EVALUATION OF TWO REFINED MANN–WHITNEY PROCEDURES

by

LAI CHOO HENG

Thesis submitted in fulfillment of the requirements
For the degree of
Doctor of Philosophy

January 2009

ACKNOWLEDGEMENTS

It is my pleasure to express my gratitude to my research advisor Associate Professor Dr Abdul Rahman Bin Othman for providing excellent guidance and assistance throughout my entire dissertation. He was most accommodative with his time and extraordinarily patient with the progress of this research.

I also like to express my sincere appreciation to my co advisor Associate Professor Dr Sharipah Soa’ad Syed Yahaya for her gracious services by offering helpful suggestions and comments.

Finally I wish to express my special appreciation and gratitude to my dear wife, Teo Ah Wan and my three children Lai Chi Qin, Lai Chi Yi and Lai Chi Ning who have endured gracefully and patiently to this seemingly never ending educational endeavour.

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION

CHAPTER 2 - LITERATURE REVIEW

CHAPTER 3 – METHODOLOGY

CHAPTER 4 - RESULTS OF SIMULATION

CHAPTER 5 - CONCLUSION AND RECOMMENDATION

APPENDICES

LIST OF TABLES

LIST OF FIGURES

PENILAIAN DUA PROSEDUR MANN–WHITNEY YANG TELAH DIBUAT PENAMBAHBAIKAN

ABSTRAK

Pengujian ke atas persamaan min dua kumpulan yang tidak bersandar merupakan satu masalah inferens yang sering berlaku dalam bidang pendidikan dan psikologi. Antara ujian yang lazim digunakan ialah ujian-ujian klasik seperti ujian *t* dan ujian Mann-Whitney. Namun demikian ujian-ujian klasik ini mempunyai kelemahan kerana prestasi mereka bergantung ke atas anggapan-anggapan tertentu. Terdapat pelbagai ujian telah direkabentuk dengan tujuan mengurangkan kesan anggapan ke atas prestasi ujian. Oleh sebab itu, pemilihan satu ujian yang sesuai merupakan satu usaha yang rumit. Kajian ini ingin mempermudahkan pemilihan ujian dengan mengenal pasti ujian yang mempunyai prestasi yang menyeluruh dan/atau menyenaraikan syarat pemilihan untuk ujian-ujian yang lazim diguna pakai. Kajian ini menggunakan pendekatan simulasi berkomputer Monte Carlo untuk menjanakan data berasaskan keadaan eksperimen untuk menilaikan suatu kaedah pengubahbaikkan Mann–Whitney yang dicadangkan oleh Babu dan Padmanabhan (2002) bersama–sama dengan ujian–ujian alternatif yang lain. Keteguhan ujian–ujian ini akan dinilai berdasarkan ralat jenis I dan kuasa ujian. Keadaan eksperimen yang diubahsuai secara sistematik terdiri daripada gabungan pelbagai bentuk taburan, homogeneiti varians dan pasangan saiz sampel manakala ujian–ujian alternatif yang dikenalpasti ialah ujian Welch (1974), ujian Mann–Whitney , ujian transformasi Johnson (1978) dan ujian transformasi Hall (1992). Hasil kajian mendapati dalam kalangan ujian–ujian yang dikenal pasti tiada yang menunjukkan prestasi terbaik merentasi semua kombinasi keadaan yang mungkin tetapi hanya prestasi terhad.

Kaedah pengubahbaikkan Mann-Whitney I didapati mampu memberi prestasi yang baik dan dapat mengekalkan ralat jenis I pada liputan kebarangkalian yang lebih luas. Pada masa yang sama, kaedah yang dicadangkan ini mampu memghasilkan kuasa ujian yang tinggi. Syarat pemilihan untuk ujian-ujian lain juga dipaparkan agar ia dapat digunakan sebagai langkah asas pemilihan dalam pendekatan adaptif.

EVALUATION OF TWO REFINED MANN-WHITNEY PROCEDURES

ABSTRACT

Testing for the equality of means across two independent groups is a common inferential problem especially in education and psychology. One of the most frequently used tests is either the classical *t* test or the Mann-Whitney test. But these classical tests are not without flaws as their performance depends on underlying assumptions. A plethora of test statistics and procedures have since appeared, designed to be less sensitive to violation of the underlying assumptions. Hence selecting the appropriate robust statistical test will be tedious. This study intends to facilitate this by identifying broader robust tests and/or providing boundary conditions of popular statistical tests. This study adopts the Monte Carlo computer simulation which generates data under experimental conditions to evaluate the small-sample behaviours of the refinement procedures proposed by Babu and Padmanabhan (2002) and its

