DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

RESAMPLING METHODS IN HYPOTHESIS TESTING

by
Uğur BİNZAT

July, 2017
İZMİR

RESAMPLING METHODS IN HYPOTHESIS TESTING

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
Uğur BİNZAT

July, 2017
İZMİR

M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled **"RESAMPLING METHODS IN HYPOTHESIS TESTING"** completed by **UĞUR BİNZAT** under supervision of **ASST. PROF. DR. ENGİN YILDIZTEPE** 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.

Asst. Prof. Dr. Engin YILDIZTEPE

Supervisor

Assoc.Prof.Dr.Atila GÖKTAŞ

(Jury Member)

Assoc.Prof.Dr.A.Fırat Özdemir

(Jury Member)

Prof.Dr. Emine İlknur CÖCEN
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor Asst. Prof. Dr. Engin YILDIZTEPE for the continuous support of my study and research, for his patience, motivation, enthusiasm, and immense knowledge.

Besides my advisor, I would like to offer my special thanks to my dissertation committee members Assoc. Prof. Dr. A. Fırat ÖZDEMİR and Assoc. Prof. Dr. Atilla GÖKTAŞ for their insightful comments and precious advices.

I am also greatly indebted to my good friend Gözde NAVRUZ for her critical comments and suggestions during the phases of thesis work.

Last but not the least, I would like to express the deepest appreciation to my parents Ahmet BİNZAT and Zeliha BİNZAT, to my brother Murat BİNZAT and my sister Merve BİNZAT for supporting me and encouraging me throughout not only in writing this thesis but also all my life with their best wishes.

Uğur BİNZAT

RESAMPLING METHODS IN HYPOTHESIS TESTING

ABSTRACT

Making inference about the comparison of the parameters of the populations is one of the most common issues in statistics. It is well known that the conventional parametric methods are highly affected by violations of normality and homoscedasticity assumptions. Non-parametric tests are the most frequently used methods in the case of non-normality and heteroscedasticity. In addition to these methods, resampling based methods become more prominent since computers spread rapidly in last decades.

In this thesis, the resampling based methods and their applications on hypothesis testing are investigated. The investigated methods for two independent and three or more independent samples and their parametric and non-parametric counterparts are compared in terms of their ability to control type I error and powers via a Monte Carlo simulation study. The performance of the methods are investigated under various theoretical distributions and real data sets. The balanced and unbalanced designs and heteroscedasticity are also considered. In general, bootstrap methods based on trimmed mean give satisfactory results when controlling the type I error. However, no superiority is obtained between the investigated methods in terms of power estimates. The most powerful methods are varied from case to case.

Keywords: Resampling, hypothesis testing, permutation test, bootstrap, Monte Carlo simulation

HİPOTEZ TESTİNDE YENİDEN ÖRNEKLEME YÖNTEMLERİ

ÖZ

Kitle parametrelerinin karşılaştırılmasına ilişkin çıkarım yapmak, istatistikteki en yaygın problemlerden biridir. Geleneksel parametrik yöntemlerin normallik ve sabit varyanslılık varsayımlarının ihlalinden oldukça etkilendiği bilinmektedir. Parametrik olmayan testler, normallik ve sabit varyanslılık varsayımlarının ihlali durumunda en sık kullanılan yöntemlerdir. Bu yöntemlere ek olarak, bilgisayarların yaygınlaşması sebebiyle yeniden örnekleme yöntemleri de ön plana çıkmıştır.

Bu tezde, yeniden örnekleme yöntemleri ve hipotez testleri üzerine uygulamaları incelenmiştir. İki bağımsız ve üç veya daha fazla bağımsız örneklem için kullanılan yeniden örneklemeye dayalı yöntemler, I. tip hata ve güç açısından parametrik ve parametrik olmayan alternatifleriyle bir Monte Carlo simülasyon çalışması ile karşılaştırılmıştır. Yöntemlerin performansı dengeli ve dengesiz tasarımlar ile farklı varyanslılık da dikkate alınarak, farklı özellikteki teorik dağılımlar ve gerçek veri setleri için incelenmiştir. Genel olarak, budanmış ortalamaya dayalı bootstrap yöntemler, I. tip hatayı kontrol etmede daha başarılı sonuçlar vermiştir. Bununla birlikte, tahmin edilen test güçlerine göre yöntemler arasında üstünlük saptanamamıştır. En güçlü yöntem incelenen duruma göre değişmektedir.

Anahtar kelimeler: Yeniden örnekleme, hipotez testi, permütasyon testi, bootstrap, Monte Carlo benzetimi

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 - HISTORY OF RESAMPLING BASED METHODS 4

CHAPTER THREE - RESAMPLING BASED INFERENCE 6

3.1 Permutation Test 6
3.1.1 A Permutation Test Procedure for One Sample 6
3.1.2 A Permutation Test Procedure for Two Independent Samples 8
3.1.3 A Permutation Test Procedure for Two Dependent Samples 9
3.1.4 A Permutation Test Procedure for Three or More Independent Samples 10
3.2 The Bootstrap 12
3.2.1 A Bootstrap Test Procedure for One-Sample 13
3.2.2 A Bootstrap Test Procedure for Two Independent Samples 15
3.2.3 A Bootstrap Test Procedure for Two Dependent Samples 17
3.2.4 A Bootstrap Test Procedure for Three or More Independent Samples ... 18

CHAPTER FOUR - APPLICATION 23

4.1 Design of the Simulation Study 23
4.2 Results 25
4.2.1 Two Independent Samples 25
4.2.2 Three or More Independent Samples 29

CHAPTER FIVE - CONCLUSION 39

REFERENCES 42

APPENDICES 45

LIST OF FIGURES

Page

Figure 4.1 Histograms of the real data sets 24

Figure 4.2 Estimated powers - Normal (0,1) (homoscedastic case, $n_1 = n_2 = n_3 = 20$) 31

Figure 4.3 Estimated powers - Logistic distribution with $\mu = 0$, $\sigma = 1$ (homoscedastic case, $n_1 = n_2 = n_3 = 20$) 33

Figure 4.4 Estimated powers - Uniform distribution with $\mu = 0$, $\sigma = 1$ (homoscedastic case, $n_1 = n_2 = n_3 = 20$) 34

Figure 4.5 Estimated powers - Gamma (1.5,0.5) ($n_1 = n_2 = n_3 = 20$) 35

LIST OF TABLES

Page

Table 1.1 Possible outcomes in hypothesis testing 1
Table 4.1 Descriptive statistics of the real data sets..... 24
Table 4.2 Abbreviations of the methods 24
Table 4.3 Theoretical distributions for two independent samples 26
Table 4.4 Actual type I error rates and estimated powers when population distributions are Logistic 26
Table 4.5 Actual type I error rates and estimated powers when population distributions are Uniform..... 27
Table 4.6 Actual type I error rates and estimated powers when population distributions are Gamma 28
Table 4.7 Actual type I error rates and estimated powers - SAT Quantitative (two independent samples) 28
Table 4.8 Actual type I error rates and estimated powers - Total AA Degrees (two independent samples) 29
Table 4.9 Actual type I error rates - Normal $(0, \sigma)$..... 30
Table 4.10 Estimated powers - Normal $(0, \sigma)$ (heteroscedastic case) 30
Table 4.11 Actual type I error rates - Logistic distribution with $\mu = 0$ 31
Table 4.12 Estimated powers - Logistic distribution with $\mu = 0$ (heteroscedastic case) 32
Table 4.13 Actual type I error rates - Uniform distribution with $\mu = 0$ 33
Table 4.14 Estimated powers - Uniform distribution with $\mu = 0$ (heteroscedastic case) 34
Table 4.15 Actual type I error rates - Gamma (1.5,0.5) 35
Table 4.16 Estimated powers - Gamma 36
Table 4.17 Actual type I error rates - SAT Quantitative (three or more independent samples)..... 36
Table 4.18 Power Estimates (Δ=65) - SAT Quantitative (three or more independent samples)..... 37

Table 4.19 Actual type I error rates – Total AA Degrees (three or more independent samples) 37
Table 5.1 Suggested methods for two independent samples..... 40
Table 5.2 Suggested methods for three or more independent samples 41

CHAPTER ONE
INTRODUCTION

The statistical hypothesis testing is a method to make conclusions about the population parameter using the information gained from a sample. Null hypothesis and alternative hypothesis are the two statements of a statistical hypothesis. The null hypothesis (H_0) is a statement which is believed to be true. The alternative hypothesis (H_1 or H_A) is constructed oppositely to the null hypothesis and only one of them can be true at the same time.

Table 1.1 Possible outcomes in hypothesis testing

	H_0 is true	H_0 is false
Do not reject H_0	Right Decision	Wrong Decision (Type II Error)
Reject H_0	Wrong Decision (Type I Error)	Right Decision

Practitioners might encounter two types of errors in a hypothesis testing procedure. Type I error (false positive) arises if the null hypothesis is rejected which is actually true and Type II error (false negative) arises if the null hypothesis is not rejected which is actually false. Type I error rate or significance level is called probability of making a type I error and it is denoted by the Greek letter "α". Statistical power means probability of rejecting the null hypothesis which is actually false. The type II error rate is denoted by Greek letter "β" and the power of a test can be computed as (1- β).

Basically, a hypothesis testing can be conducted in 5 steps as follows,

1. Construct the null and alternative hypothesis statements,
2. Choose a proper test statistic,
3. Calculate the *p*-value for the test statistic,"
4. Reject the null hypothesis if *p*-value is equal or less than the significance level otherwise the null hypothesis cannot be rejected,

5. Make a conclusion.

The *p*-value is the probability of obtaining a result equal to or "more extreme" than the test statistic when the null statement is true. It is compared with significance level to make conclusions about the population parameter. In addition, confidence interval can be used instead of *p*-value as a testing tool and they are often used interchangeably.

Hypothesis testing for one, two and three or more samples have been applied and investigated for decades. Generally, measure of locations are used for dealing with one, two and three or more sample problems. For example, when the assumptions are met, one sample *t*-test is the most popular way to test whether a sample of observations has a specific mean. Also, comparing two independent samples is a statistical problem practitioners frequently encounter. A *t*-statistic computed from the difference of the independent sample means follows Student's *t* distribution if it is assumed that the populations are normally distributed and have a common variance. Welch's *t*-test, which is an adaptation of Student's *t*-test, is used when the population variances are not assumed to be equal. The one-way analysis of variance (ANOVA) is used to determine whether there is any statistically significant difference between the means of three or more independent groups. These are the most known parametric methods in statistics, but they are highly affected by violations of normality and homoscedasticity assumptions.

In general, non-parametric methods require less assumptions than the classical parametric methods. Median test is the most known alternative of the one sample *t*-test. The Wilcoxon-Mann-Whitney test (sometimes called The Mann-Whitney U test) is the most preferred non-parametric ranked based method to test whether two samples are likely to derive from the same population. Additionally, the Kruskal-Wallis test is the most known ranked based counterpart of the ANOVA.

Resampling methods are statistical procedures that use sample data repeatedly for the purpose of statistical inference. They are computer intensive methods and can be applied to almost all statistical problems. These methods can be used to estimate the

bias of an estimator, construct confidence intervals and test statistical hypothesis about the parameter being estimated (Chernick, 2012). Permutation tests, the jackknife, the bootstrap and the cross-validation are the most known resampling methods.

Permutation tests, also known as “randomization tests” or “exact tests”, were developed by R. A. Fisher (1935) and E. J. G. Pitman (1937). They are known as the oldest resampling based methods. The jackknife, was suggested by Quenouille (1949) and extended by Tukey (1958). Another resampling based method, the bootstrap, was introduced in 1979 by Bradley Efron as an alternative method to the jackknife. Cross-validation is a resampling method to assess the adequacy of a statistical model. Basically, this approach splits the data into parts and uses one part to fit the model and predicts the accuracy of the fitted model with the other parts of the data.

In this thesis, the resampling methods and their applications on hypothesis testing are investigated. The investigated methods for two independent samples, three or more independent samples and their popular non-parametric counterparts are compared in terms of test power and their ability to control the probability of type I error via a Monte Carlo simulation study. This thesis is organized in five chapters where the history of resampling based methods are given in chapter two, the methods are described in chapter three, the design of the simulation study and its results and are given in chapter four. Finally, concluding remarks are given in chapter five.

