The Performance of Statistical Inference after Model Checking

Mohd Iqbal Shamsudheen

Thesis for Submission at University College London

Research Degree: Statistical Science

DECLARATION

I, Mohd Iqbal Shamsudheen, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signature:

Date: 23rd September 2020

To my mother,
the strongest person I know, and to whom I owe everything.

To Ayra Qadeeja,
may you always pursue your dreams.

ACKNOWLEDGEMENTS

First and foremost, Alhamdulillah, praise to almighty Allah, for without His blessings and His will, I would not have been able to complete this thesis. I would like to extend my highest, most sincere gratitude to my supervisor Christian Hennig who has suffered with good grace and humour my many, *many* stupid questions (and slow writing!). Thank you for the opportunity, guidance, advice, support and patience. I cannot imagine a better teacher to take me on this journey.

There are not enough superlatives to describe how wonderful and important the love given to me by my mother has been. I could not have gotten this far without her support. I would like to thank my closest friends, Norli and Hilyati for their love, understanding and encouragement throughout this four years and ten thousand kilometers away from each other. I would be remiss to not give thanks to the most important people in my life in London, my housemates, my family away from family; Amar, Jalal, Fakhrul, Wan Bad and Hafiz. I am eternally grateful and I hope I will be able to repay the love and delicious food (!) you lads have given me.

A special thanks goes to Wittawat and Sam for your *invaluable* help (I could not have done this without you guys, seriously!). Thank you Anna, Yin Cheng, Ain and the PhD students at UCL for the lunch sessions where we talk about everything other than our PhDs. It was a welcome distraction. I would also like to thank all the support staff in the department for their help and accommodations. Thank you also to MARA for sponsoring my PhD. Not forgetting those who have directly or indirectly helped me on this thesis, there's too many of you to thank. Last but certainly not the least, I want to thank my family, my wife Hafizah, and my sisters; Shaqilah, Shafiqah and Shaqinah for being there to always encourage me. It means so much to have all of you in my corner, thank you.

May Allah reward all of your kindness in abundance in this world and the next, insha Allah.

ABSTRACT

Most standard statistical inference procedures rely on model assumptions such as normality, independent and identically distributed and the like. Often in practice, such assumptions are formally tested before applying the inference. Such a procedure does not ensure that the model assumptions are really fulfilled because the standard theory for popular inference tests does not take into account that the data has been selected by a previous model check. Applying a misspecification test violates the very model assumption it was meant to enforce. ("misspecification paradox"). In practice it is useful to have an alternative test in the case that the misspecification test rejects the model assumption. However, this does not completely address the misspecification paradox because there is still a certain probability that the model assumption is rejected when it is in fact true, and vice versa.

This thesis is about investigating, theoretically and by simulations, the performance of such a combined procedure. A novel simulation process is proposed where samples can be randomly chosen from a situation where the model assumption is fulfilled or violated. A few combinations of distributions and statistical tests are considered and both level and power are presented and discussed. Although the levels show no strong evidence of choosing the combined procedure over the tests run without model checking, the power plots show that in certain conditions, it can be more powerful.

A theory is presented where it is shown that in a particular situation and with reasonable assumptions, the combined procedure does have a higher power compared to unconditional tests. The assumptions were relaxed a little and the same conclusions could be made. Finally, a three stage testing procedure in two different scenarios, distributional shape and linear regression significance, are presented and discussed. The same conclusions can be made from the levels and powers.

IMPACT STATEMENT

Hypothesis testing is an important part of many statistical analysis. It is useful to infer the result of a hypothesis performed on a sample from a larger population. Although there are differing opinions on how one goes about performing a statistical hypothesis test, namely Fisher versus Neyman-Pearson and frequentism versus Bayesianism, it is clear that testing a hypothesis on a sample to infer the population where the sample comes from is a powerful tool in a statistician's tool box. The proposed methods in this project seek answers to fundamental questions in the field, namely the model assumptions of a test. The new methods will have impact in at least three ways:

Hypothesis tests, like many methods in statistical analysis, requires some assumptions for the test to be valid. Checking the assumptions is a subject of quite a number of discussions. There is much confusion especially in the medical research field about whether to check model assumptions. The first impact of this project is to have a review of these discussions and present them in a comprehensive manner to help applied researchers have a better understanding of the issues. We have found that there is no agreement on whether model assumption checking should be done. The theory of model-based method usually relies on the implicit assumption that there are no data-dependent pre-selection or pre-processing. However in practice, researchers tend to use model checking methods before carrying out the final hypothesis test.

Secondly, this project introduces a new method in the simulation by randomising the samples that go into what we refer to as a combined procedure. This challenges the combined procedure to distinguish between a model constrained test or an unconstrained alternative test. This introduces a "*Bayesian flavour*" into a frequentist simulation study where there is prior distribution on the random generation of samples to be entered into the combined procedure. This could potentially open up a new area of simulation for other frequentist methods.

Most importantly, we show a positive result in favour of the combined procedure. It is shown that the combined procedure has a larger power than tests where the assumptions were not tested when both situations of fulfilling and violating the model assumptions can occur and with some realistic but strict supposition. This result is useful for researchers particularly those not from a statistical background to carry out hypothesis testing with model checking. It provides a better justification of what is

widely recommended in practice than what has been published up to now, which is helpful against the confusion.

These three impacts outlined here aims to mainly help research practitioners sort out the confusion surrounding the problem of model checking by first outlining the discussions that have taken place in literature and then proposing a situation where model checking can be helpful to help ease the confusion. This of course does help the scientific community in general.

OUTPUT

Papers in Consideration

M. I. Shamsudheen and C. Hennig. Should we test the model assumptions before running a model-based test?. Statistical Methods and Application. Pre-print available at https://arxiv.org/pdf/1908.02218.pdf

Poster Presentation

M. I. Shamsudheen and C. Hennig. Some issues on post misspecification-testing inference. Poster session presented at the ISI World Statistics Congress, Kuala Lumpur, Malaysia, August 2019

Invited Talk for a Special Session

C. Hennig and **M. I. Shamsudheen**. On whether to check the model before doing model-based inference. 12th International Conference of the ERCIM WG on Computational and Methodological Statistics (CMStatistics 2019), December 2019

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AR(1) AutoRegressive model of the first order

AU Alternative Unconstrained

CLT Central Limit Theorem

CP Combined Procedure

DW Durbin-Watson

MC Model-based Constrained

MS Misspecification

ncp non-centrality parameter

SW Shapiro-Wilk

WMW Wilcoxon-Mann-Whitney

1

INTRODUCTION

> ‘‘I almost wish I hadn’t gone down that rabbit-hole - and yet -
> and yet - it’s rather curious, you know, this sort of life!’’
> Alice (Alice in Wonderland)

When any statistical model is used, data are needed and statisticians continuously endeavour to model the link between data and how the world actually works. A model is merely a tool used to understand the intricate structure of the real world represented mathematically. This is, naturally, something that is created by the human mind and therefore, is restrictive. What goes on in the world is a domain that is distinct from what goes on in modelling, which is a human activity set up by humans (Hennig, 2010). Despite this, it does not mean that models are not useful, as explained by Box (1979):

> *Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law PV = RT relating pressure P, volume V and temperature T of an “idea” gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behaviour of gas molecules.*

To perfectly represent the world in a model is impossible, however, statistical models account for this with model assumptions. When a model is chosen, it is based on the statistician’s perception on how the world works based on knowledge that he has, namely his own personal reality, and he must always be mindful of this fact. The purpose of modelling something is to make sense of or discover the world around and there lies the constraint in the concept of statistical modelling, namely the pursuit to understand and quantify something that is nontrivial. As Hand (2014) stated:

> *In general, when building statistical models, we must not forget that the aim is to understand something about the real world. Or predict, choose an action, make a*

decision, summarize evidence, and so on, but always about the real world, not an abstract mathematical world: our models are not the reality — a point well made by George Box in his oft-cited remark that "all models are wrong, but some are useful".

1.1 MOTIVATION

Following the statement by George Box, all mathematical and statistical models are wrong because they will never be able to be reality, they are just models of reality. However, not all hope is lost. Models usually have a set of assumptions for them to be useful. From the literature reviewed, it seems that researchers sometimes overlook this or maybe do not fully understand the meaning of these assumptions. They tend to resort to just copying methods used by studies they have read.

Strasak et al. (2007a) conducted a bibliometric analysis of all original research articles published during the first half of the year 2004 in Volume 30, Numbers 1-26 of the New England Journal of Medicine (NJEM) and Volume 10, Numbers 1-6 of Nature Medicine (NMed). They reviewed the use of statistical methods used in these medical journals. At least one kind of inferential statistical method were used in 94.5% out of 91 articles in the NJEM and 82.4% out of 34 papers in NMed. Among the most frequently used methods were the t-test and non-parametric tests at 36.8% and 24.8%, respectively, out of the total number of papers. A subgroup of 53 papers (31 from NJEM and 22 from NMed) were further assessed. It was observed that 20.8% of these articles contained the use of wrong or suboptimal statistical tests resulting from the incompatibility of tests with examined data, inappropriate use of parametric methods, or using wrong statistical tests for the hypothesis under investigation. It was also observed that 63% of the papers that use the t-test fail to report whether the test assumptions were checked and 10.9% carried out improper multiple pairwise comparisons without α-level correction. Similarly, Strasak et al. (2007b) assessed 15 papers from *Wiener Klinische Wochenschrift* and 7 papers from *Wiener Medizinische Wochenschrift* and found that the practice of improper use of statistical methods and failure to validate model assumptions were also found in these Austrian medical journals. It was observed in the papers that reported usage of t-test, 41.2% failed to report whether the test assumptions were checked.

A Chinese study carried out by S. Wu et al. (2011) reviewed articles from 10 Chinese biomedical journals regarding the misuse of statistical methods in 1998 and 2008. All the original articles published, 1,335 in 1998 and 1,578 in 2008, were reviewed. Out of these, a total of 1,334 or 45.8% were reported to have incorrectly used either one of the most common statistical methods in these journals namely$t$-tests, contingency tables,

analysis of variance (ANOVA) or rank based non-parametric tests. The authors mention that the most common error committed was the inappropriate choice of statistical methods. The most common misuses of *t*-tests (1062/2913, 36.45%) are the use of multiple *t*-tests to compare means of more than two groups (282/1062, 26.55%), use of *t*-test under a non-parametric setting (149/1062, 14.03%) and the use of the *t*-test to conduct repeated-measure data analysis (133/1062, 12.52%). The errors committed when using rank transformation non-parametric tests (62/254, 24.41%) include the use of multiple pair-wise comparison for multiple groups (34/62, 54.84%) and using the wrong type of rank sum test for different study types (7/62, 11.29%). They offer the possibility that researchers did not give enough attention to the distributional characteristics of the variables and the nature of the data. Hence, it was recommended to Chinese researchers to increase the quality of writing, to raise the level of knowledge of statistical methods among clinicians and to include a statistician as a consultant.

Sridharan and Gowri (2015) studied the statistical errors committed by medical researchers in eight Indian medical and surgical journals over a period of 2 years. They collected 195 articles from 2005 and 220 articles from 2006. They found that 33.7% of these articles did not mention checking normality prior to parametric tests and 28% of the articles used multiple statistical tests, ranging from 14 to 126 times, without adjusting the *p*-value. One article even reported the *p*-value as 1.3, which is obviously wrong. Hassan et al. (2015) compared errors in statistical methods made in articles from ten Indian medical journals in 2003 and 2013 to ascertain whether the statistical methodology used in these journals has improved in one decade by analysis of the number of errors committed. They reviewed 588 articles from 2003 and 774 articles from 2013. The most used statistical methods is the *t*-test, contingency tables and ANOVA. They observed that the proportion of erroneous statistical analyses had not decreased significantly, 25% in 2003 compared to 22.6% in 2013. However, they noticed an increased use of rank based nonparametric tests in 2013, which they assume indicates that more attention were being paid to the assumptions of parametric tests.

More recently, a study was done in Egypt by Nour-Eldein (2016) that assessed statistical methodology errors in family medicine articles by authors affiliated with the Suez Canal University over 5 years. Out of the 60 papers reviewed, the author found that a quarter (25%) "*failed to report that test assumptions were not violated*" as well as a few more errors that were made by medical researchers. It should be noted that this does not mean that the assumptions were violated and the researchers used the methods anyway. These studies are limited to assessing the information reported in the publications. There is simply no way of knowing the unpublished details unless the authors were contacted and asked to show that the assumptions were not violated. It was also suggested

that some authors merely copied methods from previous work without actually knowing what is needed before using a statistical test. This could result from the fact that model checking was not reported and this practice was copied by subsequent studies.

Altman (2002) said "*once incorrect procedures become common, it can be hard to stop them from spreading through the medical literature like a genetic mutation*". This survey indicates that in the medical community at least, there is still a lack of understanding about the philosophy of hypothesis testing, either by checking model assumptions or at the very least understanding the risks of using multiple tests. Since the medical field deals in human life, a treatment plan or a new drug that is approved using an incorrect use or interpretation of a statistical test can be quite costly.

Keselman, Huberty et al. (1998) reviewed articles from 17 journals of educational and behavioural science research. The authors claimed to provide evidence that the vast majority of educational researchers conducted statistical analyses without taking into account the distributional assumptions of the procedure they were using. Out of the 411 articles reviewed, 61 had a between subjects univariate design. 13 out of the 61 did not report any cell of group standard deviations for any of the dependent variables under investigation. When the authors looked at the remaining articles it was found that the ratio of the largest to smallest standard deviation had a mean of 2.0, a median of 1.5 and a maximum of 23.8. In the articles that carried out factorial studies, the ratios has a mean of 2.8, a median of 1.7 and a maximum of 29.4. This shows that in the majority of the studies, the samples do not show variance homogeneity. Yet, tests that assumes variance homogeneity were used. Only in 12 articles were the violations of the distributional assumptions mentioned as a source of concern by the author(s).

Choi (2005) mentioned that the most common statistical errors involve "*failure to recognize the correct distribution of the data*", leading to incorrect choice of descriptive and inferential statistics. Of course, it is impossible to truly define the correct distribution of any data and also one does not need assumptions of the distribution to do descriptive statistics. According to Olsen (2003), a frequent error made in data analysis is the application of statistical tests that assumes a normal distribution on data that actually follows a skewed distribution.

Model assumptions needs to be checked before running any statistical inference test, this is taught in elementary statistics courses. However, no agreement can be reached as to how this must be done. This is presented in Rule number 8 of the Ten Simple Rules for Effective Statistical Practice by Kass et al. (2016) aptly named 'Check Your Assumptions'. The authors mention that "*every statistical inference involves assumption ... even the so-called "model-free" techniques require assumptions, albeit less restrictive assumptions*". The availability of software that can perform analyses without any attention to the assump-

tions can cause misleading results. At the very least, how well the model fits the data should be checked with visual displays such as plots of the data or residuals or using some basic techniques for assessing model fit. Linearity often works well as an indicator or as a depiction of a general trend. Another example could be to check the assumption of normality of a data set using a misspecification test. Other assumptions include, but not limited to, the assumption of independence, assumption of equal variances and assumption of normality of residuals.

However, there is no such thing as data or sample that follows a normal distribution. We argue that the normality assumption is always violated. Any measurements taken by an equipment is truncated to a range of values and has limited accuracy. As we know, the normal distribution is continuous between $(-\infty, \infty)$. This fact alone invalidates normality and subsequently, any tests for normality. Micceri (1989) and Wilcox, Charlin and Thompson (1986) observed that the data collected by educational and psychological researchers rarely, if ever, come from populations that are characterized by the normal density function or by homogeneous variances.

The thesis sets out to review practices that use model assumption checking that subsequently makes a decision about a hypothesis. We then discuss the already existing studies done to investigate the validity of these practices. Authors such as Strasak et al. (2007a), Strasak et al. (2007b), S. Wu et al. (2011), Keselman, Othman and Wilcox (2013) and Kass et al. (2016) are for testing model assumptions either with a formal test or graphical methods. Others like Bancroft (1964), Arnold (1970), C. V. Rao and Saxena (1981), Saleh and Sen (1983), Moser, Stevens and Watts (1989), Moser and Stevens (1992), Gupta and Srivastava (1993), Albers, Boon and Kallenberg (1998), Albers, Boon and Kallenberg (2000a), Albers, Boon and Kallenberg (2000b), Zimmerman (2004) and Rochon and Kieser (2011) are against model checking. However, the studies reviewed were done in a restrictive context. We aim to study some unique situations and perhaps be able to propose a general guideline to approach problems of model checking.

1.2 STRUCTURE OF THESIS

Chapter 2 follows this introduction with a full literature review by first discussing some historical background about statistical inference testing. This is followed by a discussion about model assumptions and methods of dealing with them. The misspecification paradox is introduced which forms the basis of the motivation of this study. Some practices of misspecification testing is presented testing distributional shape (equality of means), linear regression slope coefficient significance and post selection inference.

Chapter 3 presents some theoretical framework of the simulation process proposed in this thesis. Several definitions are made to help formulate the theory statement. Rigid assumptions of independence between the misspecification test and the main test are made to serve as a useful starting point to proof the theory. These assumptions are then relaxed by adding some small measure of dependence. The theory shows that in a certain situation, the combined procedure can actually be useful with more power.

Chapter 4 first presents a replication of the work done in Zimmerman (2004). This is followed by the introduction of the suggested framework for our simulation where two samples are generated randomly either from a distribution that violates the model assumption or a distribution that does not violate the model assumption. We will call this framework the combined procedure and proceed to define what it is. A few combinations of distributions and hypothesis tests are presented and discussed in terms of their level and power.

Chapter 5 then continues the simulations by looking at a three stage procedure where two model assumptions could be tested. First we look at a setup where the main null hypothesis is that two samples have equal distributions. Two model assumptions are checked namely the assumption of normality and the assumption of equal variance. These two model assumption tests choose between three tests; one where both model assumptions are violated, another when the assumption of normality is not violated but the assumption of equal variance is violated and lastly one where both model assumptions are not violated. This procedure is then repeated in a linear regression problem where we test the residuals for model assumptions and decide what the final model will be.

2

LITERATURE REVIEW

> ‘‘Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.’’ George E. P. Box

2.1 STATISTICAL INFERENCE & HYPOTHESIS TESTING

Statistical inference is the process of drawing conclusions about populations or scientific truths from data. R. A. Fisher (1922) describes the general goal of statistics as follows:

> *In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.*
>
> *This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion.*

Fisher describes here the characteristics of a statistical inference that consists of inferring using the data observed, a sample, an instance of an underlying model population. The merit of this approach is that the characteristics of this population can be described exhaustively by a small number of parameters. This obviously depends on the appropriateness of the model population selected.

Fisher then outlines the general task of statistics, specifically the reduction of data, into three types of problems. First, the problem of specification consisting of forming a model, which usually cannot be derived and requires deliberation and understanding of

the way in which the data are supposed to originate to choose the mathematical form of the distribution. Fisher attributes this step to the logic of inference, typical of his inductive inference, that is, the transition from concrete data to mathematical models. Second, the problem of estimation whose formulation requires some mathematical-statistical model. This involves the choice of method to calculate a statistic from a sample, for example the mean, to estimate the value of the parameter of the hypothesized distribution. The third problem, the problem of distribution that includes the deduction of the exact nature of the distribution of the parameter estimates derived from samples. Fisher continues:

> *As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested a posteriori. We must confine ourselves to those forms which we know how to handle, or for which any tables which may be necessary have been constructed.*

For Fisher, a model is an entire class of hypotheses, and he terms the selection of one hypothesis from this class as specification. In the significance testing approach introduced by Fisher, a null hypothesis is potentially rejected on the basis of data that is significant under its assumption, but the null hypothesis is never accepted or proved.

In 1933, Neyman and Pearson criticized the treatment of the null hypothesis put forth by Fisher in his statistical testing. Fisher started with a null hypothesis of which a test might reject the said null hypothesis. To test say another null hypothesis, another test must be carried out accordingly. Neyman and Pearson insist that the situation must be treated differently. In other words, a model should have two competing hypothesis, null versus alternative, whereby a test on the data should be able to choose which hypothesis is to be preferred. A Neyman-Pearson test, as Neyman interprets it, is a rule of inductive behaviour as opposed to the then more accepted term inductive reasoning:

> *to decide whether a hypothesis H, of a given type to be rejected or not, calculate a specific character x of the observed facts; if $x > x_0$ reject H; if $x \leq x_0$ accept H* (Neyman and Pearson (1933), p. 291).

Casella and Berger (2002) discussed the distinction between using the term accepting the null hypothesis and not rejecting the null hypothesis on a philosophical level. What Neyman and Pearson mean by the term "accept" is not to believe that null hypothesis is true, but rather to use the null hypothesis as a basis for further action. However, this is often misunderstood. The term that would be used throughout this thesis is not

rejecting the null hypothesis. This is simply interpreted as there is not enough evidence to reject the null hypothesis. This is due to the fact that Fisher did not use an alternative hypothesis; hence there was no concept of accepting an alternative in his construction of significance tests. Therefore, rejecting or not rejecting a null hypothesis is identified with making a specific decision, for example, to publish a result or to announce a new effect. The set of outcomes related to rejecting the null hypothesis is known as the rejection region as such;

> *it may often be proved that if we behave according to such rule, then in the long run we shall reject H when it is true not more, say, than once in a hundred times, and in addition we may have evidence that we shall reject H sufficiently often when it is false* (Neyman and Pearson (1933), p. 291).

