, W., et al. (2005). Applied Linear Statistical Models, 5th edition. *Analysis of Variance*, 690-699.

[2]. 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.

[3]. Field, A.P.(2009). Discovering statistics using SPSS, 3rd edition. London: Sage. *Additional material*.

[4]. Moder, K. (2010). Alternatives to F-Test in One Way ANOVA in case of heterogeneity ofvariances (a simulation study). *Psychological Test and Assessment Modeling,* Volume 52, 2010 (4), 343-353

[5]. Kruskal, Wallis (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47 (260): 583–621.

[6]. McDonald, J.H. (2014). Handbook of Biological Statistics 3rd edition. *Sparky House Publishing, Baltimore, Maryland*, 157-164.

[7]. Underwood, A. J. (1997). *Experiments in ecology – Their logical design and interpretation using analysis of variance*. Cambridge: Cambridge University Press.

[8]. “Log-normal distribution.” Wikipedia: The Free Encyclopedia. Wikimedia Foundation, Inc, 27 June 2015. Web.

[9]. Legendre, P., Borcard, D. Statistical comparison of univariate tests of homogeneity of variances. *Journal of Statistical Computation and Simulation*.

[10]. Wuensch, K. (2009). Simulating Data for a Three-Way Independent Samples ANOVA. 8 May 2015. Web.

APPENDIX A

RESULT TABLES

Table 1:

Empirical type I error rate of equal sample sizes 5, 10, 20 and 40, equal means $\mu_1 = \mu_2 = \mu_3 = 0$, ratio of $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1$, nominal α of 0.05. Best values in bold.

n1=n2=n3	F-test	Welch's F	Kruskal-Wallis
5	**0.0501**	0.0445	0.0577
10	**0.0500**	0.0493	0.052
20	**0.0544**	0.0545	0.0555
40	0.0482	0.0467	**0.0500**

Table 2:

Empirical type I error rate of equal sample sizes 5, 10, 20 and 40, equal means $\mu_1 = \mu_2 = \mu_3 = 0$, ratios of $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:2 \sim 1:4:4$, nominal α of 0.05. Best values in bold.

$\sigma_1 : \sigma_2 : \sigma_3$	n1=n2=n3	F-test	Welch's F	Kruskal-Wallis
1:1:2	5	0.0672	**0.0497**	0.0636
	10	0.0629	**0.0489**	0.0595
	20	0.0617	**0.0503**	0.0605
	40	0.0617	**0.0487**	0.0587
1:2:2	5	0.0604	**0.0519**	0.0659
	10	0.0566	**0.0512**	0.0589
	20	0.0556	**0.0465**	0.0557
	40	0.055	**0.0528**	0.0556
1:1:3	5	0.0837	**0.0502**	0.0718
	10	0.0803	**0.0526**	0.0696
	20	0.0774	**0.0538**	0.0726
	40	0.0734	**0.0523**	0.0719
1:2:3	5	0.0702	**0.0515**	0.0711
	10	0.0654	**0.0525**	0.0638
	20	0.0598	**0.0482**	0.0611
	40	0.0604	**0.0536**	0.0624
1:3:3	5	0.0647	**0.0574**	0.0707
	10	0.0577	**0.0527**	0.0651
	20	0.0568	**0.0487**	0.0622
	40	0.0575	**0.0473**	0.0606
1:1:4	5	0.0957	**0.0504**	0.0729
	10	0.082	**0.0496**	0.0767
	20	0.0843	**0.0537**	0.0793
	40	0.0809	**0.0533**	0.0734

1:2:4	5	0.0796	**0.0506**	0.0731
	10	0.0715	**0.0529**	0.0698
	20	0.0741	**0.0545**	0.0733
	40	0.0645	**0.0482**	0.064
1:3:4	5	0.0665	**0.0579**	0.0731
	10	0.0613	**0.0481**	0.0677
	20	0.0589	**0.0509**	0.0657
	40	0.0567	**0.0471**	0.0612
1:4:4	5	0.0686	**0.0582**	0.0754
	10	0.0609	**0.0557**	0.0706
	20	0.0613	**0.0501**	0.0702
	40	0.0602	**0.0502**	0.0675

Table 3:

Power of equal sample sizes 5, 10, 20 and 40, unequal means $\mu_1:\mu_2:\mu_3 = 0:0:1 \sim 0:2:2$, ratio of $\sigma_1:\sigma_2:\sigma_3 = 1:1:2$. Best values in bold.