CHAPTER TWO
HISTORY OF RESAMPLING BASED METHODS

Due to its tedious process, resampling could not be applied easily by hand and so the importance of the resampling methods has increased especially after the 1990's when computers spread rapidly.

Permutation tests have evolved from R. Fisher's and E. J. G. Pitman's works in the early 1930's and they are based on randomly rearranging the sample observations without replacement. Edgington (1995) states that the terms "Permutation test" and "Randomization test" are often used interchangeably and Ersnt (2004) briefly explains the difference of these terms.

The jackknife method was first suggested by Quenouille (1949, 1956) for bias estimation and Tukey (1958) used it to estimate the variance. Miller (1974) gives an extended review about this method. In a nutshell, the jackknife generates subsamples by discarding each observation in a sample and the desired statistic is computed from these subsamples. This process is very similar to the mechanism of the Swiss Army knife and the name was given by Tukey after he got inspired by this similarity. Though the jackknife is a currently used method for solving some statistical problems, it gets less use since the bootstrap is becoming more popular (Chernick, 2012).

The bootstrap, which has a wide range of application areas in statistics, was suggested by Efron (1979). Bootstrapping can refer to any test that relies on random sampling with replacement. The bootstrap uses synthetically generated sampling distribution of an estimated parameter to obtain confidence interval and standard error of parameters like mean, median, proportion and correlation coefficient. Different from resampling in permutation test structure, bootstrap is a method based on sampling with replacement. Thus, sampling with bootstrap produces more varied generated samples than permutation methods.

Efron (1982) publishes the first monograph on the bootstrap. After a while, Efron recognizes the implementation of bootstrap to confidence intervals, hypothesis testing and more complex problems. Efron & Gong (1983), Diaconis & Efron (1983), and Efron & Tibshirani (1986) consider these topics. In the 1990's there were a large number of studies related to bootstrap and other resampling methods. Hall & Wilson (1991) recommend two guidelines for nonparametric bootstrap hypothesis testing. The first one is about conducting the resampling in a way that reflects the null hypothesis and the second one is about using pivotal statistic, which means correcting for scale. Hall (1992) gives a detailed study about the connection between the bootstrap and Edgeworth expansion. Westfall & Young (1993) show a variety of methods and applications for resampling in multiple testing. Shao & Tu (1995) give a theoretical review of the bootstrap and the jackknife. Davison & Hinkley (1997) give an extensive overview and consider important topics on the bootstrap. Martin (2007) discusses bootstrap hypothesis testing in detail for some popular statistical problems such as hypothesis testing of correlation coefficient, variance, two sample means, regression models and ANOVA models. A wide range of applications and simulation studies of the bootstrap and permutation tests on robust methods can be found in the book by Wilcox (2017). Ozdemir et al. (2017) gives an extensive simulation study about the robust methods for three or more than independent samples.

CHAPTER THREE
RESAMPLING BASED INFERENCE

In this section, the implementations of two resampling based methods for hypothesis testing are explained.

3.1 Permutation Test

Permutation tests are based on resampling and the samples are drawn without replacement. Many non-parametric tests (independence, location, etc.) can be implemented as permutation tests. For two independent populations case, under the assumption that the null hypothesis is true (two independent population means are equal), reassigning the observations to the samples randomly between groups should not cause a problem. Thus, the desired statistic can be calculated by repeating this resampling process for all possible arrangements.

The computation of all possible permutations can be a time consuming process even for moderate sample sizes. This exact computation can be approximated by taking a large number of *B* random permutations. There is no strict rule for the amount of *B* but Davison & Hinkley (1997) indicates that 999 permutations should suffice.

3.1.1 A Permutation Test Procedure for One Sample

Assume that $(X_1,...,X_n)$ is a random sample of size *n* from a symmetric distribution *X*. The null hypothesis is $H_0:\theta=\theta_0$, against the alternative $H_1:\theta\neq\theta_0$. Say, the test statistic is mean, under the null, the mean of the deviations of θ about the θ_0 should be close to zero, because positive and negative deviations would cancel each other (Good, 2004). Maindonald & Braun (2003) describes a permutation test procedure for α significance level as below,

1. Shift the sample about θ_0,

$$d_i = X_i - \theta_0, \qquad (i = 1,...,n) \tag{3.1}$$

2. Compute the mean from the shifted observations,

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \tag{3.2}$$

3. Pool the plus and minus signs of the $|d_i|$'s and determine the number of possible permutations $(C = 2^n)$ of size *n*,

4. For each of the *C* permutations (or *B* random permutations), compute the test statistic, $\hat{\delta}^{(1)},...,\hat{\delta}^{(C)}$,

5. The exact upper *p*-value is computed as below,

$$\hat{p} = \frac{\#(\hat{\delta}^{(c)} \geq \bar{d})}{C}, \qquad (c = 1,...,C) \tag{3.3}$$

The approximate *p*-value is,

$$\hat{p} = \frac{\#(\hat{\delta}^{(b)} \geq \bar{d}) + 1}{B + 1}, \qquad (b = 1,...,B) \tag{3.4}$$

The *p*-value for two sided test is,

$$\hat{p}_t = \min(\hat{p}, 1 - \hat{p}) \tag{3.5}$$

6. Reject H_0 if $2\hat{p}_t \leq \alpha$ for two sided test.

3.1.2 A Permutation Test Procedure for Two Independent Samples

Suppose $(X_{11},...,X_{1n_1})$ and $(X_{21},...,X_{2n_2})$ are independent random samples from the unknown populations X_1 and X_2, respectively. The aim is to test whether the two population means are equal or not. The null hypothesis is $H_0 : \mu_1 = \mu_2$, against the alternative $H_1 : \mu_1 \neq \mu_2$. Wilcox (2017) describes a permutation test procedure for α significance level as below,

1. Compute the difference between the sample means, $d = \bar{X}_1 - \bar{X}_2$, where,

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}, \qquad (i = 1, 2) \tag{3.6}$$

2. Pool the data,

3. Consider any permutation of the pooled data, compute the sample mean of the first n_1 observations, then compute the sample mean using the remaining n_2 observations, finally compute the difference between these sample means,

4. Repeat the previous step for all possible C permutations (or B random permutations), and obtain C differences, $\hat{\delta}^{(1)},...,\hat{\delta}^{(C)}$,

$$C = \frac{(n_1 + n_2)!}{n_1! n_2!} \tag{3.7}$$

5. Put these C differences in ascending order, $\hat{\delta}^{(1)} \leq ... \leq \hat{\delta}^{(C)}$,

6. Reject the null hypothesis if $d < \hat{\delta}^{(\ell+1)}$ or if $d > \hat{\delta}^{(u)}$, where $\ell = \alpha C / 2$, rounded to the nearest integer and $u = C - \ell$.

In addition, *p*-value can be used to test the null hypothesis. The exact upper *p*-value is computed as below,

$$\hat{p} = \frac{\#(\hat{\delta}^{(c)} \geq d)}{C}, \qquad (c = 1, ..., C) \tag{3.8}$$

The approximate upper *p*-value is,

$$\hat{p} = \frac{\#(\hat{\delta}^{(b)} \geq d) + 1}{B + 1}, \qquad (b = 1, ..., B) \tag{3.9}$$

The p-value for two sided test is,

$$\hat{p}_t = \min(\hat{p}, 1 - \hat{p}) \tag{3.10}$$

H_0 is rejected if $2\hat{p}_t \leq \alpha$ for two sided test.

3.1.3 A Permutation Test Procedure for Two Dependent Samples

Suppose $(X_{11}, X_{21}), \ldots, (X_{1n}, X_{2n})$ are paired random samples of size *n*. The aim is to test whether the two population means are equal or not. The null hypothesis is $H_0 : \mu_1 = \mu_2$, against the alternative $H_1 : \mu_1 \neq \mu_2$. Butar & Bandulasiri (2009) describes a permutation test procedure at α significance level as below,

1. Obtain the differences from the paired observations, and compute the mean of the differences,

$$d_i = X_{1i} - X_{2i}, \qquad (i = 1, ..., n) \tag{3.11}$$

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n} \tag{3.12}$$

2. Pool the plus and minus signs of the $|d_i|$'s and determine the number of possible permutations $(C = 2^n)$ of size *n*,

3. For each of the C permutations (or B random permutations), compute the test statistic, $\hat{\delta}^{(1)},...,\hat{\delta}^{(C)}$,

4. The exact upper p-value is computed as below,

$$\hat{p} = \frac{\#(\hat{\delta}^{(c)} \geq \bar{d})}{C}, \qquad (c = 1,...,C) \tag{3.13}$$

The approximate upper p-value is,

$$\hat{p} = \frac{\#(\hat{\delta}^{(b)} \geq \bar{d}) + 1}{B + 1}, \qquad (b = 1,...,B) \tag{3.14}$$

The p-value for two sided test is,

$$\hat{p}_t = \min(\hat{p}, 1 - \hat{p}) \tag{3.15}$$

5. Reject H_0 if $2\hat{p}_t \leq \alpha$ for two sided test.

3.1.4 A Permutation Test Procedure for Three or More Independent Samples

ANOVA is a test procedure which is used to determine whether the differences of the means of three or more independent groups are statistically significant. Let Y_{ij} be the responses where $i = 1,...,k$ and $j = 1,...,n_i$ are the group numbers and sample sizes, respectively. The ANOVA-F statistic is,

$$F = \frac{\sum_{i=1}^{k} n_i \left(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet}\right)^2 \Big/ (k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(Y_{ij} - \bar{Y}_{i\bullet}\right)^2 \Big/ (N-k)} \tag{3.16}$$

where, $\bar{Y}_{i\bullet}$ is the mean of the *i*th group, $\bar{Y}_{\bullet\bullet}$ is the overall mean, and $N = n_1 + n_2 + ... + n_k$ is the total observation number. The null hypothesis is $H_0 : \mu_1 = \mu_2 = ... = \mu_k$, against the alternative,

$$H_1 : \mu_i \neq \mu_l, \quad \text{for at least one pair of } (i,l) \quad (i = 1,...,k;\ l = 1,...,k;\ i \neq l) \tag{3.17}$$

Ersnt (2004) describes a permutation test procedure for α significance level as below,

1. Compute the *F*-statistic in equation (3.16), from the Y_{ij}'s and call this statistic F,
2. Pool the data,
3. Consider the all possible permutations from the pooled data,

$$C = \frac{N!}{n_1!n_2!...n_k!}, \qquad (i = 1,...,k) \tag{3.18}$$

Then compute the test statistic from these permutations and call these test statistics $F^{*(c)}\ (c = 1,...,C)$. The star notation indicates that F^* is computed from a permuted Y_{ij},

4. The exact *p*-value is computed as below,

$$\hat{p} = \frac{\#\left(F^{*(c)} \geq F\right)}{C}, \qquad (c = 1,...,C) \tag{3.19}$$

The approximate *p*-value is,

$$\hat{p} = \frac{\#\left(F^{*(b)} \geq F\right) + 1}{B + 1}, \qquad (b = 1,...,B) \tag{3.20}$$

5. Reject H_0 if $\hat{p} \leq \alpha$.

3.2 The Bootstrap

The bootstrap can be applied as parametric bootstrap or non-parametric bootstrap, but non-parametric bootstrap is more popular than the parametric one. In this section, we only consider non-parametric bootstrap.

Suppose $x = (x_1, ..., x_n)$ is an observed random sample of size *n* from a distribution with cumulative density function (cdf) $F(x)$. If X^* is selected randomly from the *x*, that is,

$$P(X^* = x_i) = \frac{1}{n}, \qquad i = 1, ..., n \tag{3.21}$$

Bootstrap resampling generates a random sample $(X_1^*, ..., X_n^*)$ of size *n* by sampling with replacement from *x* and the star notation represents that X^* is the resampled version of *x*. The random variables X_i^* are iid (independent and identically distributed) and uniformly distributed on the set $x = (x_1, ..., x_n)$.

$F_n(x)$ is the distribution function of a random variable, where the random variable is uniformly distributed on the set $(x_1, ..., x_n)$. The empirical cumulative distribution function (ecdf) $F_n(x)$ is an estimator of $F(x)$ where $F_n(x)$ is a sufficient statistic for $F(x)$. In bootstrap there are two approximations. The ecdf F_n is the approximation to the cdf F_X and the ecdf F_n^* of bootstrap replicates is an approximation to the ecdf F_n. Resampling from the sample *x* is equivalent to generating random samples from the distribution $F_n(x)$ (Rizzo, 2007).

