DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ROBUST METHODS FOR COMPARING K INDEPENDENT GROUPS

by

İrem YILMAZ

January, 2015

İZMİR

ROBUST METHODS FOR COMPARING K INDEPENDENT GROUPS

A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Master of Science in
Statistics, Statistics Program

by
İrem YILMAZ

January, 2015
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"ROBUST METHODS FOR COMPARING K INDEPENDENT GROUPS"** completed by **İREM YILMAZ** under supervision of **ASSOC. PROF. DR. A. FIRAT ÖZDEMİR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. A. Fırat ÖZDEMİR

Supervisor

Asst. Prof. Dr. Engin YILDIZTEPE

(Jury Member)

Assoc. Prof. Dr. İlhan Karaaslan

(Jury Member)

Prof.Dr. Ayşe OKUR
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

I would like to express my deepst gratitude to my advisor, Assoc. Prof. Dr. A. Fırat ÖZDEMİR, for his excellent guidance, continuous support, patience and providing me with a perfect atmosphere to do my research. Besides my thesis I also would like to thank to him sharing his experiences during my master degree programme.

I would like to acknowledge my dissertation committee members, Asst. Prof. Dr. Engin YILDIZTEPE and Assoc. Prof. Dr. İlhan KARAKILIÇ, for their constructive comments and suggestions.

I would like to especially thank to my mother, Ayşe YILMAZ for her excellent support and limitless patience. Also I want to thank to my brother, Kadir YILMAZ, and his wife, Duygu Soy YILMAZ, for their support.

I would like to thank to Gülce ÖZEL, Gözde NAVRUZ and Bilgehan ONAT, who are as good friends, were always willing to help and their support.

I would like to express my appreciation to Mehmet Yavuz ONAT for his helping, support, patience and knowledge.

Lastly, I would like thank to anyone else who support and encourage me during my thesis process.

İrem YILMAZ

ROBUST METHODS FOR COMPARING K INDEPENDENT GROUPS

ABSTRACT

ANOVA-F test is the best known method for comparing more than two groups. However, ANOVA-F test can give misleading results when the underlying assumptions of the method are violated. In these circumstances, robust methods instead of ANOVA-F test can be preferred to get satisfying results in terms of nominal Type I error and power.

In this study, eight different methods are compared by using their Type I errors under different settings comprised of normality, non-normality, homoscedasticity, heteroscedasticity, equal sample sizes and unequal sample sizes. The methods without robust estimators are ANOVA-F test, Kruskal-Wallis test and Welch's Heteroscedastic F test while the methods with robust estimators are Welch's heteroscedastic F test with trimmed mean, Welch's heteroscedastic F test with trimmed mean and bootstrap-t, Box method, a method for comparing medians with McKean-Schrader estimate of standard error and Schrader and Hettmansperger method. After calculating Type I error rates of the methods by using a simulation study, powers are calculated for the four best methods which controlled the Type I error rate within Bradley's robustness limits. Welch's heteroscedastic F test with trimmed mean can be higly recommended to avoid the problems of assumption violation.

Keywords: ANOVA-F test, Welch test, Box method, Schrader and Hettmansperger method, McKean-Schrader estimate of the standard error, bootstrap-t, Type I error rate.

K BAĞIMSIZ GRUBU KARŞILAŞTIRMAK İÇİN DAYANIKLI METODLAR

ÖZ

İkiden fazla grup karşılaştırmalarında en çok bilinen yöntem ANOVA-F testtir. Ancak, ANOVA-F test varsayımları ihlal edildiğinde yanıltıcı sonuçlar verebilir. Bu durumlarda, iyi sonuçlar alabilmek için, ANOVA-F test yerine dayanıklı yöntemler belirlenen 1. Tip hatalarına ve güç değerlerine göre tercih edilebilir.

Bu çalışmada, sekiz farklı yöntem normalliğin ve homojen varyanslılığın ihlal edildiği ve edilmediği**,** örneklem genişliğinin eşit alındığı ve alınmadığı durumlar altında gerçekleşen 1. Tip hata oranlarına göre karşılaştırılmışlardır. Bu çalışmada dayanıklı kestiricilere sahip olmayan yöntemler; ANOVA-F test, Kruskal-Wallis test ve Welch'in heteroscedastic F testi iken dayanıklı kestiricilere sahip olan yöntemler; budanmış ortalama ile Welch'in heteroscedastic F testi, budanmış ortalama ve bootstrap-t ile Welch'in heteroscedastic F testi, Box yöntemi, McKean-Schrader standart hata kestirimi ile ortanca karşılaştırılması için bir yöntem ve Schrader ve Hettmansperger yöntemidir. Bir benzetim çalışması ile bu yöntemlere ait 1. Tip hatalar hesaplandıktan sonra, Bradley'nin daayanıklılık limitlerine gore 1. Tip hata kontrolünü sağlayan en iyi dört yöntem için bir de güç değerleri hesaplanmıştır. Tüm sonuçlara göre, varsayım ihlallerinin olduğu durumda budanmış ortalama ile Welch'in heteroscedastic F testi önerilmektedir.

Anahtar kelimeler: ANOVA-F test, Welch testi, Box yöntemi, Schrader ve Hettmansperger yöntemi, McKean-Schrader standart hata kestirimi, bootstrap-t, 1. Tip hata oranı.

CONTENTS

Page

THESIS EXAMINATION RESULT FORM ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZ v
LIST OF FIGURES viii
LIST OF TABLES ix

CHAPTER ONE - INTRODUCTION 1

CHAPTER TWO – ROBUST MEASURES AND METHODS 8

2.1 Robust Measures of Location and Scale 8
2.1.1 Winsorized Mean 8
2.1.2 Trimmed Mean 9
2.1.3 Winsorized Variance 10
2.1.4 Median 11
2.1.5 Median Absolute Deviation (MAD) 11
2.1.6 M Measures of Location 12
2.2 Robust Methods 13
2.2.1 Welch's Heteroscedastic F Method 13
2.2.2 Welch's Heteroscedastic F Method with Trimmed Mean 14
2.2.3 Welch's Heteroscedastic F Method with Trimmed Mean Based on a Bootstrap-t 16
2.2.4 Box Method 16
2.2.5 A Method for Comparing Medians with McKean-Schrader Estimate of Standard Error 18
2.2.6 Schrader & Hettmansperger Method 19

CHAPTER THREE – DESIGN OF SIMULATION AND RESULTS 21

3.1 Design of Simulation 21
3.1.1 g-h Distribution 22
3.2 Results 24
3.2.1 Actual Type I Error Rates for Normal Distribution (Symmetric) 24
3.2.2 Actual Type I Error Rates for g-h (g=0.5, h=0.2) Distribution (Asymmetric) 25
3.2.3 Actual Type I Error Rates for Real Data Sets 26
3.2.4 Power Results 28

CHAPTER FOUR - CONCLUSION 33

REFERENCES 36

LIST OF FIGURES

Page

Figure 3.1 g-h distribution with g = 0.5 23
Figure 3.2 g-h distribution with g = 0.2 23
Figure 3.3 Histograms for real data sets 27
Figure 3.4 Line plot for power results of four methods under each setting in normal distribution 29
Figure 3.5 Line plot for power results of four methods under each setting in g-h (g=0.5, h=0.2) distribution 30
Figure 3.6 Line plot for power results of four methods under each setting in minimum pitch values 31
Figure 3.7 Line plot for power results of four methods under each setting in number of pulses 32
Figure 4.1 Bar plot for number of settings in terms of actual Type I error rates of the methods 34

LIST OF TABLES

Page

Table 3.1 Variances and sample sizes.....22
Table 3.2 Actual Type I error rates for normal distribution.....25
Table 3.3 Actual Type I error rates for g-h (g=0.5, h=0.2) distribution 26
Table 3.4 Descriptive statistics of real data sets.....26
Table 3.5 Actual Type I error rates for minimum pitch values.....27
Table 3.6 Actual Type I error rates for number of pulses.....28
Table 3.7 Power results for normal distribution.....29
Table 3.8 Power results for g-h (g=0.5, h=0.2) distribution 30
Table 3.9 Power results for minimum pitch values.....31
Table 3.10 Power results for number of pulses 32

CHAPTER ONE
INTRODUCTION

Hypothesis testing procedures are used in many fields such as psychology, medicine, and biology to determine the true difference between the groups. The most known method to compare more than two independent groups is the analysis of variance (ANOVA) which was created by Sir Ronald Fisher in the 1920s. ANOVA is a statistical technique to find out the effects of factor(s) (independent variable(s)) on dependent variable. The technique is limited by the number of factor(s) in model equation. If it is analyzed with a single factor, it is called one-way ANOVA-F test.