alternatives in terms of Type I error rates and statistical power. The experimental conditions that were systematically manipulated are multiple combinations of various distribution shapes, variance heterogeneity and group sample sizes. The alternatives are Welch's (1974) test, the Mann-Whitney test, Johnson's (1978) transformation of the Welch's test and Hall's (1992) transformation of the Welch's test. The findings of this study have demonstrated that there is no statistical test that is superior to the others in all test conditions. All the identified statistical tests and procedures are specified tests. However, the proposed Refinement Procedure 1 is found to be generally more robust as it is capable of producing broader probability coverage of maintaining the Type 1 error. Furthermore the Refinement Procedure 1 is comparatively a powerful test in conditions where it is appropriate. Recommendation for the other statistical tests and procedures are made based on their respective boundary conditions discovered in this study. Statisticians will be able to utilize these boundary conditions and incorporate them into an adaptive approach of selecting a more flexible and robust statistical test.

CHAPTER 1

INTRODUCTION

1.1 Rationale

The two-sample statistical comparison is one of the most important procedures in hypothesis testing especially in educational and social behavioral sciences. Among the various statistical tests, the most often used procedure in obtaining small-sample inferences about the differences between populations especially difference in location is the *t* test. The two-sample *t* test is a test of the null hypothesis that two populations have the same mean, under the assumption that they are normally distributed with equal variances. Inferences from small-sample *t* test about the difference in their means are valid if the sampled populations deviate slightly from normality. On the other hand, when the sampled populations depart greatly from normality, then *t* test is invalid and inferences derived from the procedure are suspected. If non-normality is suspected, then there are two approaches that may be considered. They are firstly, transforming the data to promote normality and then performing *t* test or secondly, select a viable alternative test procedure (non-parametric test) which is insensitive or robust to the violation of normality. A robust test will maintain the actual Type I error rate closer to the nominal level of significance, α even when the data do not conform to the test's derivational assumptions and at the same time maintain the actual statistical power close to the theoretical power.

Transformation can be applied to correct problems of unequal dispersion. Transforming the samples to remedy non-normality often results in correcting heteroscedasticity, hence producing a comparable dispersion. A variety of transformations are available to be applied to a set of data depending on the particular type and degree of assumption violation that is present in the data. Transformations are usually chosen from the `power family' and if such transformations can be found, the transformed data may be suitable for use with *t* test. Unfortunately, applying a suitable transformation to a data is not always a simple solution and has a number of limitations. Transformation involves changing the metric in which the data are analyzed, which may make interpretation of the result difficult if the transformation is complicated. Conclusions are drawn based on the transformed scores, not the original observations.

The second approach to handling non-normality entails the selection of a test statistics that is insensitive to the deviation of normality. Non-parametric tests are those that make no assumption about the distribution of the data. They are therefore more robust when the distributions of the data are not well behaved. In such a situation, the non-parametric Mann-Whitney test is commonly used for detecting differences in location or the central tendency between samples. Even though Mann-Whitney test is a distribution free test, this test is only theoretically appropriate when the samples are drawn randomly from populations with the same second and higher –order moments. This is because the Mann-Whitney test is based on the assumption that the underlying populations from which the samples are derived are identical in shape which implies equal dispersion of data within each distribution. The shapes of the underlying population distributions, however, do not have to be normal. If the underlying population distributions are different, generally σ_u^2 is the wrong standard error for the Mann-Whitney *U* statistic and this can result in relatively poor power and unsatisfactory confidence interval for *p* (Wilcox, 2003). The value *p* is the probability that a randomly sampled observation from the first group is less than a randomly sampled observation from the second. Therefore, the Mann-Whitney test is strictly a test of the null hypothesis that the populations are identical.

Micerri (1989) concluded that out of the four hundred samples investigated, 28.4% of the distributions in the education and educational psychological fields were relatively symmetric, and that 30.7% were extremely asymmetric. With the availability of many different parametric and non-parametric statistical tests for use under different assumptions, selecting the appropriate test will be difficult.

1.2 Historical Development

The occurrences of non-normal and heterogeneous variances are fairly common in real data. The comparison of the mean of samples from populations with unknown variances has been the subject of much discussion. Several articles, e.g. by Wetherill (1960), Pratt (1964) and Zimmerman (1998) have documented these theoretical shortcomings, which unfortunately, have not always been heeded.

Wetherill (1960) investigated asymptotically the power and efficiency, as well as the level, of the *t* test and Mann-Whitney test. In his investigation, he permitted skewness as well as inequality of variances but required the two populations to approach one another as the sample sizes increased.

Pratt (1964) carried out a more comprehensive study on the effect of different population variances on the asymptotic levels of *t* test and of various non-parametric tests. The level of these tests describes the asymptotic behavior of the true level for the test when the two distributions are not equal at nominal level, α. The level of a test or unit normal deviates, K corresponding to α is computed using both the ratio of the sample sizes and the ratio of their true variances. All these procedures were investigated asymptotically which meant that the difference of the sample means divided by an estimate of its standard deviation may be treated as unit normal under the null hypothesis in really large samples from any populations (provided they have finite variances).