Later in the same paper, different notations denote the null and alternative hypotheses namely, H_0 and H_1 respectively. Subsequently, Neyman and Pearson also introduced the idea of Type-II error. Errors happen when the wrong decision can be made when choosing between two opposing hypothesis. One can commit a Type-I error where a true null hypothesis is mistakenly rejected and a Type-II error where a false null hypothesis is incorrectly accepted. The assessment of these two errors should be made an objective of research. Neyman remarked in another paper, Neyman and Iwaszkiewicz (1935), that "the fewer the errors of one kind, the more there are of the other".

These two approaches from Fisher and Neyman & Pearson are a cause of many debates and disputes. However up to the present, the way people use these testing procedures is plagued by at least one of two types of fallacies; the fallacy of acceptance where no evidence against the null hypothesis is treated as evidence for it and the fallacy of rejection where evidence against the null hypothesis is treated as evidence for the alternative hypothesis. Several problems relating to hypothesis testing are still being discussed to this day such as:

(1) how to narrow down an infinite set of all possible models that could have produced the data to one single model?

(2) how to test the adequacy of the chosen model *a posteriori*?

(3) how to address the fallacies of rejection and acceptance in practice?

2.1.1 *Model Specification*

Before Fisher, the notion of statistical modelling was to describe the properties of the distribution of the data in hand using for example, histograms and the first few sample

moments as was done by Karl Pearson. This practice is known as the application of descriptive statistics. Some statisticians would claim generality beyond the data in hand to make inferences, which is a crucial problem. Descriptive statistics can only be used to describe data succinctly or describe certain features of a distribution, however, when the results are extended to generalize the conclusions, "a new set of problems is faced" Mills (1924). Mills also discussed the assumptions necessary for the validity of statistical induction, first, there must exist a uniformity with respect to the characteristic measured from the data, and second, the population is thoroughly represented by the sample from which the characteristic was derived.

In the beginning of this chapter, Fisher's 1922 definition of the problem of specification, that is the problem of specifying the parametric model, was presented. However, his discussion on the topic in that particular paper was quite succinct where he confined the problem as "a matter for the practical statistician". Specification entails identifying a set of all models and then from that set of models, derive several plausible candidate models to represent good approximations to the information provided by the data in hand. In studies where the properties of the population of interest is known, this problem of model specification is not an issue. In other cases, Fisher suggests choosing a suitable model based on experience and testing the choice *a posteriori*. Lehmann (1990) states that "Fisher's statement implies that in his view there can be no theory of modeling, no general modeling strategies, but that instead each problem must be considered entirely on its own merits". Following this rather strong statement, Fisher suggested two modelling strategies, namely, confining to models of known form and to consider more or less elaborate forms according to the volume of data.

According to Lehmann (1990), Neyman was concerned about the practice as well as the theory of modelling and gave three comments regarding this. First, a complex phenomena, for example something in reality, is modelled by a combination of "simple building blocks" chosen through experience and some imagination. This practice makes this complex phenomenon appear familiar and simple. This comments bears some resemblance to Fisher's suggestion of using models of known form. Secondly, Neyman made a distinction between two types of models, "interpolatory formulae" and "explanatory models". Interpolatory formulae refer to a convenient and flexible family of distribution that can be used that best fits the data. The latter tries to explain the mechanism underlying the observed phenomena hence the term "mechanistic" or "theoretical" used by George Box. Thirdly, Neyman commented about developing a "genuine explanatory theory" that requires extensive knowledge of the background of the problem.

The problem of specification can be summarized as such. First, one has to acknowledge the existence of a reservoir of models which are well understood and whose prop-

erties are known or assumed. This is provided by probability and statistical theory including, but not limited to, univariate and multivariate distributions, stochastic processes, linear and generalized linear models. This list seems inexhaustible, however, the following statement from Box (1979) helps in shortening the list.

> *... there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".*

and from Box and N. R. Draper (1987)

> *Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful.*

To choose a model that is "useful" falls in the problem of specifying the class of models from which the selection is to be made, which is the second step in model specification. Henceforth, model specification will be partitioned into two components; formulation of a set of candidate models, known as model specification, and choosing a model to be used in making inference, known as model selection.

2.2 SOME ISSUES ABOUT MODEL ASSUMPTIONS AND MULTIPLE TESTING

In this section, a few approaches to checking model assumptions will be discussed as well as philosophical discussion about doing just that. Finally, an approach to correct multiple testing will also be presented.

2.2.1 *Graphical methods*

Informal graphical assessments such as certain scatterplots for independence, others for constant variance and normal quantile-quantile plots for the adequacy of the Gaussian model are usually recommended to verify the assumptions of a particular main test, for example, a Student's t-test or testing the validity of a regression model by way of the residuals.

Consider a general statistical modelling problem, specifically a linear regression model on a set of data points. The linear regression model has a set of assumptions that needs to be fulfilled in order to validate the inference made from the model. These assumptions include;

(1) the linearity between the independent and dependent variables where the expected value of the dependent variable is a straight-line function of each independent variable,

(2) the independence of the residuals where residuals are defined as the difference between an observed value of the dependent variable and the value of the dependent variable predicted from the regression line.

(3) the homoscedaticity of the residuals,

(4) the normality of the residuals distribution.

Graphical methods can be used to check model assumptions. For instance, a scatter plot of the independent variable against the dependent variable can be used to check assumption (1) where a linear trend should be observed. An error plot can be used to check assumptions (2) and (3) where a non-random pattern may suggest violations of these two assumptions. Finally, assumption (4) can be checked using the quantile-quantile plot of the residuals. Of course, there are other plots to check these assumptions, only a subset of those are mentioned here.

2.2.2 *Misspecification testing*

Even though Fisher and Neyman had their differences, it seems they agreed that checking the assumptions of a statistical model in order to ensure its adequacy is necessary. R. A. Fisher (1922) stated:

> *For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts. Once a statistic, suitable for applying such a test, has been chosen, the exact form of its distribution in random samples must be investigated, in order that we may evaluate the probability that a worse fit should be obtained from a random sample of a population of the type considered.*

Neyman (1952) outlined the construction of a mathematical model in which he emphasized testing the assumptions of the model by observation and if the assumptions are satisfied, then the model "*may be used for deductions concerning phenomena to be observed in the future*".

The idea of misspecification testing came about as early as the early 20^{th} century when Pearson introduced the Pearson's goodness of fit chi-square test. This is a misspecification test for the adequacy of a distributional assumption, however, the term test of

goodness of fit was used. The term misspecification (MS) was only seen as early as F. M. Fisher (1961) for explaining exogenous variables in economic models. This term was further expanded by Spanos (1999) where he explained a methodology of MS testing which would be informative in how to specify and validate statistical models, and how to proceed when certain statistical assumptions are violated. This thesis will use the terms goodness of fit tests and MS tests interchangeably.

Testing distribution assumptions: Spanos (1999) proceeds to mention a few misspecification tests that are available in the Fisher type tests set of models. The first and possibly most popular one is the Pearson's chi-square test. There are also tests based on the empirical cumulative distribution function such as the Kolmogorov's test and the Cramer-Von Mises statistic. These tests quantify the difference of the distances between the observed data and the hypothesized distribution. Another family of tests are those based on ordered samples. An example of tests of this kind include the Shapiro-Wilk test for testing normality. One can also carry out an MS test based on the moments using properties of the skewness and kurtosis coefficients, for instance the skewness-kurtosis test given by R. A. Fisher (1930).

Testing dependence assumptions: Some non-parametric tests can be used to test the dependence assumption including the runs test, Spearman's rho (ρ) test and Kendall's tau (τ) test. Another approach is moment based test for example the Box and Pierce (1970) and Ljung and Box (1978).

Testing homogeneity of variance assumptions: Some early non-parametric test for the homogeneity assumption was based on the signs of the differences, two such test are the Mann (1945) and Daniels (1950). Examples of parametric tests include the χ^2-test, the F-test and the Levene's test Lehmann (1960).

2.2.3 *Controversial views of model checking*

The necessity of model checking has been stressed by many statisticians for a long time, and this is what students of statistics are often taught. At first sight, model checking seems essential for two reasons. Firstly, statistical methods that a practitioner may want to use are often justified by theoretical results that require model assumptions, and secondly it is easy to construct examples for the breakdown of methods in case that model assumptions are violated in critical ways (e.g., inference based on the arithmetic mean, optimal under the assumption of normality, applied to data generated from a Cauchy distribution will not improve in performance for any number of observations compared with only having a single observation, because the distribution of the mean of $n > 1$ observations is still the same Cauchy distribution).

Regarding the foundations of statistics, checking of the model assumptions plays a crucial role in Mayo (2018) philosophy of "severe testing", in which frequentist significance tests are portrayed as major tools for subjecting scientific hypotheses to tests that they could be expected to fail in case they were wrong; and evidence in favor of such hypotheses can only be claimed in case that they survive such severe probing. Mayo acknowledges that significance tests can be misleading in case that the model assumptions are violated, but this does not undermine her philosophy in her view, because the model assumptions themselves can be tested.

A problem with preliminary model checking is that the theory of the model-based methods usually relies on the implicit assumption that there is no data-dependent preselection or pre-processing. A check of the model assumptions is a form of pre-selection. This is largely ignored but occasionally mentioned in the literature. Bancroft (1944) was probably the first to show how this can bias a model-based method after model checking. Chatfield (1995) gives a more comprehensive discussion of the issue. Hennig (2007) coined the term "goodness-of-fit paradox" (from now on called "misspecification paradox" here) to emphasize that in case that model assumptions hold, checking them in fact actively invalidates them. Assume that the original distribution of the data fulfills a certain model assumption. Given a probability $\alpha > 0$ that the MS test rejects the model assumption if it holds, the conditional probability for rejection under passing the MS test is obviously $0 < \alpha$, and therefore the conditional distribution must be different from the one originally assumed. It is this conditional distribution that eventually feeds the model-based method that a user wants to apply.

How big a problem is the misspecification paradox? Spanos (2010) argues that it is not a problem at all, because the MS test and the main test "*pose very different questions to data*". The MS test tests whether the data "*constitute a truly typical realization of the stochastic mechanism described by the model*". He argues that therefore model checking and the model-based testing can be considered separately; model checking is about making sure that the model is "*valid for the data*" (Spanos (2018)), and if it is, it is appropriate to go on with the model-based analysis.

A view opposite to Spanos' one, namely that model checking and inference given a parametric model should not be separated, but rather that the problems of finding an appropriate distributional "shape" and parameter values compatible with the data should be treated in a fully integrated fashion, can also be found in the literature for example Easterling (1976), D. Draper (1995) and Davies (2014)). Davies (2014) argues that there is no essential difference between fitting a distributional shape, an (in)dependence structure, and estimating a location (which is usually formalized as parameter of a parametric model, but could as well be defined as a nonparametric functional).

Bayesian statistics allows for an integrated treatment by putting prior probabilities on different candidate models, and averaging their contributions. Robust and nonparametric procedures may be seen as alternatives in case that model assumptions of model-based procedures are violated, but they have also been recommended for unconditional use by Hampel et al. (1986), making prior model checking supposedly superfluous. All these approaches still make assumptions; the Bayesian approach assumes that prior distribution and likelihood are correctly specified, robust and nonparametric methods still assume data to be i.i.d., or make other structural assumptions. So the checking of assumptions issue does not easily go away, unless it is claimed (as some subjectivist Bayesians do) that such assumptions are subjective assessments and cannot be checked against data.

Another potential objection to model assumption checking is that, again in the famous words of George Box, "*all models are wrong but some are useful*". It may be argued that model assumption checking is pointless, because we know anyway that model assumptions will be violated in reality in one way or another (e.g., it makes some sense to hold that in the real world no two events can ever be truly independent, and continuous distributions are obviously not "true" as models for data that are discrete because of the limited precision of all human measurement). This has been used as argument against any form of model-based frequentist inference, particularly by subjectivist Bayesians (e.g., de Finetti (1974) famous "*probability does not exist*"). Mayo (2018) however argues that "all models are wrong" on its own is a triviality that does not preclude a successful use of models, and that it is still important and meaningful to test whether models are adequately capturing the aspect of reality of interest in the inquiry, or whether the data are incompatible with the model in ways that will mislead the desired model-based inference (the latter is our own wording). We broadly agree with this position, although we note that the current practice of model checking is almost exclusively framed in terms of whether model assumptions are fulfilled (or "approximately" fulfilled, which implies that there is a true model that could be approximated) rather than whether data indicate that the specific use made of the model may be corrupted by specific violations of the model assumptions, which would seem more appropriate. A purely logical rebuttal of the view that frequentist methods of inference such as tests can only be valid if the model assumptions are fulfilled is as follows. The basis of that view is that the theoretical characteristics of the methods are derived assuming the model, but this does not imply that their characteristics are so bad as to render inferences invalid if the model does not hold.

We here investigate model assumptions that concern data generating mechanisms, and therefore they can be checked against the data. We keep an open mind regarding

whether preliminary model checking should be recommended "good practice" and even whether (frequentist) testing is advisable at all; we rather aim at "mapping" the debate than solving it.

2.2.4 *Bonferroni correction*

The Bonferroni correction is a multiple comparison correction used when several dependent or independent statistical tests are being performed simultaneously. A given significance value or Type-I error, α, may be appropriate for each individual test, however when comparing a set of statistical tests, it can lead to inferences that have a smaller power, especially if the some of the tests are correlated or dependent. In order to avoid spurious positives, the alpha value needs to be lowered to account for the number of tests being performed (Bonferroni (1935)).

Let tests $T_1, ..., T_n$ be a set of n test statistics with corresponding p-values $p_1, ..., p_n$ for testing hypotheses $H_1, ..., H_n$. In the approach of the classical Bonferroni correction for multiple tests, the alpha value for each of the n tests is equal to α/n. If any p-value is less than α/n $[p_i \leq \alpha/n,\ (i = 1, ..., n)]$, then the corresponding H_i is rejected. The Bonferroni (1936) inequality,

$$P\left\{ \bigcup_{i=1}^{n} (p_i \leq \alpha/n) \right\} \leq \alpha\ ,\ (0 \leq \alpha \leq 1)$$

ensures that the probability of rejecting at least one hypothesis when all the null hypotheses are true is at most α.

2.3 TESTING STATISTICAL HYPOTHESIS WITH MISSPECIFICATION TESTS

Often, in a introductory statistical hypothesis testing course, students are told they can use a statistical test after checking the assumptions of said test is not violated. Vardeman and Morris (2003) advised young statisticians to acknowledge a critical fact that "*statistical analysis of data can only be performed within the context of selected assumptions, models, and/or prior distributions*". The authors also advised young statisticians must fully understand what the assumptions imply and should not claim the "usual assumptions" hold because of the chosen technique's robustness to violations of the model assumption. One way to check the violation of an assumption is to use certain graphical methods as it can give an indication of the conformity to an assumption. At the same time, the student is taught some statistical tests which can seem as a way to formalise the information the

student is looking for to make a decision when using graphical methods. Hence the idea to use an MS test to check the assumptions. For example, using the quantile-quantile plot one can tell the departure of the quantiles of a sample from the normal distribution quantiles. This can provide an assessment of how "good" the data fits to the normal distribution that is graphical, rather than reducing to a numerical summary. However, this assessment requires more skill to interpret. This leads to the use of statistical tests to quantify this assessment. Some examples are given in Section 2.3.1.

Suppose a hypothesis of substantial interest is to be tested by a certain test we shall refer to as a "main test". To see whether the use of a restricted and possibly incorrect model or a more general and less precise model is needed, an MS test can be performed on the adequacy of the restricted model. The MS test has the null hypothesis that a certain model assumption holds. If this MS test fails to reject, then the restricted model or "model-based constrained (MC) test" is chosen. Otherwise, an alternative main test or "alternative unconstrained (AU) test" is used which can be more general which does not rely on the rejected model assumption.

The idea of the above procedure is indeed appealing; if the restricted model is incorrect, the assumptions of the MC test is not valid and therefore MC test should not be used. However, always using the general and typically less powerful test would be a waste of power if the restricted model can be used instead. The MC test might also be preferred because it is simpler to use or explain. As it is not known beforehand whether the restricted model is applicable, it seems very natural to settle this simply through an MS test. However, like any other test, the MS test, which should help to decide which of the two main tests is most appropriate to test the main hypothesis, will make errors of first and second kind. This may lead to application of the MC test when in fact the restricted model is inadequate, or to application of the AU test when it is not necessary.

A "combined procedure (CP)" consists of the complete decision rule involving MS test, MC test and AU test (if specified). We generally assume that the MS test is carried out on the same data as the main test. Some of the issues discussed below can be avoided by checking the model on independent data, however such data may not be available, or this approach may not be preferred for reasons of potential waste of information and lack of power (in case the "independent" data are obtained by splitting the available dataset, see Chatfield (1995) for a discussion of this). In any case it would leave open the question whether the data used for MS testing are really independent of the data used for the main test, and whether they do really follow the same model.

The general setup we are interested in here is as follows. Given is a statistical model defined by some model assumptions Θ,

$$M_\Theta = \{P_\theta, \theta \in \Theta\} \subset M,$$

where $P_\theta, \theta \in \Theta$ are distributions over a space of interest, indexed by a parameter θ. M_Θ is written here as a parametric model, but we are not restrictive about the nature of Θ. M_Θ may even be the set of all i.i.d. models for n observations, in which case Θ would be very large. However, in the literature, M_Θ is usually a standard parametric model with $\Theta \subseteq \mathbb{R}^m$ for some m. There is a bigger model M containing distributions that do not require one or more assumptions made in M_Θ, but for data from the same space.

Given some data z, we want to test a parametric null hypothesis $\theta \in \Theta_0$, which has some suitably chosen "extension" $M^* \subset M$ so that $M^* \cap M_\Theta = M_{\Theta_0}$, against the alternative $\theta \notin \Theta_0$ corresponding to $M \setminus M^*$ in the bigger model.

In the simplest case, there are three tests involved, namely the MS test Φ_{MS}, the MC test Φ_{MC} and the AU test Φ_{AU}. Let α_{MS} be the level of Φ_{MS}, i.e., $Q(\Phi_{MS}(z) = 1) \leq \alpha_{MS}$ for all $Q \in M_\Theta$. Let α be the level of the two main tests, i.e., $P_\theta(\Phi_{MC}(z) = 1) \leq \alpha$ for all $P_\theta, \theta \in \Theta_0$ and $Q(\Phi_{AU}(z) = 1) \leq \alpha$ for all $Q \in M^*$. To keep things general, for now we do not assume that type I error probabilities are uniformly equal to α_{MS}, α, respectively, and neither do we assume tests to be unbiased (which may not be realistic considering a big nonparametric M).

The combined procedure is defined as

$$\Phi_{CP}(z) = \begin{cases} \Phi_{MC}(z) & : \quad \Phi_{MS}(z) = 0, \\ \Phi_{AU}(z) & : \quad \Phi_{MS}(z) = 1. \end{cases}$$

This allows the analysis of the characteristics of Φ_{CP}, particularly its effective level (which is not guaranteed to be $\leq \alpha$) and power under P_θ with $\theta \in \Theta_0$ or not, or under distributions from M^* or $M \setminus M^*$. General results are often hard to obtain without making restrictive assumptions, although some exist which will be discussed in Chapter 3. At the very least, simulations are possible picking specific P_θ or $Q \in M$, and in many cases results may generalize to some extent because of invariance properties of model and test.

Also of potential interest are $P_\theta\left(\Phi_{CP}(z) = 1 | \Phi_{MS}(z) = 0\right)$, i.e., the type I error probability (size) under M_{Θ_0} or the power under M_Θ in case the model was in fact passed by the MS test, $Q\left(\Phi_{CP}(z) = 1 | \Phi_{MS}(z) = 0\right)$ for $Q \in M \setminus M_\Theta$, i.e., the situation that the model M_Θ is in fact violated but was passed by the MS test, and whether Φ_{CP} can com-

pete with Φ_{AU} in case that $\Phi_{MS}(z) = 1$ (M_Θ rejected). These are investigated in some of the literature, see below.

For example, many researchers have found that the use of an MS test influences the size of the main test, meaning that $P_\theta\left(\Phi_{CP}(z) = 1 | \Phi_{MS}(z) = 0\right)$ can be substantially different from $P_\theta\left(\Phi_{MC}(z) = 1\right)$.

In Chapter 4 we look at the performance of Φ_{CP} in case there is a "hyperprobability" of having data generated from either $P_\theta \in M_\Theta$ or $Q \in M \setminus M_\Theta$; such a situation in which both satisfied and violated model assumptions can occur and Φ_{MS} has some distinction work to do has to our knowledge not yet been analyzed in the literature, which therefore may give a too pessimistic picture of the performance of the combined procedure.

Many other researchers have found that the use of an MS test to test some model assumption influences the size of the main test, for example Bancroft (1964), Arnold (1970), C. V. Rao and Saxena (1981), Saleh and Sen (1983), Moser, Stevens and Watts (1989), Moser and Stevens (1992), Gupta and Srivastava (1993), Albers, Boon and Kallenberg (1998), Albers, Boon and Kallenberg (2000a) and Albers, Boon and Kallenberg (2000b).