Let X_{ij} be the j th observation from the i th group $i = 1,...,k$ and $j = 1,...,n_i$, where n_i is the sample size of i th group, k is the number of groups and N is the total sample size. The hypothesis that is tested is

$$H_0 : \mu_1 = ... = \mu_k = \mu \tag{1.1}$$

The one-way ANOVA-F test statistic is

$$F = \frac{(N-k)\sum_{i=1}^{k} (\bar{X}_i - \bar{X})^2 n_i}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2} \tag{1.2}$$

where

$$N = \sum_{i=1}^{k} n_i \tag{1.3}$$

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i} \tag{1.4}$$

$$\bar{X} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}}{N} \tag{1.5}$$

To get significant results from ANOVA-F test, there are some assumptions that should be satisfied. These assumptions are normality, homoscedasticity,

randomization, and independence of observations. Eisenhart (1947) summarized the details of these assumptions. Randomization and independence assumptions are about sampling technique, while normality and homoscedasticity are about population distributions

Normality assumption has two properties: symmetry and kurtosis. When the fifty percent of observations stays at the right side of mean and the other fifty percent stays at the left side of mean, the distribution is symmetric. The symmetry measurement is the third moment about the mean of a distribution and it is called skewness statistic (γ_3). It equals to zero in normal distribution. On the other hand, kurtosis is the forth moment of a distribution (γ_4) and it equals to 3 in normal distribution. In fact, distributions are almost never normal. Micceri (1989) pointed out when Kolmogrov-Smirnow test of normality was used on a sample of 440 distributions from published research, 100% of the distributions were non-normal at the nominal Type I error of 0.01. Distributions can be quite skewed and they can have heavy tails. When samples are generated from quite skewed and/or heavy tailed distributions, these samples can have outliers. Heavy tails and outliers can cause really serious practical problems because of inflated standard error of the sample mean. Fisher (1922) realized that the sample mean could be inadequate when there are slight changes from normality.

In one-way designs, even if normality assumption is satisfied with equal sample sizes, but unequal variances, there are some problems about controlling Type I error rate. For example, for four independent groups with equal sample sizes of 50, Type I error rate of ANOVA-F test roughly equals to 0.09 at the 0.05 level (Wilcox, Charlin, & Thompson, 1986). In other respect, when assumption of homoscedasticity is not satisfied, but if heterogeneity of variances is very little, ANOVA-F test can deal with this violation (Box, 1954). However, when the sample sizes are not equal the effect of variance heterogeneity increases. Moreover, if there is positive pairing between variances and sample sizes, actual Type I error of ANOVA-F test is rather smaller than nominal Type I error, so ANOVA-F gives conservative result. If there is negative pairing of variances and sample sizes, actual Type I error is quite larger than

nominal one that means ANOVA-F has liberal result (Glass, Peckham, &Sanders 1972; Keselman, Rogan, & Feir-Walsh, 1977; Krutchkoff, 1988; Lee & Ahn, 2003).

When determining the appropriate analysis technique, the most important point is that the method can control Type I error. The nominal Type I error rate (α) is selected among all levels of α, while the actual Type I error rate ($\hat{\alpha}$) is the estimate of the Type I error for a method. The methods must keep close the actual Type I error to nominal one. Otherwise, it can cause of taking a wrong decision.

When any statistical hypothesis is tested, one other important concern is power as well as Type I error. When a method cannot control Type I error, power of the test cannot be interpreted, because if the actual Type I error is too larger or smaller than the nominal Type I error, calculated power of this test is not reliable. Power is an ability of a test which can detect the actual difference of populations. The traditional methods (ANOVA-F for more than two groups, student-t test for two groups) cannot control Type I error rate and they give very poor power when the underlying assumptions are violated. Wilcox (2012) gave an illustration that when sampling is from normal distribution; student-t test has a power of 0.957 to detect $\mu_1 - \mu_2 = 1$ with $n_1 = n_2 = 25$ and α of 0.05. However, when sampling is from contaminated normal distribution in that 10 percent of the observations are generated from a normal distribution with variance of 10 instead of 1, the power reduces to 0.028. Power is really affected with this very slight departure from the normal distribution. There are two important problems about why ANOVA-F test is affected so much even if there are only little changes from normal distribution. The first one is sensitivity of the population mean and variance to little changes in a distribution and the second one is the standard errors that are also sensitive to slight changes. Mathematical foundation of the problems are examine by Huber (1981); Hampel, Ronchetti, Rousseuw, & Stahel (1986); Staudte & Sheather (1990).

When the assumptions are violated, non-parametric methods instead of ANOVA-F test can be used. In order to use non-parametric methods, it is mostly enough to satisfy independence and randomization assumptions. However, there is a lack of

information when applying non-parametric methods, and that results in poor power. In particular, if heterogeneity of variances is high, it is not recommended (Krutchkoff, 1988). The most used non-parametric method as an alternative of ANOVA-F is Kruskal-Wallis test.

There has been a long history behind unequal variances when testing for mean equality since Behrens (1929) and Fisher (1935). When two independent populations are normal, but their variances are not equal, there is a problem of hypothesis testing for the difference of means and it is called Behrens-Fisher problem. The first test that aims to solve this problem was derived by Welch (1947). Welch test is a test whose degrees of freedom depend on both sample size and variance. Also, Welch developed this test for comparing k independent populations (Welch, 1951). Although Welch test is robust to heteroscedasticity, it behaves so sensitive to non-normality because of the population mean and variance. In this case, more robust measures are needed to avoid from sensitivity to these violations.

In general, if a measure is insensitive to departures from assumptions, it can be said that the measure is robust. However, μ and σ and their estimators ($\bar{X}$ and s) are not robust if there are only little changes from normality. The particular idea of robustness theory was developed by Hampel (1968) and Huber (1964).

Though many robust estimators have been recommended in the literature, use of the trimmed mean and the winsorized variance are more common due to good theoretical properties and computational simplicity (Wilcox, 1995a). It was mentioned that the Welch (1951) test is robust to heteroscedasticity but, not robust to non-normality. If a robust measure of location (e.g. trimmed mean) and robust measure of scale (e.g. winsorized variance) are used instead of the sample mean and the sample variance, Welch method can be robust to both violations (Wilcox, 2012). Besides Welch test, Box (1954) test gives good results when it is used with the sample trimmed mean and the sample winsorized variance. When Box (1954) and Welch (1951) methods are used with trimmed mean, they can control Type I error under non-normality and heteroscedasticity. The test statistics of Welch method with

trimmed mean and Box method similarly control Type I error. When comparing two groups, both of them give exactly same results (Wilcox, 2012). For more than two groups, Box method can control Type I error better than Welch method with trimmed mean in some situations. Lix and Keselman (1998) considered that the test statistic of Box method gives a Type I error which is less probably exceeding 0.075 at the 0.05 level, but it generally has less power than the test statistic of Welch method with trimmed mean.

The methods based on trimmed mean can be combined with bootstrap methods in order to get better Type I error control (Westfall & Young, 1993). When using 20% trimmed mean with bootstrap-t method in Welch test, Type I error is controlled well (Wilcox, 2012).

Median is another common robust measure which belongs to a class of trimmed means and it is estimated with the most percentage of trimming (50%). The McKean and Schrader estimate of the standard error is used in conjunction with Welch tests for comparing population medians (McKean & Schrader, 1984). The critical value of this method is calculated by using bootstrap-t. The method can control Type I error rate well unless there are tied values and if the sample sizes are larger than 20 (Wilcox, 2012).

M measures of location are a large class of location measures and the population mean is a member of this class. M measure of location is estimated with an iterative estimation method called Newton-Raphson method. Furthermore, one iteration yields one-step M estimator. If only there is a single iteration to estimate M measure, asymptotic properties of the estimator are good (Serfling, 1980). Schrader and Hettmansperger (1980) derived a method that performs well when using one-step M estimator. When comparing one-step M-estimators, the critical value is calculated by bootstrap-t as well as the method of comparing medians.

When the effect of non-normality is investigated, there are three different type of non-normality; symmetric heavy tailed, asymmetric light tailed, and asymmetric

heavy tailed. The effects of departure from normality can be observed by using g-h distribution where g parameter controls skewness and h parameter controls kurtosis. In the simulation part of this study, observations were generated from normal distribution and g-h distributions (g=0.5, h=0.2).

Eight different methods were used for a simulation study. These methods are ANOVA-F test which is traditional method, Kruskal-Wallis test that is a non-parametric alternative to ANOVA-F test, Welch's heteroscedastic F test (Welch, 1951) as an alternative to ANOVA-F test, Welch method with trimmed mean, Welch method with trimmed mean based on bootstrap-t (Wilcox, 2012), Box method (Box-1951), a method for comparing medians with McKean-Schrader estimate of standard error (McKean & Schrader, 1984) and Schrader and Hettmansperger method (Schrader & Hettmansperger, 1980).

The purpose of this study is to compare these methods in terms of their Type I error rates under different settings which are combination of normality, non-normality, homoscedasticity, heteroscedasticity, equal sample sizes, and unequal sample sizes. Furthermore, the methods that controlled the Type I error rate are compared in terms of their power results under these settings.

The second chapter of the study is divided into two parts: Robust measures and scale and robust methods. In the first part, robust measures of location and scale and their estimators used in this study were described. These location estimators are sample winsorized mean with $\gamma = 0.2$, sample trimmed mean with $\gamma = 0.2$, sample median, and one-step M estimator. The estimators of scale are winsorized variance and normalized median absolute deviation. In the second part, descriptions and test statistics of six methods were given.

The third chapter has two parts: design of simulation and results. The first part is about how the study design was set with the combinations of non-normality, heteroscedasticity and unequal sample sizes, while there are the Type I error and

power results for theoretical distributions (normal and g-h distribution) and real data sets in second part.

The last chapter is conclusion which consists of criticism about the important points of the study.

CHAPTER TWO
ROBUST MEASURES AND METHODS

In this chapter, there are two basic parts which are robust measures and methods. The first part comprises of descriptions, properties and estimators of robust measures which are used in the robust methods. In second part, each robust method compared in the simulation study is examined circumstantially.