In his article, Zimmerman (1998) provided counterexamples to some commonly held generalizations about the benefits of non-parametric tests. The article is about a simulation study where the two assumptions of parametric statistical significance test, i.e. normality and homogeneity of variance, were concurrently violated. The findings reveal that non-parametric methods were not always acceptable substitutes for parametric methods when parametric assumptions were not satisfied. Multiple violations of assumptions can produce anomalous effects not observed in separate violations.

It is therefore acknowledged that standard distribution free tests for two sample location problem require that the populations be of the same shape so as to maintain the nominal significance level under the null hypothesis. Subsequent efforts are focused on modifying these tests so that they can be used with fewer assumptions on the shape of their populations. Several procedures have been proposed. As documented by Wang (1971), the first “exact” solution to the Behrens-Fisher problems was given by Behrens and was extended by Fisher as a correct fiducial solution. Exact test is a test where all assumptions which the derivation of the distribution of the test statistics is based are met, as opposed to the approximate test in which the approximate maybe as close as desired by making the sample size large enough. Weerahandi (1987) developed an exact test to deal with statistical testing problems with nuisance parameter and also when it is difficult to find a non trivial test with some optimal properties. Tsui and Weerahandi (1989) introduced the concept of generalized *p*-value method which is useful for developing hypothesis test. Tsui and Weeranhandi (1989) also established that the generalized *p*-value method is numerically equivalent and computationally more efficient formula for the *p*-value.

Welch provided an approximate degree of freedom solution as well as asymptotically series solution as an approximate *t* test for the problem. These two Welch’s tests are known as Welch APDF (Approximate Degree of Freedom) (Aspin, 1948) and Aspin-Welch tests (Aspin, 1949) and are recommended only when the data are normal; sample sizes are small and variances heterogeneous. Yuen (1974) introduced the modification to the Welch’s test, incorporating trimmed means (involving censoring or removing extreme observations in the tail of the distribution) and Winsorized variances (replacing most extreme observation with less extreme value in the distribution). The rationale of substituting these robust measures of location and scale for the usual mean and variance, respectively in the Welch’s statistic is to ensure a test statistic that is insensitive to nonnormality can be obtained. When handling non-normal data due to extreme observations, the standard error of the trimmed mean is less affected than the usual mean. Furthermore, the Winsorized variance compliments the corresponding trimmed mean as it is a consistent estimator of the variance of the trimmed mean.

Keselman, Cribbie and Zumbo (1997) pitted several modified test with the usual Mann-Whitney test. The modified tests highlighted in the article are the two versions of the Yuen's (1974) modification of the Welch's test and a modified Mann-Whitney test (RSKEW) presented by Randles and Wolfe (1979). The article recommended the non-parametric approach particularly the usual Mann-Whitney test because it is more powerful. Furthermore, to benefit from the modified tests, one has to know the shape of the distribution. Subsequently, these modified tests are known as specialized tests, favoring only known distribution. For heteroscedastic data that cannot be normally transformed, then alternative tests which are more robust are viable options.

Further efforts in handling heteroscedastic data focused on developing robust nonparametric tests that were intended to increase the ability of the standard nonparametric test to detect the difference between populations when the underlying distributions were asymmetrical.

Potthoff (1963) presented a conservative technique for utilizing the Wilcoxon test for the two-sample problem to test null hypotheses, like that encountered in the Behrens-Fisher problem. He recommended that no matter what the two populations were, the usual Wilcoxon test (Wilcoxon, 1945) with its variance $(m+n+1)/12mn$ replaced by 1/4[minimum(*mn*)], may be used to test the null hypothesis of the equality of the medians of two symmetrical (continuous) distribution with emphasis that the populations are of the same form even though they have different or unknown scale parameters. On the whole, the test still works for testing the equality of the medians of any two symmetrical distributions.