2.3.1 *The problem of whether to pool variances*

Historically the first problem for which preliminary MS testing and combined procedures were investigated was whether to assume equal variances for comparing the means of two samples. Until now this is the problem for which most work investigating combined procedures exists. The Behrens-Fisher problem, named after statisticians Walter Behrens and Ronald Fisher, generally presents a problem of assessing the equality of location parameters of samples that come from two populations of the same location-scale family of distributions where the scale parameter is unknown and not necessarily equal. It has been demonstrated that the two sample t-test is robust against violations of equality of variances when sample sizes are equal as shown by P. L. Hsu (1938), Scheffé (1970), Posten, Yeh and Owen (1982) and Zimmerman (2006). When both variances and sample size are unequal, the probability of the Type-I error exceeds the nominal significance level if the larger variance is associated with the smaller sample size and vice versa, see Zimmerman (2006), Wiedermann and Alexandrowicz (2007), and Moder (2010). In this case, Welch's t-test presented in Welch (1938), Satterthwaite (1946) and Welch (1947) is recommended as a adequate alternative, see also Rasch, Kubinger and Moder (2011). Scheffé (1970) discussed some other practical solutions to the Behrens-Fisher problem like the d-solutions, Welch-Aspin solution, Behrens-Fisher solution and Student solution.

Bancroft (1944) investigated the bias when estimating the variance for an analysis of variance test. By using an F test as an MS test to test the homogeneity of the variance, he decides to use either a pooled estimate $(n_1 s_1^2 + n_2 s_2^2)/(n_1 + n_2)$ as an estimate of σ_1^2 or s_1^2 as an estimate of σ_1^2 depending on the decision made by the F test. Bancroft concluded that the lowest bias can be had by always pooling the estimate and not use the MS test to check the model assumption.

Starting from Bancroft's work, from the end of the 1940s, a good amount of research was done on the problem of pooling variances, much of which concerned the estimation of means and the corresponding mean squared errors, but some work also dealt with combined testing procedures. Bancroft and Han (1977) published a comprehensive bibliography, also including other problems of preliminary assumption testing. One reason for the popularity of the variance pooling problem in early work is that, as long as normality is assumed, only the ratio of the variances needs to be varied to cover the case of violated model assumptions, which makes it easier to achieve theoretical results without computer-intensive simulations.

C. A. Markowski and E. P. Markowski (1990) evaluated the setup of having a MS test of homogeneity, the F-test, before doing a t-test for various combinations of sample size and significance level. The samples were drawn from normal distributions, a contaminated normal distribution with a higher frequency of outliers, the exponential distribution and the chi-squared distribution. For data with non-normal distributions, the results supports those of Box (1979) where Box strongly discourages the use of the F-test as an MS test. For situation with data generated from the normal distribution, the MS test was either unnecessary or ineffective as an MS test to alert the researcher that a t-test may be inappropriate. For equal sample sizes, no MS test is needed as the t-test is robust enough. However, for unequal sample sizes, the t-test is not so robust and the authors note that a more effective MS test would be desirable.

Zimmerman (2004) investigated the rejection rates of a two-stage procedure consisting of an MS test, specifically the Levene test on samples of different sizes with equal and unequal variances followed by either a pooled-variance Student t-test or a separate-variance Welch t-test. Two samples from the normal distribution were generated and put through the unconditional Student t-test, Welch t-test and the two-stage procedure. The simulation results strengthen his views against MS testing for equality of variances. The final recommendation is to use the Welch t-test unconditionally, especially when the sample sizes are unequal.

Also, there are considerable evidence that the separate-variance Welch t-test is superior to a pooled-variance t-test when variances are unequal as discussed by Cohen (1974), Zimmerman and Zumbo (1993), Overall, Atlas and Gibson (1995) and Ruxton (2006).

Example of an MS test of variance homogeneity: The Levene's test In the case where two groups are involved, the Levene's test tests the hypothesis that both groups have equal variance, namely

$$H_0 : \sigma_1 = \sigma_2 \qquad \text{against} \qquad H_1 : \sigma_1 \neq \sigma_2.$$

The version of Levene's test considered to be the best in Brown and Forsythe (1974) and Conover, M. E. Johnson and M. M. Johnson (1981) is the one-way analysis of variance F-test based on $z_{ij} = |y_{ij} - \tilde{y}_i|$, where $\tilde{y}_i$ is the median of $\{y_{ij} : i = 1, ..., k, j = 1, ..., n_i\}$. The test statistic is

$$F = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} n_i (\bar{z}_{i\cdot} - \bar{z}_{\cdot\cdot})^2}{\sum_{i=1}^{k} \sum_{j=1}^{n_i} (z_{ij} - \bar{z}_{i\cdot})^2}$$

where

$$\bar{z}_{i\cdot} = \sum_{j=1}^{n_i} \frac{z_{ij}}{n_i} \quad \text{and} \quad \bar{z}_{\cdot\cdot} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{z_{ij}}{N}.$$

This test statistic is not exactly distributed as the usual F-distribution with $k-1$ and $N-k$ degrees of freedom. However Lehmann (1960) showed by simulation that the usual F statistic provides a good approximation (Gastwirth, Gel and Miao (2009)).

2.3.2 *Tests of normality in one-sample and two-sample problem*

In any introductory statistical methods text, one of the most basic topics is statistical hypothesis testing, especially concerning the inference of a population mean, μ. The primary tool to test an hypothesis about a population mean is the Student t-test. Ever since the work of Gosset ("Student") (1908) and R. A. Fisher (1925) on statistical inference about differences in means, specifically the Student's t-test, a good deal of research focused on the properties of the t statistic. Some assumptions were needed to be made in order for the two sample Student's t-test to perform optimally for the comparison of means from independent samples. The assumptions are of normality, homoscedasticity and independence of the observations made. When these assumptions are met, the two sample Student's t-test was shown to perform optimally for the comparison of means of two samples as shown in Hodges and Lehmann (1956) and Randles and Wolfe (1979). However, in empirical data, violations of one or more assumptions always exists, and the robustness properties of significance tests are of great interest.

Early theoretical findings suggest that the two sample t-test is fairly robust against violations of the normality assumption shown by Bartlett (1935). Bartlett (1935) concluded that the theoretical results are "incomplete and not perhaps of much quantitative value", however the t-test may still be used for moderate departures from normality particularly when the two samples have equal number of observations. Keselman, Othman and Wil-

cox (2013) discussed different types of MS testing for normality and presented some literature about how in the not too distant past, it was claimed that violations of normality would not jeopardise scientific findings. In the situation where the F-test is used in an ANOVA procedure to test the similarity of means across groups, T. C. Hsu and Feldt (1969) claim that some moderate skewness or kurtosis has little effect on the Type-I error of the F-test and Lunney (1970) even investigated how dichotomous variables affect the ANOVA test that requires the normality assumption. However, the opinion seems to have shifted to the opposite as shown in numerous simulation studies, for example, Boneau (1960), Neave and Granger (1968), Posten (1978), Posten (1984) and Rasch and Guiard (2004). Although the two sample t-test is able to protect the nominal significance level α under non-normality, considerable evidence exists that the non-parametric Wilcoxon-Mann-Whitney U-test is robust and even more powerful compared to the t-test under non-normal distributions as discussed by Hodges and Lehmann (1956), Neave and Granger (1968), Randles and Wolfe (1979) and Sawilowsky and Blair (1992).

Most software packages provide optional test results for the Gaussian (normality) assumption and homogeneity of variance. The Gaussian distribution is the most well-known and widely used distribution in many fields such as engineering, statistics and physics. One of the major reasons why the Gaussian distribution has become so prominent is because of the Central Limit Theorem. Especially when there is no information about the distribution of observations, the Gaussian assumption appears as the most reasonable choice (S. Park, Serpedin and Qaraqe (2013)). This is true when the sample size is sufficiently large.

As with informal assessments, the interpretation of the results is very much subjective. Therefore, some formal tests are sometimes used in place of informal assessments. Examples for tests of normality include, but not limited to, the Kolmogorov-Smirnov test, the Shapiro-Wilk test, the Pearson chi-square test, the Lilliefors test, the Jarque-Bera test and the Anderson-Darling test. Razali and Wah (2011) concluded that Shapiro-Wilk has the best power for a given level, when comparing the Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefor and Anderson-Darling tests. This result concurs with Mendes and Pala (1990), Farrell and Rogers-Stewart (2006) and Keskin (2006). The Anderson-Darling test was a close second.

However, normality can never be proven. Observations obtained in any experiment are limited in their precision. Most measurements taken are truncated numbers. It is a fact that the normal distribution is symmetric about its mean and is non-zero over the *entire* real line. Hence, no values measured or obtained from an experiment is ever truly normal.

Take for example a one-sample Student's t-test that is used when an inference about the population is made for a sample of n independent observations $X_1, X_2, ..., X_n$ from a distribution $\mathcal{F}$. The t-test assumes that the underlying population distribution is normal. This assumption is often checked by an MS test. The two-stage procedure is as follows: if the MS test is not significant, the one sample t-test is applied; if the MS test rejects the null hypothesis of normality, a non-parametric test is done. The idea behind this procedure is clear: in all practical situations, the researcher does not know in advance whether or not the *assumption* of normality is satisfied; this issue is usually decided on after inspection of the sample data. Consequently, preliminary testing for normality before proceeding with the final statistical test is very common for example, as discussed in Wilcox (1998) and Schoder, Himmelmann and Wilhelm (2006).

Easterling and Anderson (1978) provides objective evidence in support of MS testing for goodness of fit. They did this because they believed that the practice of this two-stage procedure is not only conventional, but also good. To do this they considered various distributions such as normal, uniform, exponential, two central and two non-central t distributions. They only considered sample sizes 10 and 20. They drew a random sample from the distribution that is being studied and tested for normality at 10% significance. The samples that were not rejected by the MS test is categorised under "normality significant at 10%" and the samples that were rejected were categorised under "normality not significant at 10%" until 1000 samples are obtained in each category. They used both the Anderson-Darling and the Shapiro-Wilk MS tests. After obtaining those samples, the empirical distribution of the 1000 t values were compared to the expected frequencies from the Student's t distribution. As expected, there were no issues when the samples were drawn from the normal distribution. However, for symmetrical non-normal distributions, the results were mixed and for situations where the distributions were asymmetric, the results were not in the favour of the model checking before the main test because the distribution of the t values do not resemble a Student's t distribution. They offer the following as possible reasons this is so:

> *There are various reasons why the distributions of the t ratio in the cases considered might not follow a Student's t distribution —the nonnormality of the numerator, the nonzero expectation of the numerator, the nonchi-squareness of the square of the denominator, and lack of independence of numerator and denominator. For the asymmetric sampling distributions, the empirical distributions of t (not shown in this paper) suggest that the preliminary goodness of fit test causes a shift in mean. In order to obtain a sample from such a distribution which would pass a test for normality (which includes symmetry as a property) that sample would have to have fewer observations in the elongated tail than are expected.*

To investigate this, they adjusted the numerator by replacing the true mean with the mean of the 1000 sample mean to adjust for a shift in mean. This did not improve the results. They then continued to analyse the rejection rates of the empirical t values. Again, the same results hold, the distribution of the t values does not resemble the empirical t values based on the chi-square statistic. A discussion followed that MS testing for normality is not the proper thing to do when estimating the normal theory interval or difference of the means. A non-parametric estimation approach was proposed. If a probability model is to be used as a reporting device to discover and describe patterns of variability, then MS testing is recommended. Therefore, it was also proposed that goodness of fit and estimation be done simultaneously by finding parameter regions for which an MS test statistic for a completely specified distribution is smaller than some percentage point on the null distribution of that statistic (Easterling (1976)).

Schucany and Ng (2006) investigated the Type I error rate of the one sample t-test given that the sample has passed the MS test, the Shapiro-Wilk test for normality, named the conditional Type I error rate. Data were sampled from normal, uniform, exponential and Cauchy populations. The simulation study showed that, for the uniform distribution, screening of samples by an MS test for normality leads to a more conservative conditional Type I error rate than application of the one-sample t-test without MS testing. In contrast, for the exponential distribution, the conditional Type I error rate is even more elevated than the Type I error rate of the t-test without MS testing (i.e. the unconditional Type I error rate) which is already above the nominal level. Furthermore, larger sample sizes and more liberal significance levels of the MS test shift the conditional Type I error rate even further away from the unconditional Type I error rate of the t-test and also from the nominal level, leading to either more conservative of more liberal test decisions. This common feature of the uniform and exponential distributions is especially interesting to note as, in both cases, the t-test without MS testing show an acceptable Type I error rate at least for sample sizes of $n = 50$.

Rochon and Kieser (2011) investigated the reasons behind the characteristics of the one sample t-test with MS testing for normality. Samples were drawn from the exponential, lognormal, uniform, Student's t with 2 degrees of freedom and standard normal distributions that had passed the pretest. The Shapiro-Wilk test and the Lilliefors modification of the Kolmogorov Smirnov test was used. However, it was found that the results from the two MS tests were similar, therefore only results from the Shapiro-Wilk test were presented. For the exponential and lognormal distributions, the Type I error rate is elevated for samples tested without model checking and it is further increased by the MS test for normality. The inspection of the densities of the samples that pass the MS test shows that the closer the underlying population distribution is to the normal,

the less important an MS test is. Consequently, the further away the population distribution is from normal, the MS screening in fact selects samples that look like normal and thus can no longer be considered representative of the true underlying population. They concluded that formal MS testing for normality cannot be recommended. Alternatives such as the unconditional *t*-test relying on the normal approximation of reasonably large sample sizes by way of the Central Limit Theorem taking into account that the one sample *t*-test is more sensitive to skewness than to heaviness or lightness of the tails (P. V. Rao (1998)). If it is at least suspected or assumed that the underlying population distribution is symmetric, a non-parametric application such as the Wilcoxon-Mann-Whitney signed-rank test could be considered. In any case, it is recommended that checking the model assumption must be derived from external data sources and not from the data set at hand.

Rochon, Gondan and Kieser (2012) more recently examined the reasons behind the characteristics of the one-sample *t*-test with MS testing for normality. Data were sampled from the exponential, uniform and normal distributions. Two strategies were used. The first strategy is the usual way of MS testing where both a two sample *t*-test is conducted if both samples had passed the Shapiro-Wilk test for normality. Otherwise, the Mann-Whitney $\mathcal{U}$ test is performed. In the second strategy, the MS test is done once on the collapsed set of residuals from both samples. The conditional and unconditional Type I error were calculated. They concluded from a formal perspective, that MS testing for normality is incorrect and therefore should be avoided. They recommended that the assumption of normality must come from extra-data sources such as results of earlier trials or pilot studies (Lewis (1999)). From a practical perspective however, MS testing does not seem to cause much harm in the cases they have considered. The worse that can be said about MS testing is that it is unnecessary.

Keselman, Othman and Wilcox (2013) discussed different types of MS testing for normality. They used the Kolmogorov-Smirnov, Cramer-von Mises and Anderson-Darling MS tests, 26 different shape distribution (14 distributions with different skewness and kurtosis values, 8 contaminated normal mixture models and 4 multinomial models), 3 sample sizes and 4 different level of significance. They concluded that the Anderson-Darling test is the most effective at detecting non-normality and they suggested that the MS test be carried out at significance level larger than 0.05, for example 0.15 or 0.20 to increase power. According to Razali and Wah (2011), where the authors made power comparisons of the Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, they concluded that the Shapiro-Wilk test has the largest power to reject normality which is slightly better compared to the Anderson-Darling test.

Example of an MS test of normality: The Shapiro-Wilk test The Shapiro-Wilk test tests the null hypothesis that a given sample comes from a population that is normally distributed against the alternative that the sample does not come from a normally distributed population. The original test was proposed in Shapiro and Wilk (1965) and was limited for sample sizes between 3 and 50. They claimed that this test is sensitive to outliers and Althouse, Ware and Ferron (1998) claims that the Shapiro-Wilk test was the first test for normality that was able to detect departures due to either skewness or kurtosis, or both. Then, J. P. Royston (1982) extended the range of sample sizes up to 2000 and later extended the sample size restriction to 5000 in P. Royston (1992). P. Royston (1995) provided an algorithm called AS R94 to provide approximate p-values for $3 \leq n \leq 5000$ where the calculation of the p-value is exact for $n = 3$ and approximations are used separately for $4 \leq n \leq 11$ and $n \geq 12$. The Shapiro-Wilk test statistic is given below

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_i\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2}$$

where $y_1 < y_2 < ... < y_n$ is an ordered sample of size n, $\bar{y}$ is the sample mean and $\mathbf{a} = (a_1, a_2, ..., a_N)^T$ is such that $(n-1)^{-\frac{1}{2}} \sum a_i y_i$ is the best linear unbiased estimate of the standard deviation of y_i, assuming the null hypothesis of normality.

2.3.3 *More than one misspecification test*

Rasch, Kubinger and Moder (2011) assessed the statistical properties of a three-stage procedure including testing for normality and for homogeneity of the variances. They considered 5 distributions with different location, spread, skewness and kurtosis parameters. Various sample sizes, equal and unequal, and different ratios of the standard deviation were considered. They considered three main statistical tests, the Student's t-test, the Welch's t-test and the Wilcoxon-Mann-Whitney U-test. For the MS testing, they used the Kolmogorov-Smirnov test for testing normality and the Levene's test for testing the homogeneity of the variances of the two samples that were generated. If normality was rejected by the Kolmogorov-Smirnov test, the Wilcoxon-Mann-Whitney U-test was used. If normality was not rejected, the Levene's test was run and if homogeneity was rejected, the Welch's t-test was used and if homogeneity was not rejected, the standard t-test was used. The authors presented the rejection rates and the power of the procedure and compared it with the tests when the model assumption were not checked. The authors concluded that assumptions underlying the two sample t-test should not be

pre-tested because "*pre-testing leads to unknown final Type I and Type II risks if the respective statistical tests are performed using the same set of observations*".

To our knowledge this is the only investigation of a combined procedure involving more than one MS test. Spanos (2018) proposed a "probabilistic reduction" approach in order to systematise the process of model building involving MS testing of various assumptions, but he did not define a fully automatised procedure that could be investigated by means of theory or simulation.

2.3.4 *Regression*

In standard linear regression,

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + e_i,\ i = 1, \ldots, n,$$

with response $Y = (y_1, \ldots, y_n)$ and explanatory variables $X_j = (x_{j1}, \ldots, x_{jn})$, $j = 1, \ldots, p$. $e_1, \ldots, e_n$ are in the simplest case assumed i.i.d. normally distributed with mean 0 and equal variances.

The regression model selection problem is the problem to select a subset of a given set of explanatory variables $\{X_1, \ldots, X_p\}$. This can be framed as a model misspecification test problem, because a standard regression assumes that all variables that systematically influence the response variable are in the model. If it is of interest, as main test problem, to test $\beta_j = 0$ for a specific j, the MS test would be a test of null hypotheses $\beta_k = 0$ for one or more of the explanatory variables with $k \neq j$. The MC test would test $\beta_j = 0$ in a model with X_k removed, and the AU test would test $\beta_j = 0$ in a model including X_k. This problem was mentioned as a second example in Bancroft (1944) seminal paper on preliminary assumption testing.

Traditional model selection approaches such as forward selection and backward elimination are often based on such tests and have been analyzed (and criticized) in the literature. We will not review this literature here. There is sophisticated and innovative literature on post-selection inference in this problem. Berk et al. (2013) propose a procedure in which main inference is adjusted for simultaneous testing taking into account all possible submodels that could have been selected. Efron (2014) uses bootstrap methods to do inference that takes the model selection process into account. Both approaches could also involve other MS testing such as of normality, homoscedasticity, or linearity assumptions, as long as combined procedures are fully specified. For specific model selection methods there now exists work allowing for exact post-selection inference, see Lee et al. (2016). For a critical perspective on these issues see Leeb and

Pötscher (2005) and Leeb and Pötscher (2015). In econometrics, David Hendry and co-workers developed an automatic modeling system that involves MS testing and conditional subsequent testing with adjustments for decisions in the modeling process, see for example, Hendry and Doornik (2014). Earlier, some authors such as Saleh and Sen (1983) analyzed the effect of preliminary testing on later conditional main testing.

Godfrey (1988) listed a plethora of MS tests to test the various assumptions of linear regression. However, no systemic way to apply these tests was discussed. In fact, Godfrey noted that the literature left more questions open rather than answered. Some of these questions are: (i) the choice among different MS tests, (ii) whether to use nonparametric or parametric tests, (iii) what to do when any of the model assumptions are invalid as well as (iv) some potential problems with MS testing such as repeated use of data, multiple testing and pre-test bias. Godfrey (1996) discussed destructive and constructive value of MS tests. He concluded that efforts should be made to develop 'attractive', useful and simple combined procedures, keeping in mind that the combination of tests must be well-behaved. One suggestion was to use the Bonferroni correction for each test as "*the asymptotic dependence of test statistics is likely to be the rule, rather than the exception, and this will reduce the constructive value of individual checks for misspecification*".