2.1 Robust Measures of Location and Scale

Distributions are characterized by measures. The most known measures of location and scale are mean (μ) and variance (σ^2). As it was indicated in chapter one, they can be affected so easily when there are small departures from the assumptions as well as their estimators which are the sample mean ($\bar{X}$) and variance (s^2). If it is supposed to be more clarified, one of the important reasons of this sensitivity is their breakdown points. For instance, the finite sample breakdown point of mean is $1/n$ and the asymptotic breakdown point is zero. Therefore, the mean is extremely affected from each observation (Wilcox, 2012).

Robustness means the insensitivity of measures to violations of assumptions such as non-normality. There are three different ways to explain the robustness properties of estimators: The breakdown point, influence function, and the continuity of descriptive measure. Staudte and Sheather (1990) described these three tools as quantitative robustness, infinitesimal robustness, and qualitative robustness.

2.1.1 Winsorized Mean

As it was indicated, the mean is affected from each value, especially outliers. One solution of this problem is to give less weighted to values in the tails. This can be achieved by winsorized the distribution.

Let F be any distribution, and x_{γ} and $x_{\gamma-1}$ be the γ and $\gamma-1$ quantiles. A $\gamma-$ Winsorized analog of F is

$$F_w(x) = \begin{cases} 0, & \text{if} \quad x < x_{\gamma} \\ \gamma, & \text{if} \quad x = x_{\gamma} \\ F(x), & \text{if} \quad x_{\gamma} < x < x_{\gamma-1} \\ 1, & \text{if} \quad x \geq x_{\gamma-1} \end{cases} \tag{2.1}$$

The mean of winsorized distribution is

$$\mu_w = \int_{x_{\gamma}}^{x_{\gamma-1}} x dF(x) + \gamma(x_{\gamma} + x_{\gamma-1}) \tag{2.2}$$

Winsorized mean pays less attention to values in the tails of distribution. However, it is not a qualitatively robust. The influence function of Winsorized mean is bounded, but not smooth and the breakdown point is the amount of γ (Wilcox, 2012).

In order to estimate winsorized mean, the first step is to choose γ which represents the proportion of trimming from each tail of distribution. $\lambda = [\gamma n]$, where $[\gamma n]$ is the largest integer less than equal or equal to $[\gamma n]$ that n is the sample size (Wilcox, 2012).

$$Y_i = \begin{cases} X_{(\lambda+1)}, & \text{if} \quad X_i < X_{(\lambda+1)} \\ X_i, & \text{if} \quad X_{(\lambda+1)} < X_i < X_{(n-\lambda)} \\ X_{(n-\lambda)}, & \text{if} \quad X_i \geq X_{(n-\lambda)} \end{cases} \tag{2.3}$$

The sample winsorized mean is

$$\bar{X}_{win} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{2.4}$$

where Y_i is given by 2.3

2.1.2 Trimmed Mean

The trimmed mean is another solution for reducing the effects of observations on the tails, and it removes the observations on tails by specified proportion. The $\gamma-$ trimmed mean is

$$\mu_t = \frac{1}{2-2\gamma}\int_{x_\gamma}^{x_{\gamma-1}} x\,dF(x) \tag{2.5}$$

The influence function is bounded, but not smooth like winsorized mean. However, it is qualitatively robust and it has the breakdown point of γ. A value of γ must be chosen between 0 and 0.5. If γ is selected 0.5, the breakdown point achieves the highest value. However, if the trimming proportion is chosen too far from zero, some practical problems could arise especially when sample size is not large. In general, the optimal amount of trimming is chosen between 0 and 0.25 to minimize standard error (Hill & Dixon, 1982; Wu, 2002). Moreover, Wilcox (1995a) suggests that 20% trimming proportion.

Let $X_1 \ldots X_n$ be a random sample, given steps are followed to estimate the trimmed mean.

- The observations are written in ascending order ($X_{(1)} \leq \ldots \leq X_{(n)}$).
- The amount of trimming is chosen within the interval of $0 \leq \gamma \leq 0.5$.
- $\lambda = [\gamma n]$ is calculated and rounded to the nearest integer.
- The effective sample size is calculated ($h = n - 2\lambda$).
- The sample trimmed mean is computed by using equation 2.13.

$$\bar{X}_t = \frac{1}{h}\sum_{\lambda+1}^{n-\lambda} X_i \tag{2.6}$$

2.1.3 Winsorized Variance

The γ − Winsorized variance is

$$\sigma_\omega^2 = \int_{x_\gamma}^{x_{1-\gamma}} (x-\mu_\omega)^2 dF(x) + \gamma[(x_\gamma - \mu_\omega)^2 + (x_{1-\gamma} - \mu_\omega)^2] \tag{2.7}$$

σ_ω^2 is equal to 0.4129 for a standard normal distribution. Its influence function is bounded. The winsorized variance is involved by influence function of a large class of L-measures of scale (Welsh & Morrison, 1990).

To estimate the winsorized variance,

$$s_{\omega}^{2} = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{X}_{\omega})^2 \tag{2.8}$$

Where $\bar{X}_{\omega}$ is the sample winsorized mean that is indicated in 2.1.1. Y_i is given by equation 2.3

2.1.4 Median

Another robust measure of location is median. The median, which belongs to a class of trimmed means, represents a maximum amount of trimming. It is qualitative robust and its breakdown point is 0.5

2.1.5 Median Absolute Deviation (MAD)

The most common robust measure of scale is MAD (ω). It is given by

$$P(|X - x_{0.5}| < \omega) = \frac{1}{2} \tag{2.9}$$

$x_{0.5}$ is the population median and ω is the 0.5 quantile of distribution of $|X - x_{0.5}|$. When sampling is from a standard normal distribution, $x_{0.5}$ is equal to zero and ω is determined by $P(-\omega \le Z \le \omega)$ where Z has a standard normal distribution and ω is the 0.75 quantile of the standard normal distribution. It is approximately equal to 0.6745.

The standard estimate of ω is given as follows,

$$MAD = MED|X_1 - M|,...,|X_n - M| \tag{2.10}$$

Where M is the sample median and MAD is the sample median of $|X_1 - M|,...,|X_n - M|$. The finite sample breakdown point of MAD is approximately equal to 0.5.

If sampling is from a normal distribution, MAD doesn't estimate the population standard deviation. In this case, normalized median absolute deviation is recommended, because MADN is more efficient than MAD (Wilcox, 2012).

$$MADN = \frac{MAD}{Z_{0.75}} \approx \frac{MAD}{0.6745} \tag{2.11}$$

2.1.6 M Measures of Location

M measures are a huge class of measures of location. The population mean is a member of this class.

μ_m is estimated with an iterative estimation method which is the Newton Raphson method. To compute the M estimator of location μ_m,

Set $k = 0$, $\mu_k = M$, the sample median, and choose a value for k. Generally k is chosen 1.28.

Step1. Let

$$A = \sum \psi\left(\frac{X_i - \mu_k}{MADN}\right) \tag{2.12}$$

Huber's ψ function,

$$\psi(x) = \max\{-k, \min(k, x)\} \tag{2.13}$$

Step2. Let

$$B = \sum \psi'\left(\frac{X_i - \mu_k}{MADN}\right) \tag{2.14}$$

ψ' is the derivative of ψ,

$$\psi'(x) = \begin{cases} 1, & \text{if} \quad -k \le x \le k \\ 0, & \text{if} \quad \text{otherwise} \end{cases} \tag{2.15}$$

B is the number of observations X_i and it satisfies $-k \le \left(\frac{X_i - \mu_k}{MADN}\right) \le k$.

Step3. Set

$$\mu_{k+1} = \mu_k + \frac{MADN \times A}{B} \tag{2.16}$$

Step4. If $\left|\hat{\mu}_{k+1} - \hat{\mu}_k\right| < 0.0001$, stop and set $\hat{\mu}_m = \hat{\mu}_{k+1}$. Otherwise, k increments by one and from 1 to 4 steps are repeated.

One iteration yields one-step M estimator. Here, we can explain to estimate one-step M estimator with more clear way.

$$\hat{\mu}_{os} = \hat{\mu}_m + \frac{MADN\sum \psi\left(\frac{X_i - \hat{\mu}_k}{MADN}\right)}{\sum \psi'\left(\frac{X_i - \hat{\mu}_k}{MADN}\right)} = \frac{1.28MADN(i_1 - i_2) + \sum_{i=i_1+1}^{n-i_2} X_i}{n - i_1 - i_2} \qquad (2.17)$$

Where $X_{(1)} \le ... \le X_{(n)}$ are the order statistics, $\hat{\mu}_k$ is the sample median, i_1 is the number of observations satisfying $\frac{(X_i - \hat{\mu}_k)}{MADN} < -1.28$ and i_2 is the number of observations satisfying $\frac{(X_i - \hat{\mu}_k)}{MADN} > 1.28$. MADN was given by section 2.1.6

2.2 Robust Methods

2.2.1 Welch's Heteroscedastic F Test

The heteroscedatic alternative to the ANOVA-F test is Welch method that is robust to heteroscedasticity but, not robust to non-normality (Welch, 1951).

$$H_0 : \mu_1 = ... = \mu_k = \mu \qquad (2.18)$$

X_{ij} : The j th observation of i th group

$\bar{x}_i$: The sample mean of i th group

$\bar{x}$: The sample mean

n_i : The number of observations of i th group

k : Number of groups

The test statistic is

$$F_w = \frac{\sum_{i=1}^{k} w_i(\bar{x}_i - \bar{x})^2}{k-1}\left[1 + \frac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right)\left(1 - \frac{w_i}{\sum_{i=1}^{k} w_i}\right)^2\right]^{-1} \quad (2.19)$$

where

$$w_i = \frac{n_i}{s_i^2} \quad (2.20)$$

$$s_i^2 = \frac{\sum_{j=1}^{n_i}(X_{ij} - \bar{x}_i)^2}{n_i - 1} \quad (2.21)$$

When the null hypothesis is true, F_w has approximately an F distribution with $k-1$ and ν_w degrees of freedom and the critical value is $F_{k-1,\nu_w;\alpha}$.

$$\nu_w = \frac{k^2 - 1}{3\sum_{i=1}^{k}\left(\frac{1}{n_i - 1}\right)\left(1 - \left(\frac{1}{\sum_{i=1}^{k} w_i}\right)\right)^2} \quad (2.22)$$

When $F_w \geq F_{k-1,\nu_w;\alpha}$, H_0 is rejected.