An approach to comparing groups based on median that currently seems to have practical value is the *T* statistics (Wilcox, 2003, page 252). In comparing two groups, the *T* statistics takes the form of

$$T = \frac{M_1 - M_2}{\sqrt{S_1^2 + S_2^2}},$$

with M be the usual sample median from the respective group and S^2 is some estimate of the standard error of the sample median M . There are many estimates of the standard error of *M* that have been proposed. Many of these proposed estimates have been studied and compared by Price and Bonett (2001). Subsequently, Bonett and Price (2002) approximated the null hypothesis of *T* with the standard normal distribution using an estimate of the standard error simply known as the Price-Bonett estimate. Wilcox (2003, 2006) suggested a similar strategy but rather than the Price-Bonett estimate of the standard error, an estimator derived by McKean and Schrader (1984) was used. The McKean-Schrader estimate of the standard error of *M* is very simple. Initially, compute

$$k = \frac{n+1}{2} - z_{.995}\sqrt{\frac{n}{4}},$$

where k is rounded to the nearest integer and $z_{.995}$ is the .995 quantile of a standard normal distribution. Next, the observed values are arranged in ascending order to form $X_{(1)} \le \cdots \le X_{(n)}$. Hence the McKean-Schrader estimate of the standard error of M is

$$\left(\frac{X_{(n-k+1} - X_{(k)}}{2z_{.995}}\right)^2$$

Hettmansperger (1973) and Hettmansperger and Malin (1975) have also proposed similar conservative tests. The former paper suggested a conservative test based on Mathisen's (1943) median test that requires no shape assumption of the populations but caution that a nominal 0.05 test can in fact be extremely conservative and thus the power of the test may be quite depressed. On the other hand, Hettmansperger and Malin (1975) have proposed asymptotically distribution-free tests based on Mood's (1954) median test.

Fligner and Pollicello (1981) developed a closely related non-parametric test, the robust rank-order test (also known as Fligner-Pollicello test) to correct some of the theoretically shortcomings of the commonly used Mann-Whitney test. The robust rank-order test was much less sensitive to the population distribution assumptions and substantially outperformed the Mann-Whitney test when the sample sizes were small or very large. For medium-sized samples, the test was likely to give false positive results but this was more a shortcoming of the normal approximation than the test itself.

Even though the robust rank-order test retained all the desirable properties of the original Mann-Whitney test statistic irrespective of the populations being identical or not, Feltovich (2003) discovered that there were disadvantages. First, when the population distributions are asymmetric, the test itself suffered from many of the same problems as the Mann-Whitney test; it performed inconsistently and was sensitive to sample sizes. Second, even when the population distributions were symmetrical, little information is available about the distribution of the robust rank-order test statistic (such as its critical values for some common levels of significance).

The second shortcoming of the robust rank-order test can be remedied by additional information concerning the distribution of the significance level. In his paper, Feltovich (2005) expand the number of critical values available to the robust rank-order test. Until now the usage of the robust rank-order test has been limited, partly due to the limited availability of exact critical values. These are available for small sample sizes. The first shortcoming, however, can only be completely overcome by looking at alternative statistical tests or techniques. Subsequently, a refined procedure based on Mann-Whitney test was proposed by Babu and Padmanabhan (2002). This refined Mann-Whitney test actually consists of two procedures known simply as Refinement Procedure 1 and Refinement Procedure 2 which highlighted the use of a resampling method namely bootstrapping.

In developing and identifying a robust statistical procedure to handle comparison of unbalanced design, one has the option of selecting a central tendency measure that is robust in response to a variety of distribution shape. One such measure is the median (i.e. the 50th percentile). In comparison with mean and other central tendency of location (e.g. trimmed means), median performs well as a point estimator because it reduces the impact of outliers (Wilcox, 1997, 1998). In many instances, the median is less subjected to sampling variability and provides a measure of central tendency that is closer to the bulk of the data as compared with the mean. Outliers can dramatically influence the variability of the data. They can be responsible for the heterogeneity of variances between two or more samples. In addition, outliers can have a dramatic impact on the value of a sample mean. Median is resistant to outliers, hence it is expected that the median test would provide a highly robust inferential test in response to varying distributional characteristic. The proposed refined Mann-Whitney test evaluates the sample differences with respect to their median values.

In the refinement procedures, the hypothesis testing is done using bootstrap, which is similar to that of randomization tests. The refinement procedures here used resampling with replacement (bootstrap) instead of the usual replication of data by all possible combinations in the randomization tests. The adaptation of bootstrap hypothesis test in the refinement procedures eliminates the question of data randomness and the concern regarding the population distribution. When determining whether two samples; sample *X* of size *m* and sample *Y* of size *n*, have been drawn from population distributions with the same central tendency, the usual Mann-Whitney test can only be employed if the populations are symmetrical. Under the null hypotheses the Mann-Whitney test implies the γ = $P(X_i \leq Y_i)$ = 0.5 and hence $E(U) = 0.5mn$ even when their scales, σ_X and σ_Y are unequal. If the symmetrical assumption is violated, then the value of γ differs from 0.5 and its value depends on the unknown distribution function of the populations. In such a case, the Mann-Whitney statistics, U is not centered correctly at 0.5*mn*. Consequently, the performance of the Mann-Whitney test displays a huge variation, depending on the distribution assumption; in some cases, it is conservative, in others, extremely liberal. To overcome this, the refinement procedures center (U/mn) at a bootstrapped estimator $\hat{\gamma} = P(X_i \le Y_i)$ and also employs the bootstrap percentile method to obtain critical values for decision making.