D. E. A. Giles and J. A. Giles (1993) reviewed the substantial amount of work done in econometrics regarding preliminary testing in regression up to that time, a limited amount of which is about MC and/or AU tests conditionally on MS tests. This involves pre-testing of a known fixed variance value, homoscedasticity, and independence against an autocorrelation alternatives. The cited results are mixed. King and D. E. A. Giles (1984) comment positively on a combined procedure in which absence of autocorrelation is tested first by a Durbin-Watson or t-test. Conditionally on the result of that MS test, either a standard t-test of a regression parameter is run (MC test) or a test based on an empirically generalized least squares estimator taking autocorrelation into account (AU test). In simulations the combined procedure performs similar to the MC test and better than the AU test in absence of autocorrelation, and similar to the AU test and better than the MC test in presence of autocorrelation. Also here it is recommended to run the MS test at a level higher than the usual 5%. Most related post-1993 work in econometrics seems to be on estimation after pre-testing, and regression model selection. Ohtani and Toyoda (1985) propose a combined procedure for testing linear hypotheses in regression conditionally on testing for known variance. Toyoda and Ohtani (1986) test the equality or different regressions conditionally on testing for equal variances. In both papers power gains for the combined procedure are reported, which are sometimes but not always accompanied with an increased type I error probability.

2.3.5 *MS testing in applied research*

Kim (2015) and Kim and J. H. Park (2019) suggests using the Shapiro-Wilk (SW) or Kolmogorov-Smirnov test to test the normality assumption of a t-test, if normality assumption is rejected, the WMW test should be used. Kim (2015) then suggests to use the Levene's or Barlett's test to check the equal variance assumption. If the assumption of equal variance is rejected, the Welch's t-test is performed.

Shan, Hwang and Wong (2017) took an empirical survey of 30 Singapore based construction companies. After collecting the data, they carried out a SW test to test for normality. The Kruskal-Wallis and WMW tests were used as SW rejected the normality assumption for the variables of interest. Hwang et al. (2017) conducted a survey of residents of houses that have been green retrofitted. The SW test was used to test normality and since all the variables of interest had the normality assumption rejected, the Kruskal-Wallis test was used for further analysis.

Kokosińska et al. (2018) studied heart rate variability in six groups of patients where one of the groups are healthy patients. The study used a SW test to test the normality of each group of patients. If normality is rejected, they used a WMW test to test the mean difference in heart rate variability between groups. If normality is not rejected, they used a standard t-test.

W. Wu et al. (2019) aimed to compare design review and assessment of a tiny house between novice students and professional experts using virtual reality and mixed reality technologies. The data collected was in count form which are numeric and discrete. A SW test was performed to determine the normality of the samples in order to select "what t-test should be used". The SW test did not reject normality and the standard t-test was used.

These studies show that the combined procedure is still used implicitly without being defined formally in a wide range of research areas.

3

SOME THEORETICAL RESULTS

"That's all well and good in practice, but how does it work in theory?" Shmuel Weinberger

3.1 THE SETUP

Let P_θ be a distribution that fulfills the model assumptions of the MC test, and $Q \in M \setminus M_\Theta$ a distribution that violates these assumptions. For considerations of power, let the null hypothesis of the main test be violated, i.e., $\theta \notin \Theta_0$ and $Q \notin M^*$ (an analogous setup is possible for considerations of size). The general setup we are interested in here is as follows. Given is a statistical model defined by some model assumptions Θ,

$$M_\Theta = \{P_\theta, \theta \in \Theta\} \subset M,$$

where $P_\theta, \theta \in \Theta$ are distributions over a space of interest, indexed by a parameter θ. M_Θ is written here as a parametric model, but we are not restrictive about the nature of Θ. M_Θ may even be the set of all i.i.d. models for n observations, in which case Θ would be very large. However, in the literature, M_Θ is usually a standard parametric model with $\Theta \subseteq \mathbb{R}^m$ for some m. There is a bigger model M containing distributions that do not require one or more assumptions made in M_Θ, but for data from the same space.

Given some data z, we want to test a parametric null hypothesis $\theta \in \Theta_0$, which has some suitably chosen "extension" $M^* \subset M$ so that $M^* \cap M_\Theta = M_{\Theta_0}$, against the alternative $\theta \notin \Theta_0$ corresponding to $M \setminus M^*$ in the bigger model.

In the simplest case, there are three tests involved, namely the MS test Φ_{MS}, the MC test Φ_{MC} and the AU test Φ_{AU}. Let α_{MS} be the level of Φ_{MS}, i.e., $P_\theta(\Phi_{MS}(z) = 1) \leq \alpha_{MS}$ for all $P \in M_\Theta$. Let α be the level of the two main tests, i.e., $P_\theta(\Phi_{MC}(z) = 1) \leq \alpha$ for all $P_\theta, \theta \in \Theta_0$ and $Q(\Phi_{AU}(z) = 1) \leq \alpha$ for all $Q \in M^*$. To keep things general, for now we do not assume that type I error probabilities are uniformly equal to α_{MS}, α, respectively,

and neither do we assume tests to be unbiased (which may not be realistic considering a big nonparametric M).

The combined test is defined as

$$\Phi_{CP}(z) = \begin{cases} \Phi_{MC}(z) & : \quad \Phi_{MS}(z) = 0, \\ \Phi_{AU}(z) & : \quad \Phi_{MS}(z) = 1. \end{cases}$$

Assume that a dataset is with probability $\lambda \in [0,1]$ generated from P_θ and with probability $(1-\lambda)$ from Q (we stress that as opposed to standard mixture models, λ governs the distribution of the whole dataset, not every single observation independently). The cases $\lambda = 0$ and $\lambda = 1$ are those that have been treated in the literature, but only if $\lambda \in (0,1)$ the ability of the MS test to inform the researcher whether the data are more likely from P_θ or from Q is actually required. This setup challenges the MS test to distinguish between these two situations, which is different than the setup treated in the literature.

Several simulations were ran in Chapter 4. Looking at the nominal levels in the simulation, it is hard to choose a "winner" between MC, AU or CP as the levels were very often, though not always, respected under H_0 (namely for both P_θ, $\theta \in \Theta_0$ and $Q \in M^*$). However, in the power simulations where H_0 was violated, a pattern emerged. For $\lambda = 0$ (model assumption violated), the AU test was best and the MC test was worst. For $\lambda = 1$, the MC test was best and the AU test was worst. The CP was in between which was mostly the case in our simulations. The consequence of this is that the CP performs clearly better than both MC and AU over the the best part of the range of λ.

For all the tests mentioned hereafter, we are in the situation where the null hypotheses are violated. The events of rejection of the respective H_0 are denoted $R_{MS} = \{\Phi_{MS}(z) = 1\}$, $R_{MC} = \{\Phi_{MC}(z) = 1\}$, $R_{AU} = \{\Phi_{AU}(z) = 1\}$, $R_{CP} = \{\Phi_{CP}(z) = 1\}$. In the case that we are in P_θ and Φ_{MC} is used, the probability of rejecting the null hypothesis is $P_\theta(R_{MC})$. In the case that we are in P_θ and Φ_{AU} is used, the probability of rejecting the null hypothesis is $P_\theta(R_{AU})$. In the case that we are in Q and Φ_{MC} is used, the probability of rejecting the null hypothesis is $Q(R_{MC})$. In the case that we are in Q and Φ_{AU} is used, the probability of rejecting the null hypothesis is $Q(R_{AU})$.

We define the differences between the powers of the tests in either P_θ or Q as below;

(a) $\Delta_{MC} = P_\theta(R_{MC}) - Q(R_{MC})$

(b) $\Delta_{AU} = P_\theta(R_{AU}) - Q(R_{AU})$

(c) $\Delta_{P_\theta} = P_\theta(R_{MC}) - P_\theta(R_{AU})$

(d) $\Delta_Q = Q(R_{AU}) - Q(R_{MC})$.

We assume that both Δ_{P_θ}, $\Delta_Q > 0$ because $0 \leq P_\theta(R_{MC}), P_\theta(R_{AU}), Q(R_{MC}), Q(R_{MC}) \leq 1$ and also in the case that we are in P_θ, Φ_{MC} is expected to perform better compared to Φ_{AU}. Conversely, Φ_{AU} is expected to perform better in Q compared to Φ_{MC}.

We define T_1 meaning that P_θ was selected by the Bernoulli (λ)-experiment selecting the distribution, and T_2 is when Q was selected. We further define $P(T_1) = \lambda$, $P(T_2) = 1 - \lambda$, $P(R_{MC}|T_1) = P_\theta(R_{MC})$, $P(R_{MC}|T_2) = Q(R_{MC})$, $P(R_{AU}|T_1) = P_\theta(R_{AU})$ and $Q(R_{MC}|T_2) = Q(R_{AU})$.

The powers of the MC and AU test in the setup where it could be in either situation T_1 or situation T_2 is defined as follows;

$$\begin{aligned} P_\lambda(R_{MC}) &= P(R_{MC} \cap T_1) + P(R_{MC} \cap T_2) \\ &= P(R_{MC}|T_1)P(T_1) + P(R_{MC}|T_2)P(T_2) \\ &= P_\theta(R_{MC})P(T_1) + Q(R_{MC})P(T_2) \\ &= \lambda P_\theta(R_{MC}) + (1-\lambda)Q(R_{MC}) \\ &= \lambda\left[P_\theta(R_{MC}) - Q(R_{MC})\right] + Q(R_{MC}) \\ &= \lambda\Delta_{MC} + Q(R_{MC}). \end{aligned} \tag{1}$$

Similarly,

$$\begin{aligned} P_\lambda(R_{AU}) &= P(R_{AU} \cap T_1) + P(R_{AU} \cap T_2) \\ &= P(R_{AU}|T_1)P(T_1) + P(R_{AU}|T_2)P(T_2) \\ &= P_\theta(R_{AU})P(T_1) + Q(R_{AU})P(T_2) \\ &= \lambda P_\theta(R_{AU}) + (1-\lambda)Q(R_{AU}) \\ &= \lambda\left[P_\theta(R_{AU}) - Q(R_{AU})\right] + Q(R_{AU}) \\ &= \lambda\Delta_{AU} + Q(R_{AU}). \end{aligned} \tag{2}$$

This shows that the powers of the MC and AU are linear on λ with slopes Δ_{MC} and Δ_{AU} respectively as shown in Equation 1 and Equation 2.

Let's then consider a combined procedure where the model assumption is tested using an MS test. Depending on the outcome of the MS test (Φ_{MS}), either Φ_{MC} or Φ_{AU} is used. Let the MS test have Type I error (α_{MS}) where Q is assumed when we are in fact in P_θ and Type II error ($1 - \alpha^*_{MS}$) where P_θ is assumed when in fact we are in Q. This combined procedure would have eight possible outcomes, four of which are rejection; (i) in P_θ, MS testing not rejecting model assumption and using Φ_{MC} to test the main hypothesis, (ii) in P_θ, MS testing rejecting the the model assumption and using Φ_{AU} to test the main hypothesis, (iii) in Q, MS testing rejecting the model assumption and using Φ_{AU} to

test the main hypothesis and (iv) in Q, MS testing not rejecting the model assumption and using Φ_{MC} to test the main hypothesis. The results from these four situations are summarised in a 'global' rejection rate for the combined procedure $P_\lambda(R_{CP})$. Note that R^c_{MS} is the non-rejection of the MS test. From the law of total probability,

$$\begin{aligned}
P_\lambda(R_{CP}) &= P(R_{AU} \cap R_{MS} \cap T_1) + P(R_{MC} \cap R^c_{MS} \cap T_1) + P(R_{AU} \cap R_{MS} \cap T_2) \\
&\quad + P(R_{MC} \cap R^c_{MS} \cap T_2) \\
&= P(R_{AU}|R_{MS} \cap T_1)P(R_{MS} \cap T_1) + P(R_{MC}|R^c_{MS} \cap T_1)P(R^c_{MS} \cap T_1) \\
&\quad + P(R_{AU}|R_{MS} \cap T_2)P(R_{MS} \cap T_2) + P(R_{MC}|R^c_{MS} \cap T_2)P(R^c_{MS} \cap T_2) \\
&= P(R_{AU}|R_{MS} \cap T_1)P(R_{MS}|T_1)P(T_1) + P(R_{MC}|R^c_{MS} \cap T_1)P(R^c_{MS}|T_1)P(T_1) \\
&\quad + P(R_{AU}|R_{MS} \cap T_2)P(R_{MS}|T_2)P(T_2) + P(R_{MC}|R^c_{MS} \cap T_2)P(R^c_{MS}|T_2)P(T_2).
\end{aligned}$$

For simplicity, $P(T_1) = \lambda$, $P(T_2) = 1 - \lambda$, $P(R_{MS}|T_1) = \alpha_{MS}$, $P(R^c_{MS}|T_1) = 1 - \alpha_{MS}$, $P(R_{MS}|T_2) = \alpha^*_{MS}$, $P(R^c_{MS}|T_2) = 1 - \alpha^*_{MS}$, $P(R_{AU}|R_{MS} \cap T_1) = P_\theta(R_{AU}|R_{MS})$, $P(R_{MC}|R^c_{MS} \cap T_1) = P_\theta(R_{MC}|R^c_{MS})$, $P(R_{AU}|R_{MS} \cap T_2) = Q(R_{AU}|R_{MS})$ and $P(R_{MC}|R^c_{MS} \cap T_2) = Q(R_{MC}|R^c_{MS})$.

$$\begin{aligned}
P_\lambda(R_{CP}) &= \lambda\left[P_\theta(R_{AU}|R_{MS})\alpha_{MS} + P_\theta(R_{MC}|R^c_{MS})(1 - \alpha_{MS})\right] \\
&\quad + (1 - \lambda)\left[Q(R_{AU}|R_{MS})\alpha^*_{MS} + Q(R_{MC}|R^c_{MS})(1 - \alpha^*_{MS})\right]. \qquad (3)
\end{aligned}$$

This shows that the rejection probability of the combined procedure is a weighted mean of rejection probabilities of the model-based and the alternative procedure. One may wonder why, if this is the case, the combined procedure can have a rejection probability that is higher than both of these. The reason is that the two rejection probabilities here are not unconditional, but conditional on the decision of the misspecification test, and if the misspecification test does a good job, one should expect that the conditional probabilities are larger than the unconditional ones. Therefore the combined procedure can indeed be better than both unconditional rejection probabilities of the model-based and alternative procedure. Equation (3) will be used in the next section as a starting point for two lemmas that will be presented.

3.2 A POSITIVE RESULT FOR COMBINED PROCEDURES

In this section we present a point of view and a result that makes us think somewhat more positively about combined procedures and the impact of preliminary model testing. A characteristic of the literature analyzing combined procedures is that they compare the combined procedure with unconditional MC or AU tests both in situations where the model assumption of the MC test is fulfilled, or not fulfilled. However, they do not investigate a situation in which the MS test can do what it is supposed to do, namely to distinguish between these situations. From hereonafter, we require four assumptions:

(I) $\Delta_{P_\theta} = P_\theta(R_{MC}) - P_\theta(R_{AU}) > 0$,

(II) $\Delta_Q = Q(R_{AU}) - Q(R_{MC}) > 0$,

(III) $\alpha^*_{MS} = Q(R_{MS}) > \alpha_{MS} = P_\theta(R_{MS})$,

(IV) Both R_{MC} and R_{AU} are independent of R_{MS} under both P_θ and Q.

Keep in mind that this is about power, i.e., the H_0 of the main test is violated for both P_θ and Q. Assumption (I) means that the MC test has the better power under P_θ, (II) means that the AU test has the better power under Q. Assumption (III) means that the MS test has some use, i.e., it has a certain (possibly weak) ability to distinguish between P_θ and Q. All these are essential requirements for preliminary model assumption testing to make sense. Assumption (IV) though is very restrictive. It asks that rejection of the main null hypothesis by both main tests is independent of the decision made by the MS test. This is unrealistic in most situations. Approximate independence of the MS test and the main tests is an important desirable feature of a combined test, and it should not come as a surprise that a condition of this kind is required.

Lemma 1 Assuming (I) - (IV), $\exists\lambda \in (0,1)$ such that $P_\lambda(R_{CP}) > P(R_{MC}), P_\lambda(R_{CP}) > P(R_{AU})$ in the situation that the null hypothesis of the main test is violated.

Proof. By (I), for $\lambda = 1: \ P_\lambda(R_{MC}) > P_\lambda(R_{AU})$ and, by (II), for $\lambda = 0: \ P_\lambda(R_{AU}) > P_\lambda(R_{MC})$. As $P_\lambda(R_{MC})$ and $P_\lambda(R_{AU})$ are linear functions of λ, there must be $\lambda^* \in (0,1)$ so that $P_{\lambda^*}(R_{AU}) = P_{\lambda^*}(R_{MC})$. Obtain

$$\begin{aligned}
P_{\lambda^*}(R_{MC}) &= P_{\lambda^*}(R_{AU}) \\
\lambda^*\Delta_{MC} + Q(R_{MC}) &= \lambda^*\Delta_{AU} + Q(R_{AU}) \\
\lambda^*(\Delta_{MC} - \Delta_{AU}) &= \Delta_Q \\
\lambda^*(\Delta_{P_\theta} + \Delta_Q) &= \Delta_Q \qquad \text{(from (a), (b), (c) \& (d))} \\
\lambda^* &= \frac{\Delta_Q}{\Delta_{P_\theta} + \Delta_Q} \in (0,1). \qquad (4)
\end{aligned}$$

Using assumption (IV), (3) now becomes

$$\begin{aligned}
P_\lambda(R_{CP}) &= \lambda\big[P_\theta(R_{AU})\alpha_{MS} + P_\theta(R_{MC})(1-\alpha_{MS})\big] \\
&\quad + (1-\lambda)\big[Q(R_{AU})\alpha^*_{MS} + Q(R_{MC})(1-\alpha^*_{MS})\big] \\
&= \lambda\big[P_\theta(R_{AU})\alpha_{MS} + P_\theta(R_{MC}) - P_\theta(R_{MC})\alpha_{MS})\big] \\
&\quad + (1-\lambda)\big[Q(R_{AU})\alpha^*_{MS} + Q(R_{MC}) - Q(R_{MC})\alpha^*_{MS})\big] \\
&= \lambda\big[-\alpha_{MS}\Delta_{P_\theta} + P_\theta(R_{MC})\big] + (1-\lambda)\big[\alpha^*_{MS}\Delta_Q + Q(R_{MC})\big] \\
&= \lambda\big[-\alpha_{MS}\Delta_{P_\theta} + P_\theta(R_{MC}) - \alpha^*_{MS}\Delta_Q - Q(R_{MC})\big] + \alpha^*_{MS}\Delta_Q + Q(R_{MC}) \\
&= Q(R_{MC}) + \lambda\Delta_{MC} + \lambda\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q.
\end{aligned}$$

From (1),

$$P_\lambda(R_{CP}) = P_\lambda(R_{MC}) + \lambda\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q. \qquad (5)$$

Plugging (4) into (5) we now have the rejection rate of CP, $P_{\lambda^*,1}(R_{CP})$ where λ^* comes from (4) and 1 is because this is the proof for Lemma 1,

$$\begin{aligned}
P_{\lambda^*,1}(R_{CP}) &= P_{\lambda^*}(R_{MC}) + \lambda^*\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q \\
&= P_{\lambda^*}(R_{MC}) + \frac{\Delta_Q}{\Delta_{P_\theta} + \Delta_Q}\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q \\
&= P_{\lambda^*}(R_{MC}) + \Delta_Q\left[\frac{-\alpha_{MS}\Delta_{P_\theta}}{\Delta_{P_\theta} + \Delta_Q} - \frac{\alpha^*_{MS}\Delta_Q}{\Delta_{P_\theta} + \Delta_Q} + \alpha^*_{MS}\right]
\end{aligned}$$

$$= P_{\lambda^*}(R_{MC}) + \Delta_Q \left[\frac{-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q + \alpha^*_{MS}\Delta_{P_\theta} + \alpha^*_{MS}\Delta_Q}{\Delta_{P_\theta} + \Delta_Q} \right]$$

$$= P_{\lambda^*}(R_{MC}) + \frac{\Delta_{P_\theta}\Delta_Q}{\Delta_{P_\theta} + \Delta_Q} \left[\alpha^*_{MS} - \alpha_{MS} \right]. \tag{6}$$

Let $\tau = \frac{\Delta_{P_\theta}\Delta_Q}{\Delta_{P_\theta}+\Delta_Q}[\alpha^*_{MS} - \alpha_{MS}]$ Equation (6) now becomes

$$P_{\lambda^*,1}(R_{CP}) = P_{\lambda^*}(R_{MC}) + \tau. \tag{7}$$

Note that (7) is in the case where $P_{\lambda^*}(R_{MC}) = P_{\lambda^*}(R_{AU})$, therefore

$$P_{\lambda^*,1}(R_{CP}) = P_{\lambda^*}(R_{AU}) + \tau. \tag{8}$$

$\tau > 0$ by (I) - (III) so $P_{\lambda^*,1}(R_{CP})$ is greater than both $P_{\lambda^*}(R_{MC})$ and $P_{\lambda^*}(R_{AU})$. □

The independence assumption in the previous proof is then relaxed by introducing some small values as a measure of dependence, δ_1, δ_2, δ_3 & δ_4. Here we assume that by adding a small value, say δ_i, the powers of the MC and AU that are not conditional on the MS test but still conditional to whether or not the model assumption is violated will be equal to the power of the MC and AU test conditional on the rejection or non-rejection of the model assumption by the MS test. Note that the small value can be a positive or a negative number. The relationship between the power of the main tests conditional on the rejection or non-rejection of the model assumption and the power of the main test not conditional on the MS test is still unclear. Therefore, we assume an arithmetic relationship. We do not discard the possibility that the relationship could be more complex. A new set of definitions is as follows,

(e) $P_\theta(R_{AU}|R_{MS}) = P_\theta(R_{AU}) + \delta_1$

(f) $P_\theta(R_{MC}|R^c_{MS}) = P_\theta(R_{MC}) + \delta_2$

(g) $Q(R_{AU}|R_{MS}) = Q(R_{AU}) + \delta_3$

(h) $Q(R_{MC}|R^c_{MS}) = Q(R_{MC}) + \delta_4$.