2.2.2 Welch's Heteroscedastic F Test with Trimmed Mean

As it is mentioned in 2.2.1, Welch method is just robust to unequal variances, not to non-normality. If a robust measure of location (e.g. trimmed mean) and scale (e.g. winsorized variance) are used instead of the sample mean and the sample variance, Welch method could be robust for both violations (Wilcox, 2012).

$\bar{X}_{ti}$: The sample trimmed mean of i th group as it was given by equation 2.6

$\bar{X}_{wini}$: The sample winsorized mean of i th group as it was given by equation 2.4

s_{wi}^2 : The sample winsorized variance of i th group as it was given by equation 2.17

The hypothesis is tested that is

$$H_0 : \mu_{t1} = ... = \mu_{tk} = \mu_t \quad (2.23)$$

The test statistic is

$$F_t = \frac{A}{B+1} \tag{2.24}$$

where

$$A = \frac{1}{k-1}\sum_{i=1}^{k} w_i(\bar{X}_{ti} - \tilde{X})^2 \tag{2.25}$$

and

$$B = \frac{2(k-1)}{k^2-1}\sum_{i=1}^{k}\frac{\left(1-\frac{w_i}{u}\right)^2}{h_i - 1} \tag{2.26}$$

where

$$\tilde{X} = \frac{\sum_{i=1}^{k} w_i \bar{X}_{ti}}{U} \tag{2.27}$$

$$U = \sum_{i=1}^{k} w_i \tag{2.28}$$

$$w_i = \frac{1}{d_i} \tag{2.29}$$

$$d_i = \frac{(n_i - 1)s_{wi}^2}{h_i(h_i - 1)} \tag{2.30}$$

When the null hypothesis is true F_t has approximately an F distribution with $k-1$ and ν_{wt} degrees of freedom.

$$\nu_{wt} = \left(\frac{3}{k^2-1}\sum_{i=1}^{k}\frac{\left(1-\frac{w_i}{U}\right)^2}{h_i} \right)^{-1} \tag{2.31}$$

The critical value is $F_{k-1,\nu_{wt};\alpha}$ and H_0 is rejected if $F_t \geq F_{k-1,\nu_{wt};\alpha}$.

2.2.3 Welch's Heteroscedastic F Test with Trimmed Mean Based on a Bootstrap-t

When 20% trimmed mean with bootstrap-t are used in Welch method, the probability of Type I error can be controlled well (Westfall & Young, 1993; Wilcox, 2012).

$$H_0 : \mu_{t1} = ... = \mu_{tk} = \mu_t \quad (2.32)$$

For this method, the test statistic F_t is calculated as shown by 2.24.

The steps of bootstrap-t are

- Determine a bootstrap sample $(X^*_{i1}, X^*_{i2}, ..., X^*_{in})$ from each group.
- Next, set $C^*_{ij} = X^*_{ij} - \bar{X}_{ti}$, $j = 1, 2, ..., n_i$, where $\bar{X}_{ti}$ is the sample trimmed mean of i th original group. The C_{ij} values are the empirical distribution of the i th group centered. Therefore, the empirical trimmed mean is zero and hypothesis of equal trimmed means among these k distributions is true.
- Calculate F^*_{ti} with C^*_{ij} values and F^*_{ti} is the test statistic of i th bootstrap sample.
- Repeat this process B times.
- All statistics which are obtained with bootstrapping, are written in ascending order ($F^*_{t(1)} \leq ... \leq F^*_{t(B)}$).
- An estimate of α critical value is $F^*_{t(U)}$ where, $U = (1-\alpha)B$ rounded to nearest integer.
- F_t is the test statistic of original sample and if $F_t \geq F^*_{t(U)}$, the null hypothesis is rejected.

2.2.4 Box Method

The hypothesis is

$$H_0 : \mu_{t1} = ... = \mu_{tk} = \mu_t \quad (2.33)$$

$\bar{X}_t$: The sample trimmed mean

h_i : The effective sample size of i th group

s^2_{wi} : The sample winsorized variance of i th group

The test statistic is

$$F_b = \frac{\sum_{i=1}^{k} h_i (\bar{X}_{ti} - \bar{X}_t)^2}{\sum_{i=1}^{k} 1 - \left(\frac{h_i}{H}\right) S_i^2} \tag{2.34}$$

where

$$H = \sum_{i=1}^{k} h_i \tag{2.35}$$

$$\bar{X}_t = \sum_{i=1}^{k} \frac{h_i \bar{X}_{ti}}{H} \tag{2.36}$$

and

$$S_i^2 = \frac{(n_i - 1)s_{wi}^2}{h_i - 1} \tag{2.37}$$

When the null hypothesis is true, F_b has approximately an F distribution with ν_1 and ν_2 degrees of freedom.

$$\nu_1 = \frac{\left[\sum_{i=1}^{k} (1 - f_i) S_i^2\right]^2}{\left(\sum_{i=1}^{k} f_i S_i^2\right)^2 + \sum_{i=1}^{k} S_i^4 (1 - 2f_i)} \tag{2.38}$$

and

$$\nu_2 = \frac{\left[\sum_{i=1}^{k} (1 - f_i) S_i^2\right]^2}{\frac{\sum_{i=1}^{k} S_i^4 (1 - f_i)}{(h_i - 1)}} \tag{2.39}$$

where

$$f_i = \frac{h_i}{H} \tag{2.40}$$

H_0 is rejected if $F_b \geq F_{\nu_1, \nu_2, \alpha}$ (Box, 1954; Wilcox, 2012).

2.2.5 A Method for Comparing Medians with McKean-Schrader Estimate of Standard Error

The McKean-Schrader estimate of the standard error (McKean & Schrader, 1984) is used in conjunction with Welch test. Estimation of the standard error of the sample median could be incorrect when there are tied values even with large samples. Therefore, this method is not recommended when the sample has tied values (Wilcox, 2012).

In order to calculate the McKean-Schrader estimate of the squared standard error, k must be computed by using equation 2.41

$$k = \frac{n+1}{2} - z_{0.995}\sqrt{\frac{n}{4}} \tag{2.41}$$

k is rounded to the nearest integer The observations are written in ascending order. Then the McKean-Schrader estimate of the squared standard error of median is

$$\left(\frac{X_{(n-k+1)} - X_{(k)}}{2z_{0.995}}\right)^2 \tag{2.42}$$

M_i : The sample median of i th group

S_i^2 : The Mc-Kean and Schrader estimate of the squared standard error of M_i

The hypothesis is equivalance of the population medians and the test statistic of the method is

$$F_m = \frac{A}{1+B} \tag{2.43}$$

where

$$A = \frac{1}{k-1}\sum_{i=1}^{k} w_i (M_i - \bar{M})^2 \tag{2.44}$$

and

$$B = \frac{2(k-2)}{k^2-1}\sum_{i=1}^{k}\frac{\left(1-\frac{w_i}{U}\right)}{n_i-1} \tag{2.45}$$

where

$$\bar{M} = \frac{1}{U}\sum_{i=1}^{k} w_i M_i \tag{2.46}$$

$$U = \sum_{i=1}^{k} w_i \tag{2.47}$$

$$w_i = \frac{1}{S_i^2} \tag{2.48}$$

When the null hypothesis is true, F_m has approximately an F distribution with $\nu_1 = k-1$ and $\nu_2 = \infty$ degrees of freedom. $F_{k-1,\infty;\alpha}$ is the critical value and if $F_m \geq F_{k-1,\infty;\alpha}$, H_0 is rejected.

2.2.6 Schrader and Hettmansperger Method

The method described here performs well when one-step M-estimator is used. This method was derived by Schrader and Hettmansperger (1980) and studied by He, Simpson, and Portnoy (1990). The goal is to test

$$H_0 : \theta_1 = \ldots = \theta_k = \theta \tag{2.49}$$

θ_i : A population M-measure of location for i th group.

$\hat{\theta}_i$: One-step M estimator for i th group

The test statistic is

$$H = \frac{1}{N}\sum_{i=1}^{k} n_i(\hat{\theta}_i - \bar{\theta})^2 \tag{2.50}$$

where

$$N = \sum_{i=1}^{k} n_i \tag{2.51}$$

and

$$\bar{\theta} = \frac{1}{k}\sum_{i=1}^{k} \hat{\theta}_i \tag{2.52}$$

The critical value can be determined when the empirical distributions of each group is shifted. Hence, the estimated of measure of location is zero. The steps are

- Generate bootstrap samples from each group.
- Compute the test statistic based on bootstrap samples yiealding H^*.
- Repeat the process B times.
- $H_1^*, ..., H_B^*$ are test statistics for each bootstrap sample. They are written in ascending order ($H_{(1)}^* \leq ... \leq H_{(B)}^*$).
- $H_{(U)}^*$ is the critical value, where $U = (1-\alpha)B$, is rounded to the nearest integer.
- H_0 is rejected if $H > H_{(U)}^*$.