Diagram below shows the two approximations,

$$F \rightarrow X \rightarrow F_n \tag{3.22}$$

$$F_n \rightarrow X^* \rightarrow F_n^* \tag{3.23}$$

A bootstrap sample of $x = (x_1, ..., x_n)$ can be generated in two steps below,

1. Generate n random integers $i_1, ..., i_n$ uniformly distributed on $1, ..., n$,

2. Select the bootstrap sample based on the random integers,

$$x^* = (x_{i_1}, ..., x_{i_n}) \tag{3.24}$$

Suppose θ is the parameter of interest, and the estimator of θ is $\hat{\theta}$. The bootstrap estimate of the distribution of $\hat{\theta}$ is obtained as follows,

1. Generate the bootstrap sample of $x = (x_1, ..., x_n)$, that is, $x^* = (x_1^*, ..., x_n^*)$,

2. Compute the $\hat{\theta}$ from the bootstrap sample,

3. Repeat the previous steps *B* times and obtain a set of estimated $\hat{\theta}$ from the bootstrap samples. The empirical distribution of the replicates $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(B)}$ is the bootstrap estimate of $F_{\hat{\theta}}(.)$.

There is no strict rule for the amount of *B*. Davison & Hinkley (1997) indicates that 999 bootstrap resamples should suffice.

3.2.1 A Bootstrap Test Procedure for One-Sample

Suppose $(X_1, ..., X_n)$ is a random sample of size *n* from an unknown population *X*. The null hypothesis is $H_0 : \theta = \theta_0$, against the alternative $H_1 : \theta \neq \theta_0$. Say, θ is the unknown population mean. Efron & Tibshirani (1993) describes a bootstrap test procedure for α significance level as below,

1. Compute the *t*-statistic from the random sample,

$$t = \frac{\bar{X} - \theta_0}{s/\sqrt{n}} \tag{3.25}$$

where,

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{3.26}$$

$$s^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2}{n-1} \tag{3.27}$$

2. Transform the sample mean to make it equal to the population mean in order to satisfy the null hypothesis,

$$X_i' = X_i - \bar{X} + \theta_0, \qquad (i = 1, ..., n) \tag{3.28}$$

The prime notation indicates that X' is the transformed form of *X*,

3. Bootstrap the sample $X_1', ..., X_n'$ and compute the *t* statistic from the bootstrap sample X'^*, that is,

$$t^* = \frac{\bar{X}'^* - \theta_0}{s_{X'^*}/\sqrt{n}} \tag{3.29}$$

where,

$$\bar{X}'^* = \frac{\sum_{i=1}^{n} X_i'^*}{n} \tag{3.30}$$

$$s_{X'^*}^2 = \frac{\sum_{i=1}^{n} \left(X_i'^* - \bar{X}'^* \right)^2}{n-1} \tag{3.31}$$

4. Repeat the step-3 *B* times and obtain a set of test statistics, $t^{*(1)},...,t^{*(B)}$,

5. The upper *p*-value is computed as below,

$$\hat{p} = \frac{\#\left(t^{*(b)} \geq t\right)+1}{B+1}, \qquad (b=1,...,B) \tag{3.32}$$

The *p*-value for two sided test is,

$$\hat{p}_t = \min(\hat{p}, 1-\hat{p}) \tag{3.33}$$

6. Reject H_0 if $2\hat{p}_t \leq \alpha$ for two sided test.

3.2.2 A Bootstrap Test Procedure for Two Independent Samples

Suppose $(X_1,...,X_n)$ and $(Y_1,...,Y_m)$ are independent random samples of size *n* and *m* from the unknown populations *X* and *Y*, respectively. In general, the aim is to test whether the two population means are equal ($D_0 = 0$) or not. The null hypothesis is $\mathrm{H}_0 : \mu_X - \mu_Y = D_0$, against the alternative $\mathrm{H}_1 : \mu_X - \mu_Y \neq D_0$. Hall & Martin (1988) describes a bootstrap test procedure for α significance level as below,

1. Compute the *T*-statistic from the random samples,

$$T = \frac{\left(\bar{X} - \bar{Y} - D_0\right)}{\sqrt{\left(n^{-1}s_X^2 + m^{-1}s_Y^2\right)}} \tag{3.34}$$

Which is also known as the Welch *t*-statistic and it is suggested to remedy Behrens-Fisher problem,

2. Obtain the two collections, $Y' = \{Y_1 - \bar{Y} + \bar{X} - D_0, ..., Y_m - \bar{Y} + \bar{X} - D_0\}$ and $X' = \{X_1, ..., X_n\}$ in order to satisfy the null hypothesis and to preserve the characteristics of the samples,

3. Bootstrap X' and Y' separately, and calculate the *T*-statistic, that is,

$$T^* = \frac{\left(\bar{X}'^* - \bar{Y}'^* - D_0\right)}{\sqrt{\left(n^{-1}s^2_{X'^*} + m^{-1}s^2_{Y'^*}\right)}} \tag{3.35}$$

4. Repeat the previous step *B* times and obtain a set of test statistics, $T^{*(1)}, ..., T^{*(B)}$,

5. The upper *p*-value is computed as below,

$$\hat{p} = \frac{\#\left(T^{*(b)} \geq T\right) + 1}{B + 1}, \qquad (b = 1, ..., B) \tag{3.36}$$

The *p*-value for two sided test is,

$$\hat{p}_t = \min(\hat{p}, 1 - \hat{p}) \tag{3.37}$$

6. Reject H_0 if $2\hat{p}_t \leq \alpha$ for two sided test.

In addition to the above method, trimmed mean can be used to compare two independent samples. The trimmed mean of the sample $(X_1, ..., X_n)$ can be computed in three steps. First, sort the sample in ascending order, $X_{(1)} \leq ... \leq X_{(n)}$. The indices in parentheses represent the orders of the observations. The amount of trimming, γ, can be between 0 and 0.5. Let,

$$g = [\gamma n] \tag{3.38}$$

Where, g is rounded down to the nearest integer. The trimmed mean of sample $(X_1, ..., X_n)$ can be computed as below,

$$\bar{X}_t = \frac{X_{(g+1)} + ... + X_{(n-g)}}{n-2g} \tag{3.39}$$

The aim is to compare the trimmed means of the independent samples where the null hypothesis is $H_0 : \mu_{tX} = \mu_{tY}$, against the alternative $H_1 : \mu_{tX} \neq \mu_{tY}$. Wilcox (2017) describes a bootstrap test procedure for α significance level as below,

1. Let $\hat{\theta}_X^*$ the sample trimmed mean computed from the bootstrap sample of $(X_1,...,X_n)$ and $\hat{\theta}_Y^*$ the sample trimmed mean computed from the bootstrap sample of $(Y_1,...,Y_m)$, and calculate, $D^* = \hat{\theta}_X^* - \hat{\theta}_Y^*$,

2. Repeat the previous step *B* times and obtain a set of differences, $D^{*(1)},...,D^{*(B)}$,

3. The upper *p*-value is computed as below,

$$\hat{p} = \frac{\#\left(D^{*(b)} > 0\right)}{B}, \qquad (b = 1,...,B) \tag{3.40}$$

The *p*-value for two sided test is

$$\hat{p}_t = \min(\hat{p}, 1-\hat{p}) \tag{3.41}$$

4. Reject H_0 if $2\hat{p}_t \leq \alpha$ for two sided test.

3.2.3 A Bootstrap Test Procedure for Two Dependent Samples

Suppose $(X_{11}, X_{21}),\ldots, (X_{1n}, X_{2n})$ are paired random samples of size *n*. Let θ_j be the parameter of interest which is obtained from *j*th group (*j*=1,2). The null hypothesis is $\mathrm{H}_0 : \theta_1 - \theta_2 = 0$, against the alternative $\mathrm{H}_1 : \theta_1 - \theta_2 \neq 0$. Wilcox (2017) describes a bootstrap test procedure for α significance level as below,

1- Bootstrap the sample pairs and obtain bootstrapped random pairs,

$$\begin{matrix} (X_{11}^{*}, X_{21}^{*}) \\ \vdots \\ (X_{1n}^{*}, X_{21}^{*}) \end{matrix} \tag{3.42}$$

2. Calculate the $\hat{\theta}_1^*$ from the first group generated by bootstrap and calculate the $\hat{\theta}_2^*$ from the second group generated by bootstrap, and then calculate the test statistic,

$$d^* = \hat{\theta}_1^* - \hat{\theta}_2^* \tag{3.43}$$

3. Repeat the previous steps *B* times and obtain a set of test statistic, $d^{*(1)},...,d^{*(B)}$,

4. The upper *p*-value is computed as below,

$$\hat{p} = \frac{\#\left(d^{*(b)} > 0\right)}{B}, \qquad \left(b = 1,...,B\right) \tag{3.44}$$

The *p*-value for two sided test is,

$$\hat{p}_t = \min(\hat{p}, 1-\hat{p}) \tag{3.45}$$

4. H_0 is rejected if $2\hat{p}_t \le \alpha$ for two sided test.

3.2.4 A Bootstrap Test Procedure for Three or More Independent Samples

Let Y_{ij} be the responses where $i = 1,...,k$ and $j = 1,...,n_i$ are the group numbers and sample sizes, respectively. The null hypothesis is $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ against the alternative,

$$H_1 : \mu_i \neq \mu_l, \quad \text{for at least one pair of } (i,l) \quad (i = 1,...,k;\ l = 1,...,k;\ i \neq l) \tag{3.46}$$

Martin (2007) describes a bootstrap test procedure for α significance level as below,

1. Compute the *F*-statistic in equation (3.16) from the Y_{ij}'s and call this statistic F,

2. Shift the observations around the overall mean, so that resampling in this way satisfies the null hypothesis,

$$Y_{ij}^{'} = Y_{ij} - \bar{Y}_{i\bullet} + \bar{Y}_{\bullet\bullet}, \qquad \left(i = 1,...,k;\ j = 1,...,n_i\right) \tag{3.47}$$

where, $\bar{Y}_{i\bullet}$ is the mean of the *i*th group and $\bar{Y}_{\bullet\bullet}$ is the overall mean,

3. Bootstrap the first transformed sample $Y_{1j}^{'}$, and call this bootstrap sample, $Y_{1j}^{'*}$. Perform this process for the other groups separately and obtain the bootstrap resamples $Y_{ij}^{'*}$,

4. Compute the *F*-statistic from the resampled data,

5. Repeat the step-3 and step-4 *B* times, and obtain a set of test statistics, $F^{*(1)},...,F^{*(B)}$,

6. Compute the *p*-value,

$$\hat{p} = \frac{\#\left(F^{*(b)} \geq F\right) + 1}{B + 1}, \qquad \left(b = 1,...,B\right) \tag{3.48}$$

7. H_0 is rejected if $\hat{p} \leq \alpha$.

In addition, trimmed mean can be used to compare three or more independent groups. Let Y_{ij} be the responses where $i = 1,...,k$ and $j = 1,...,n_i$ are the group numbers and sample sizes, respectively. The null hypothesis is $H_0 : \mu_{t1} = \mu_{t2} = ... = \mu_{tk}$, against the alternative,

$$H_1 : \mu_{ti} \neq \mu_{tl}, \quad \text{for at least one pair of } (i,l) \quad (i = 1,...,k;\ l = 1,...,k;\ i \neq l) \tag{3.49}$$

The Winsorized mean and variance are some robust measures of location and scale, respectively. The Winsorized variance is required in computation of equation (3.53) and Winsorization of a sample $(X_1,...,X_n)$ can be done as below,

$$W_p = \begin{cases} X_{(g+1)}, & if\ X_p \leq X_{(g+1)} \\ X_p, & if\ X_{(g+1)} < X_p < X_{(n-g)} \\ X_{(n-g)}, & if\ X_p \geq X_{(n-g)} \end{cases}, \qquad p = 1,...,n \tag{3.50}$$

The computation of g is given in equation (3.38). Differently from trimming, the trimmed observations are replaced by the last observations in each side. The Winsorized mean and the variance of the sample given above is computed as follows,

$$\bar{X}_w = \sum_{p=1}^{n} \frac{W_p}{n} \tag{3.51}$$

$$s_w^2 = \sum_{p=1}^{n} \frac{\left(W_p - \bar{X}_w\right)^2}{n-1}, \qquad \left(p = 1,...,n\right) \tag{3.52}$$