Let there be a $\delta > 0$ so that $\delta = max\{|\delta_1|, |\delta_2|, |\delta_3|, |\delta_4|\}$. For the following Lemma, assumption (I) to (III) will still be required, but assumption (IV) will be replaced by assumption (V) as follows;

(V) Both R_{MC} and R_{AU} under P_θ and Q are dependent on R_{MS} with a small enough value depending on the involved probabilities $\delta = max\{|\delta_1|, |\delta_2|, |\delta_3|, |\delta_4|\} > 0$ where δ_i for i = (1, 2, 3, 4) is given in (e) - (h).

A new definition is introduced, τ where

$$\begin{aligned}\tau &= \lambda^*\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q \\ &= \frac{\Delta_{P_\theta}\Delta_Q}{\Delta_{P_\theta}+\Delta_Q}[\alpha^*_{MS} - \alpha_{MS}].\end{aligned} \tag{9}$$

Note that the definition of τ only depends on assumptions (I) - (III).

Lemma 2 Assuming (I) - (III) and (V), $\exists\lambda \in (0,1)$ such that $P_{\lambda^*}(R_{CP}) > P(R_{MC})$, $P_{\lambda^*}(R_{CP}) > P(R_{AU})$ in the situation that the null hypothesis of the main test is violated.

Proof. By (I) and (II), there must be $\lambda^* \in (0,1)$ so that $P_{\lambda^*}(R_{AU}) = P_{\lambda^*}(R_{MC})$. Obtain

$$\lambda^* = \frac{\Delta_Q}{\Delta_{P_\theta}+\Delta_Q} \in (0,1).$$

Equation (3) is updated with the definitions given above in (e), (f), (g) and (h) and where 2 in the index of the power of the CP, $P_{\lambda,2}(R_{CP})$ is because this is the proof for Lemma 2,

$$\begin{aligned}P_{\lambda,2}(R_{CP}) &= \lambda\big[(P_\theta(R_{AU})+\delta_1)\alpha_{MS} + (P_\theta(R_{MC})+\delta_2)(1-\alpha_{MS})\big] \\ &\quad + (1-\lambda)\big[(Q(R_{AU})+\delta_3)\alpha^*_{MS} + (Q(R_{MC})+\delta_4)(1-\alpha^*_{MS})\big] \\ &= \lambda\alpha_{MS}P_\theta(R_{AU}) + \lambda\alpha_{MS}\delta_1 + \lambda P_\theta(R_{MC}) + \lambda\delta_2 - \lambda\alpha_{MS}P_\theta(R_{MC}) \\ &\quad - \lambda\alpha_{MS}\delta_2 + \alpha^*_{MS}Q(R_{AU}) + \alpha^*_{MS}\delta_3 + Q(R_{MC}) + \delta_4 - \alpha^*_{MS}Q(R_{MC}) \\ &\quad - \alpha^*_{MS}\delta_4 - \lambda\alpha^*_{MS}Q(R_{AU}) - \lambda\alpha^*_{MS}\delta_3 - \lambda Q(R_{MC}) - \lambda\delta_4 \\ &\quad + \lambda\alpha^*_{MS}Q(R_{MC}) + \lambda\alpha^*_{MS}\delta_4 \\ &= Q(R_{MC}) + \lambda\big[P_\theta(R_{MC}) - Q(R_{MC})\big] + \lambda\big[-\alpha_{MS}(P_\theta(R_{MC}) - P_\theta(R_{AU})) \\ &\quad - \alpha^*_{MS}(Q(R_{AU}) - Q(R_{MC}))\big] + \alpha^*_{MS}\big[Q(R_{AU} - Q(R_{MC})\big] \\ &\quad + \lambda\alpha_{MS}(\delta_1-\delta_2) + \lambda\alpha^*_{MS}(-\delta_3+\delta_4) + \lambda(\delta_2-\delta_4) + \alpha^*_{MS}(\delta_3-\delta_4) + \delta_4.\end{aligned}$$

Using the definitions (a), (c) and (d) given in the beginning of this section,

$$\begin{aligned}P_{\lambda,2}(R_{CP}) &= Q(R_{MC}) + \lambda\Delta_{MC} + \lambda\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q \\ &\quad + \lambda\alpha_{MS}(\delta_1-\delta_2) + \lambda\alpha^*_{MS}(-\delta_3+\delta_4) + \lambda(\delta_2-\delta_4) + \alpha^*_{MS}(\delta_3-\delta_4) + \delta_4.\end{aligned}$$

From (1), $Q(R_{MC}) + \lambda\Delta_{MC} = P_\lambda(R_{MC})$,

$$\begin{aligned}P_{\lambda,2}(R_{CP}) &= P_\lambda(R_{MC}) + \lambda\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q + \lambda\alpha_{MS}(\delta_1-\delta_2) \\ &\quad + \lambda\alpha^*_{MS}(-\delta_3+\delta_4) + \lambda(\delta_2-\delta_4) + \alpha^*_{MS}(\delta_3-\delta_4) + \delta_4.\end{aligned}$$

Substitute $\lambda^* = \frac{\Delta_Q}{\Delta_{P_\theta}+\Delta_Q}$ and note that $\lambda^*\big[-\alpha_{MS}\Delta_{P_\theta} - \alpha^*_{MS}\Delta_Q\big] + \alpha^*_{MS}\Delta_Q = \tau$ as defined in (9),

$$\begin{aligned} P_{\lambda^*,2}(R_{CP}) = P_{\lambda^*}(R_{MC}) + \tau + \lambda^*\alpha_{MS}(\delta_1 - \delta_2) + \lambda^*\alpha^*_{MS}(-\delta_3 + \delta_4) \\ + \lambda^*(\delta_2 - \delta_4) + \alpha^*_{MS}(\delta_3 - \delta_4) + \delta_4. \end{aligned}$$

By (V),

$$\begin{aligned} P_{\lambda^*,2}(R_{CP}) &\geq P_{\lambda^*}(R_{MC}) + \tau + \lambda^*\alpha_{MS}(-\delta - \delta) + \lambda^*\alpha^*_{MS}(-\delta - \delta) \\ &\quad + \lambda^*(-\delta - \delta) + \alpha^*_{MS}(-\delta - \delta) - \delta \\ &\geq P_{\lambda^*}(R_{MC}) + \tau - \delta(2\lambda^*\alpha_{MS} + 2\lambda^*\alpha^*_{MS} + 2\lambda^* + 2\alpha^*_{MS} + 1) \end{aligned}$$

It is known that $\alpha_{MS}, \alpha^*_{MS} \in [0,1]$ and $\lambda^* \in (0,1)$, therefore $(2\lambda^*\alpha_{MS} + 2\lambda^*\alpha^*_{MS} + 2\lambda^* + 2\alpha^*_{MS} + 1) < 9$,

$$P_{\lambda^*,2}(R_{CP}) > P_{\lambda^*}(R_{MC}) + \tau - 9\delta. \tag{10}$$

Note that (10) is in the case where $P_{\lambda^*}(R_{MC}) = P_{\lambda^*}(R_{AU})$, therefore

$$P_{\lambda^*,2}(R_{CP}) > P_{\lambda^*}(R_{AU}) + \tau - 9\delta. \tag{11}$$

Clearly $\tau > 0$ and by definition $\delta > 0$. If we assume $\tau > 9\delta$, then $P_{\lambda^*,2}(R_{CP})$ is larger than both $P_{\lambda^*}(R_{MC})$ and $P_{\lambda^*}(R_{AU})$. □

4

SIMULATING A TWO-STAGE MS TESTING PROCEDURE

''In God we trust; all others must bring data.'' W. Edwards Deming

In the literature reviewed, the simulation studies to test the performance when using MS testing were carried out in restricted situations. Restricted here means that the samples used to test MS testing procedure all come from a single source. For example, in the case of the one sample test, the samples are generated from one distribution and in the case of two sample testing, both samples are generated from one distribution.

This simulation 'models' a situation in which both fulfilled and (in a specific way) violated model assumptions can happen. As such this can hardly be criticised as less realistic as simulations that only simulate one side of this. However, this is not free from criticism as the reality is not really like this. We feel this setup is at least as realistic as pretty much every other simulation or even theorem based on parametric models. One must note is that although in reality we should not ever believe that any statistical model assumption is fulfilled, in reality in fact there are situations in some data sets that are pretty close to a specific model assumption and some that are further away. So one will find situations that are very close to the parametric model (and depending on how the situation exactly looks like, one could suspect that the simulation results for the parametric case are very close to this) and other situations that are close to the second distribution that is being looked at.

Obviously in reality there will be more than two candidate distributions, however looking at a mixture of two is the simplest and therefore logical next step from looking at only one (which is what is normally done). The aim is to understand what goes on when looking at more than one possible model.

4.1 LEVENE'S TEST AS AN MS TEST FOR EQUALITY OF VARIANCES

This section replicates the work done in Zimmerman (2004) to investigate the performance of an MS testing procedure to test equality of variances and depending on the

rejection of the MS test, either carry out a standard two sample t-test Gosset ("Student") (1908) or the Welch's modified t-test Welch (1938) Welch (1947) Satterthwaite (1946) to test the main null hypothesis that both samples come from same distribution. This combined procedure is denoted by CP. This combined procedure is then compared to the unconditional tests, the standard t-test and the Welch's t-test denoted by Model-based Constrained (MC) and Alternative Unconstrained (AU) respectively.

First, an independent random sample of size n_1 was generated from the standard normal distribution. The second sample n_2 was generated with the ratio of the standard deviation having a predetermined value ranging from 1 to 2.5 in increments of 0.5. The total sample size $n_1 + n_2$ was fixed at 30 or 60 and the ratio n_1/n_2 ranged from 0.2 to 5. For every sample n_1 and n_2 generated, the two sample Student's t-test and the Welch's t-test were applied without testing the homogeneity condition at the 0.01 and 0.05 significance level on both samples. Then, the Levene's test of equality of variances (see Section 2.3.2) was performed at the 0.01 and 0.05 significance level. If this test does not reject the hypothesis of equal variances on both samples, the usual two sample Student t-test based on pooled variances was performed at the same significance level. If the Levene's test rejected the hypothesis of equal variances on either or both samples, the Welch's modified t-test based on separate variances was performed also at the same significance level. This procedure is repeated 10000 times and the rejection rates were calculated.

The results shown in Table 1 is consistent with the results presented in Zimmerman (2004). The unconditional two sample t-test's or MC rejection rate depends on the sample size and the magnitude of the ratio, that is when a large variance ratio is associated with a large sample size ratio, the rejection rate falls below the nominal level. Conversely, when a large variance ratio is associated with a small sample size ratio, the rejection rate increases above the nominal level. The rejection rates when the unconditional Welch's t-test or AU remains consistently around the nominal level. The CP rejection rates are better than the MC but not as good as the AU. Therefore the author recommends always using the Welch's t-test when testing the main hypothesis of equality of means.

n_1	n_2	$\frac{\sigma_1}{\sigma_2}$	$\alpha = 1\%$			$\alpha = 5\%$		
			MC	AU	CP	MC	AU	CP
50	10	1.0	1.01	1.14	1.03	4.95	5.17	5.32
		1.5	0.11	1.11	0.28	1.19	4.93	2.80
		2.0	0.01	1.15	0.43	0.41	4.96	3.94
		2.5	0.00	1.11	0.60	0.10	5.15	4.67
40	20	1.0	0.92	0.95	0.95	5.03	4.97	5.10
		1.5	0.37	0.98	0.55	2.55	5.03	4.04
		2.0	0.17	1.09	0.73	1.66	4.72	4.33
		2.5	0.19	0.82	0.71	1.29	5.11	5.03
20	40	1.0	1.06	0.96	1.04	4.90	4.92	4.94
		1.5	2.15	0.94	1.76	8.49	4.91	6.25
		2.0	3.55	1.01	1.76	10.98	5.14	5.67
		2.5	4.82	0.99	1.25	13.00	4.99	5.14
10	50	1.0	1.01	1.29	1.04	4.99	5.13	5.21
		1.5	4.54	1.06	3.73	13.16	4.89	9.30
		2.0	9.24	1.07	5.01	20.83	5.22	8.52
		2.5	13.05	0.90	3.51	26.33	5.24	6.75
25	5	1.0	0.96	1.72	0.99	5.03	5.42	5.53
		1.5	0.12	1.58	0.21	1.16	5.72	2.75
		2.0	0.03	0.97	0.07	0.43	5.02	2.41
		2.5	0.01	0.90	0.13	0.23	5.34	3.17
20	10	1.0	1.18	1.18	1.18	4.83	4.89	4.89
		1.5	0.34	0.84	0.39	2.80	5.08	3.69
		2.0	0.29	1.14	0.46	1.91	4.94	3.81
		2.5	0.12	0.92	0.34	1.52	5.20	4.35
10	20	1.0	1.04	1.22	1.04	4.96	5.04	5.07
		1.5	2.53	1.15	2.43	8.67	5.15	7.42
		2.0	3.76	1.17	2.98	11.45	5.08	7.41
		2.5	5.14	1.28	3.23	13.26	5.01	6.34
5	25	1.0	1.02	1.53	1.03	5.00	5.58	5.51
		1.5	4.28	1.71	4.10	13.76	5.70	11.65

2.0	8.87	1.48	7.04	20.15	5.17	12.90
2.5	13.45	1.32	8.66	25.71	5.51	12.00

Table 1: Rejection rates of the null hypothesis (%) for various combinations of sample sizes, significance levels, ratios of standard deviations and the different methods.

4.2 SIMULATION SETUP

The motivation for this section is to challenge the MS test by randomising the distribution of samples being put through the procedures. A simulation study was designed to investigate the performance of a statistical inference test with and without model checking.

Consider a parametric test with some assumptions and call this MC. Additionally, consider a non-parametric test where one or more of the assumptions of MC are not needed. We call this AU. We also consider a combined procedure where an MS test is carried out to test a certain model assumption. Depending on the outcome of the MS test, we either choose to do MC when the MS test decides that the model assumptions is not violated or AU when the MS test decides that the model assumption is violated. We call this combined procedure CP. The null hypothesis of MC and AU can be either fulfilled to investigate the rejection rates or violated to investigate the power. All of these tests were carried out at a significance level α. Figure 1 shows the flow chart of the simulation process involving MC, AU and CP.

A two-sample situation with sample sizes n_1 and n_2 respectively is considered. Both samples are generated from either one of two distributions P_θ or Q randomly. The choice to generate both samples from either P_θ or Q is determined by a random number generator from the Bernoulli distribution with λ probability of generating from P_θ. As opposed to standard mixture models, λ governs the distribution of the whole dataset, not every single observation independently.

MC is applied without any conditions and the decision to reject or not reject the MC test's null hypothesis is noted. Next, AU is applied without any conditions and the decision to reject or not reject the AU test's null hypothesis is noted. Then, CP is applied after model checking using an MS test. The MS test is done on both samples and if both the samples do not reject the model assumption, the MC test is used. If either one or both the samples reject the model assumption, the AU test is used. The decision to reject or not reject the main null hypothesis is noted. Finally, CP_{adj} is applied with an adjusted

value of α_{MS} for the MS test. This is repeated M times and the rejection rates of all the four methods is noted. All the simulations were done in R.

4.3 TESTING THE MAIN NULL HYPOTHESIS OF EQUAL DISTRIBUTIONS

The simulation process outlined in Figure 1 was carried out to investigate the effect of MS testing for normality. Let's consider sample sizes $n = 8, 27$ and 125 for both samples generated. These sample sizes were chosen because they are the cube of the first three prime numbers and they are representative of a sufficiently small sample size, a moderately sized sample and a fairly large sample size. The simulation above was repeated for $M = 100,000$ times. All simulations in this thesis were done in R.

The t-test is chosen as the MC test because this is one of the most popular tests that is used in research today. It was developed to monitor the stout quality in a brewery. There are two types of t-tests; one sample and two samples. The one sample t-test is used to test the null hypothesis that the mean of the population is equal to a certain value, while the two sample t-test is used to compare the mean values of both samples. The two sample t-tests may be conducted on independent samples, paired samples or overlapping samples. As mentioned, it can be used as a quality control tool. Furthermore, another important use of t-tests is in the medical industry where the paired t-test is used widely to study the impact of a particular treatment on a sample of patients before and after the medication. For these reasons, the two sample t-test was chosen to be studied here. Two important assumption of the t-test is that the means of the samples are normally distributed and both samples have equal variance. The null hypothesis of the two sample t-test is that both samples have equal means therefore it is implied that both distributions are equal. This is so because it is posited that if two samples have the same means and the same variance, therefore they must be the same. The alternative hypothesis is that the means are not equal hence the distributions are not equal. The Welch's t-test is a modification of the standard t-test that does not assume equal variances of both samples.

We also have the Wilcoxon-Mann-Whitney (WMW) test. This test also tests for equality of distributions in two samples. The WMW does not have the assumption that the sample means are normally distributed and also that both samples have equal variances. This leads researchers to believe that the WMW is an alternative to the t-test. Hence, we choose this as the AU test. The null hypothesis of the non-parametric Wilcoxon-Mann-Whitney test is two distributions are equal. The alternative hypothesis is that one distribution is stochastically larger that the other. Although Mann and Whitney developed their method with the aforementioned hypotheses, there are many other ways

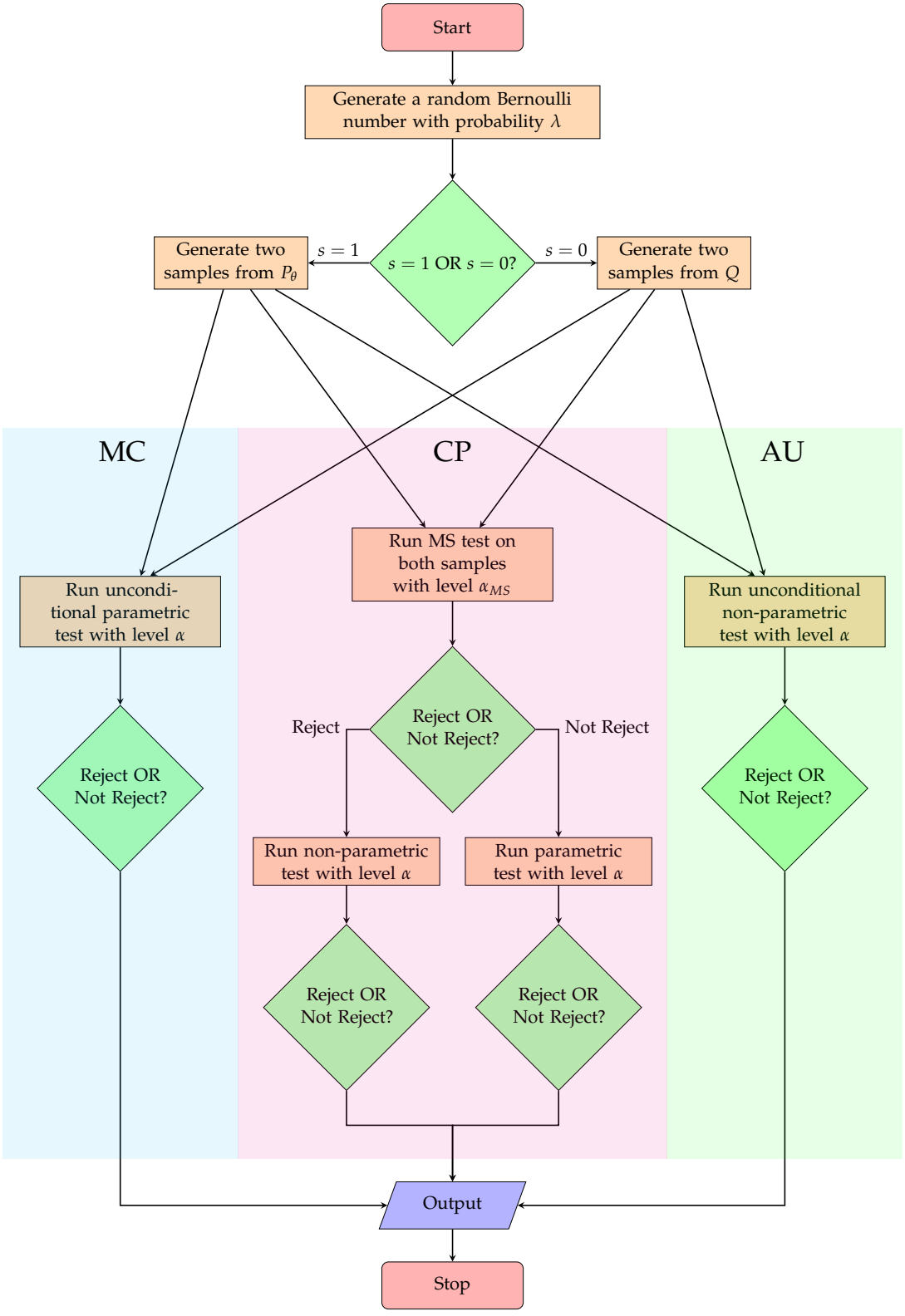


Figure 1: A flow chart of the simulation process involving MC, AU and CP. A Bernoulli random variable is used to decide between generating from a situation where the model assumption is fulfilled or violated. After generating the sample, it is put through three procedures, namely the MC, AU and CP, to calculate the level or power of the testing procedures. MC and AU procedures do not involve MS testing to check model assumptions.

to formulate the hypotheses such that the Wilcoxon-Mann-Whitney U-test will still give a valid test, see Fay and Proschan (2010) for some examples of formulation of the hypotheses.