$\mu_1:\mu_2:\mu_3$	n1=n2=n3	F-test	Welch's F	Kruskal-Wallis
0:0:1	5	**0.1923**	0.1175	0.1672
	10	**0.3174**	0.2124	0.2653
	20	**0.564**	0.4311	0.4788
	40	**0.8467**	0.7429	0.7695
0:0:2	5	**0.4978**	0.3153	0.4220
	10	**0.8161**	0.6715	0.7315
	20	**0.9856**	0.9592	0.9654
	40	**1.0000**	**1.0000**	**1.0000**
0:1:1	5	0.1758	**0.2034**	0.1983
	10	0.3240	**0.4697**	0.4083
	20	0.6342	**0.8217**	0.7449
	40	0.9469	**0.9891**	0.9764
0:1:2	5	**0.4085**	0.3532	0.4076
	10	0.7275	**0.7323**	0.7315
	20	0.9662	**0.9765**	0.9656
	40	0.9998	0.9999	**1.0000**
0:2:2	5	0.5389	**0.6537**	0.6076
	10	0.9113	**0.9771**	0.9579
	20	0.9997	**1.0000**	**1.0000**
	40	**1.0000**	**1.0000**	**1.0000**

Table 4:

Power of equal sample sizes 5, 10, 20 and 40, unequal means $\mu_1 : \mu_2 : \mu_3 = 0:0:1 \sim 0:2:2$, ratio of $\sigma_1 : \sigma_2 : \sigma_3 = 1:2:2$. Best values in bold.

$\mu_1 : \mu_2 : \mu_3$	n1=n2=n3	F-test	Welch's F	Kruskal-Wallis
0:0:1	5	0.1318	0.1067	**0.1325**
	10	**0.2248**	0.1938	0.2155
	20	**0.4262**	0.3962	0.4120
	40	**0.7305**	0.7174	0.7094
0:0:2	5	**0.3582**	0.2864	0.3440
	10	**0.6912**	0.6420	0.6648
	20	**0.9575**	0.9456	0.9434
	40	**0.9997**	**0.9997**	0.9996
0:1:1	5	0.1244	0.1326	**0.1412**
	10	0.2218	**0.2929**	0.2546
	20	0.4477	**0.5901**	0.5030
	40	0.8101	**0.9026**	0.8405
0:1:2	5	0.2876	0.2938	**0.3094**
	10	0.5896	**0.6582**	0.6156
	20	0.9201	**0.9573**	0.9242
	40	0.9985	**0.9994**	0.9982
0:2:2	5	0.3748	**0.4357**	0.4256
	10	0.7586	**0.8527**	0.7961
	20	0.9874	**0.9962**	0.9895
	40	0.9998	0.9998	**1.0000**

Table 5:

Empirical type I error rate of log-normal data, equal sample sizes 5, 10, 20 and 40, Equal means $\mathrm{E}(\bar{Y_1}) = \mathrm{E}(\bar{Y_2}) = \mathrm{E}(\bar{Y_3}) = 1$, $\sigma(Y_1){:}\,\sigma(Y_2){:}\,\sigma(Y_3) = 0.25{:}\,0.25{:}\,0.25{\sim}0.25{:}\,0.87{:}\,0.87$ Ratios of scale parameter $b_1{:}\,b_2{:}\,b_3 = 0.25{:}\,0.25{:}\,0.25{\sim}0.25{:}\,0.75{:}\,0.75$, nominal α of 0.05. Best values in bold.