Wilcox (2017) describes a bootstrap test procedure for trimmed means as below,

The test statistic is,

$$d_i = \frac{\left(n_i - 1\right) s_{wi}^2}{h_i \left(h_i - 1\right)}, \qquad \left(i = 1,...,k\right) \tag{3.53}$$

The weights w_i are evaluated by using h_i and s_{wi}^2, where h_i is the number of observations remain after trimming and s_{wi}^2 is the Winsorized variance of the *i*th group.

$$w_i = \frac{1}{d_i} \tag{3.54}$$

$$U = \sum_{i=1}^{k} w_i \tag{3.55}$$

$$\tilde{Y} = \frac{1}{U} \sum_{i=1}^{k} w_i \bar{Y}_{ti\bullet} \tag{3.56}$$

Where $Y_{ti\bullet}$ is the trimmed mean of *i*th group and $\tilde{Y}$ is the weighted overall trimmed mean.

$$A = \frac{1}{k-1} \sum_{i=1}^{k} w_i \left(\bar{Y}_{ti\bullet} - \tilde{Y} \right)^2 \tag{3.57}$$

$$B = \frac{2(k-2)}{k^2 - 1} \sum_{i=1}^{k} \frac{\left(1 - \frac{w_i}{U}\right)^2}{h_i - 1} \tag{3.58}$$

The test statistic is,

$$F_t = \frac{A}{1+B} \tag{3.59}$$

F_t is approximately follows *F* distribution with degrees of freedom,

$$v_1 = k - 1 \tag{3.60}$$

$$v_2 = \left[\frac{3}{k^2 - 1} \sum_{i=1}^{k} \frac{\left(1 - \frac{w_j}{U}\right)^2}{h_i - 1} \right]^{-1} \tag{3.61}$$

When there is no trimming this test procedure is equivalent to Welch's (1951) adjusted degrees of freedom method for means.

The bootstrap procedure is given below,

1. Bootstrap the first group, Y_{1j}, and call this bootstrapped sample Y_{1j}^*. Perform this process for the other samples separately and obtain the bootstrap resamples Y_{ij}^*,

2. In order to satisfy the null hypothesis, set,

$$C_{ij}^* = Y_{ij}^* - \bar{Y}_{ti\bullet}, \qquad i = 1,...,k \tag{3.62}$$

So that, $C_{i1}^*,...,C_{in_i}^*$ represents the *i*th bootstrap sample from a distribution with trimmed mean equals zero. Compute the test statistic in equation (3.59), from the C_{ij}^* values and call this statistic F_t^*.

3. Repeat the step-1 and step-2 *B* times, and obtain a set of test statistics, $F_t^{*(1)},...,F_t^{*(B)}$,

4. Sort the resampled test statistics, $F_t^{*(1)},...,F_t^{*(B)}$, in ascending order, so that, $F_t^{*(1)} \le ... \le F_t^{*(B)}$.

5. $F_t^{*(u)}$ is critical value, where $u = (1-\alpha)B$, rounded to the nearest integer,

6. Reject the null hypothesis if $F_t > F_t^{*(u)}$.

In addition, the *p*-value below can be used the test the null hypothesis,

$$\hat{p} = \frac{\#\left(F_t^{*(b)} \ge F_t\right)+1}{B+1}, \qquad (b = 1,...,B) \tag{3.63}$$

Reject the null hypothesis if $\hat{p} \le \alpha$.

CHAPTER FOUR
APPLICATION

In this section, design of the simulation study and corresponding results are given. The methods for two and three or more independent samples and their parametric and non-parametric counterparts are compared in terms of their ability to control type I error rate and powers when testing at 0.05 significance level. The simulations are investigated under both homoscedasticity and heteroscedasticity. Note that, the hypotheses for two independent samples are constructed as two sided.

4.1 Design of the Simulation Study

The simulations are performed with 10,000 replications and 1,500 bootstrap samples for each *n*. Instead of computing all possible permutations, 1,500 permutation samples are generated as an approximate to exact computation. The sample sizes are set between 10 and 50 where balanced and unbalanced designs are conducted for various sample size settings. Regarding the methods based on trimmed mean, 20% symmetric trimming is used and the simulation study is conducted using R 3.1.0.

The random samples are generated from skewed, symmetric heavy tailed and symmetric short tailed theoretical distributions which are Gamma, Logistic and Uniform, respectively. Additionally, Normal distribution is added to the simulation for comparing three or more independent samples in terms of giving variety. The real data sets Total AA Degrees and SAT Quantitative scores from the Carnegie Foundation's 2010 Classification of Universities (http://carnegieclassifications.iu.edu/) are also used. The data sets can be downloaded from the links http://kisi.deu.edu.tr//engin.yildiztepe/datastore/Total_AA_Degrees.dat and http://kisi.deu.edu.tr//engin.yildiztepe/datastore/SAT_Quantitative.dat. Histograms and descriptive statistics of the data sets are given in Figure 4.1 and Table 4.1, respectively.

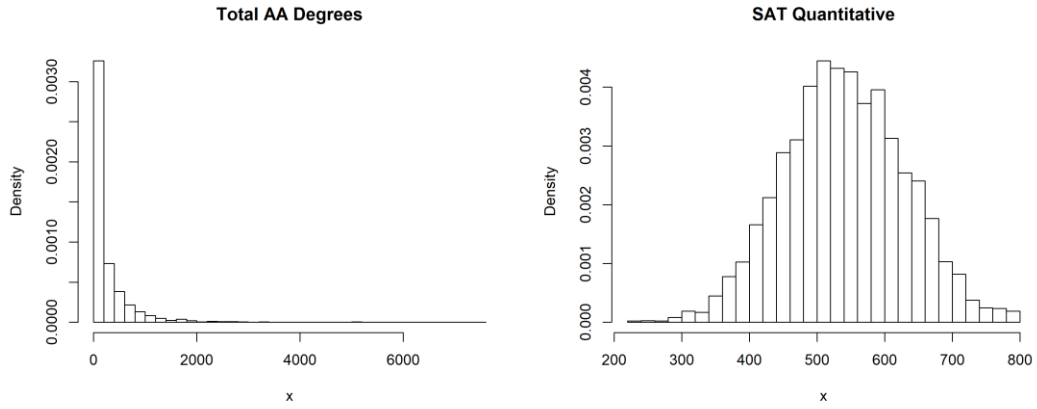


Figure 4.1 Histograms of the real data sets

Table 4.1. Descriptive statistics of the real data sets

Data Set	N	Min	Max	Median	Mean	Trimmed Mean (0.2)	Standard Deviation	Skewness	Kurtosis
Total AA Degrees	2859	1	7489	114	268.98	140.83	451.96	4.98	47.31
SAT Quantitative	15168	230	800	540	545.3	544.62	89.49	0.026	2.86

Table 4.2 Abbreviations of the methods

# of samples	Method	Abbreviation
two	The Wilcoxon-Mann-Whitney test	WMW
	The Permutation test procedure	PT
	The bootstrap test procedure for means	BT
	The bootstrap test procedure for trimmed means	BT-trim
three or more	ANOVA-*F* test	F
	The Kruskal-Wallis test	KW
	The Permutation test procedure	PT
	The bootstrap test procedure for means	BT
	The bootstrap test procedure for trimmed means	BT-trim

In addition to the given methods, the Wilcoxon-Mann-Whitney test, the Kruskal-Wallis test and ANOVA-*F* test are used as reference methods. The abbreviations of

the methods for two independent samples and three or more independent samples are given in the Table 4.2.

4.2 Results

The actual type I error rates and power estimates of the methods are given in this section. The seriousness of a type I error may vary according to circumstances (e.g. departure from normality). Bradley (1978) suggests that the actual type I error rate should be between 0.045 and 0.055 when testing at 0.05 significance level and can be stretched up to 0.025 and 0.075 in extreme situations. In this study the limits are determined as 0.040 and 0.060 ($\alpha \pm 0.2\alpha$). The actual type I error rate of a test must fall in the stated interval to satisfy the determined criterion above. Additionally, tests with actual type I error rates less than the lower limit (0.040) will be called "conservative" and more than the upper limit (0.060) will be called "liberal". The actual type I error rates in the tables that did not fall within the determined limits are in boldface and the highest estimated powers highlighted in grey. Note that, the estimated powers of methods that could not control the type I error rate are not considered.

4.2.1 Two Independent Samples

In this section, results of the methods for two independent samples are given. For the convenience of representation, the theoretical distributions used in this section are given in Table 4.3. Delta (Δ) values in the Table 4.3 represent the effect size to compute estimated power.

Table 4.4 shows the actual type I error rates and estimated powers when the underlying population distributions are Logistic. The actual type I error rates satisfied the type I error criterion under homoscedasticity for all methods except one case. The actual type I error rates of WMW and PT did not satisfy the criterion for unbalanced design under heteroscedasticity. Also, BT could not control type I error in some cases. For the homoscedastic situation, WMW gave the highest estimated power in most

cases. WMW gave the highest power with equal sample sizes under heteroscedasticity, whereas BT-trim and as a second choice BT gave the highest estimated powers with unequal sample sizes. There were some instances where PT was more powerful than the other methods but due to the poor performance over controlling the type I error with unequal sample sizes, the estimated powers of PT were not highlighted.

Table 4.3 Theoretical distributions for two independent samples

Result	Estimation	Standard Deviations	Distribution 1	Distribution 2	Δ
Table 4.4	Actual Type I Error Rate	$\sigma_1 = \sigma_2$	Logistic(0,1)	Logistic(0,1)	-
		$\sigma_1 < \sigma_2$	Logistic(0,1)	Logistic(0,2)	
	Power	$\sigma_1 = \sigma_2$	Logistic(0,1)	Logistic(1,1)	1
		$\sigma_1 < \sigma_2$	Logistic(0,1)	Logistic(1,2)	
Table 4.5	Actual Type I Error Rate	$\sigma_1 = \sigma_2$	Uniform(0,1)	Uniform(0,1)	-
		$\sigma_1 < \sigma_2$	Uniform(0,1)	Uniform(-0.5,1.5)	
	Power	$\sigma_1 = \sigma_2$	Uniform(0,1)	Uniform(0.5,1.5)	1
		$\sigma_1 < \sigma_2$	Uniform(0,1)	Uniform(0,2)	
Table 4.6	Actual Type I Error Rate	$\sigma_1 = \sigma_2$	Gamma(1.5,0.5)	Gamma(1.5,0.5)	-
		$\sigma_1 < \sigma_2$	Gamma(2,1.5)	Gamma(1.5,2)	
	Power	$\sigma_1 = \sigma_2$	Gamma(1,1)	Gamma(4,0.5)	1
		$\sigma_1 < \sigma_2$	Gamma(1,1)	Gamma(1,2)	

Table 4.4 Actual type I error rates and estimated powers when population distributions are Logistic

		Actual Type I Error Rates				Power Estimates			
n_1, n_2	σ_1, σ_2	WMW	PT	BT	BT-trim	WMW	PT	BT	BT-trim
10,10	$\sigma_1 = \sigma_2$	0.046	0.053	0.059	0.049	0.211	0.228	0.232	0.223
10,20		0.054	0.055	**0.061**	0.054	0.294	0.290	0.292	0.292
20,20		0.045	0.048	0.052	0.045	0.421	0.402	0.401	0.413
20,30		0.047	0.049	0.056	0.048	0.474	0.462	0.455	0.476
30,30		0.046	0.050	0.055	0.048	0.593	0.567	0.564	0.581
30,50		0.049	0.050	0.056	0.051	0.686	0.658	0.648	0.672
50,50		0.052	0.054	0.057	0.051	0.814	0.785	0.781	0.806
10,10	$\sigma_1 < \sigma_2$	0.054	0.059	**0.061**	0.053	0.119	0.132	0.128	0.123
10,20		**0.036**	**0.024**	**0.061**	0.052	0.125	0.091	0.175	0.167
20,10		**0.080**	**0.105**	0.060	0.055	0.180	0.213	0.137	0.136
20,20		0.051	0.048	0.053	0.047	0.210	0.197	0.196	0.194
20,30		**0.039**	**0.030**	0.056	0.046	0.224	0.180	0.248	0.250
30,20		**0.074**	**0.093**	**0.063**	0.056	0.262	0.275	0.215	0.218
30,30		0.055	0.052	0.057	0.050	0.303	0.281	0.280	0.285
30,50		**0.037**	**0.023**	0.054	0.049	0.333	0.262	0.371	0.376
50,30		**0.076**	**0.094**	0.057	0.052	0.370	0.395	0.296	0.304
50,50		0.060	0.054	0.060	0.051	0.449	0.415	0.413	0.427

Table 4.5 shows the actual type I error rates and estimated powers when the underlying population distributions are Uniform. For BT there were five instances where the type I error criterion was not satisfied and whereas BT-trim had three unsatisfied results. Both BT and PT gave the highest estimated powers under homoscedasticity. When the sample sizes and standard deviations were positively paired, BT had the highest estimated power and for the other settings PT had the highest estimated power under heteroscedasticity.