This method appears to give good results when using an M-estimator (with Huber's ψ) as well as the Harrel-Davis of the median, but it is not recommended when comparing means or even trimmed means (Wilcox, 2012).

CHAPTER THREE
DESIGN OF SIMULATION AND RESULTS

This chapter consists of design of simulation, the results of Type I error rate and power of the methods for theoretical distributions and real data sets.

3.1 Design of Simulation

Type I error rates of the methods were investigated under different conditions. Furthermore, power was calculated for the methods which controlled the Type I error. 28 settings were the combination of homoscedasticity and heteroscedasticity, equal and unequal sample sizes with positive and negative pairing of variances, normal and non-normal distributions. Nominal significance level was set as 0.05. 10000 replications of each condition were performed for all methods and B was chosen 599 for the methods based on bootstrap-t. All simulations were performed with statistical programming language R (version 3.1.2)

The methods with their abbreviated names are shown below.

- Anova-F test: F
- Kruskal Wallis test: K
- Welch's heteroscedastic F test: W
- Welch's heteroscedastic F test with trimmed mean: WTR
- Welch's heteroscedastic F test with trimmed mean based on bootstrapping: WTRB
- Box Method: B
- Comparing medians with McKean-Schrader standard error: Mc-S
- Schrader & Hettmansperger method to compare one-step M estimators: SH

Each comparison was done for three independent groups. The sample sizes and variances were paired positively and negatively. Positive pairing is where the smallest sample size is generated from the population which has the smallest

variance, while the largest sample size is paired with the largest variance. On the other hand, negative pairing is where the smallest sample size is paired with the largest variance, while the largest sample size is paired with the smallest variance. All sample sizes and variances are indicated in Table 3.1.

Table 3.1 Variances and sample sizes

Sample Size	Equal Variance	Unequal Variance
20, 20, 20	1, 1, 1	4, 25, 81
30, 30, 30	1, 1, 1	4, 25, 81
20, 30, 40	1, 1, 1	4, 25, 81; 81, 25, 4

3.1.1 g-h Distribution

A value X is generated from a g (-1 < g < 1) and h (0 <h < 1) distribution by first generating Z from standard normal distribution. Then,

when $g \neq 0$,

$$X = \frac{(\exp(gZ) - 1)\exp\left(\frac{hZ^2}{2}\right)}{g} \tag{3.1}$$

when $g = 0$

$$X = Z\exp\left(\frac{hZ^2}{2}\right) \tag{3.2}$$

Data are generated by using these transformations (Hoaglin, 1985). Some properties of g-h distribution are

- When g=h=0, X has standard normal distribution.
- When g=0, X has a symmetric distribution.
- As g increases, the distribution becomes more skewed.
- As h increases, the tails of distribution get heavier.

- When g=1 and h=0, X has log-normal distribution.

There are examples about the shape of distributions when g and h takes different values in Figure 3.1 and 3.2.

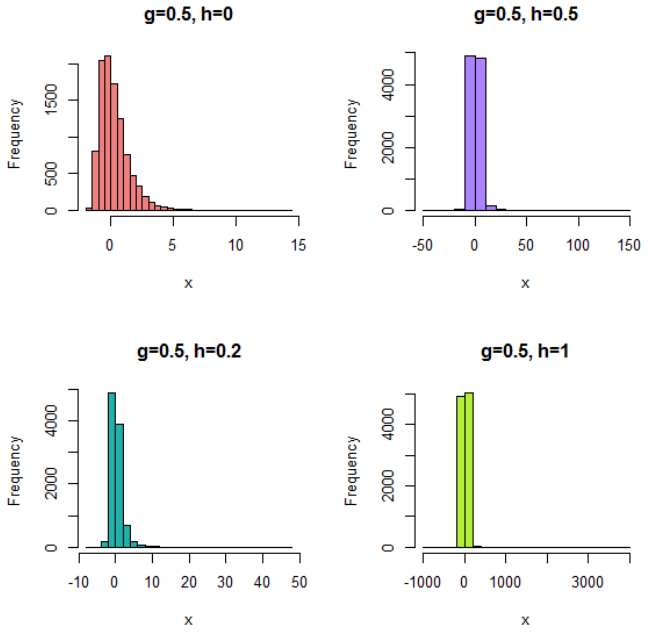


Figure 3.1 g-h distribution with g = 0.5

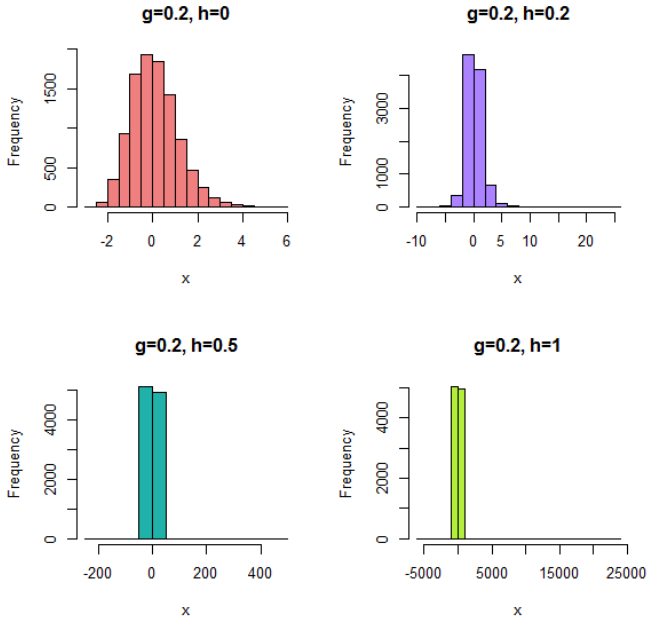


Figure 3.2 g-h distribution with g = 0.2

3.2 Results

In this chapter, Type I error results were interpreted for normal and g-h distributions besides two real data sets about Parkinson patients. Bradley (1978) presented two robustness criterions named by liberal ($0.5\alpha \leq \alpha \leq 1.5\alpha$) and conservative ($0.9\alpha \leq \alpha \leq 1.1\alpha$) criterions to be able to determine which hypothesis testing procedure works sufficiently. A criterion between liberal and conservative criterion ($\alpha - 0.2\alpha$ and $\alpha + 0.2\alpha$) were used in this study. A test considered to be robust when Type I error rate of the method is within 0.040 and 0.060 when $\alpha = 0.05$. After Type I error rates were calculated, power was computed for the methods that controlled the Type I error rates within 0.04 and 0.06.

Table 3.2 and 3.3 show actual Type I error rate obtained from normal distribution and g-h distribution respectively. Moreover, the Type I error rates of real data sets (minimum pitch values and number of pulses) are indicated in Table 3.5 and 3.6. All results in the tables are divided by 10000. In each table, cells colored yellow represent that the result is liberal whereas cells colored grey show that the result is conservative. Table 3.7 and 3.8 show power results for normal and g-h distribution respectively while Table 3.9 and 3.10 show power results for minimum pitch values and number of pulses.

3.2.1 Actual Type I Error Rates for Normal Distribution (Symmetric)

According to Table 3.2, all methods with robust estimator controlled Type I error rate under each setting. As it was mentioned before, W is robust to heteroscedasticity, but not robust to non-normality. Therefore, it controlled Type I error for all settings under normality. On the other hand, F and K had liberal results when there was heteroscedasticity except for the setting of positive pairing. Also, F gave one conservative result under positive pairing. If it is supposed to summarize the results for normal distribution, results of F and K were considerably affected when variances were unequal, while robust methods had no problem to control Type I error rate for settings of normal distribution.

Table 3.2 Actual Type I error rates for normal distribution

Normal Distribution										
Groups	n	σ	F	K	W	WTR	WTRB	B	Mc-S	SH
k=3	20 20 20	1 1 1	0.0496	0.0463	0.0492	0.0505	0.0437	0.0458	0.0487	0.0485
	20 20 20	2 5 9	0.0686	0.0664	0.0499	0.0547	0.0471	0.0537	0.0564	0.0599
	30 30 30	1 1 1	0.0506	0.0504	0.0519	0.0538	0.0508	0.0508	0.0497	0.0539
	30 30 30	2 5 9	0.0699	0.0697	0.0534	0.0572	0.0508	0.0531	0.0548	0.0576
	20 30 40	1 1 1	0.0511	0.0513	0.0516	0.0503	0.0460	0.0482	0.0487	0.0518
	20 30 40	2 5 9	0.0318	0.0428	0.0545	0.0532	0.0517	0.0542	0.0503	0.0585
	20 30 40	9 5 2	0.1398	0.0965	0.0477	0.0518	0.0437	0.0511	0.0541	0.0557

3.2.2 Actual Type I Error Rates for g-h (g=0.5, h=0.2) Distribution (Asymmetric)

In this study, g-h distribution was used with g=0.5 and h=0.2 paramaters. That means the distribution is quite skewed. According to results of Table 3.3, WTR, WTRB, and B had the best control on the Type I error rate among all methods when the distribution was skewed. They had only one conservative result under equal sample sizes (20, 20, 20) and equal variances. On the other hand, F, K, and W gave liberal results under heteroscedasticity. As it can be seen, W did not perform well in contrast to its performance under normality. Furthermore, F had two conservative results besides its liberal results. If it is looked at the results of the methods for comparing medians and one-step M estimators (Mc-S and SH), they generally gave conservative results, especially Mc-S had an attidute conservatively for almost all settings of g-h (g=0.5, h=0.2) distribution.