4.3.1 *t-test versus Wilcoxon-Mann-Whitney*

MC is the two sample t-test, AU is the Wilcoxon-Mann-Whitney U-test and the MS test is the Shapiro-Wilk test (see Section 2.3.2). Each of the tests was carried out at $\alpha = \alpha_{MS} = 5\%$ significance level.

Two distributions were considered and the probability of generating samples from one of the distributions are given by λ. The values for P_θ, Q, and λ are given by Table 2. Both samples generated from P_θ or Q have the same sample size and also have the same parameters, i.e. same mean and variance in the case of P_θ and the same degree of freedom in the case of Q. The t distribution with 3 degrees of freedom was chosen because it is the smallest degree of freedom where a mean and a variance exists to describe the density curve. The t distribution with 4 degrees of freedom was chosen because the variance for this distribution is $\frac{v}{v-2} = 2$ and distribution P_θ was also set to have a variance of 2. The two sample t-test has an equal variance assumption therefore the variances are made to be identical. The t distribution with 2 degrees of freedom was chosen as it only has a mean value and has heavier tails, which may be able to challenge the two step MS procedure. Since this is a multiple test situation, the Bonferroni correction is also applied to the MS test in a separate procedure where the significance level is changed to α_{adj} where $\alpha_{adj} = \frac{\alpha_{MS}}{2}$ using the Bonferroni correction to check the model assumptions of both samples. We will call this combined procedure with adjusted level CP_{adj}.

Situation	Distribution P_θ	Distribution Q	λ
(1)	Normal, $\mu = 0, \sigma^2 = 1$	t, $df = 3$	0.5
(2)	Normal, $\mu = 0, \sigma^2 = 1$	t, $df = 3$	0.25
(3)	Normal, $\mu = 0, \sigma^2 = 1$	t, $df = 3$	0.75
(4)	Normal, $\mu = 0, \sigma^2 = 2$	t, $df = 4$	0.5
(5)	Normal, $\mu = 0, \sigma^2 = 2$	t, $df = 4$	0.25
(6)	Normal, $\mu = 0, \sigma^2 = 2$	t, $df = 4$	0.75
(7)	Normal, $\mu = 0, \sigma^2 = 1$	t, $df = 2$	0.5
(8)	Normal, $\mu = 0, \sigma^2 = 1$	t, $df = 2$	0.25
(9)	Normal, $\mu = 0, \sigma^2 = 1$	t, $df = 2$	0.75

Table 2: Distributions and lambda values examined in the simulation study. Both samples are generated from either P_θ or Q with the given parameters

4.3.1.1 *Main null hypothesis is fulfilled*

The situation where the main null hypothesis of equal distributions were looked at first to study the Type-I error rates. The level of the MC, AU and CP tests were carried out at the nominal level 5%. The levels of the main tests in CP_{adj} were carried out at 5%, the difference is the level of the MS test was adjusted using the Bonferroni correction where $\alpha_{adj} = \frac{\alpha_{MS}}{2}$.

Situation	n	Method			
		MC	AU	CP	CP_{adj}
(1)	8	4.616 (0.066)	4.968 (0.069)	5.082 (0.069)	4.999 (0.069)
	27	4.711 (0.067)	4.704 (0.067)	4.955 (0.069)	4.922 (0.068)
	125	4.798 (0.068)	4.977 (0.069)	5.011 (0.069)	4.995 (0.069)
(2)	8	4.406 (0.065)	4.899 (0.068)	4.987 (0.069)	4.899 (0.068)
	27	4.534 (0.066)	4.690 (0.067)	4.826 (0.068)	4.834 (0.068)
	125	4.813 (0.068)	4.942 (0.069)	4.989 (0.069)	4.985 (0.069)
(3)	8	4.849 (0.068)	5.068 (0.069)	5.183 (0.070)	5.098 (0.070)
	27	4.975 (0.069)	4.943 (0.069)	5.168 (0.070)	5.156 (0.070)
	125	4.957 (0.069)	4.974 (0.069)	5.069 (0.069)	5.036 (0.069)
(4)	8	4.742 (0.067)	4.935 (0.068)	5.064 (0.069)	5.005 (0.069)
	27	4.854 (0.068)	4.781 (0.067)	5.011 (0.069)	5.001 (0.069)
	125	4.958 (0.069)	5.019 (0.069)	5.103 (0.070)	5.098 (0.070)
(5)	8	4.521 (0.066)	4.817 (0.068)	4.929 (0.068)	4.848 (0.068)
	27	4.730 (0.067)	4.790 (0.068)	4.968 (0.069)	4.975 (0.069)
	125	5.051 (0.069)	4.974 (0.069)	4.998 (0.069)	4.980 (0.069)
(6)	8	4.840 (0.068)	4.962 (0.069)	5.091 (0.070)	5.042 (0.069)
	27	4.964 (0.069)	4.797 (0.068)	5.088 (0.069)	5.091 (0.070)
	125	4.988 (0.069)	4.999 (0.070)	5.085 (0.070)	5.056 (0.069)
(7)	8	4.361 (0.065)	5.086 (0.069)	5.108 (0.070)	5.039 (0.069)
	27	4.502 (0.066)	4.880 (0.068)	5.017 (0.069)	5.003 (0.069)
	125	4.581 (0.066)	4.933 (0.068)	5.018 (0.069)	5.001 (0.069)
(8)	8	3.995 (0.062)	4.956 (0.069)	4.984 (0.069)	4.912 (0.068)
	27	4.339 (0.064)	4.759 (0.067)	4.918 (0.068)	4.908 (0.068)
	125	4.453 (0.065)	4.977 (0.069)	5.004 (0.069)	5.000 (0.069)

	8	4.699 (0.067)	5.127 (0.070)	5.183 (0.070)	5.132 (0.070)
(9)	27	4.744 (0.067)	4.885 (0.068)	5.086 (0.069)	5.072 (0.069)
	125	4.919 (0.068)	5.102 (0.070)	5.172 (0.070)	5.132 (0.069)

Table 3: Rejection rates of the null hypothesis (%) and standard errors (in parentheses)(%) for various sample sizes. The MC test is the standard t-test and the AU test is the WMW test. Values underlined were rejected by the proportion test as significantly different than 5%

Table 3 shows the values of the rejection rates and their respective standard errors when running the simulations described above. As expected, most of the values of the rejection rates are close to the nominal level of 5%. The rejection rates are then tested with the proportion test with the null hypothesis that the rejection rate is equal to 5%. The 95% confidence interval for the test is $(4.865, 5.135)$ using the following expression

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{M}}. \tag{12}$$

The values that were rejected by the proportion test is underlined in Table 3. The results of the proportion test show that the unconditional parametric test constantly underperforms at level 5%. However, the performances of the AU, CP and CP_{adj} are quite similar and no clear advantage can be seen in favour of one method. The CP_{adj} seems to perform the best with only 3 instances where the rejection rates are significantly different than 5% but a definitive conclusion cannot be made without studying the power of the procedures. The Bonferroni correction does not seem to improve the rejection rates. This could be due to the combined procedure having an alternative situation in the case that the model assumption is not rejected.

4.3.1.2 *Main null hypothesis is violated*

We then consider investigating the power of these three approaches (MC, AU and CP) to try and get a clearer picture of how they perform when the main null hypothesis is violated. The main null hypothesis is that the distributions of both samples are equal. The simulation study from the Section 4.2 is repeated to calculate the power of rejecting a false null. Hence, out of the two samples generated, the second sample is generated with the mean shifted at three different degrees of violation. This shift is referred to as the non-centrality parameter (*ncp*). Three *ncp*s were considered, namely 0.5, 1 and 2. Eleven different values of $\lambda = [0, 1]$ were considered to study how the power changes with different mixture of samples. Although only two values of λ is needed to plot a

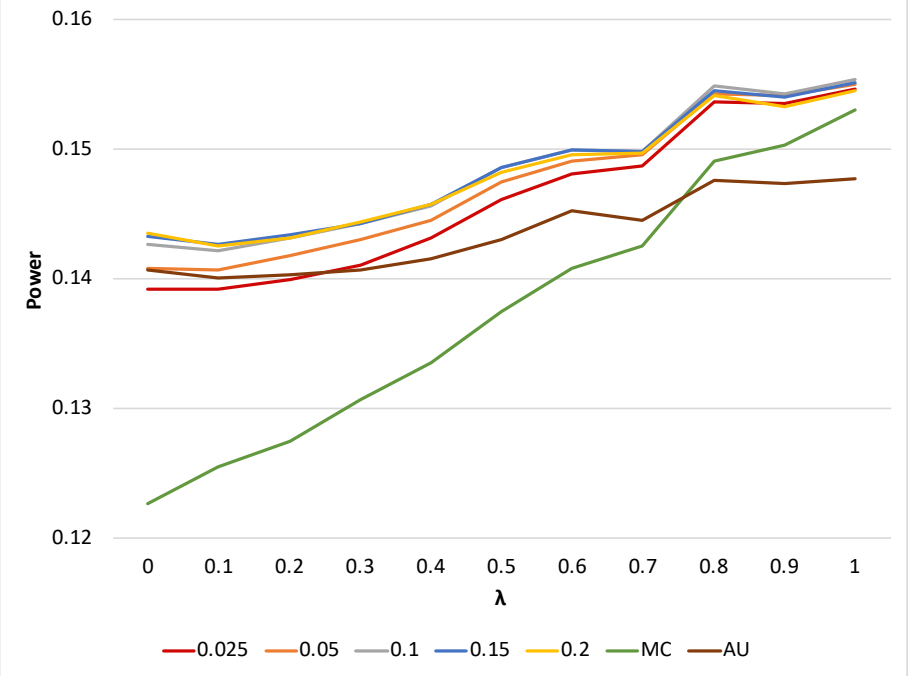


Figure 2: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 8$ and $ncp = 0.5$. The values on the legend for the red, orange, grey, blue and yellow lines are the level on which the model assumptions are tested. Levels of the main tests are 0.05.

linear line as proven in Lemma 1 in Chapter 4, simulating over the range of λ will give a clearer picture of how the power behaves with random variation in the simulation.

Keselman, Othman and Wilcox (2013) recommended that the MS test be carried out at a higher significance level to increase the power to "*detect effects and concomitantly reduce the probability of falsely accepting the null hypothesis that data are normally distributed*". Therefore, 5 different levels of α_{MS} were considered, $\alpha_{MS} = \{0.025, 0.05, 0.1, 0.15, 0.2\}$. The results of these simulations are shown in the figures below.

Figures 2 - 10 shows the power analysis of different sample sizes across different λ values considering different α_{MS} levels and three procedures of testing the null hypothesis that two samples come from the same distribution. Interestingly, the MS testing procedure has consistently larger power compared to the MC and AU testing. When $\lambda = 0$, the MC or in this case, the unconditional t-test, has a lower power to detect difference in means between two samples drawn from the t distribution. The AU, in this case the Wilcoxon-Mann-Whitney U-test, always has a larger power than the MC when $\lambda = 0$. In certain situations for example in Figures 4 - 9 and when $\lambda = 1$, namely when all the samples were drawn from the normal distribution, the MC or the t-test performs just as well as the MS testing procedure.

Looking at the different α_{MS} levels, nothing conclusive can be said about which levels give the best power. However, a case can be made for $\alpha_{MS} = 0.1$ (grey line). In Figure

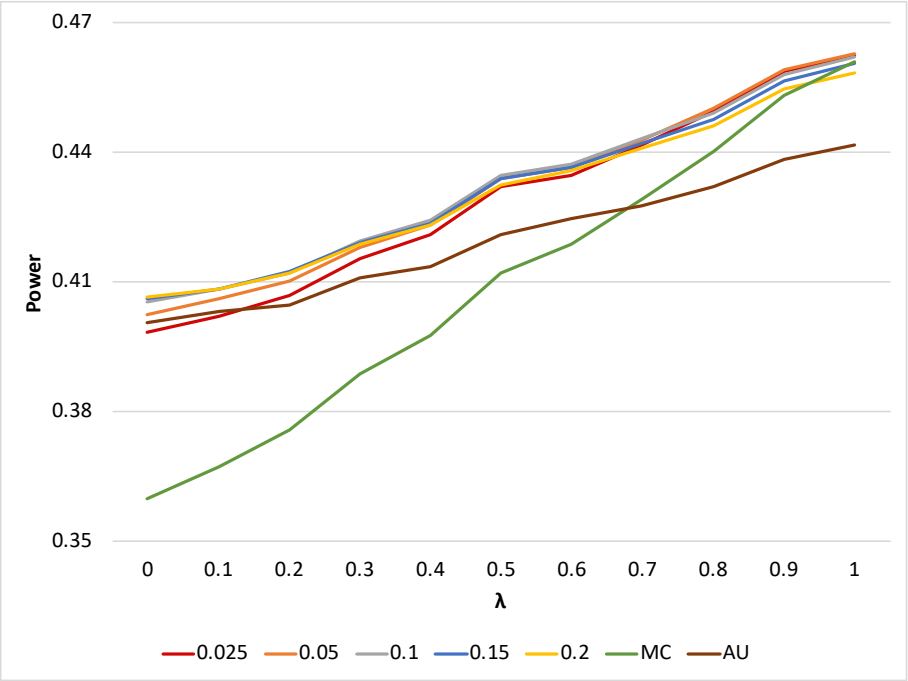


Figure 3: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 8$ and $ncp = 1$

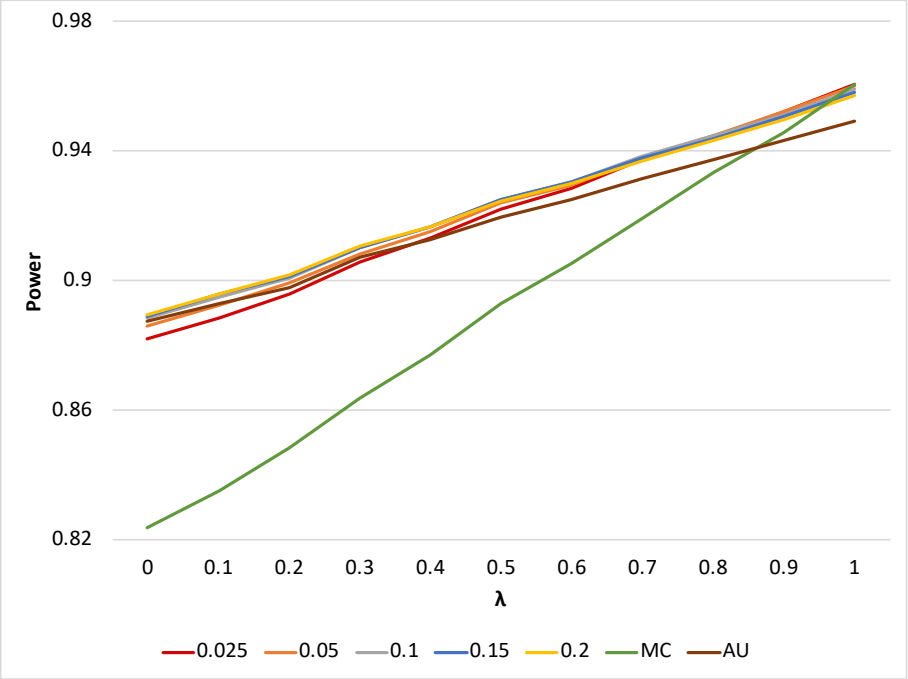


Figure 4: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 8$ and $ncp = 2$

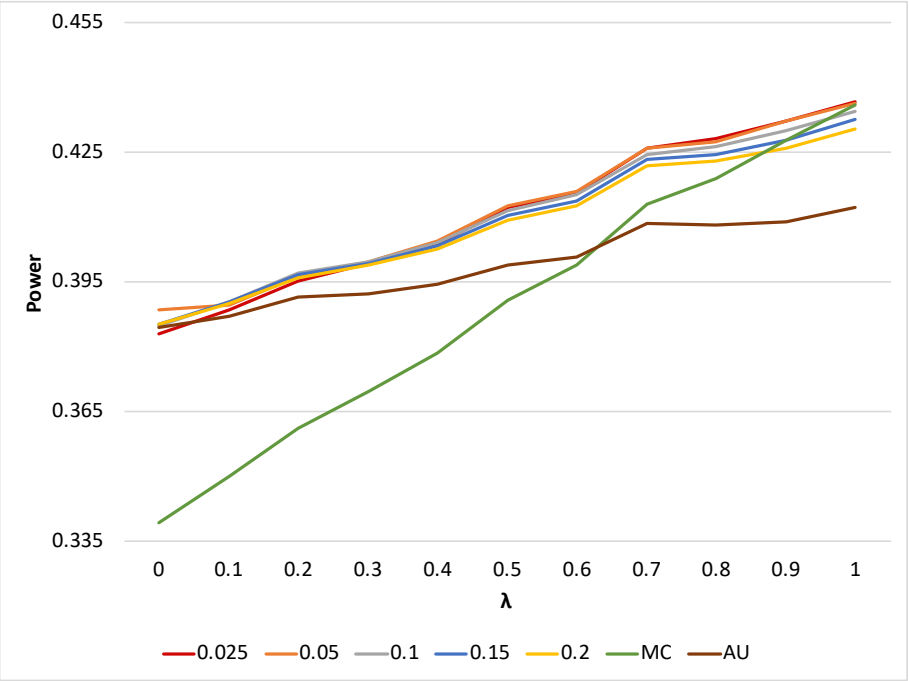


Figure 5: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 27$ and $ncp = 0.5$

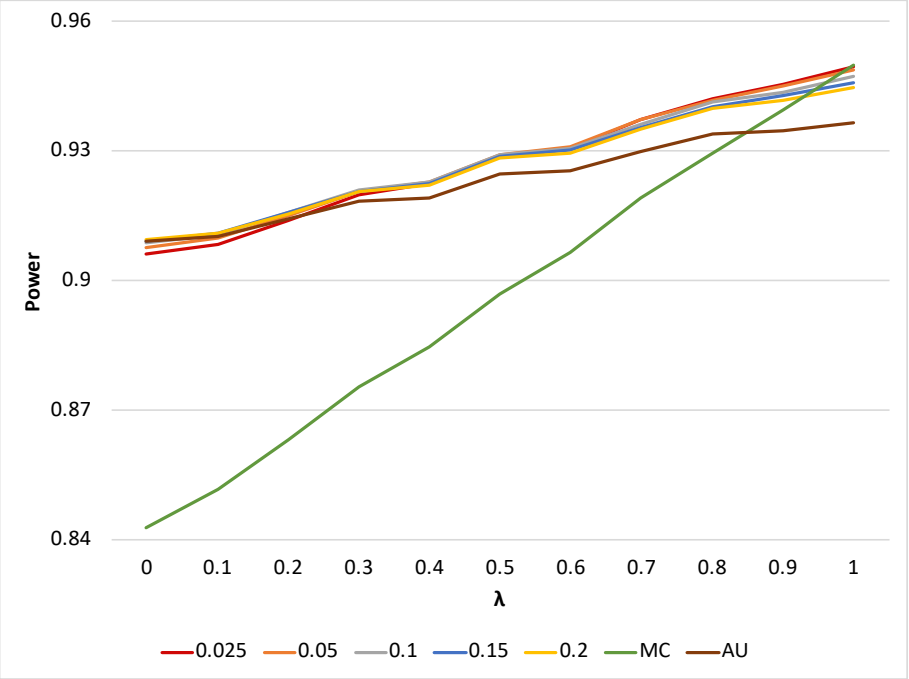


Figure 6: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 27$ and $ncp = 1$

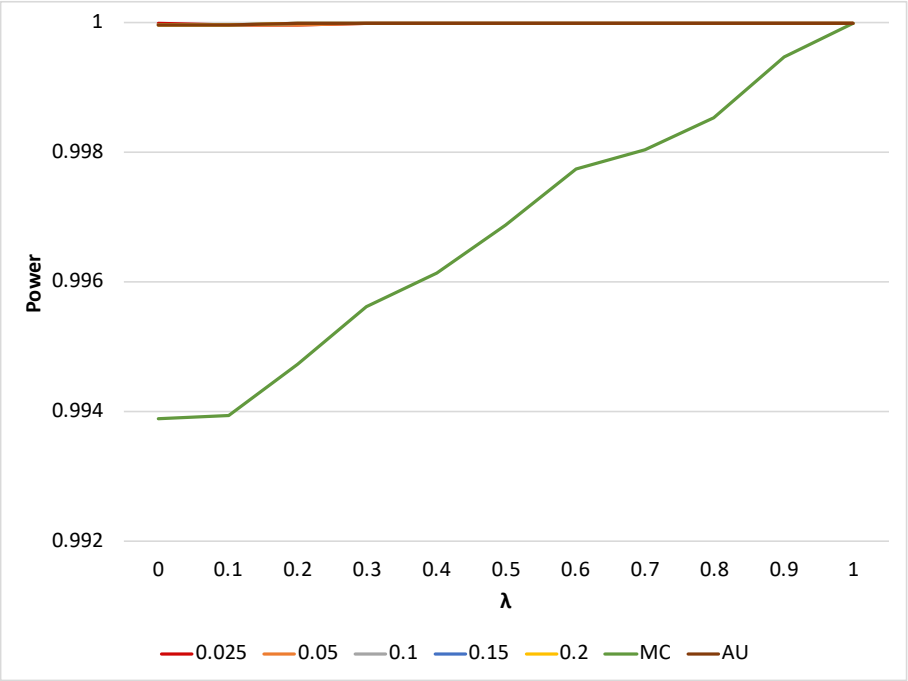


Figure 7: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 27$ and $ncp = 2$

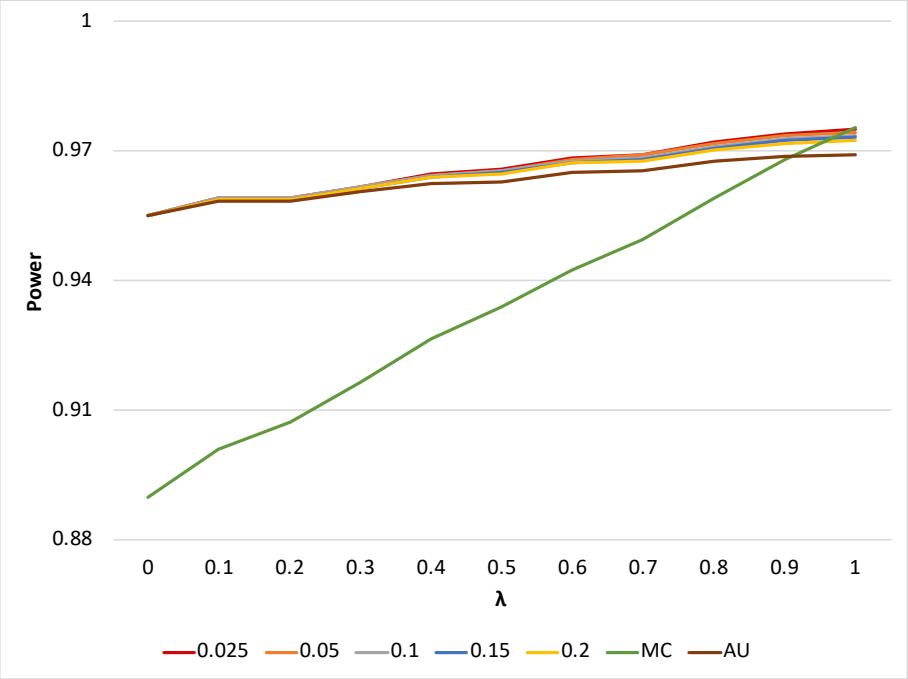


Figure 8: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 125$ and $ncp = 0.5$