As acknowledged earlier, existing procedures used to handle the Behrens-Fisher problems relied on the theoretical distribution of its population, which was usually met by a large-sample size. Therefore in this refined Mann-Whitney test, we incorporate bootstrap procedures when faced with situations where the population is ill defined or when one is skeptical about the underlying theoretical distribution. In short, through combination of a robust point estimator (i.e. the median) with a flexible inferential procedure (i.e. bootstrapping), the refinement procedures are free of mathematical assumptions and this makes it a good alternative when confronted by Behrens-Fisher problems. But the refinement procedures also have their share of shortcomings. Refinement Procedure 1 and Refinement Procedure 2 produced contradictory outcomes with the former being conservative test and the latter as a liberal test as reported by Babu and Padmanabhan (2002).

1.3 Purpose of the Study

The standard statistical tests for two sample location problem were designed to test the null hypothesis when the populations were identical. Their usage for testing a broader type of null hypothesis similar to that encountered in the generalized Behrens-Fisher problem required very restrictive assumptions regarding the populations.

The primary purpose of this study is to investigate the Type I error and the power properties of several identified statistical tests that may be appropriate for testing a broader type of null hypothesis with fewer assumptions on the variances and distributional shapes of the populations. The findings of this study will provide researchers with useful information about the boundary conditions and the utility of the selected statistical test procedures. The robust statistical test highlighted in this study is a refinement of the conventional Mann-Whitney test. These Mann-Whitney refinement procedures or simply called as Refinement Procedures comprise of two procedures namely Refinement Procedure 1 and Refinement Procedure 2. The secondary purpose of this study is to review the performance of these new procedures in handling problems of unequal variances and different shapes of the populations.

1.4 Criteria and Strategy Employed In the Study

The two criteria generally employed to evaluate the performance of a statistical test are the robustness and the power of the test. The robustness of a statistical test is the ability of the test to maintain its Type I error rate. Hence, for a statistical test to be robust, the test's actual significance level must remain very close to the nominal significance level.

As for the power of the statistical test, this is an equally important criterion that will indicate how effective the test is in detecting treatment differences which in actual fact existed. The power of a statistical test can be viewed as the probability that a decision made is correct.

Armed with these two criteria, the validity of a statistical test can then be evaluated and compared. Both the null and non null experimental test with various study conditions can be modeled through Monte Carlo simulations. The study conditions are usually distributional assumptions that the statistical test is expected to be appropriate. These extreme conditions are usually conditions that violate the assumptions of the statistical test. The proportion of rejections by the statistical test is then tabulated. Under the null condition, this proportion of rejections is an estimate of the Type I error rate for the given experiment and is recorded as $\hat{\alpha}$. When the non null conditions are modeled, then this proportion of rejections will represent the empirical power of the experiment.

The choice statistical test will be the statistical test that maintains its $\hat{\alpha}$ within an accepted interval of α based on Bradley's (1978) criterion and its empirical power closer or higher than that of the predetermined power rate. A more elaborate discussion on the Monte Carlo evaluation procedure adopted in this study will be disclosed in Chapter 3.

1.5 Implication of the Study

The results of the evaluation, which is incorporated in Chapter Four, will provide some idea of the strength and weakness of the selected statistical tests. The result will also review the relative performance of the modified tests with the standard statistical test and whether it is worth the initiative. It will also help researchers to determine which statistical test they should adopt under specific conditions of concern.

1.6 Limitation of the Study

In comparing two groups we adopted the approach of comparing robust measures of location and scale. Despite this, there is very little attention focused on global comparisons of two distributions. The basis of global comparisons is that if two distributions differ, they might do so in many complicated and interesting ways that might not be reviewed by the difference between the single measures of location or scale (Wilcox, 2005). This approach in which the entire distributions might be compared is called shift function. This was basically developed by Doksum (1974, 1977) and also Doksum and Sievers (1976). Shift function measures how much the control group must be shifted so that it is comparable to the experimental group at a particular quantile.

In situation, where there are two different distributions with equal means and variances, it will be more appropriate instead to adapt the shift function approach by comparing the quantiles of the two groups. This situation usually arises if the two distributions differ and are skewed in the opposite direction. The distributions considered in this study are from similar distribution skewed in the same direction.