Table 3.3 Actual Type I error rates for g-h (g=0.5, h=0.2) distribution

g-h (g=0.5, h=0.2)Distribution										
Groups	n	σ	F	K	W	WTR	WTRB	B	Mc-S	SH
k=3	20 20 20	1 1 1	0.0376	0.0433	0.0404	0.0374	0.0372	0.0334	0.0302	0.0276
	20 20 20	2 5 9	0.0770	0.0625	0.0850	0.0503	0.0463	0.0498	0.0391	0.0438
	30 30 30	1 1 1	0.0442	0.0514	0.0487	0.0463	0.0453	0.0435	0.0346	0.0384
	30 30 30	2 5 9	0.0757	0.0675	0.0854	0.0519	0.0489	0.0492	0.0382	0.0463
	20 30 40	1 1 1	0.0430	0.0507	0.0503	0.0476	0.0480	0.0407	0.0374	0.0375
	20 30 40	2 5 9	0.0393	0.0406	0.0710	0.0477	0.0482	0.0476	0.0344	0.0442
	20 30 40	9 5 2	0.1476	0.0992	0.0947	0.0562	0.0482	0.0491	0.0419	0.0464

3.2.3 Actual Type I Error Rates for Real Data Sets

In this part of the study, there are two applications about Parkinson disease that obtained from http:/www.archive.ics.uci.edu /ml/. The data set is collected voice samples from a set of speaking exercises of Parkinson patients. Two features of voice samples named as minimum pitch values and number of pulses are used in the simulation study. Table 3.4 shows some descriptive statistics for minimum pitch values and number of pulses of Parkinson patients. Sets are given by Figure 3.3.

Table 3.4 Descriptive statistics of real data sets

Data	n	Mean	Median	μ_{os}	S	Skewness	Kurtosis
Minium Pitch	1040	27.058	12.774	15.661	36.673	2.469	9.798
Number of Pulses	1040	234.876	199 886	207.018	121.541	1.154	3.056

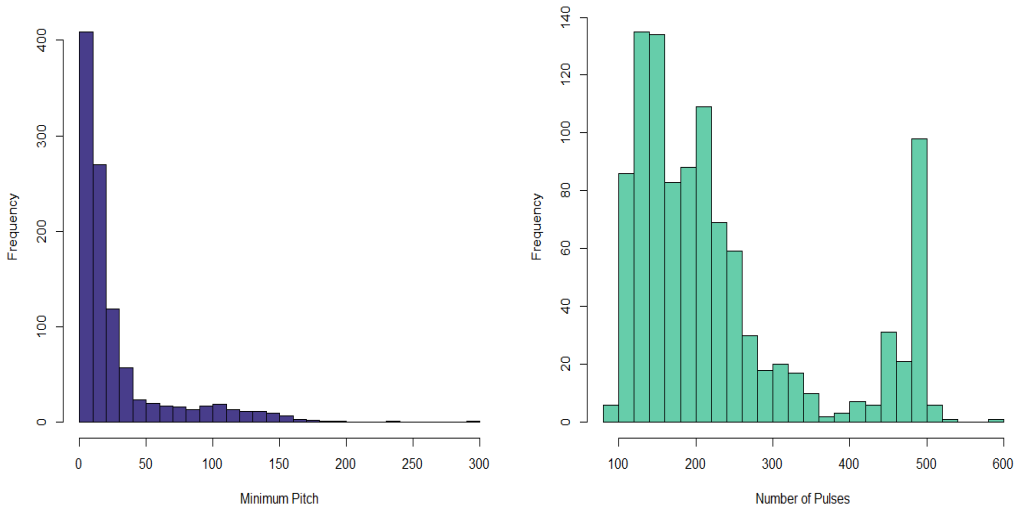


Figure 3.3 Histograms for real data sets

According to Table 3.5, the methods without robust estimator generally gave liberal results while the methods with robust estimators generally gave conservative results. Most of methods were adversely affected because of higly skewed distribution. However, F, K, WTRB, and B controlled Type I error better than the others for this data set. Unlike the results was obtained from the theoretical distributions, surprisingly, WTR had conservative results when there was homoscedasticity while it had liberal results under heteroscedasticity. Mc-S and SH gave conservative results under all settings for this data. These results show that Mc-S and SH tend to give conservative results when the distribution is highly skewed.

Table 3.5 Actual Type I error rates for minimum pitch values

Minimum pitch values										
Groups	n	σ	F	K	W	WTR	WTRB	B	Mc-S	SH
k=3	20 20 20	1 1 1	0.0410	0.0429	0.0589	0.0270	0.0218	0.0161	0.0172	0.0099
	20 20 20	2 5 9	0.0976	0.0878	0.1364	0.0757	0.0607	0.0447	0.0269	0.0260
	30 30 30	1 1 1	0.0434	0.0486	0.0599	0.0260	0.0201	0.0152	0.0168	0.0095
	30 30 30	2 5 9	0.0878	0.1032	0.1069	0.0648	0.0554	0.0403	0.0304	0.0287
	20 30 40	1 1 1	0.0401	0.0441	0.0677	0.0285	0.0241	0.0150	0.0171	0.0092
	20 30 40	2 5 9	0.0425	0.0612	0.0820	0.0489	0.0446	0.0310	0.0221	0.0220
	20 30 40	9 5 2	0.1583	0.1487	0.1284	0.0873	0.0671	0.0525	0.0355	0.0378

In Table 3.6, W had the best control on Type I error. It gave only two liberal results under heteroscedasticity. F and K gave liberal results again under heteroscedasticity. When comparing robust methods WTR had the best results and it could control Type I error rate under heteroscedasticity.

Table 3.6 Actual Type I error rates for number of pulses

Number of pulses										
Groups	n	σ	F	K	W	WTR	WTRB	B	Mc-S	SH
k=3	20 20 20	1 1 1	0.0485	0.0472	0.0531	0.0346	0.0257	0.0222	0.0262	0.0269
	20 20 20	2 5 9	0.0734	0.0791	0.0725	0.0494	0.0382	0.0403	0.0416	0.0710
	30 30 30	1 1 1	0.0514	0.0487	0.0555	0.0321	0.0244	0.0241	0.0303	0.0275
	30 30 30	2 5 9	0.0685	0.0865	0.0584	0.0470	0.0388	0.0390	0.0434	0.0424
	20 30 40	1 1 1	0.0500	0.0480	0.0594	0.0346	0.0290	0.0264	0.0326	0.0310
	20 30 40	2 5 9	0.0283	0.0518	0.0534	0.0389	0.0337	0.0302	0.0322	0.0332
	20 30 40	9 5 2	0.1399	0.1233	0.0698	0.0537	0.0394	0.0461	0.0555	0.0498

3.2.4 Power Results

As it was indicated in the beginning of 3.2, Bradley's criterion (0.04 - 0.06) was used for interpretting Type I error results. Eventually, power was calculated for W, WTR, WTRB, and B due to the performance of controlling Type I error rates. Furthermore, the effect size Δ was chosen 1 for normal and g-h distribution while it was chosen 25 and 100 for minimum pitch values and number of pulses respectively. Table 3.7, 3.8, 3.9, and 3.10 show the power results. Moreover, Figure 3.4, 3.5, 3.6, and 3.7 are line plots of the power results. In figures, settings in each line of the tables were numbered as 1, 2, 3, 4, 5, 6, 7 respectively on x axis.

In Table 3.7, W and WTR had the best power results among the methods. That can be seen in Figure 3.6. W and WTR generally gave closer results to 1 than the others. However, no method gave high power results under heteroscedasticity.

Table 3.7 Power results for normal distribution

Normal distribution								
Groups	Settings	n	σ	Δ	W	WTR	WTRB	B
k=3	1	20 20 20	1 1 1	1	0.8971	0.8221	0.7931	0.8291
	2	20 20 20	2 5 9	1	0.0972	0.0949	0.0808	0.0613
	3	30 30 30	1 1 1	1	0.9795	0.9555	0.9475	0.9574
	4	30 30 30	2 5 9	1	0.1310	0.1206	0.1095	0.0735
	5	20 30 40	1 1 1	1	0.9793	0.9536	0.9440	0.9566
	6	20 30 40	2 5 9	1	0.1246	0.1153	0.1106	0.0764
	7	20 30 40	9 5 2	1	0.1286	0.1190	0.0991	0.0749

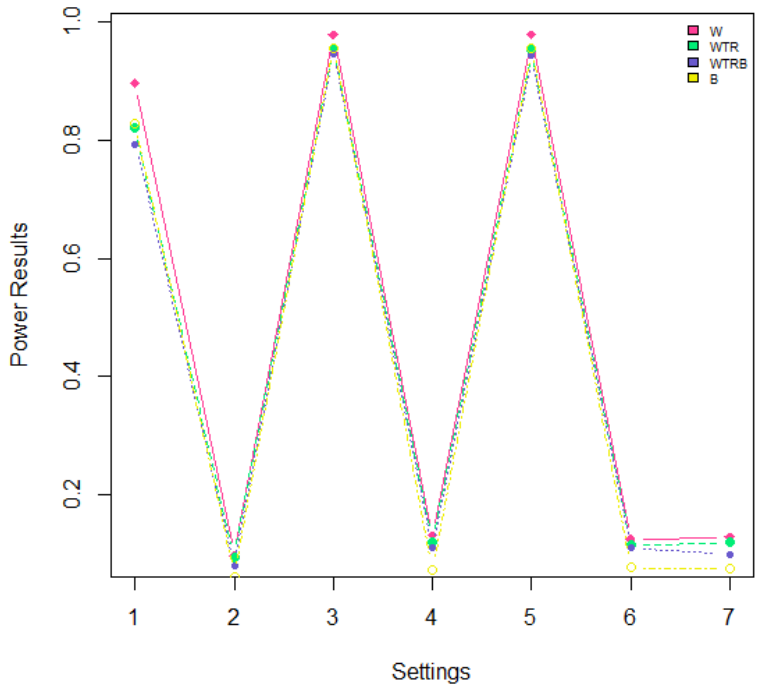


Figure 3.4 Line plot for power results of four methods under each setting in normal distribution

According to Table 3.8, WTR had the highest power results among these four robust methods. WTR, WTRB and B had higher results than W because it does not have a robust estimator. As it can be seen from Figure 3.5, WTR gave the best results.