$\sigma_1{:}\sigma_2{:}\sigma_3$	n1=n2=n3	F-test	Welch's F	Kruskal-Wallis
1/4:1/4:1/4	5	**0.0472**	0.043	0.054
	10	**0.0503**	0.0476	0.0519
	20	**0.0487**	0.048	0.0523
	40	0.0482	**0.0496**	0.0506
1/4:1/4:1/2	5	0.0774	**0.0583**	0.0778
	10	0.0721	**0.0608**	0.0877
	20	0.0649	**0.0543**	0.113
	40	0.0648	**0.0554**	0.1744
1/4:1/2:1/2	5	**0.062**	0.0655	0.0756
	10	**0.0557**	0.0652	0.0788
	20	**0.052**	0.0603	0.1095
	40	**0.0565**	0.0583	0.1702
1/4:1/4:3/4	5	0.1239	**0.0933**	0.1129
	10	0.1201	**0.0932**	0.1721
	20	0.1132	**0.0882**	0.2811
	40	0.0969	**0.0713**	0.4862
1/4:1/2:3/4	5	**0.0825**	0.098	0.1033
	10	**0.0842**	0.0947	0.1389
	20	**0.0759**	0.0827	0.2243
	40	**0.0734**	0.0736	0.4187
1/4:3/4:3/4	5	**0.0746**	0.1207	0.109
	10	**0.0681**	0.1253	0.159
	20	**0.0652**	0.1091	0.2869
	40	**0.0591**	0.0861	0.5192

Table 6:

Empirical type I error rate of unequal sample sizes, equal means $\mu_1 = \mu_2 = \mu_3 = 0$, the ratios of $\sigma_1 : \sigma_2 : \sigma_3 = 1:1:1 \sim 1:4:4$, nominal α of 0.05. Best values in bold.

$\sigma_1 : \sigma_2 : \sigma_3$	n1:n2:n3	F-test	Welch's F	Kruskal-Wallis
1:1:1	15:15:30	0.0524	**0.0487**	0.0520
	15:20:25	0.0537	**0.0528**	0.0566
	30:20:10	0.0494	0.0491	**0.0501**
	25:20:15	**0.0510**	0.0516	0.0522
	30:15:15	**0.0503**	0.0513	0.0522
1:1:2	15:15:30	0.0194	**0.0469**	0.0339
	15:20:25	0.0381	**0.0509**	0.0475
	30:20:10	0.0625	**0.0528**	0.0616
	25:20:15	0.0937	**0.0477**	0.0705
	30:15:15	0.0968	**0.0527**	0.0750
1:2:2	15:15:30	0.0407	**0.0507**	0.0476
	15:20:25	0.039	**0.0478**	0.0426
	30:20:10	0.0546	**0.0483**	0.056
	25:20:15	0.0776	**0.0478**	0.0709
	30:15:15	0.1058	**0.0498**	0.0820
1:1:3	15:15:30	0.0202	**0.0486**	0.0365
	15:20:25	0.0375	**0.0502**	0.0495
	30:20:10	0.0708	**0.0509**	0.0646
	25:20:15	0.1241	**0.0510**	0.0840
	30:15:15	0.1241	**0.0531**	0.0888
1:2:3	15:15:30	0.0248	**0.0544**	0.0393
	15:20:25	0.0369	**0.0533**	0.0448
	30:20:10	0.0618	**0.0528**	0.0635
	25:20:15	0.1014	**0.0484**	0.0827
	30:15:15	0.1174	**0.0482**	0.0982
1:3:3	15:15:30	0.0386	**0.0522**	0.0458
	15:20:25	0.0384	**0.0496**	0.0444
	30:20:10	0.0567	**0.0477**	0.0639
	25:20:15	0.0911	**0.0499**	0.0873
	30:15:15	0.1262	**0.0490**	0.1076
1:1:4	15:15:30	0.016	**0.0484**	0.0363
	15:20:25	0.0425	**0.0507**	0.0543
	30:20:10	0.0808	**0.0485**	0.0728
	25:20:15	0.1409	**0.053**	0.0973
	30:15:15	0.1476	**0.0547**	0.1002
1:2:4	15:15:30	0.0202	**0.0530**	0.0363
	15:20:25	0.0348	**0.0496**	0.0425
	30:20:10	0.0693	**0.0493**	0.0659
	25:20:15	0.1229	**0.0516**	0.0924