Table 4.5 Actual type I error rates and estimated powers when population distributions are Uniform

		Actual Type I Error Rates				Power Estimates			
n_1, n_2	σ_1, σ_2	WMW	PT	BT	BT-trim	WMW	PT	BT	BT-trim
10,10	$\sigma_1 = \sigma_2$	0.046	0.054	**0.034**	0.055	0.367	0.423	0.405	0.347
10,20		0.054	0.055	0.042	**0.061**	0.509	0.569	0.533	0.428
20,20		0.045	0.049	**0.038**	0.049	0.694	0.762	0.768	0.571
20,30		0.047	0.050	0.040	0.051	0.770	0.834	0.839	0.637
30,30		0.046	0.048	0.042	0.050	0.865	0.915	0.919	0.742
30,50		0.049	0.051	0.045	0.054	0.925	0.959	0.963	0.823
50,50		0.052	0.052	0.050	0.053	0.975	0.989	0.991	0.915
10,10	$\sigma_1 < \sigma_2$	0.060	**0.061**	**0.035**	**0.061**	0.153	0.202	0.148	0.171
10,20		0.054	0.055	0.042	**0.061**	0.163	0.186	0.304	0.235
20,10		0.047	0.048	**0.037**	0.053	0.232	0.342	0.144	0.182
20,20		0.045	0.049	**0.038**	0.049	0.276	0.384	0.355	0.267
20,30		0.047	0.050	0.040	0.051	0.306	0.398	0.486	0.345
30,20		0.050	0.055	0.045	0.055	0.337	0.487	0.377	0.291
30,30		0.046	0.048	0.042	0.050	0.394	0.550	0.546	0.371
30,50		0.049	0.051	0.045	0.054	0.454	0.597	0.720	0.492
50,30		**0.093**	**0.096**	0.040	0.054	0.473	0.682	0.569	0.400
50,50		0.052	0.052	0.050	0.053	0.584	0.779	0.781	0.552

Table 4.6 shows the actual type I error rates and estimated powers when the underlying population distributions are Gamma. The actual type I error rates of BT did not satisfy the type I error criterion for almost all settings. Due to the poor performance of BT over controlling the type I error, the estimated powers of BT were discarded. WMW gave the highest power under homoscedasticity and PT gave the highest power under heteroscedasticity.

Simulation results for the real data set SAT Quantitative are given in Table 4.7. The actual type I error rates satisfied the type I error criterion under homoscedasticity for all methods and both PT and BT gave the highest estimated powers under this setting. The actual type I error rates of WMW and PT did not satisfy the criterion for

unbalanced design under heteroscedasticity. WMW, PT and BT gave the highest estimated powers with equal sample sizes under heteroscedasticity whereas BT and as a second choice BT-trim gave the highest estimated powers with unequal sample sizes.

Table 4.6 Actual type I error rates and estimated powers when population distributions are Gamma

		Actual Type I Error Rates				Power Estimates			
n_1, n_2	σ_1, σ_2	WMW	PT	BT	BT-trim	WMW	PT	BT	BT-trim
10,10		0.046	0.054	**0.063**	0.049	0.664	0.600	-	0.626
10,20		0.046	0.047	**0.067**	0.048	0.808	0.741	-	0.694
20,20		0.049	0.050	**0.063**	0.047	0.949	0.861	-	0.916
20,30	$\sigma_1 = \sigma_2$	0.049	0.052	**0.066**	0.051	0.971	0.911	-	0.932
30,30		0.049	0.049	**0.060**	0.050	0.992	0.952	-	0.982
30,50		0.047	0.049	**0.060**	0.047	0.997	0.980	-	0.989
50,50		0.054	0.058	**0.070**	0.054	1.000	0.996	-	0.999
10,10		0.045	0.051	0.059	0.050	0.218	0.272	-	0.247
10,20		0.050	0.042	**0.064**	0.053	0.294	0.341	-	0.329
20,10		**0.062**	**0.062**	**0.070**	0.056	0.320	0.410	-	0.320
20,20		0.058	0.056	**0.066**	0.056	0.446	0.539	-	0.456
20,30	$\sigma_1 < \sigma_2$	0.049	0.045	**0.061**	0.049	0.507	0.616	-	0.534
30,20		0.059	0.060	**0.067**	0.053	0.514	0.635	-	0.521
30,30		0.059	0.054	**0.063**	0.052	0.616	0.728	-	0.623
30,50		0.059	0.045	**0.063**	0.059	0.720	0.819	-	0.733
50,30		0.060	0.058	0.058	0.054	0.711	0.829	-	0.705
50,50		**0.062**	0.052	0.059	0.056	0.837	0.924	-	0.839

Table 4.7 Actual type I error rates and estimated powers - SAT Quantitative (two independent samples)

		Actual Type I Error Rates				Power Estimates (Δ=70)			
n_1, n_2	σ_1, σ_2	WMW	PT	BT	BT-trim	WMW	PT	BT	BT-trim
10,10		0.043	0.054	0.048	0.049	0.333	0.381	0.355	0.350
10,20		0.043	0.051	0.047	0.047	0.459	0.502	0.465	0.444
20,20		0.046	0.051	0.048	0.049	0.637	0.683	0.672	0.614
20,30	$\sigma_1 = \sigma_2$	0.047	0.052	0.050	0.048	0.727	0.764	0.756	0.700
30,30		0.049	0.052	0.050	0.050	0.822	0.850	0.844	0.791
30,50		0.049	0.053	0.051	0.050	0.898	0.918	0.913	0.868
50,50		0.044	0.047	0.044	0.045	0.965	0.976	0.975	0.950
10,10		0.049	0.053	0.048	0.051	0.189	0.194	0.176	0.184
10,20		**0.031**	**0.021**	0.047	0.048	0.210	0.166	0.274	0.259
20,10		**0.086**	**0.108**	0.052	0.059	0.264	0.303	0.176	0.191
20,20		0.059	0.055	0.050	0.052	0.334	0.333	0.324	0.294
20,30	$\sigma_1 < \sigma_2$	0.044	**0.031**	0.052	0.050	0.412	0.356	0.454	0.406
30,20		**0.077**	**0.086**	0.050	0.054	0.410	0.441	0.343	0.322
30,30		0.057	0.052	0.051	0.050	0.470	0.463	0.457	0.413
30,50		**0.037**	**0.021**	0.049	0.046	0.573	0.513	0.636	0.567
50,30		**0.082**	**0.092**	0.049	0.055	0.563	0.605	0.489	0.450
50,50		0.060	0.052	0.050	0.051	0.690	0.688	0.686	0.617

Simulation results for the real data set Total AA Degrees are given in Table 4.8. Under homoscedasticity, all methods could control type I error except BT and it had liberal behavior over controlling the type I error. In addition, WMW gave the highest power under homoscedasticity. In general, only BT-trim could control type I error rate under both heteroscedasticity. However, the results of PT also satisfied the type I error criterion when the sample sizes and standard deviations are positively paired. The power estimates of other methods except BT-trim are not highlighted due to poor performance over controlling the type I error under heteroscedasticity.

Table 4.8 Actual type I error rates and estimated powers – Total AA Degrees (two independent samples)

		Actual Type I Error Rates				Power Estimates (Δ=170)			
n_1, n_2	σ_1, σ_2	WMW	PT	BT	BT-trim	WMW	PT	BT	BT-trim
10,10		0.043	0.052	**0.082**	0.044	0.487	0.275	-	0.291
10,20		0.046	0.049	**0.106**	0.047	0.615	0.366	-	0.303
20,20		0.048	0.046	**0.088**	0.043	0.807	0.358	-	0.510
20,30	$\sigma_1 = \sigma_2$	0.043	0.045	**0.092**	0.041	0.856	0.427	-	0.523
30,30		0.048	0.050	**0.091**	0.047	0.936	0.427	-	0.678
30,50		0.047	0.050	**0.090**	0.049	0.963	0.526	-	0.695
50,50		0.047	0.051	**0.086**	0.046	0.994	0.559	-	0.880
10,10		**0.068**	**0.116**	**0.125**	0.050	0.317	0.234	-	0.233
10,20		0.043	0.050	**0.094**	0.045	0.443	0.310	-	0.288
20,10		**0.102**	**0.177**	**0.146**	0.058	0.434	0.296	-	0.345
20,20		**0.081**	**0.091**	**0.109**	0.046	0.618	0.403	-	0.436
20,30		**0.061**	0.057	**0.094**	0.042	0.701	0.464	-	0.493
30,20	$\sigma_1 < \sigma_2$	**0.102**	**0.123**	**0.109**	0.047	0.717	0.465	-	0.534
30,30		**0.098**	**0.080**	**0.095**	0.048	0.803	0.539	-	0.595
30,50		**0.081**	0.043	**0.085**	0.051	0.873	0.631	-	0.674
50,30		**0.130**	**0.120**	**0.093**	0.050	0.890	0.620	-	0.579
50,50		**0.125**	**0.072**	**0.085**	0.045	0.958	0.742	-	0.822

4.2.2 Three or More Independent Samples

In this section, results of the methods for three or more independent samples are given and the simulation study is conducted for only three samples. Delta (Δ) values in the tables and figures represent the effect size to compute estimated power. To be more precise, delta values specify the amount of shift for the third sample. Due to the poor performances of F, KW and PT over controlling the type I error under heteroscedasticity, the estimated powers of these methods are discarded from the relative tables.

Table 4.9 Actual type I error rates – Normal $(0, \sigma)$

Sample Sizes	Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1,1,1	0.048	0.043	0.049	0.045	0.054
20,20,20		0.050	0.050	0.047	0.044	0.048
10,20,30		0.052	0.048	0.053	0.053	0.059
30,30,30		0.050	0.049	0.049	0.048	0.047
20,30,50		0.052	0.049	0.051	0.052	0.049
50,50,50		0.053	0.050	0.051	0.050	0.050
10,10,10	1,2,4	**0.077**	**0.063**	**0.081**	0.048	0.059
10,15,20	1,2,4	**0.033**	**0.039**	**0.033**	0.049	0.055
10,15,20	4,2,1	**0.143**	**0.094**	**0.148**	0.056	0.050
30,30,30	1,2,4	**0.068**	**0.065**	**0.066**	0.047	0.046
20,30,50	1,2,4	**0.017**	**0.031**	**0.018**	0.049	0.049
20,30,50	4,2,1	**0.164**	**0.109**	**0.165**	0.053	0.045

Table 4.10 Estimated powers - Normal $(0, \sigma)$ (heteroscedastic case)

Sample Sizes	Standard Deviations	Δ	BT	BT-trim
10,10,10	1,2,4	1.25	0.130	0.106
		2.50	0.384	0.229
		3.75	0.704	0.430
		5.00	0.921	0.637
10,15,20	1,2,4	1	0.165	0.121
		2	0.500	0.340
		3	0.839	0.647
		4	0.978	0.885
10,15,20	4,2,1	0.7	0.090	0.111
		1.4	0.211	0.333
		2.1	0.480	0.650
		2.8	0.770	0.873
30,30,30	1,2,4	0.7	0.148	0.107
		1.4	0.424	0.281
		2.1	0.758	0.566
		2.8	0.945	0.825
20,30,50	1,2,4	0.6	0.154	0.105
		1.2	0.485	0.344
		1.8	0.811	0.657
		2.4	0.965	0.895
20,30,50	4,2,1	0.4	0.075	0.105
		0.8	0.158	0.330
		1.2	0.371	0.665
		1.6	0.659	0.903

Table 4.9 shows the actual type I error rates when the underlying population distributions are Normal. Under homoscedasticity, all methods could control the type I error. But under heteroscedasticity, the actual type I error rates of F, KW and PT did not satisfy the type I error criterion. Figure 4.2 shows the power estimates when the sample sizes are 20. Under homoscedasticity, F and KW gave slightly higher estimated power. Table 4.10 shows the estimated powers under heteroscedasticity. When the sample sizes and standard deviations are negatively paired BT-trim, for the other settings BT gave the highest power.