Table 3.8 Power results for g-h (g=0.5, h=0.2) distribution

g-h (g=0.5, h=0.2) distribution								
Groups	Settings	n	σ	Δ	W	WTR	WTRB	B
k=3	1	20 20 20	1 1 1	1	0.5354	0.7527	0.7315	0.7031
	2	20 20 20	2 5 9	1	0.5952	0.8584	0.7987	0.4885
	3	30 30 30	1 1 1	1	0.6981	0.9257	0.9178	0.8987
	4	30 30 30	2 5 9	1	0.7728	0.9774	0.9655	0.6922
	5	20 30 40	1 1 1	1	0.6902	0.9232	0.9103	0.8872
	6	20 30 40	2 5 9	1	0.7699	0.9712	0.9646	0.7368
	7	20 30 40	9 5 2	1	0.7884	0.9803	0.9509	0.6342

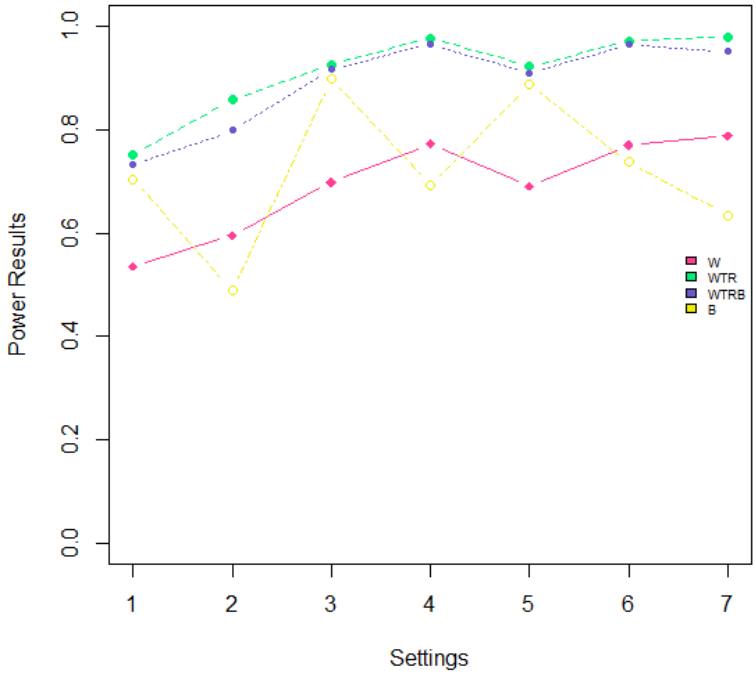


Figure 3.5 Line plot for power results of four methods under each setting in g-h (g=0.5, h=0.2) distribution

For the power results minimum pitch values indicated in Tablo 3.9, WTR had the best results among the methods. It gave higher power than the others even when the distribution is considerably skewed. Furthermore, WTRB had good power results as well.

Table 3.9 Power results for minimum pitch values

Minimum pitch values								
Groups	Settings	n	σ	Δ	W	WTR	WTRB	B
k=3	1	20 20 20	1 1 1	25	0.6721	0.9441	0.8763	0.8184
	2	20 20 20	2 5 9	25	0.1148	0.1027	0.0826	0.1008
	3	30 30 30	1 1 1	25	0.8406	0.9894	0.9687	0.9460
	4	30 30 30	2 5 9	25	0.1160	0.1398	0.1017	0.1275
	5	20 30 40	1 1 1	25	0.8840	0.9895	0.9575	0.9338
	6	20 30 40	2 5 9	25	0.0816	0.1356	0.1176	0.1186
	7	20 30 40	9 5 2	25	0.1185	0.1505	0.1011	0.1314

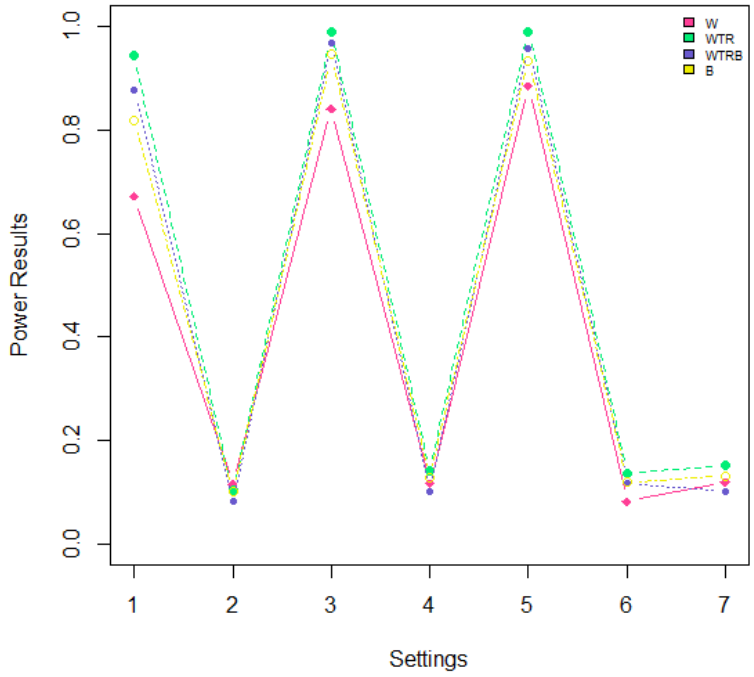


Figure 3.6 Line plot for power results of four methods for under setting in minimum pitch values

As it can be seen from Table 3.10, the results of W and WTR were higher than the others for number of pulses. Nevertheless, each method did not give a high power under heteroscedasticity.

Table 3.10 Power results for number of pulses

Number of pulses								
Groups	Settings	n	σ	Δ	W	WTR	WTRB	B
k=3	1	20 20 20	1 1 1	100	0.7797	0.7974	0.7220	0.6420
	2	20 20 20	2 5 9	100	0.0781	0.0628	0.0496	0.0685
	3	30 30 30	1 1 1	100	0.9368	0.9290	0.8977	0.8365
	4	30 30 30	2 5 9	100	0.0948	0.0763	0.0598	0.0832
	5	20 30 40	1 1 1	100	0.9331	0.9237	0.8793	0.8112
	6	20 30 40	2 5 9	100	0.0867	0.0745	0.0637	0.0777
	7	20 30 40	9 5 2	100	0.0906	0.0725	0.0486	0.0746

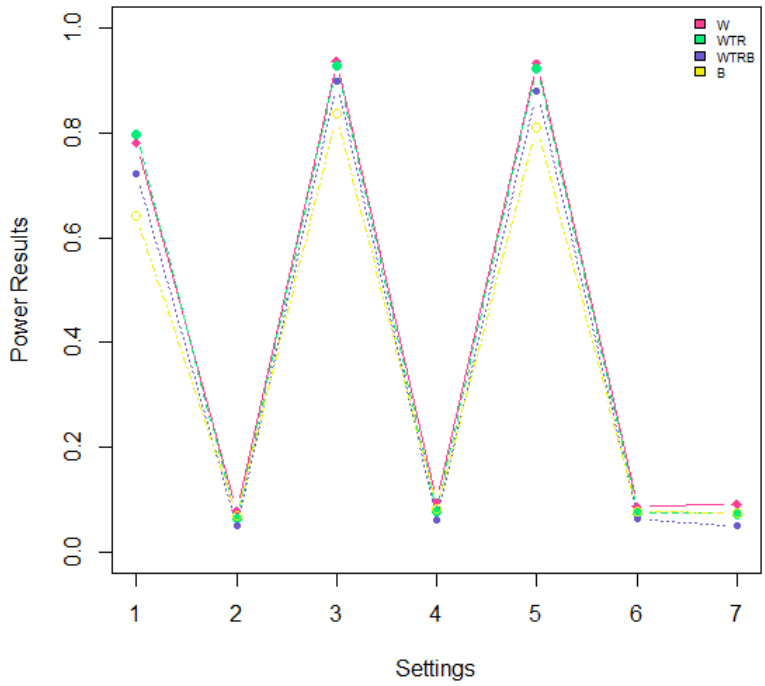


Figure 3.7 Line plot for power results of four methods under each setting in number of pulses

When comparing all power results of the methods, WTR generally had higher power than the others. However, the methods could not be so successful to give high power especially under heteroscedasticity.