	30:15:15	0.1336	**0.0522**	0.1056
1:3:4	15:15:30	0.0260	**0.0530**	0.0404
	15:20:25	0.0356	**0.0507**	0.0431
	30:20:10	0.0600	**0.0513**	0.0654
	25:20:15	0.1021	**0.0526**	0.0904
	30:15:15	0.1299	**0.0476**	0.1120
1:4:4	15:15:30	0.0393	**0.0496**	0.0464
	15:20:25	0.0379	**0.0527**	0.0475
	30:20:10	0.0585	**0.0465**	0.0650
	25:20:15	0.0890	**0.0535**	0.0904
	30:15:15	0.1285	**0.0514**	0.1186

APPENDIX B

SAS CODE FOR THE SIMULATION STUDY

SAS CODE

```
/* Balanced Normal TypeI error rate*/
/*Simulating the data*/
%macro ODSOff(); /* Call prior to BY-group processing */
   ods graphics off;  ods exclude all;  ods noresults;
%mend;

%macro ODSOn(); /* Call after BY-group processing */
   ods graphics on;  ods exclude none;  ods results;
%mend;

%macro simu (r=, n1=, n2=, n3=, mu1=, mu2=, mu3=, sigma1=, sigma2=, sigma3=,
times=);
data Simulation(drop=i);
do SampleID = 1 to &Times;
   do i = 1 to &n1;
      group=1;  y=rand("normal",&mu1,&sigma1);/* Group 1: x ~ N(mu1,sigma1^)
*/
      output;
   end;
   do i = 1 to &n2;
      group=2;  y=rand("normal",&mu2,&sigma2);       /* Group 2: x ~
N(mu2,sigma2^) */
      output;
   end;
      do i = 1 to &n3;
      group=3;  y=rand("normal",&mu3,&sigma3);       /* Group 3: x ~
N(mu3,sigma2^) */
      output;
   end;
end;
run;

/* 1. Analyzing Classic F-test */
 %ODSOff ;
proc glm data=Simulation;
   by SampleID;
   class group;
   model y=group;
   ods output ModelANOVA=OutDataF;
   ods graphics off;
   ods noresults;
run;
ods graphics on;
ods results;
%ODSOn

/* Construct indicator var for obs that reject H0 at 0.05 significance */
data ResultF_&n1;
   set OutDataF; if HypothesisType=3;
   ConcludeH0 = (ProbF > 0.05);
run;

/* Compute proportion: (# that conclude H0)/Times and CI */
proc freq data=ResultF_&n1;
```

```
      title "Result of Classic F";
    tables ConcludeH0 / nocum;* binomial(level='1');
      ods output OneWayFreqs=finalF;
run;

/* 2. Analyzing Welch F-test */
%ODSOff
proc glm data=Simulation;
   by SampleID;
   class group;
      model y=group;
      means group/welch;
   ods output Welch=OutWelch;
run;
%ODSOn

/* Construct indicator var for obs that reject H0 at 0.05 significance */
data ResultWelch_&n1;
   set OutWelch; if DF=2;
   ConcludeH0 = (ProbF > 0.05);
run;

/* Compute proportion: (# that conclude H0)/Times and CI */
proc freq data=ResultWelch_&n1 ;
    tables ConcludeH0 / nocum;* binomial(level='1');
ods output OneWayFreqs=finalWelch;
run;

/* 3. Analyzing Kruskal-Wallis */
/*First get the ranked data*/
proc sort data=Simulation out=Rank;
by SampleID y;
run;

data Rank1;
set Rank;
rank+1;
by SampleID;
if first.SampleID then rank=1;
run;

/*Second do the analysis*/
%ODSOff
proc glm data=rank1;
      by SampleID;
      class group;
      model rank=group;
      ods output ModelANOVA=OutKruskal;
run;
%ODSOn

/* Construct indicator var for obs that reject H0 at 0.05 significance */
data ResultKruskal_&n1;
   set OutKruskal; if HypothesisType=3;
   ConcludeH0 = (ProbF > 0.05);
run;
/* Compute proportion: (# that conclude H0)/Times and CI */
```

```
proc freq data=ResultKruskal_&n1 ;
    tables ConcludeH0 / nocum; * binomial(level='1');
ods output OneWayFreqs=finalKruskal;
run;

data final;
set finalF finalWelch finalKruskal;
if ConcludeH0=0 then delete;
truepercent=1-Frequency/&times;
run;
title;

data final;
set final;
obs=_N_;
run;

proc transpose data=final out=table;
id obs;
var truepercent;
run;

data initial;
set initial table;
run;
proc print data=initial; run;
%mend;

data initial;
input Obs _NAME_ $ _1 _2 _3 ;
cards;
0 NA . . .
;
run;