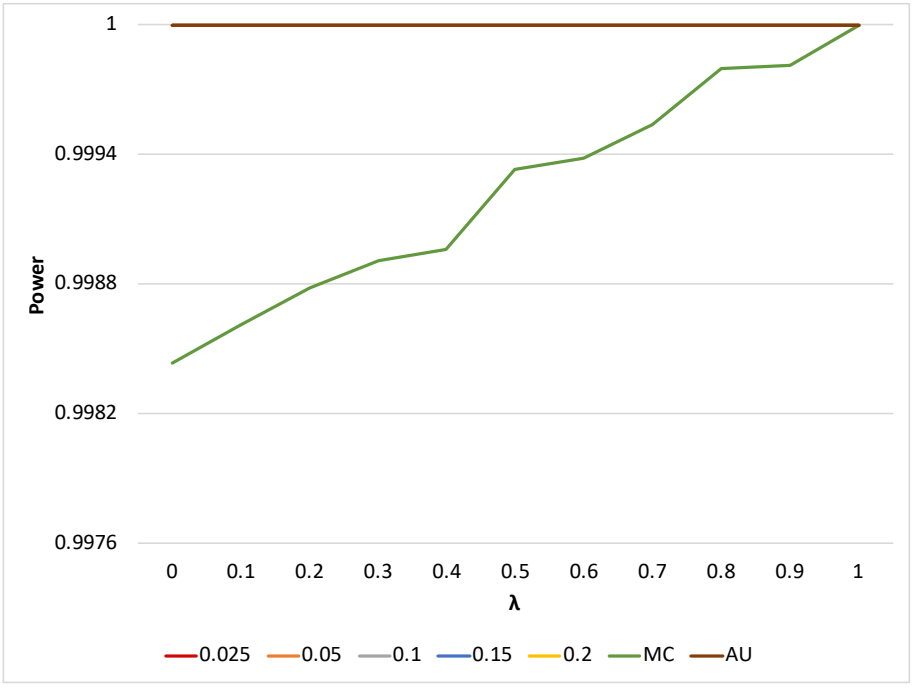


Figure 9: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 125$ and $ncp = 1$

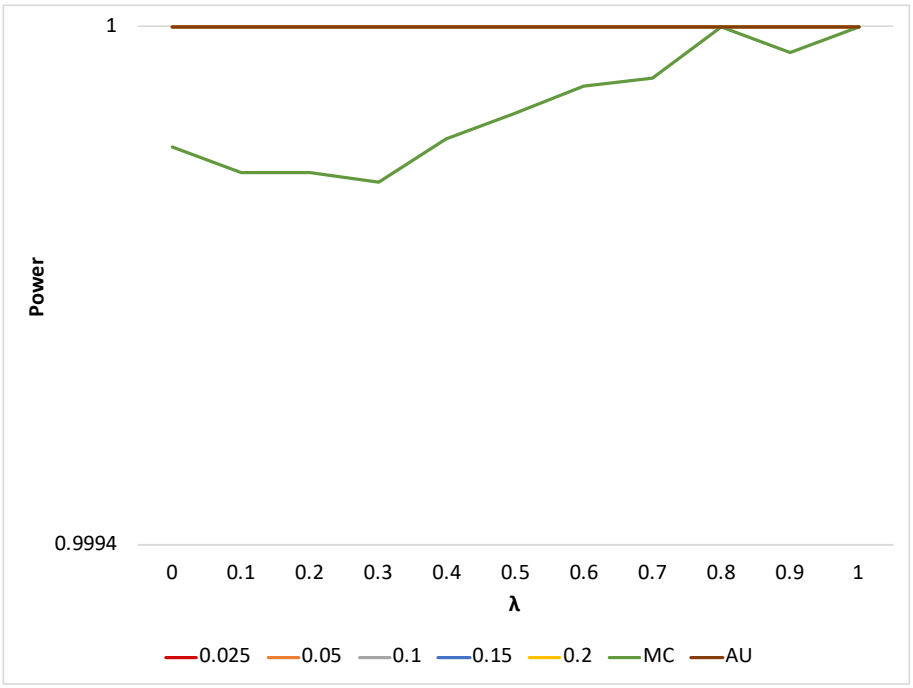


Figure 10: Power levels for 5 different values of α_{MS}, the MC and AU methods across different λ values for $n = 125$ and $ncp = 2$

2, where a small sample size ($n = 8$) and a minor violation of the null hypothesis ($ncp = 0.5$) is studied, when $\lambda \geq 0.4$, the power when $\alpha_{MS} = 0.1$ is the highest although not by a huge magnitude. When $\lambda < 0.4$, the power when $\alpha_{MS} = 0.1$ is in the middle of the pack. In the other plots, the trend seems to be that at smaller λ values, again when there are more samples drawn from the non-normal distribution, larger α_{MS} give a slightly higher power. This means that the MS test is able to reject more samples where the model assumptions were not fulfilled and pushing these samples to the non-parametric test which has a more relaxed assumption leading to a higher power. As the λ gets larger and there are more samples drawn from the normal distribution, the MS test procedure benefits from a smaller α_{MS} and pushing more samples to the parametric test where normality is assumed. A good compromise then would be to set $\alpha_{MS} = 0.1$. When the sample size is sufficiently large, $n = 125$, the power of AU and the MS testing is quite similar. The MC still has the lowest power at smaller λ values.

4.3.2 *Welch's t-test versus Wilcoxon-Mann-Whitney*

The simulations in the previous section is repeated replacing the MC t-test with the Welch's t-test. In the literature, the Welch's t-test is suggested to be more robust and should always be used in favour of the standard t-test, see Zimmerman (2004), Ruxton (2006) and Delacre, Lakens and Leys (2017). The AU remains the Wilcoxon-Mann-Whitney. P_θ is the normal distribution and Q is the t distribution with 3 degrees of freedom.

4.3.2.1 *Main null hypothesis is fulfilled*

n	MC	AU	CP	CP_{adj}
8	4.112 (0.063)	4.899 (0.068)	4.758 (0.067)	4.649 (0.067)
27	4.481 (0.065)	4.690 (0.067)	4.808 (0.068)	4.814 (0.068)
125	4.809 (0.068)	4.942 (0.069)	5.001 (0.069)	4.991 (0.069)

Table 4: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes and $\lambda = 0.25$

n	MC	AU	CP	CP_{adj}
8	4.328 (0.064)	4.968 (0.069)	4.859 (0.068)	4.751 (0.067)
27	4.680 (0.067)	4.704 (0.067)	4.941 (0.069)	4.905 (0.068)
125	4.793 (0.068)	4.977 (0.069)	5.059 (0.069)	5.039 (0.069)

Table 5: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes and $\lambda = 0.5$

n	MC	AU	CP	CP_{adj}
8	4.602 (0.067)	5.068 (0.069)	4.972 (0.069)	4.877 (0.068)
27	4.939 (0.069)	4.943 (0.069)	5.145 (0.070)	5.132 (0.070)
125	4.953 (0.069)	4.974 (0.069)	5.096 (0.070)	5.105 (0.070)

Table 6: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes and $\lambda = 0.75$

Tables 4 - 6 shows the value of the rejection rates and their respective standard errors across different values of λ. About half of all the rejection rates shown are significantly different than 5% according to the proportion test given in (12). These values are underlined in Tables 4 - 6. The results show that in small sample sizes the Welch's t-test underperforms. In this particular setup, the Wilcoxon-Mann-Whitney test performs quite similarly to the combined procedure.

4.3.2.2 *Main null hypothesis is violated*

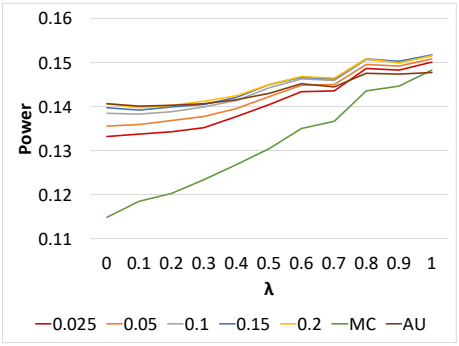
The simulation in Section 4.3.1.2 is repeated with one change, MC is now the Welch's t-test. The main null hypothesis is violated with three ncps. The results are shown below in Figure 11.

The power levels for the unconditional testing and the combined procedures are shown in Figure 11. The results are quite similar with the ones presented in Section 4.3.1.2. The same conclusions can be made. There seems to be one difference, the point where the powers of the MC and AU meet is now closer to $\lambda = 1$ in general. The Welch's t-test could be expected to be more powerful compared the standard t-test in cases where the variances or sample size are not equal as was concluded in Zimmerman (2004).

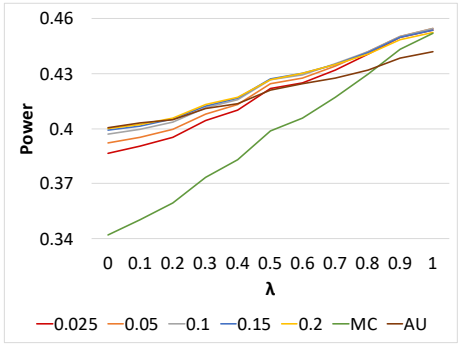
The power analysis is then repeated with the variances now being unequal i.e. the ratio of the variances of samples is 1.5 ($\frac{\sigma_2}{\sigma_1} = 1.5$). The powers of the unconditional MC and AU tests as well as the CP across λ are shown in Figure 12. An obvious difference can be seen here. The power of all the tests considered now have a downward trend as λ increases. There are exceptions to this pattern when the sample size and ncp is large. Then the MC test has a larger power when $\lambda = 1$ compared to when $\lambda = 0$. The CP test still has the highest power for most of the λ range compared to MC and AU.

4.3.3 *t-test versus Wilcoxon-Mann-Whitney with the uniform distribution*

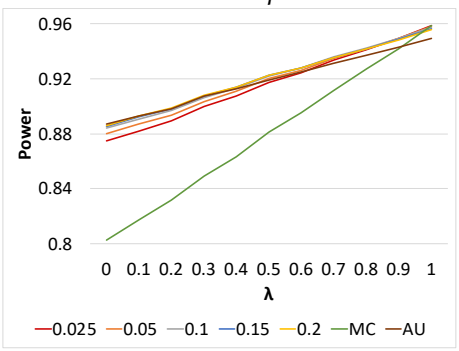
The simulations in the previous section are repeated. P_θ is the standard normal distribution and Q is the uniform distribution with minimum and maximum of $\pm\sqrt{3}$,



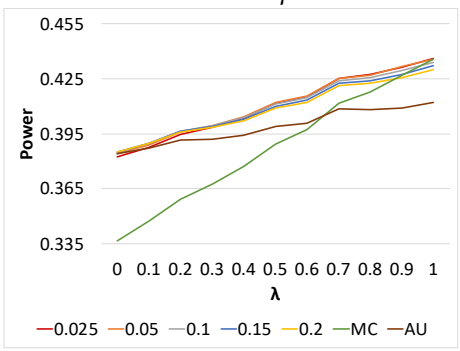
$n = 8$ and $ncp = 0.5$



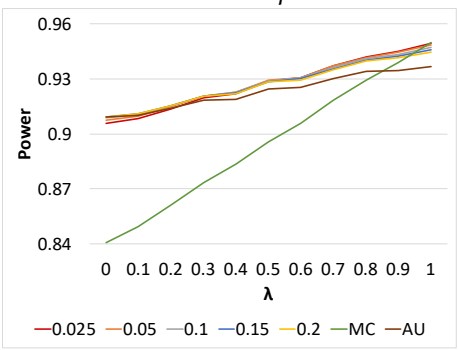
$n = 8$ and $ncp = 1$



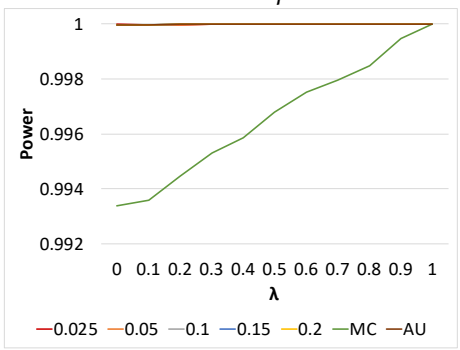
$n = 8$ and $ncp = 2$



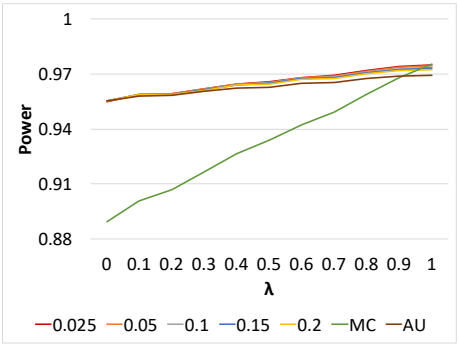
$n = 27$ and $ncp = 0.5$



$n = 27$ and $ncp = 1$



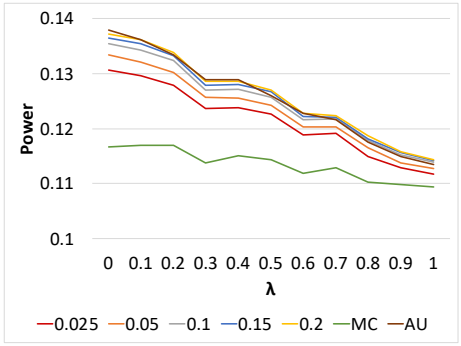
$n = 27$ and $ncp = 2$



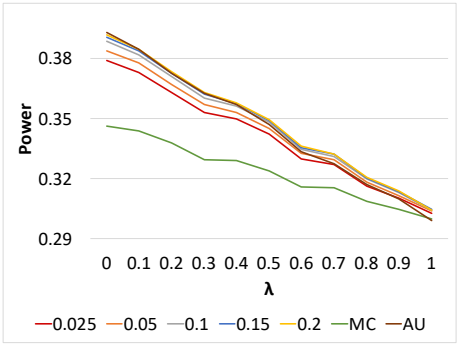
$n = 125$ and $ncp = 0.5$

Figure 11: Power levels for 5 different values of α_{MS}, the MC (Welch's t-test) and AU methods across different λ values, different sample sizes and different ncps. *Note: the axes and line labels are the same as Figures 2 - 10.*

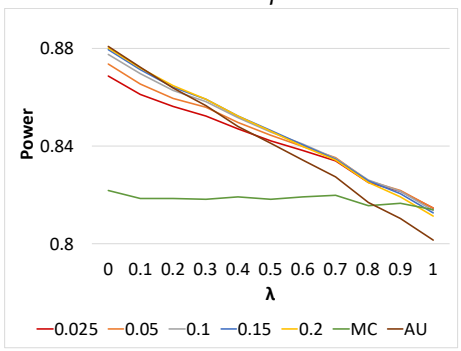
$Y \sim Unif(-\sqrt{3}, \sqrt{3})$. Choosing such minimum and maximum would make the Y be comparable with the standard normal distribution. The mean for Y is $\frac{a+b}{2} = \frac{-\sqrt{3}+\sqrt{3}}{2} = 0$ and the standard deviation for Y is $\frac{(b-a)^2}{12} = \frac{(\sqrt{3}-(-\sqrt{3}))^2}{12} = 1$.



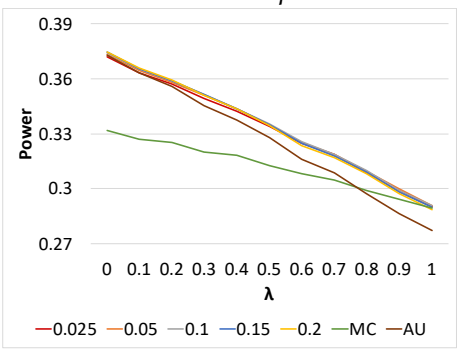
$n = 8$ and $ncp = 0.5$



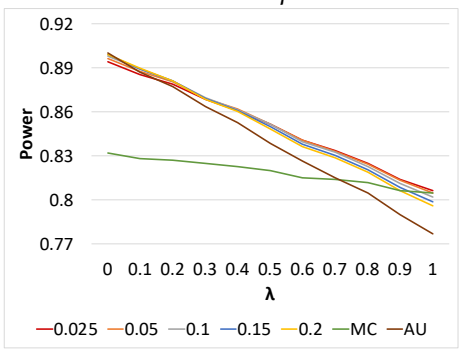
$n = 8$ and $ncp = 1$



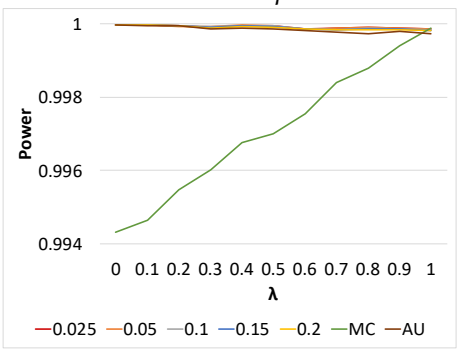
$n = 8$ and $ncp = 2$



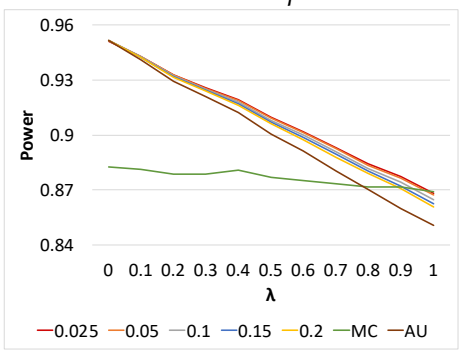
$n = 27$ and $ncp = 0.5$



$n = 27$ and $ncp = 1$



$n = 27$ and $ncp = 2$



$n = 125$ and $ncp = 0.5$

Figure 12: Power levels for 5 different values of α_{MS}, the MC (Welch's t-test) and AU methods across different λ values, different sample sizes, different ncps and ratio of samples' variances is $\sigma_2/\sigma_1 = 1.5$. *Note: the axes and line labels are the same as Figures 2 - 10.*

4.3.3.1 *Main null hypothesis is fulfilled*

Tables 7 - 9 shows the value of the rejection rates and their respective standard errors across different values of λ. Some of the rejection rates shown are significantly different

n	MC	AU	CP	CP_{adj}
8	5.087 (0.070)	4.918 (0.068)	5.212 (0.070)	5.191 (0.070)
27	4.939 (0.069)	4.670 (0.067)	4.881 (0.068)	4.913 (0.068)
125	5.030 (0.069)	4.985 (0.069)	5.010 (0.069)	5.004 (0.069)

Table 7: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes and $\lambda = 0.25$

n	MC	AU	CP	CP_{adj}
8	5.143 (0.070)	4.996 (0.069)	5.263 (0.071)	5.244 (0.071)
27	4.977 (0.069)	4.753 (0.067)	4.967 (0.069)	4.985 (0.069)
125	4.788 (0.068)	4.800 (0.068)	4.864 (0.069)	4.842 (0.068)

Table 8: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes and $\lambda = 0.5$

n	MC	AU	CP	CP_{adj}
8	5.088 (0.070)	4.975 (0.069)	5.260 (0.071)	5.223 (0.070)
27	4.982 (0.069)	4.701 (0.067)	5.017 (0.069)	5.020 (0.069)
125	4.966 (0.069)	4.943 (0.069)	5.058 (0.069)	5.029 (0.069)

Table 9: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes and $\lambda = 0.75$

than 5% according to the proportion test given in (12). These values are underlined in Tables 7 - 9. When λ is 0.25 or 0.75, the levels of MC, AU and CP are quite stable around 5%. However, when $\lambda = 0.5$, the levels of all thre procedure drops significantly below 5% with the exception of a few situations when $n = 8$ as shown in Table 8.