1.7 Organization of the Thesis

In brief, Chapter One provided an introduction of the study which included the rationale, historical development, the purpose, significance of the study, criteria and strategy proposed and finally the limitation of the study. Chapter Two introduces the review of literature of the two sample problem, test description and the evaluation of the test’s robustness and power. Chapter Three proposes the research method and testing framework in this study. This chapter also outlines the procedure of generating and manipulating selected distributions based on various violations of test assumptions. Chapter Four presents the findings and results of the Monte Carlo simulation study of the robustness and power of the two sample tests. This chapter contains the characterization of Type I error and statistical power for each test across multiple violations. Chapter Five summarizes the findings and discusses both the relative strengths and implications of the tests. Recommendations are also made for each test with regard to its general and specific robustness.

CHAPTER 2

LITERATURE REVIEW

2.1 Organization of Chapter

This chapter presents a review of literature related to this study. It includes a review of the two–sample test under the non-directional alternative hypothesis in Section 2.2. Section 2.3 contains the description of the statistical tests which will include test assumptions, test procedures and their sampling properties. Section 2.3 also contained literature related to the performances of these tests. The statistical tests identified for investigation in this study are Welch’s test, Mann-Whitney test, Johnson’s transformation, Hall’s transformation and the two proposed refinement procedures, RF1 and RF2.

The objective of this study is to investigate the robustness and statistical power of these statistical tests in conditions such as sample size combinations, variance ratios, and various degrees of skewness and kurtosis of the populations. Section 2.4 presents a framework on the evaluation of these statistical tests in terms of Type I error rates and statistical power. Section 2.4 also presents related literature regarding the standard for statistical significance employed in the testing procedure.

2.2 The Two-Sample Problem

Statistical tests have been developed to permit comparisons regarding the degree to which qualities of one group of data differ from those of another group. Each statistical test is based on certain assumptions about the population(s) from which the data are drawn. If a particular statistical test is used to analyze data collected from a sample that does not meet the expected assumptions, then the conclusions drawn from the results of the test will be flawed. The statistical comparison to determine the difference between two samples usually starts with the formulation of a null hypothesis, H_0 to the effect that both samples come from identical populations against the alternative, H_1 which indicates a difference between both samples. The test procedures that follow suit are influenced by the alternative hypothesis of the two-sample problem being contemplated.

R. A. Fisher proposed a method for testing a hypothesis which is related to the maximum likelihood estimators. The likelihood ratio test is a statistical test with a likelihood ratio test statistic denoted as Λ, in which the numerator corresponds to the maximum probability of an observed result under the null hypothesis. The denominator of the likelihood ratio corresponds to the maximum probability of an observed result under the alternative hypothesis. The value of Λ can be used to make decision between the null hypothesis and alternative hypothesis. If the distribution of the likelihood ratio corresponding to a particular null and alternative hypothesis can be explicitly determined, then the ratio Λ can be directly used to form decision. Unfortunately, the likelihood ratio method does not always produce test statistic with known distribution. The remedy will be then to transform the likelihood ratio into log-likelihood ratio. For a large *n*, the transformed log-likelihood ratio –2log(Λ) has approximately a χ^2 distribution with $r_o - r$ degrees of freedom. When determining the degree of freedom, r_o denotes the number of free parameters in the parameter subset specified by the null hypothesis and r denotes the number of free parameters specified in the parameter space. Hence, the likelihood ratio procedure provides a general method of developing statistical test. Many common test statistics such as *Z*-test and *F*-test can be phased as log-likelihood ratios. The likelihood ratio test requires that the distribution of the sampled populations must be known otherwise the likelihood functions cannot be determined and the method cannot be applied.

The goal of a hypothesis test is to decide, based on samples from populations, which of the two complementary hypotheses is true. Therefore these hypotheses must be formulated in terms of some reasonable easily interpreted measure of difference. Let the first sample consists of *m* independent observations; $\mathrm{X}_1, \mathrm{X}_2, \mathrm{K}, \mathrm{X}_\mathrm{m}$ on a random variable *X* with distribution $G(X)$ and the other sample with *n* independent observations; $Y_1, Y_2, \mathrm{K}, Y_n$ on Y with distribution function $H(Y)$. The two-sample problem involve pitting the null hypothesis, $G(X) = H(Y)$ against the alternative hypothesis, $G(X) \neq H(Y)$. The alternative hypothesis must be formulated in terms of some reasonable easily interpreted measure of difference.

Among such measurement of difference, one of the simplest and most easily interpreted is the difference in location of distribution that is otherwise identical. This measurement of difference models on $G(X) = F(X - \theta_x)$and

$H(Y) = F(Y - \theta_y)$ where the c.d.f. F is continuous and is symmetrical about the origin. Thus the parameter Δ expressed as $\Delta = \theta_x - \theta_y$, represents a shift in location between the two distributions. θ_x and θ_y represent the medians of the distribution of $G(X)$and $H(Y)$ respectively, or equivalently as $\Delta = \mu_x - \mu_y$ where, provided they exist, μ_y and μ_x are means of $G(X)$ and $H(Y)$. If there is a difference between the two population distribution functions then that difference is reflected and realized in a difference in the location of the distribution. Hence in the location problems, the null hypothesis can be reformulated as $H_0 : \Delta = 0$.