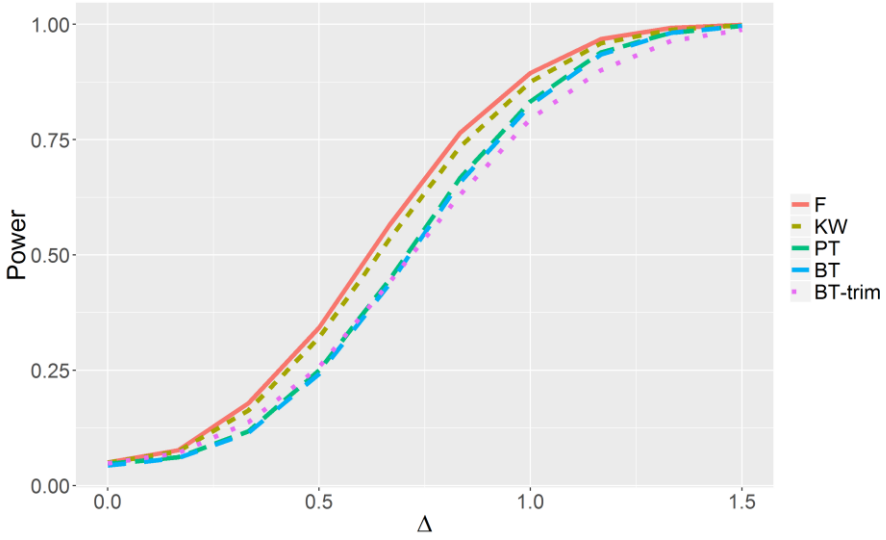


Figure 4.2 Estimated powers – Normal (0,1) (homoscedastic case, $n_1 = n_2 = n_3 = 20$)

Table 4.11 Actual type I error rates – Logistic distribution with $\mu = 0$

Sample Sizes	Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1,1,1	0.052	0.047	0.052	0.046	0.056
20,20,20		0.050	0.048	0.050	0.049	0.048
10,20,30		0.054	0.051	0.054	0.053	0.059
30,30,30		0.049	0.049	0.051	0.048	0.047
20,30,50		0.050	0.050	0.050	0.048	0.050
50,50,50		0.051	0.048	0.048	0.049	0.046
10,10,10	1,2,4	**0.069**	**0.062**	**0.075**	0.043	0.058
10,15,20	1,2,4	**0.029**	0.041	**0.032**	0.047	0.051
10,15,20	4,2,1	**0.141**	**0.090**	**0.149**	0.047	0.050
30,30,30	1,2,4	**0.067**	**0.063**	**0.068**	0.046	0.049
20,30,50	1,2,4	**0.018**	**0.028**	**0.018**	0.045	0.049
20,30,50	4,2,1	**0.161**	**0.107**	**0.161**	0.048	0.043

Table 4.12 Estimated powers – Logistic distribution with $\mu = 0$ (heteroscedastic case)

Sample Sizes	Standard Deviations	Δ	BT	BT-trim
10,10,10	1,2,4	1.25	0.138	0.109
		2.5	0.407	0.281
		3.75	0.720	0.519
		5	0.902	0.721
10,15,20	1,2,4	1	0.164	0.137
		2	0.511	0.406
		3	0.836	0.743
		4	0.971	0.929
10,15,20	4,2,1	1.25	0.071	0.103
		2.5	0.166	0.304
		3.75	0.370	0.613
		5	0.624	0.848
30,30,30	1,2,4	0.7	0.141	0.114
		1.4	0.438	0.345
		2.1	0.768	0.671
		2.8	0.941	0.888
20,30,50	1,2,4	0.6	0.152	0.131
		1.2	0.474	0.400
		1.8	0.814	0.762
		2.4	0.965	0.942
20,30,50	4,2,1	0.4	0.072	0.122
		0.8	0.171	0.409
		1.2	0.382	0.763
		1.6	0.6631	0.952

Table 4.11 shows the actual type I error rates when the underlying population distributions are Logistic. Under homoscedasticity, all methods could control the type I error. Besides, F, KW and PT did not satisfy the type I error criteria under heteroscedasticity. Figure 4.3 shows the estimated powers under homoscedasticity when the sample sizes are 20. Under homoscedasticity, F and KW gave slightly higher estimated power. Table 4.12 gives the estimated powers under heteroscedasticity. When the sample sizes and standard deviations are negatively paired BT-trim, for the other settings BT gave the highest power.

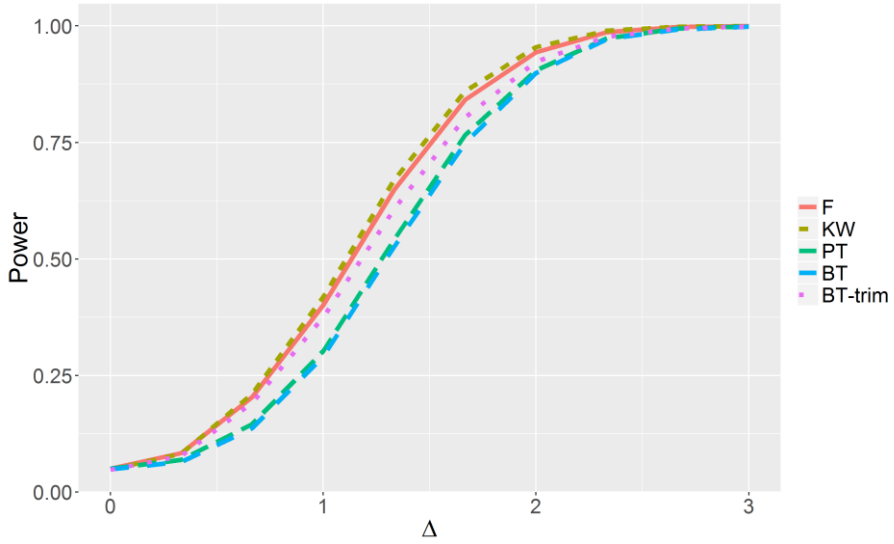


Figure 4.3 Estimated powers - Logistic distribution with $\mu = 0$, $\sigma = 1$ (homoscedastic case, $n_1 = n_2 = n_3 = 20$)

Table 4.13 shows the actual type I error rates when the underlying population distributions are Uniform. Under homoscedasticity, all methods could control the type I error. Besides F, KW and PT did not satisfy the type I error criteria under heteroscedasticity. Figure 4.4 shows the estimated powers under homoscedasticity when the sample sizes are 20. Under homoscedasticity F and KW gave slightly higher estimated power. Table 4.14 shows the estimated powers under heteroscedasticity. BT gave the highest power under heteroscedasticity except one case.

Table 4.13 Actual type I error rates – Uniform distribution with $\mu = 0$

Sample Sizes	Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1,1,1	0.052	0.048	0.051	0.049	0.054
20,20,20		0.052	0.050	0.051	0.049	0.048
10,20,30		0.051	0.049	0.051	0.055	0.059
30,30,30		0.052	0.050	0.050	0.048	0.047
20,30,50		0.053	0.051	0.052	0.053	0.046
50,50,50		0.051	0.049	0.052	0.053	0.051
10,10,10	1,2,4	**0.080**	**0.072**	**0.080**	0.051	0.059
10,15,20	1,2,4	**0.032**	**0.039**	**0.031**	0.046	0.050
10,15,20	4,2,1	**0.147**	**0.106**	**0.149**	0.058	0.049
30,30,30	1,2,4	**0.072**	**0.073**	**0.071**	0.050	0.047
20,30,50	1,2,4	**0.018**	**0.029**	**0.017**	0.048	0.044
20,30,50	4,2,1	**0.171**	**0.125**	**0.171**	0.058	0.045

Table 4.14 Estimated powers – Uniform distribution with $\mu = 0$ (heteroscedastic case)

Sample Sizes	Standard Deviations	Δ	BT	BT-trim
10,10,10	1,2,4	1.2	0.113	0.080
		2.4	0.330	0.140
		3.6	0.668	0.250
		4.8	0.917	0.396
10,15,20	1,2,4	1.25	0.138	0.109
		2.50	0.407	0.280
		3.75	0.720	0.519
		5.00	0.902	0.721
10,15,20	4,2,1	0.9	0.098	0.105
		1.8	0.309	0.295
		2.7	0.744	0.612
		3.6	0.972	0.858
30,30,30	1,2,4	0.7	0.140	0.078
		1.4	0.415	0.181
		2.1	0.756	0.366
		2.8	0.950	0.604
20,30,50	1,2,4	0.5	0.120	0.075
		1.0	0.346	0.165
		1.5	0.653	0.317
		2.0	0.888	0.539
20,30,50	4,2,1	0.5	0.081	0.099
		1.0	0.226	0.307
		1.5	0.580	0.648
		2.0	0.900	0.900

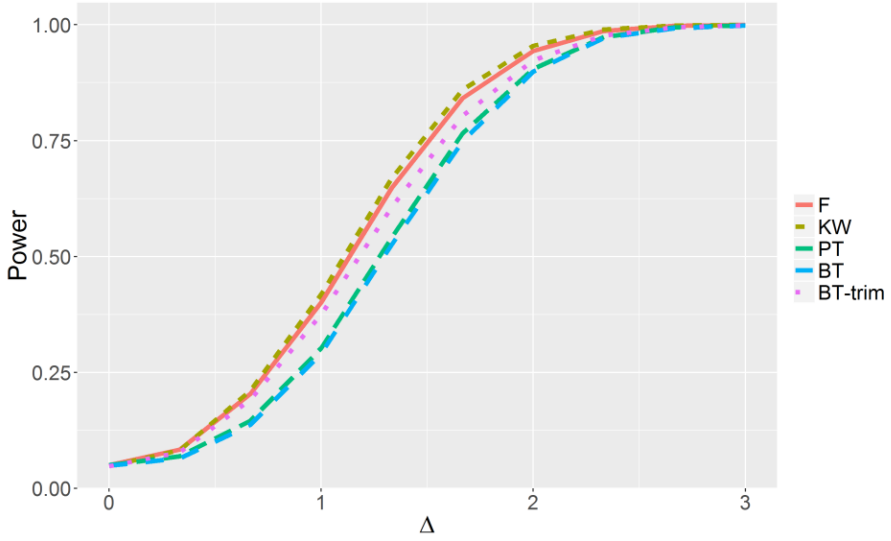


Figure 4.4 Estimated powers – Uniform distribution with $\mu = 0$, $\sigma = 1$ (homoscedastic case, $n_1 = n_2 = n_3 = 20$)

Table 4.15 shows the actual type I error rates when the underlying populations are Gamma. Under homoscedasticity, all methods controlled the type I error. Figure 4.5 shows the estimated powers when the sample sizes are 20 and delta in this figure gives the shift of the shape parameter for the third sample. F and KW gave slightly higher power than the others. Table 4.16 gives the estimated powers under heteroscedasticity. When the sample sizes and standard deviations are negatively paired BT-trim, for the other settings BT gave the highest power.