CHAPTER FOUR
CONCLUSION

The most common method is ANOVA-F test for comparing k independent groups. However, the assumptions of ANOVA-F test should be satisfied in order to get realiable results.

In this study, seven different methods were investigated to find out some alternatives to be a solution for the problems of ANOVA-F test. ANOVA-F test and these seven methods were compared in terms of their Type I error rates under seven different settings for normal, g-h distribution and two real data sets. In other words, 28 different situations were examined for each method.

When the distribution was normal, most of methods could control Type I error rates. However, when there were unequal variances between groups, ANOVA-F and Kruskal-Wallis tests could not save the nominal Type I error.

For g-h (g=0.5 and h=0.2) distribution, the methods with trimmed mean (WTR, WTRB and B) could control Type I error rate better than the others. Also, W gave liberal results like F and K under heteroscedasticity when the distribution was more skewed.

The first real data set has considerably skewed and heavy tailed distribution. Hence, each method was extremely affected. However, F, K, WTRB and B controlled Type I error rate better than the others. Furthermore, B had the best control on Type I error under heteroscedasticity in spite of overly skewness.

The second real data set is skewed but not as much as the first one. Therefore, methods were not affected so much. The best method on controlling Type I error was W. When comparing only robust methods, WTR also had better results than the others.

According to Figure 4.1, W, WTR, and B could control Type I error 17, 17, and 18 times respectively under 28 settings. F (12), K (13), and W (11) had more liberal results than the others. Also, Mc-S (17) and SH (14) mostly gave conservative results under 28 settings. It can be said that the robust methods tend to give conservative results while the methods without robust estimator tend to give liberal results.

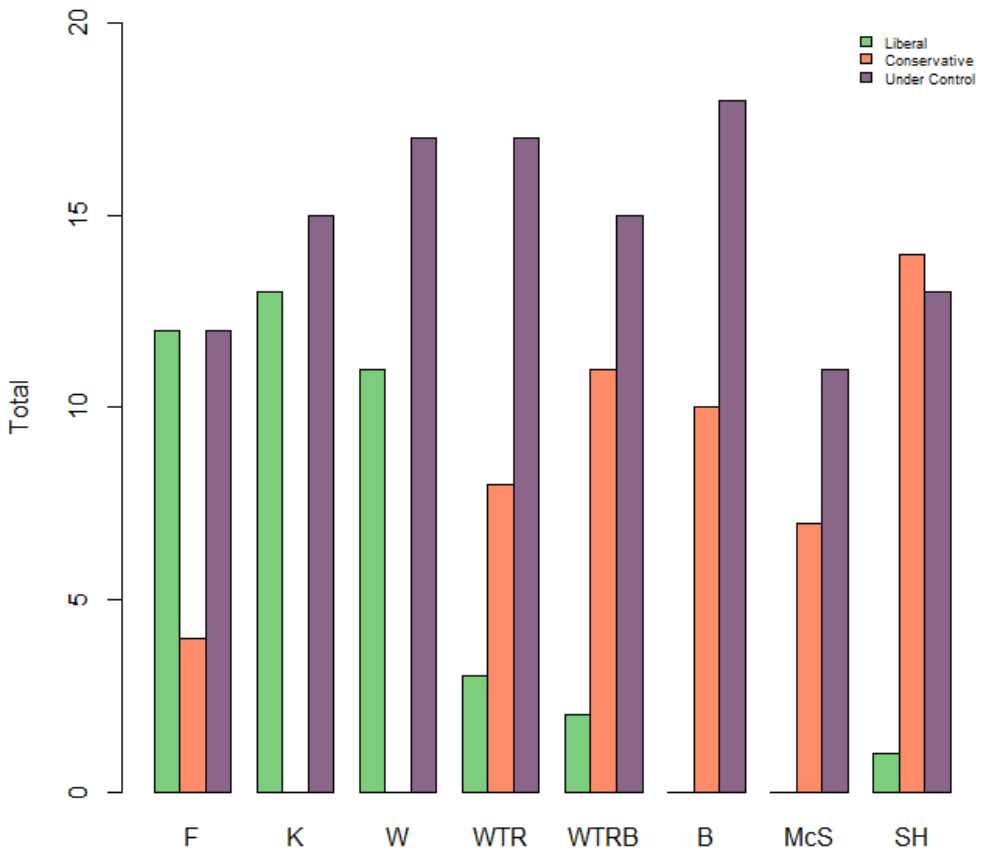


Figure 4.1 Bar plot for number of settings in terms of actual Type I error rates of the methods

According to power results, generally WTR had the best results among all methods. W and WTR gave higher power than the other methods under normality and also, they had good power results again for the data of number of pulses. On the other hand, W could not provide high power when the underlying normality assumption was violated. Minimum pitch values had quite higher skewness than number of pulses. Also g-h (g=0.5, h=0.2) distribution is highly skewed. Hence, its

power results were poor. However, WTR gave high power results even if the distribution was considerably skewed. Lastly, these four methods generally gave poor power results under heteroscedasticity.

If it is supposed to generalize the results and make a deduction, W (17), WTR (17) and B (18) controlled Type I error better than the other methods as it can be seen in Figure 4.1. Besides Type I error rates, WTR had better power results than W and B. As a result, WTR can be the most recommended method among the eight methods due to its performance of the Type I error rates and power under the settings for this study.

For further study, Schrader and Hettmansperger method can be studied with different robust estimators like Wilcox's modified one-step M estimator (Wilcox, 2012). Simulation design can be extended for larger number of groups and distributions with non-identical shapes.

REFERENCES

Behrens, W.V. (1929). Ein beitrag zur fehlerberechnung bei wenigen beobachtungen. *Landwirtsch Jahrbucher, 68,* 807-837.

Box, G.E.P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25,* 290-302.

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

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics, 3,* 1–21.

Erdoğdu, S. B., Isenkul, M., Sakar, C. O., Sertbaş, A., Gürgen, F., Delil, S., et al. (2013). *Collection and analysis of a parkinson speech dataset with multiple types of sound recordings.* Retrieved September 02, 2014 from http:/www.archive.ics.uci.edu /ml/

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Astronomical Society of London, Series, 222,* 309–368.

Fisher, R. A. (1935). The fiducial argument in statistical inference. *Annals of Eugenics,6,* 198-391.

Glass, G., Peckham, P., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analysis of variance and covariance. *American Educational Research Assocation, 42,* 237-288.

Hampel, F. R. (1968). *Contributions to the theory of robust estimation.* Unpublished PHD dissertation, Department of Statistics, University of California, Berkeley.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (1986). *Robust statistics*. New York: Wiley.

He, X.,Simpson, D. G., & Portnoy, S. L. (1990). Breakdown robustness of tests. *Journal of theAmerican Statistical Association, 85,* 446–452.

Hill, M., & Dixon, W. J. (1982). Robustness in real life: A study of clincial laboratory data. *Biometrics, 38,* 377–396.

Hoaglin, D. C., Mosteller, F., & Tukey, J. W. (1985). *Exploring data tables, trends and shapes.* New York: Wiley.

Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics, 35,* 73-101.

Huber, P. J. (1981). *Robust statistics*. New York: Wiley.

Keselman, H. J.,Rogan, J. C., & Feir-Walsh, B. J. (1977). An evaluation of somenon-parametric and parametric tests for location equality. *British Journal of Mathematical and Statistical Psychology, 30,* 213–221.

Krutchkoff, R. G. (1988). One-way fix effects analysis of variance when the error variances may be unequal. *Journal of Statistical Computation and Simulation, 30,* 259–271.

Lee, S., & Ahn, C.H. (2003). Modified ANOVA for unequal variances, *Communications in Statistics Simulations, 32,* 987-1004.

Lix, M. L., & Keselman, H. J. (1998). To trim or not to trim: Tests of location equality under heteroscedasticty and non-normality. *Educational and Psychological Measurement, 58,* 409-429.

McKean, J.W., & Schrader, R.M., (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics-Simulation and Computation. 13*, 751–773.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105(1),* 156-166.

Schrader, R. M., & Hettmansperger, T. P. (1980). Robust analysis of variance. *Biometrika, 67*, 93–101.

Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.

Staudte, R., G. & Sheather, S. J.,(1990). *Robust estimation and testing*. New York: Wiley.

Welch, B.L. (1947). The generalization of student's problem when several different population variances are involved. *Biometrika, 34*, 28-35.

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

Welsh, A. H., & Morrison, H. L. (1990). Robust L estimation of scale with an application in astronomy. *Journal of the American Statistical Association, 85,* 729-743.

Westfall, P. H., & Young, S. S. (1993). *Resampling based multiple testing.* New York: Wiley.

Wilcox, R. R. (1995a). ANOVA: A paradigm for low power and misleading measures of effect size? *American Educational Research Association, 65,* 51-77.

Wilcox, R. R., Charlin, V. L., & Thompson, K. L. (1986). New Monte Carlo results on the robustness of the ANOVA F, W, and *F*_ statistics. *Communications in Statistics–Simulation and Computation, 15*, 933–944.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing (3rd ed).* San Diago, California: AcademicsPress.

Wu, P. C. (2002). *Central limit theorem and comparing means, trimmed means one-step M-estimators and modified one-step M–estimators under non-normality*. Unpublished Doctoral Dissertation, Department of Education, University of Southern California.