A statistical test is generally conducted by means of a test statistic for which the probability distribution is determined on the assumption that the null hypothesis is true. This assumed distribution is known as null distribution of the test statistic. Hence, when calculating the test statistic, which is purely a function of data, its probability distribution should be calculated under the assumption that H_0 is true. The usual assumptions of the null distribution are that it is normal or at least symmetrical and homoscedastic. The test will suffer from distorted Type I error and loses its statistical power when the data is not normal and/or when heteroscedasticity is present.

For example, the usual *t* statistic for small sample test is calculated based on the pooled variances, applicable for moderately large samples and when the sampled populations which are approximately normal. The *t* statistic is of the form

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{\sigma}^2_{pooled}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$\hat{\sigma}^2_{pooled} = \frac{(n_1 - 1)\hat{\sigma}_1^2 + (n_2 - 1)\hat{\sigma}_2^2}{n_1 + n_2 - 2}$$

with $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ being the two sample variances (Sincich, 1993).

This *t* statistic is still valid when testing for the difference in the means of two populations even though the variances of the sampled populations are unequal provided the sample sizes are the same. In cases where the sample sizes and population variances are not equal, an approximate test for the difference in the means of the populations can be performed by modifying the degrees of freedom associated with the *t* distribution.

When conducting statistical test, the decision on the choice of parametric or non-parametric test is perhaps one of the oldest fundamental analysis decision confronting researchers in the field of psychology and education. Making the right choice is of utmost importance because its implication will affect both statistical and substantive inference. Despite the implications of this important decision, many researchers unerringly employ tests by overlooking or violating assumptions of the test. With the advent of the computer and subsequently more powerful computer, the computer is used to simulate various samples of distributions. These simulated distributions are then systematically manipulated so as to examine the sensitivity of standard parametric and nonparametric test to varying degrees of violations to the assumptions of these standard tests. Further more the versatility of the computer simulation enables the possibility of examining multiple violations to the assumptions commonly encountered.

This study will seek a robust testing procedure based on the ability of the statistical test to maintain its Type I error and at the same time a powerful test in the face of assumption violations.

2.3 Statistical Tests Description

When the required assumptions for the usual parametric test are violated, there are alternative strategies for the testing procedure. The usual strategies are robust procedures, non-parametric tests and resampling procedures (see Figure 2.1). The tests described in this section are the common tests for each strategy and this research seeks to identify a general robust test to handle the Behrens-Fisher problems. A statistical test is considered robust if it is not affected by violation of assumptions that justify it.

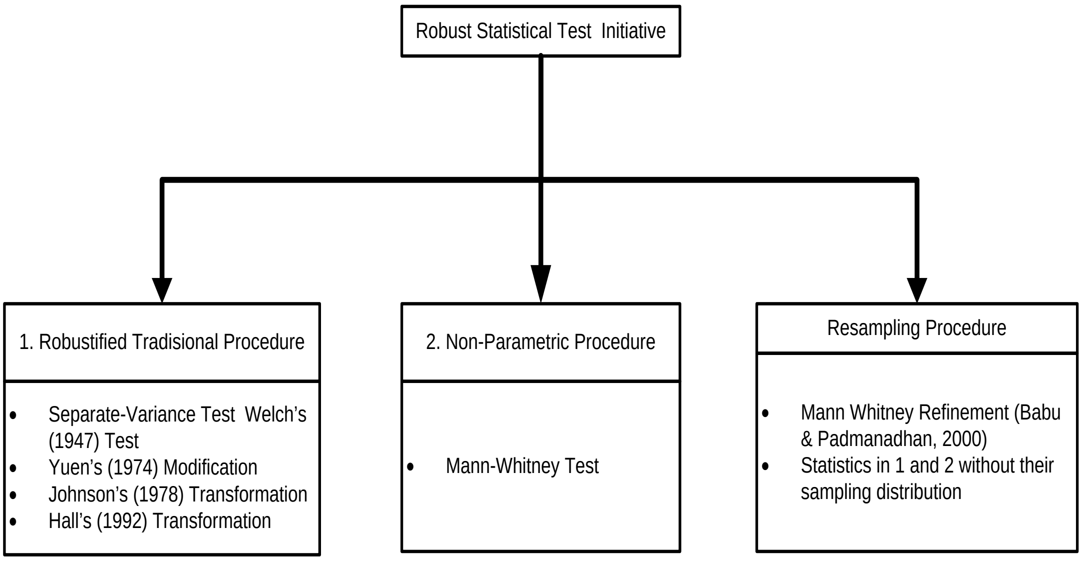
Robust Statistical Test Initiative
1. Robustified Tradisional Procedure
• Separate-Variance Test Welch's (1947) Test
• Yuen's (1974) Modification
• Johnson's (1978) Transformation
• Hall's (1992) Transformation
2. Non-Parametric Procedure
• Mann-Whitney Test
Resampling Procedure
• Mann Whitney Refinement (Babu & Padmanadhan, 2000)
• Statistics in 1 and 2 without their sampling distribution

Figure 2.1: Robust test initiative

2.3.1 Robustified traditional procedures.

Robustified tests are based on parametric test statistics in which the estimates of the parameters like means or standard deviation are replaced by robust estimates like trimmed means and Winsorized variances.