Table 4.15 Actual type I error rates - Gamma (1.5,0.5)

Sample Sizes	Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1,1,1	0.046	0.046	0.051	0.040	0.048
20,20,20		0.048	0.049	0.055	0.046	0.044
10,20,30		0.053	0.049	0.054	0.041	0.054
30,30,30		0.049	0.052	0.048	0.044	0.049
20,30,50		0.046	0.048	0.046	0.040	0.048
50,50,50		0.047	0.053	0.049	0.048	0.047
10,10,10	1,2,4	**0.084**	**0.069**	**0.088**	0.059	0.064
10,15,20	1,2,4	0.040	0.040	0.040	0.048	0.052
10,15,20	4,2,1	**0.158**	**0.106**	**0.166**	**0.077**	0.055
30,30,30	1,2,4	**0.075**	**0.074**	**0.077**	0.055	0.051
20,30,50	1,2,4	**0.022**	0.040	**0.022**	0.045	0.052
20,30,50	4,2,1	**0.018**	**0.121**	**0.180**	**0.065**	0.046

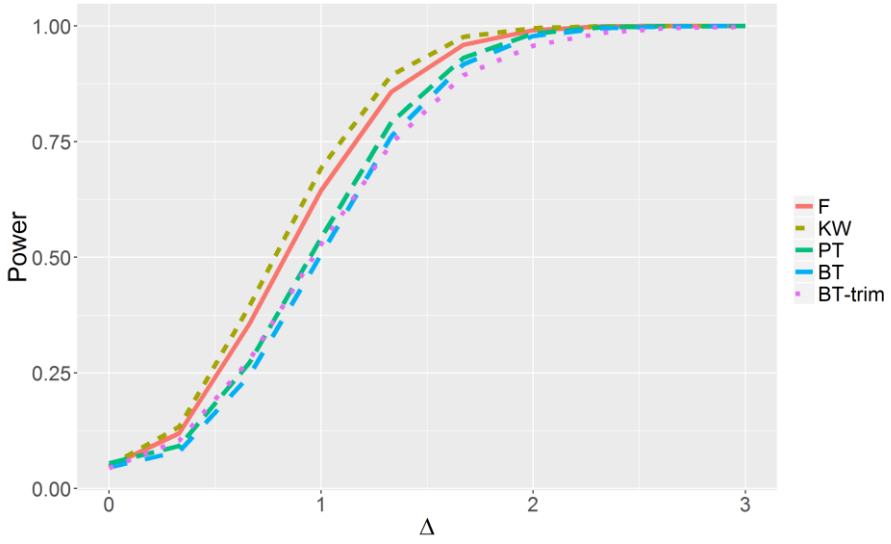


Figure 4.5 Estimated powers - Gamma (1.5,0.5) ($n_1 = n_2 = n_3 = 20$)

Table 4.16 Estimated powers – Gamma

Sample Sizes	Standard Deviations	Δ	BT	BT-trim
10,10,10	1,2,4	0.8	0.063	0.082
		1.6	0.339	0.258
		2.4	0.771	0.560
		3.2	0.992	0.780
10,15,20	1,2,4	0.5	0.065	0.079
		1.0	0.275	0.252
		1.5	0.649	0.584
		2.0	0.914	0.870
10,15,20	4,2,1	1.6	0.138	0.169
		3.2	0.264	0.401
		4.8	0.457	0.664
		5.4	0.642	0.837
30,30,30	1,2,4	0.6	0.168	0.155
		1.2	0.712	0.627
		1.8	0.985	0.960
		2.4	1.000	1.000
20,30,50	1,2,4	0.4	0.133	0.129
		0.8	0.510	0.479
		1.2	0.901	0.889
		1.6	0.994	0.994
20,30,50	4,2,1	1.4	0.149	0.260
		2.8	0.360	0.683
		4.2	0.664	0.941
		5.6	0.873	0.993

Table 4.17 Actual type I error rates - SAT Quantitative (three or more independent samples)

Sample Sizes	The Ratios of Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1:1:1	0.047	0.045	0.046	0.040	0.057
20,20,20		0.049	0.049	0.049	0.046	0.048
10,20,30		0.048	0.045	0.047	0.051	0.057
30,30,30		0.050	0.050	0.050	0.049	0.050
20,30,50		0.049	0.048	0.048	0.050	0.046
50,50,50		0.052	0.050	0.052	0.051	0.050
10,10,10	1:2:4	**0.073**	**0.064**	**0.076**	0.048	0.053
10,15,20	1:2:4	**0.028**	0.040	**0.028**	0.045	0.059
10,15,20	4:2:1	**0.147**	**0.093**	**0.150**	0.054	0.043
30,30,30	1:2:4	**0.071**	**0.070**	**0.070**	0.048	0.050
20,30,50	1:2:4	**0.018**	**0.032**	**0.018**	0.049	0.049
20,30,50	4:2:1	**0.166**	**0.113**	**0.168**	0.054	0.042

Simulation results for the real data set SAT Quantitative are given in Table 4.17 and Table 4.18. Under homoscedasticity, all methods controlled the type I error. Besides F, KW and PT did not satisfy the type I error criterion under heteroscedasticity. Under homoscedasticity F, PT and BT gave slightly higher estimated powers than the others. When the sample sizes and standard deviations are negatively paired BT-trim, for the other settings BT gave the highest estimated power under heteroscedasticity.

Table 4.18 Power Estimates (Δ=65) - SAT Quantitative (three or more independent samples)

Sample Sizes	The Ratios of Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1:1:1	0.332	0.299	0.325	0.302	0.232
20,20,20		0.636	0.601	0.633	0.625	0.492
10,20,30		0.683	0.651	0.679	0.676	0.474
30,30,30		0.822	0.793	0.820	0.817	0.704
20,30,50		0.905	0.884	0.903	0.901	0.796
50,50,50		0.968	0.958	0.967	0.967	0.923
10,10,10	1:2:4	-	-	-	0.077	0.072
10,15,20	1:2:4	-	-	-	0.101	0.081
10,15,20	4:2:1	-	-	-	0.089	0.113
30,30,30	1:2:4	-	-	-	0.147	0.097
20,30,50	1:2:4	-	-	-	0.205	0.145
20,30,50	4:2:1	-	-	-	0.138	0.273

Table 4.19 Actual type I error rates - Total AA Degrees (three or more independent samples)

Sample Sizes	The Ratios of Standard Deviations	F	KW	PT	BT	BT-trim
10,10,10	1:1:1	**0.034**	0.047	0.050	**0.011**	**0.023**
20,20,20		**0.038**	0.048	0.050	**0.014**	**0.026**
10,20,30		**0.034**	0.044	0.047	**0.012**	**0.030**
30,30,30		**0.038**	0.047	0.048	**0.017**	**0.033**
20,30,50		0.044	0.051	0.051	**0.022**	0.043
50,50,50		**0.036**	0.046	0.044	**0.022**	**0.038**
10,10,10	1:2:4	**0.131**	**0.076**	**0.146**	**0.086**	**0.065**
10,15,20	1:2:4	**0.075**	0.053	**0.079**	**0.063**	0.053
10,15,20	4:2:1	**0.190**	**0.114**	**0.210**	**0.107**	**0.077**
30,30,30	1:2:4	**0.094**	**0.101**	**0.103**	**0.066**	0.057
20,30,50	1:2:4	0.042	**0.063**	0.045	0.051	0.049
20,30,50	4:2:1	**0.196**	**0.161**	**0.211**	**0.097**	**0.063**

Simulation results for the real data set Total AA Degrees are given in Table 4.19. F, BT and BT-trim showed conservative behavior over controlling the type I error under homoscedasticity. KW and PT had no problems over controlling the type I error in this case. Under heteroscedasticity, F, KW, PT and BT could not control type I error except one case. On the other hand, especially when the sample sizes and standard deviations are positively paired, BT-trim satisfied the type I error criterion in three cases. The power estimates for this data set are not given due to poor performances of methods over controlling the type I error.

CHAPTER FIVE
CONCLUSION

In this thesis, resampling methods on common statistical problems are studied. The resampling methods for two independent and three or more independent samples and their parametric and non-parametric counterparts are compared in terms of their ability to control type I error and powers via a Monte Carlo simulation study. The performance of the methods are investigated under both various theoretical distributions and real data sets. The sample sizes are set between 10 and 50, since it is known that classical approaches perform well when the sample size is fairly large. Heteroscedasticity, which has great impact over type I error and test power, is also included in the simulation study.

For two independent samples, under homoscedasticity all methods performed well when controlling the type I error. But when the data was skewed, BT had liberal behavior. According to estimated powers under homoscedasticity, PT and BT are suggested for short tailed data and WMW is suggested for both heavy tailed and skewed data. With the aim of controlling type I error under heteroscedasticity, WMW and PT gave poor performance when the data was heavy tailed and BT had liberal behavior when the data was skewed. However, BT-trim had no problem with controlling the type I error. According to the estimated powers under heteroscedasticity, BT and BT-trim are suggested for heavy tailed data, BT and PT are suggested for short tailed data and PT is suggested for skewed data.

To sum up for the two independent samples, in out of the fifty-one, WMW and PT had nine, BT had twenty-two and BT-trim had three instances where the type I error criterion was not satisfied. Hence, when the aim is controlling type I error BT-trim is suggested. There were no superiority between the methods according to their estimated powers. Thereby, it is suggested to use the most powerful method according to circumstance.

For three or more independent samples, under homoscedasticity all methods gave satisfactory results when controlling the type I error. Additionally, the commonly used methods F and KW gave slightly better estimated powers. Under heteroscedasticity, BT and BT-trim could be able to control type I error for all sample sizes. BT was more powerful than BT-trim in two-thirds of the heteroscedastic cases. The classical methods F and KW are recommended under homoscedasticity whereas BT and BT-trim are recommended under heteroscedasticity.

As a summary, Table 5.1 and Table 5.2 gives the suggested methods under various simulation settings for two independent samples and three or more independent samples, respectively.

Table 5.1 Suggested methods for two independent samples

Distribution	n_1, n_2	σ_1, σ_2	WMW	PT	BT	BT-trim
Logistic	$n_1 = n_2$	$\sigma_1 = \sigma_2$	✓	✓	✓	✓
	$n_1 < n_2$		✓	✓	✓	✓
	$n_1 = n_2$	$\sigma_1 < \sigma_2$	✓	✓	✓	✓
	$n_1 < n_2$		✗	✗	✓	✓
	$n_2 < n_1$		✗	✗	✓	✓
Uniform	$n_1 = n_2$	$\sigma_1 = \sigma_2$	✗	✓	✓	✗
	$n_1 < n_2$		✗	✓	✓	✗
	$n_1 = n_2$	$\sigma_1 < \sigma_2$	✗	✓	✗	✗
	$n_1 < n_2$		✗	✗	✓	✗
	$n_2 < n_1$		✗	✓	✗	✗
Gamma	$n_1 = n_2$	$\sigma_1 = \sigma_2$	✓	✗	✗	✗
	$n_1 < n_2$		✓	✗	✗	✗
	$n_1 = n_2$	$\sigma_1 < \sigma_2$	✗	✓	✗	✗
	$n_1 < n_2$		✗	✓	✗	✗
	$n_2 < n_1$		✗	✓	✗	✗

Table 5.2 Suggested methods for three or more independent samples

Distribution	n_1, n_2, n_3	$\sigma_1, \sigma_2, \sigma_3$	F	KW	PT	BT	BT-trim
Normal	$n_1 = n_2 = n_3$	$\sigma_1 = \sigma_2 = \sigma_3$	✓	✓	×	×	×
	$n_1 < n_2 < n_3$		✓	✓	×	×	×
	$n_1 = n_2 = n_3$	$\sigma_1 < \sigma_2 < \sigma_3$	×	×	×	✓	×
	$n_1 < n_2 < n_3$		×	×	×	✓	×
	$n_3 < n_2 < n_1$		×	×	×	×	✓
Logistic	$n_1 = n_2 = n_3$	$\sigma_1 = \sigma_2 = \sigma_3$	✓	✓	×	×	×
	$n_1 < n_2 < n_3$		✓	✓	×	×	×
	$n_1 = n_2 = n_3$	$\sigma_1 < \sigma_2 < \sigma_3$	×	×	×	✓	×
	$n_1 < n_2 < n_3$		×	×	×	✓	×
	$n_3 < n_2 < n_1$		×	×	×	×	✓
Uniform	$n_1 = n_2 = n_3$	$\sigma_1 = \sigma_2 = \sigma_3$	✓	✓	×	×	×
	$n_1 < n_2 < n_3$		✓	✓	×	×	×
	$n_1 = n_2 = n_3$	$\sigma_1 < \sigma_2 < \sigma_3$	×	×	×	✓	×
	$n_1 < n_2 < n_3$		×	×	×	✓	×
	$n_3 < n_2 < n_1$		×	×	×	✓	✓
Gamma	$n_1 = n_2 = n_3$	$\sigma_1 = \sigma_2 = \sigma_3$	✓	✓	×	×	×
	$n_1 < n_2 < n_3$		✓	✓	×	×	×
	$n_1 = n_2 = n_3$	$\sigma_1 < \sigma_2 < \sigma_3$	×	×	×	✓	×
	$n_1 < n_2 < n_3$		×	×	×	✓	×
	$n_3 < n_2 < n_1$		×	×	×	×	✓

In this thesis, the resampling methods for comparing independent samples are considered. Clearly, there is no method that performs best in all conditions since the results can be affected by the shape of data or heteroscedasticity. Nevertheless, in this thesis the resampling based methods performed considerably well in most of the cases. Furthermore, there exist some other important topics in statistics such as comparing dependent samples or multiple comparison and they can be of interest for further study.