*******************************Table 1*********************************
* scenario 1;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=1,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=1,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=1,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=1,
times=10000);

* scenario 2;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=2,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=2,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=2,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=2,
times=10000);

* scenario 3;
```

```
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=2,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=2,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=2,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=2,
times=10000);

* scenario 4;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=3,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=3,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=3,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=3,
times=10000);

* scenario 5;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=3,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=3,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=3,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=3,
times=10000);


* scenario 6;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=3,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=3,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=3,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=3,
times=10000);

* scenario 7;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=4,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=4,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=4,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=1, sigma3=4,
times=10000);

* scenario 8;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=4,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=4,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=4,
times=10000);
```

```
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=2, sigma3=4,
times=10000);

* scenario 9;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=4,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=4,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=4,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=3, sigma3=4,
times=10000);

* scenario 10;
%simu (n1=5, n2=5, n3=5, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=4, sigma3=4,
times=10000);
%simu (n1=10, n2=10, n3=10, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=4, sigma3=4,
times=10000);
%simu (n1=20, n2=20, n3=20, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=4, sigma3=4,
times=10000);
%simu (n1=40, n2=40, n3=40, mu1=0, mu2=0, mu3=0, sigma1=1, sigma2=4, sigma3=4,

times=10000);

/* Balanced Log-normal TypeI error rate*/
/*Simulating the data*/
%macro ODSOff(); /* Call prior to BY-group processing */
   ods graphics off;  ods exclude all;  ods noresults;
%mend;

%macro ODSOn(); /* Call after BY-group processing */
   ods graphics on;  ods exclude none;  ods results;
%mend;

%macro simu (r=, n1=, n2=, n3=, mu1=, mu2=, mu3=, sigma1=, sigma2=, sigma3=,
times=);
data Simulation(drop=i);
*call streaminit(321);
do SampleID = 1 to &Times;
   do i = 1 to &n1;
      group=1;  y = exp(rand("normal",&mu1,&sigma1));      /* Group 1: x ~
N(mu1,var1^) */
      output;
   end;
   do i = 1 to &n2;
      group=2;  y = exp(rand("normal",&mu2,&sigma2));      /* Group 2: x ~
N(mu2,var2^) */
      output;
   end;
      do i = 1 to &n3;
      group=3;  y = exp(rand("normal",&mu3,&sigma3));      /* Group 3: x ~
N(mu3,var3^) */
      output;
   end;
end;
run;
```