Separate – variance *t* test. The separate-variance *t* test introduced by Welch (1938, 1947) and Scatterthwaite (1946) is one of the widely used and best known procedures for testing the difference in the means of two populations when both their variances and sample sizes are unequal. The separate-variance *t* test or Welch's test is calculated from an unpooled error term and the degrees of freedom are modified to determine the rejection region of *t*. The statistic,

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}}} \qquad (2.1)$$

is calculated based on an unpooled error term, and the degrees of freedom are modified as,

$$df = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{\hat{\sigma}_2}{n_2}\right)^2}{n_2 - 1}} \qquad (2.2)$$

With computational software we can get *t* and *F* values with rational degrees of freedom. The Welch's *t* test also served as a model for other approaches.

When sampling from a skewed population with small sample sizes, the usual group means and variances are greatly influenced by the presence of extreme observations in the population distribution. The standard error of the usual mean becomes seriously inflated when the underlying distribution has heavy tails. Lix & Keselman (1995) noted that the Welch's test is generally robust only to variance heterogeneity under normality. To obtain a test statistic that is insensitive to non-normality, the usual mean and variance is substituted by a robust measure of location and scale. To deal with the effect of extreme values, one of the strategies is to give less weight to these extreme values at the tails and instead focus more on those values near or around the centre of the distribution. This is usually implemented by either removing these extreme values or pulling them in nearer to the centre of the distribution. From the wide range of robust estimators (Gross, 1976; Lind & Zumbo, 1993), the trimmed mean and Winsorized variance are most appealing due to their computational simplicity and good theoretical properties (Wilcox, 1995).

Yuen's modification. Yuen (1974) suggested that trimmed means and Winsorized variance be used in conjunction with Welch's (1938) statistics. Hence the suggested test is known as the Yuen-Welch test or just the Yuen's test. The Yuen's test is for testing the hypothesis that two independent groups have equal trimmed means.

$$H_0 : \mu_{t1} = \mu_{t2}$$

The Yuen's test is designed to allow unequal Winsorized variances. The standard Welch's test is incorporated into the Yuen's test. In situations where trimming is not required $(\gamma = 0)$, the Yuen's test is reduced to the Welch's test which is meant for comparing means that allows unequal variances.

In trimmed means, outliers in both tails are simply omitted. Let $Y_{(1)j} \leq Y_{(2)j} \leq \text{L} \leq Y_{(n_j)j}$ represent the ordered observations associated with the *j*th group, (*j* = 1, 2). Let $g_j = \lfloor \gamma n_j \rfloor$ be the number of observations trimmed for each tail. The symbol $[\ ]$ operates on γn_j gives the nearest integer less than or equal to γn_j. The value γ is the proportion of the observations to be trimmed from each of the tail of the distribution. After trimming, the effective sample size for the *j*th group becomes $h_j = n_j - 2g_j$. The respective sample trimmed means are computed from these trimmed samples using:

$$\hat{\mu}_{tj} = \frac{1}{h_j} \sum_{i=g_j+1}^{n_j-g_j} Y_{(i)j} \ .$$

Similarly, to compute the Winsorized variance, the outliers in the distribution are identified. Instead of trimming off the tails of the distribution, they are replaced with the maximum and minimum observations respectively from the trimmed data as shown below.

$$X_{ij} = \begin{cases} Y_{(g_j+1)j} \ \ if \ \ Y_{ij} \leq Y_{(g_j+1)j} \\ Y_{ij} \ \ if \ \ Y_{(g_j+1)j} < Y_{ij} < Y_{(n_j-g_j)j} \\ Y_{(n_j-g_j)j} \ \ if \ \ Y_{ij} \geq Y_{(n_j-g_j)j} \end{cases}$$

Foremost, the Winsorized mean, which is an integral portion in computing Winsorized variance, is determined. The Winsorized mean is computed as

$$\hat{\mu}_{wj} = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ij}$$

The sample Winsorized variance is then computed by using

$$\hat{\sigma}^2_{wj} = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \left(X_{ij} - \hat{\mu}_{wj} \right)^2$$