REFERENCES

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

Butar, F. B., & Bandulasiri, A. (2009). Comparison of the power of the paired samples using permutation tests. *Journal of Mathematical Sciences & Mathematics Education, 4* (2), 19-31.

Chernick, M. R. (2012). Resampling methods. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2* (3), 255-262.

Davison, A. C., & Hinkley, D. V. (1997). *Bootstrap methods and their application* Cambridge: Cambridge University Press.

Diaconis, P., & Efron, B. (1983) Computer-intensive methods in statistics. *Scientific American, 248*, 116-130.

Edgington, E. S. (1995). *Randomization tests* (3rd ed.). New York: Dekker.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The annals of Statistics, 7* (1), 1-26.

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*. Vermont: Society for industrial and applied mathematics.

Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician, 37* (1), 36-48.

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 54-75.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. New York: Chapman & Hall.

Ernst, M. D. (2004). Permutation methods: A basis for exact inference. *Statistical Science, 19* (4), 676-685.

Fisher, R. A. (1935). *The design of experiments*. New York: Oliver and Boyd.

Good, P. (2013). *Permutation tests: A practical guide to resampling methods for testing hypotheses* (3rd ed.). Springer Science & Business Media.

Hall, P. (1992). *The bootstrap and edgeworth expansion*. Berlin: Springer.

Hall, P., & Martin, M. (1988). On the bootstrap and two-sample problems. *Australian & New Zealand Journal of Statistics, 30* (1), 179-192

Hall, P., & Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics, 47* (2), 757-762.

Maindonald, J., & Braun, J. (2003). *Data analysis and graphics using R: An example-based approach* (3rd ed.). New York: Cambridge University Press.

Martin, M. A. (2007). Bootstrap hypothesis testing for some common statistical problems: A critical evaluation of size and power properties. *Computational Statistics & Data Analysis, 51* (12), 6321-6342.

Miller, R. G. (1974). The jackknife-a review. *Biometrika, 61* (1), 1-15.

Ozdemir, A. F., Wilcox, R., & Yildiztepe, E. (2017). Comparing j independent groups with a method based on trimmed means. *Communications in Statistics-Simulation and Computation*.

Pitman, E. J. (1937). Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society, 4* (1), 119-130.

Quenouille, M. H. (1949). Approximate tests of correlation in time-series. *Journal of the Royal Statistical Society, Series B (Methodological), 11* (1), 68-84.

Rizzo, M. L. (2007). *Statistical computing with R*. Chapman & Hall/CRC.

Shao, J., & Tu, D. (2012). *The jackknife and bootstrap*. New York: Springer Science & Business Media.

Tukey, J. W. (1958). Bias and confidence in not-quite large samples. *Annals of Mathematical Statistics, 29* (2), 614-614.

Welch, B. L. (1951). On the comparison of several mean values: An alternative approach. *Biometrika, 38* (3/4), 330–336.

Westfall, P. H., & Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*. New York: John Wiley & Sons.

Wilcox, R. R. (2017). *Introduction to robust estimation and hypothesis testing* (4th ed.). Academic Press.

APPENDICES

```
# For the tests PT and BT-trim (2 Groups) and BT-trim (3 groups)
# Wilcox's R source file (Rallfun-v30) were used.
# Wilcox's R source file version 30 file can be downloaded from the link below
# https://dornsife.usc.edu/assets/sites/239/docs/Rallfun-v30.txt


#######################################################
###### FUNCTIONS FOR TWO INDEPENDENT SAMPLES#######
#######################################################

### A Bootstrap Test Procedure for Two Independent Samples (BT) ###

#Transformation
y.n <- function(x,y,d=0){
 mx <- mean(x);my <- mean(y)
 for(i in 1:length(y)){
  y[i] <- y[i]-my+mx-d}
 y
}


BT2 <- function(x,y,B=1499,d=0){
 tval <- t.test(x = x,y = y,mu = d)$statistic
 y_new <- y.n(x,y,d=d)
 tboot <- numeric(B)
 for(i in 1:B){
  x_b <- sample(x,replace = T)
  y_b <- sample(y_new,replace = T)
  tboot[i] <- t.test(x = x_b,y = y_b,mu = d)$statistic
 }
```

```
  p<-(sum(tboot>=tval)+1)/(B+1)
  p<-2*min(p,(1-p))
  p
}

#########################################################################
   #######
#Computing the actual type I error rates (A case for logistic distribution)

source("D:\\Rallfun-v30.txt") #This folder contains the functions of wilcox

N=10000
alpha <- 0.05

WMW <- numeric(N);PT <- numeric(N);BT <- numeric(N);BTtrim <- numeric(N)

for(i in 1:N) {

  x1 <- rlogis(n = 20,location = 0,scale = 1)
  x2 <- rlogis(n = 20,location = 0,scale = 1)

  WMW[i]  <- wilcox.test(x1,x2)$p.value #Wilcoxon Mann Whitney test (Base
    package)

  PT[i] <- permg(x1,x2,alpha = alpha,nboot = 1500)$reject==TRUE #PT (Wilcox's R
    function)

  BT[i] <- BT2 (x = x1,y = x2,B = 1499) #BT

  BTtrim[i] <- pb2gen(x = x1,y = x2,nboot = 1499,est = mean,tr =
    0.2,SEED=FALSE)$p.value
  #BT-trim (Wilcox's R function)
```

```
}

result <- c(mean(wmw<alpha),mean(pt),mean(bt<alpha),mean(bttrim<alpha))
data.frame(methods=c("WMW","PT","BT","BT-TRIM"),result=result)


#####################################################################
# Computing the Estimated Powers (A case for logistic distribution)

source("D:\\Rallfun-v30.txt")

N=10000
alpha <- 0.05
WMW <- numeric(N);PT <- numeric(N);BT <- numeric(N);BTtrim <- numeric(N)

for(i in 1:N) {

 x1 <- rlogis(n = 20,location = 0,scale = 1)
 x2 <- rlogis(n = 20,location = 1,scale = 1)

 WMW[i]  <- wilcox.test(x1,x2) $p.value
 #Wilcoxon Mann Whitney test (Base package)

 PT[i] <- permg(x1,x2,alpha = alpha,nboot = 1500)$reject==TRUE
 #PT (Wilcox's R function)

 BT[i] <- BT2 (x = x1,y = x2,B = 1499) #BT

 BTtrim[i] <- pb2gen(x = x1,y = x2,nboot = 1499,est = mean,tr =
   0.2,SEED=FALSE)$p.value
 #BT-trim (Wilcox's R function)
}
```

```
result <- c(mean(wmw<alpha),mean(pt),mean(bt<alpha),mean(bttrim<alpha))
data.frame(methods=c("WMW","PT","BT","BT-TRIM"),result=result)

#####################################################################
###### FUNCTIONS FOR three or more  INDEPENDENT SAMPLES#######
#####################################################################

### A Bootstrap Test Procedure for three or more Samples (BT) ###

# Transformation
null <- function(u){
  u_n <- numeric(length = length(u))
  for(i in 1:length(u)){
    u_n[i] <- u[i]-mean(u)
  }
  u_n
}

BT3 <- function(x1,x2,x3,B){
  y <- c(x1,x2,x3);ym <- mean(y)
  x1n <- null(x1)+ym;x2n <- null(x2)+ym;x3n <- null(x3)+ym
  group = rep(1:3, c(length(x1),length(x2),length(x3)))
  data = data.frame(y = y, group = factor(group))
  f_obs <- anova(lm(data[,1]~data[,2]))[["F value"]][1]
  f_boot <- numeric(B)
  for(i in 1:B){
    x1y <- sample(x1n,replace = T)
    x2y <- sample(x2n,replace = T)
    x3y <- sample(x3n,replace = T)
    y_new <- c(x1y,x2y,x3y)
    databoot = data.frame(y = y_new, group = factor(group))
```

```
   f_boot[i] <- anova(lm(databoot[,1]~databoot[,2]))[["F value"]][1]
  }
  p<-((sum(f_boot>=f_obs))+1)/(B+1)
  p
}


### A Permutation Test Procedure for 3 Groups (PT3) ###

PT3 <- function(x1,x2,x3,B){
  y <- c(x1,x2,x3)
  n1 <- length(x1);n2 <- length(x2);n3 <- length(x3)
  group = rep(1:3, c(n1,n2,n3))
  data = data.frame(y = y, group = factor(group))
  f_obs <- anova(lm(data[,1]~data[,2]))[["F value"]][1] #gözlenen f deðeri
  f_perm <- numeric(B)
  for(i in 1:B){
    y_new <- sample(y,replace = F)
    dataperm = data.frame(y = y_new, group = factor(group))
    f_perm[i] <- anova(lm(dataperm[,1]~dataperm[,2]))[["F value"]][1]
  }
  p<-((sum(f_perm>=f_obs))+1)/(B+1)
  p

}

#########################################################################
#Computing the actual type I error rates (A case for logistic distribution)

install.packages("MASS")
library(MASS)
source("D:\\Rallfun-v30.txt")
```

```
N <- 10000
alpha <- 0.05

f <- numeric(N);KW <- numeric(N);PT <- numeric(N);BT <- numeric(N);BTtrim <-
    numeric(N);

for(i in 1:N){

  x1 <- rlogis(n = 20,location = 0,scale = 1)
  x2 <- rlogis(n = 20,location = 0,scale = 1)
  x3 <- rlogis(n = 20,location = 0,scale = 1)

  group = rep(1:3, c(length(x1),length(x2),length(x3)))
  data3 = data.frame(y = c(x1,x2,x3), group = factor(group))

  #ANOVA F TEST (Base Package)
  f[i] <- anova(lm(data3[,1]~data3[,2]))[[5]][1]

  #kruskal wallis test (Base Package)
  KW[i] <- kruskal.test(list(x1,x2,x3))$p.value

  #PT
  PT[i] <- PT3(x1,x2,x3,B=1499)

  #BT
  BT[i] <- BT3(x1,x2,x3,B=1499)

  #BT-trim (Wilcox's R function)
  datalist <- list(x1,x2,x3)
  BTtrim[i] <- t1waybt(x = datalist,tr = 0.2,nboot = 1500)$p.value
}
```

```
result <-
    c(mean(f<=alpha),mean(KW<=alpha),mean(PT<=alpha),mean(BT<=alpha),mean
    (BTTRIM<=alpha))
data.frame(methods=c("F","KW","PT","BT","BT-TRIM"),result=result)

#########################################################################
# Computing the Estimated Powers (A case for logistic distribution)

install.packages("MASS")
library(MASS)
source("D:\\Rallfun-v30.txt")

N <- 10000
alpha <- 0.05

f <- numeric(N);KW <- numeric(N);PT <- numeric(N);BT <- numeric(N);BTtrim <-
    numeric(N);

for(i in 1:N){

  x1 <- rlogis(n = 20,location = 0,scale = 1)
  x2 <- rlogis(n = 20,location = 0,scale = 1)
  x3 <- rlogis(n = 20,location = 0,scale = 1)

  group = rep(1:3, c(length(x1),length(x2),length(x3)))
  data3 = data.frame(y = c(x1,x2,x3), group = factor(group))

  f[i] <- anova(lm(data3[,1]~data3[,2]))[[5]][1]
  #ANOVA F TEST (Base Package)
```

```
  KW[i] <- kruskal.test(list(x1,x2,x3))$p.value
  #kruskal wallis test (Base Package)

  PT[i] <- PT3(x1,x2,x3,B=1499)
  #PT

  BT[i] <- BT3(x1,x2,x3,B=1499)
  #BT

  datalist <- list(x1,x2,x3)
  BTtrim[i] <- t1waybt(x = datalist,tr = 0.2,nboot = 1500)$p.value
  #BT-trim (Wilcox's R function)
}

result <-
    c(mean(f<=alpha),mean(KW<=alpha),mean(PT<=alpha),mean(BT<=alpha),mean
    (BTTRIM<=alpha))
data.frame(methods=c("F","KW","PT","BT","BT-TRIM"),result=result)
```