4.3.3.2 *Main null hypothesis is violated*

The simulation in Section 4.3.1.2 is repeated. The main null hypothesis is violated with three *ncp*s. The results are shown below in Figure 13.

The power levels for the unconditional testing and the combined procedures are shown in Figure 13. The results with the alternative distribution being the uniform distribution differs than what has been shown so far. In small sample sizes ($n = 8$) with small violation of the main null hypothesis ($ncp = 0.5$), it is clear that the CP has larger power across all λ. As the violations get larger ($ncp = 1$), the power of MC get higher than the CP on all α_{MS} levels except when $\alpha_{MS} = 0.025$. Otherwise the MC test without any MS testing has the higher power compared to both CP and AU. AU has the lowest power in all cases. This is similar to the findings in Rochon, Gondan and Kieser (2012). This clearly violates the normality assumption of the t-test (despite being asymptotically still correct), and will be picked up by many normality tests. Still it would be a bad decision to use the WMW test instead, even though its assumptions are fulfilled.

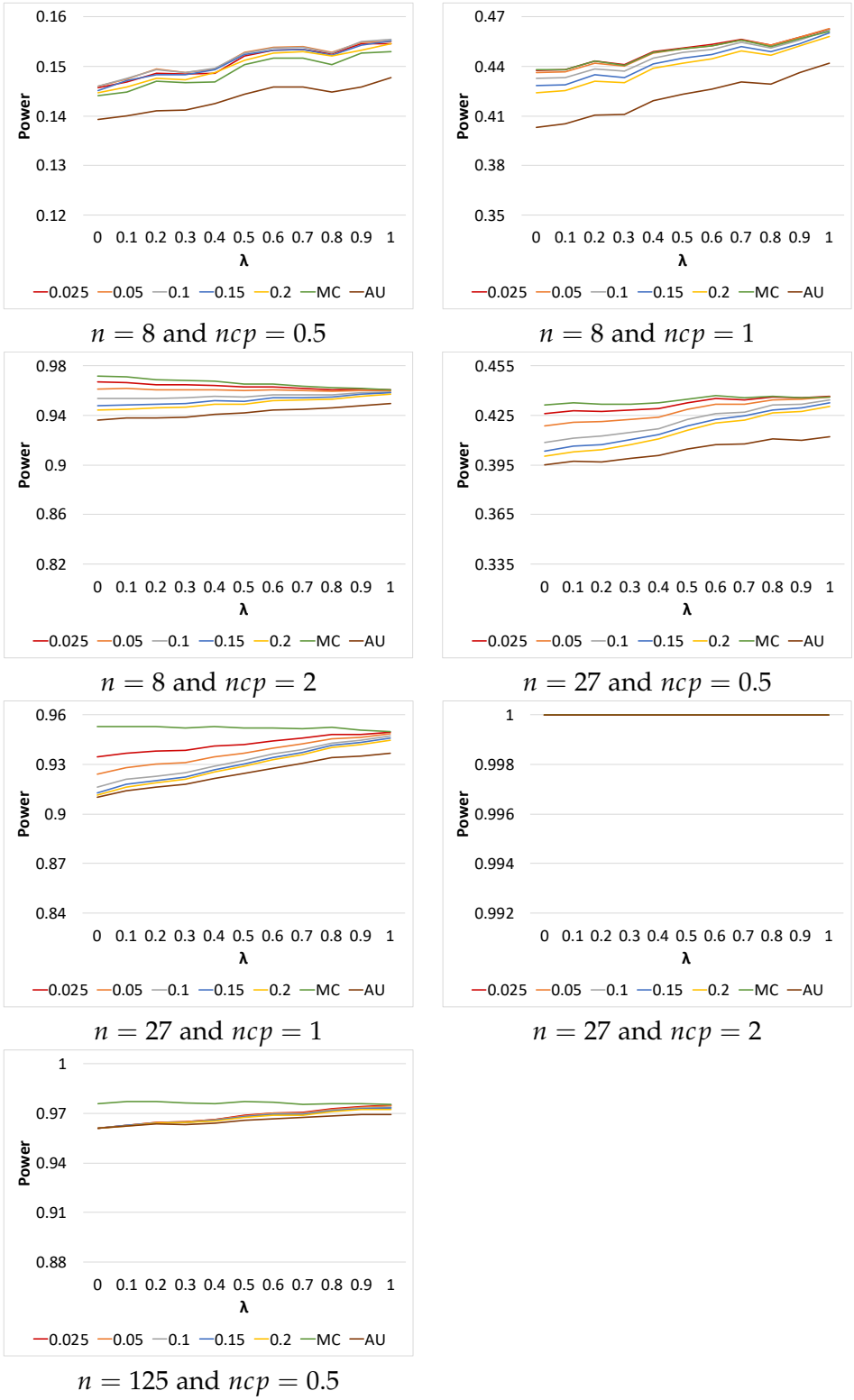


Figure 13: Power levels for 5 different values of α_{MS}, the MC (t-test) and AU methods across different λ values, different sample sizes and different ncps. *Note: the axes and line labels are the same as Figures 2 - 10.*

4.3.4 *t-test versus Wilcoxon-Mann-Whitney with skewed distributions*

The simulation was run again to investigate the effect of skewed distributions on the combined procedure. P_θ and Q are now generated from a skewed standard normal

distribution and skewed t distribution with 3 degrees of freedom respectively, both with a slant parameter of 0.5. The reason for this is to investigate the effect a small skewness has on all of the procedures. Figure 14 shows the density plot of the standard normal, t distribution with 3 degrees of freedom, skewed normal with slant parameter 0.5 and a skewed t distribution with 3 degrees of freedom and slant parameter (Azzalini (2013, p. 24)) 0.5 when $n = 10000$ and it shows that not much are separating these four distributions except that the two t distributions have heavier tails.

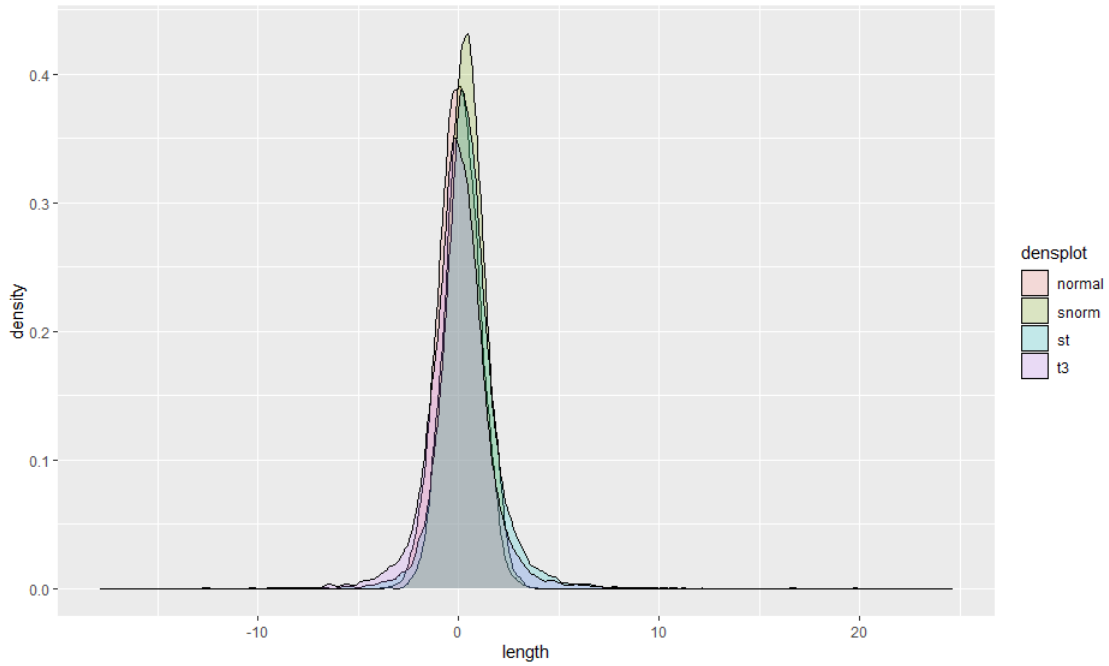


Figure 14: Density plot of the standard normal, t distribution with 3 degrees of freedom, skewed normal with slant parameter 0.5 and a skewed t distribution with 3 degrees of freedom and slant parameter 0.5 when $n = 10000$

4.3.4.1 *Main null hypothesis is fulfilled*

In this section the MC is the t-test and the AU is the WMW test. The simulations were run 100000 times and the rejection rates of the main null hypothesis is given below in Tables 10 - 12 for different values of λ.

n	MC	AU	CP	CP_{adj}
8	4.368 (0.065)	4.943 (0.069)	5.020 (0.069)	4.925 (0.068)
27	4.775 (0.067)	4.907 (0.068)	5.128 (0.070)	5.099 (0.070)
125	4.859 (0.068)	4.974 (0.069)	4.991 (0.069)	4.964 (0.069)

Table 10: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes, skewed distributions and $\lambda = 0.25$

The values that are underlined are the rejection rates that are statistically significantly different than the expected 5% according to the proportion test given in (12). For the

n	MC	AU	CP	CP_{adj}
8	4.602 (0.066)	4.992 (0.069)	5.079 (0.069)	5.017 (0.069)
27	4.830 (0.068)	4.833 (0.068)	5.060 (0.069)	5.068 (0.069)
125	4.854 (0.068)	4.943 (0.069)	5.048 (0.069)	5.037 (0.069)

Table 11: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes, skewed distributions and $\lambda = 0.5$

n	MC	AU	CP	CP_{adj}
8	4.757 (0.067)	4.973 (0.069)	5.106 (0.070)	5.027 (0.069)
27	4.978 (0.069)	4.879 (0.068)	5.164 (0.070)	5.122 (0.070)
125	4.874 (0.068)	4.843 (0.068)	4.936 (0.069)	4.953 (0.069)

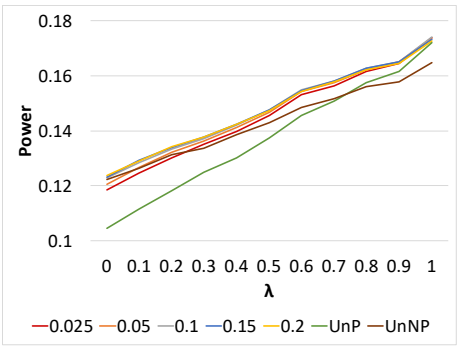
Table 12: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various samples sizes, skewed distributions and $\lambda = 0.75$

cases where a large proportion of the generated samples are from the skewed t distribution, the t-test has a low Type-I error, statistically significantly lower than 5%. The WMW seems to be much better in controlling the Type-I error in a setup with no MS test. However, the combined procedure does appear to a better job than the unconditional tests. The Type-I error is controlled well, with small standard errors.

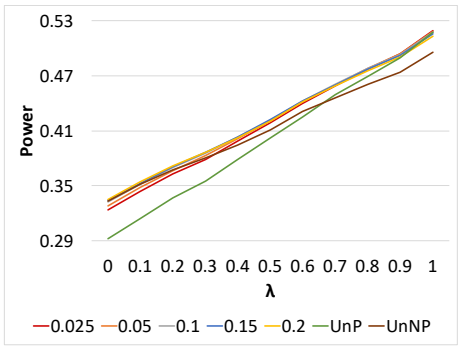
4.3.4.2 *Main null hypothesis is violated*

The simulation in Section 4.3.1.2 is repeated. MC is the standard t-test and the AU is the WMW test. The main null hypothesis is violated non-centrality parameters (ncps) representing a small departure, a medium departure and a large departure from the hypothesised mean. The results are shown in Figure 15.

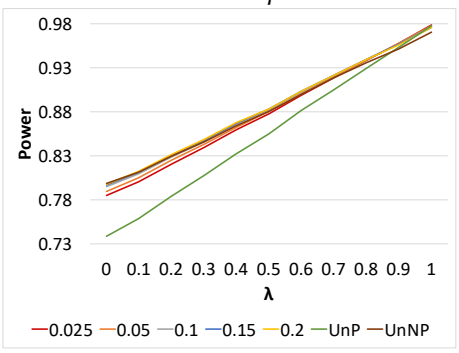
Comparing the results in Figure 15 with the ones in Section 4.3.1.2, we can see that in the case of $\lambda = 0$, the power of all three approaches in the case of skewed distributions is lesser than the case of the non-skewed distribution. The opposite is true when $\lambda = 1$, the power of all three approaches is larger than the case of non-skewed normal and t distribution. This is evident from the gradients of the powers compared to Figures 2 - 10. This could suggest that the WMW test is not very good at rejecting the main null hypothesis of equal distributions when the sample is generated from the skewed t distribution when the means were not equal. This is the same conclusion made by Rasch, Kubinger and Moder (2011). Interestingly, the standard t-test is quite good at rejecting the main null hypothesis when the samples are generated from the skewed normal distribution. This suggests the the t-test may be quite robust to a relatively small skew as is discussed in Micceri (1989) and Sawilowsky and Blair (1992). Generally, the same conclusions can be made as Section 4.3.1.2 & Section 4.3.2.2.



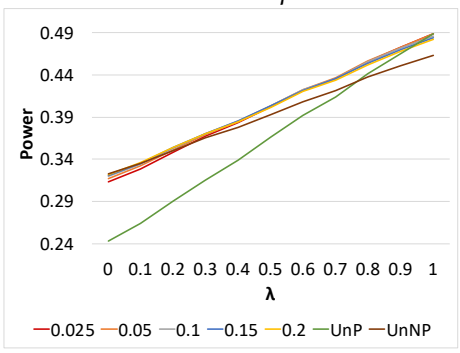
$n = 8$ and $ncp = 0.5$



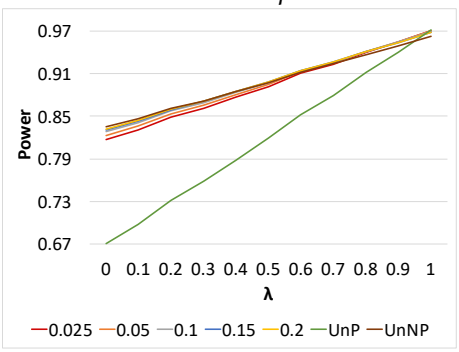
$n = 8$ and $ncp = 1$



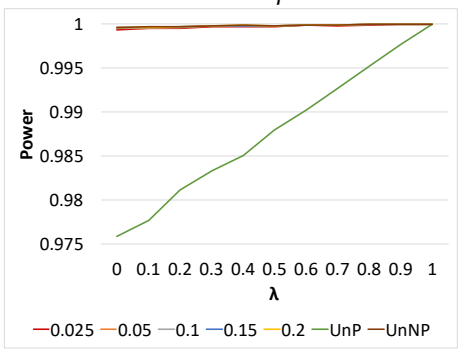
$n = 8$ and $ncp = 2$



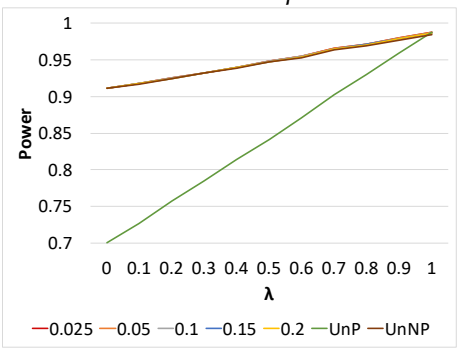
$n = 27$ and $ncp = 0.5$



$n = 27$ and $ncp = 1$



$n = 27$ and $ncp = 2$



$n = 125$ and $ncp = 0.5$

Figure 15: Power levels for 5 different values of α_{MS}, the MC (standard t-test) and AU (WMW test) methods across different λ values, different sample sizes and different ncps using skewed distributions

4.4 MS TEST LEVELS IN THE COMBINED PROCEDURE

Simulations in Section 4.3.1.2 are referred to. We are now interested to study how the MS test, namely the Shapiro-Wilk test in this particular simulation, performs. Figure 16 and Figure 17 show the rejection rates of the MS test across different levels of λ. The blue lines in both figures show the levels of the MS test as this is in the situation where the model assumption is fulfilled and the orange lines show the power of the MS test as this is in the situation where the model assumption is violated. The grey lines are the combined procedure's MS test levels. Figure 16 show the rejection rates when the MS test level is set at 5% or 0.05. The blue line shows an almost horizontal line across the axis at the rejection rate is 0.1. This indicates the MS test rejects more model assumption than it should going into the combined procedure.

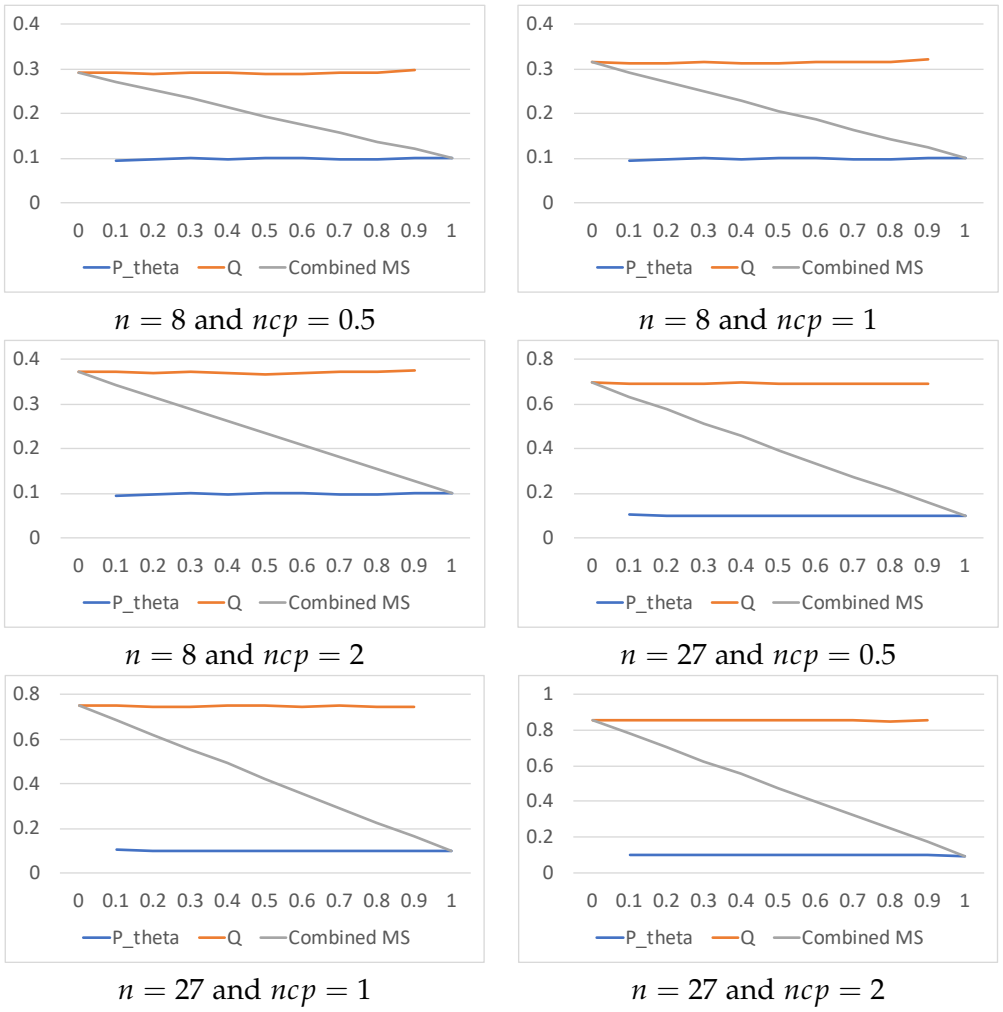


Figure 16: Rejection rates of the MS test at level 5%. The horizontal axis is λ, the vertical axis is the rejection rates. P_θ is the standard normal distribution, Q is the t_3 distribution, MC is the standard t-test and AU is the WMW test.

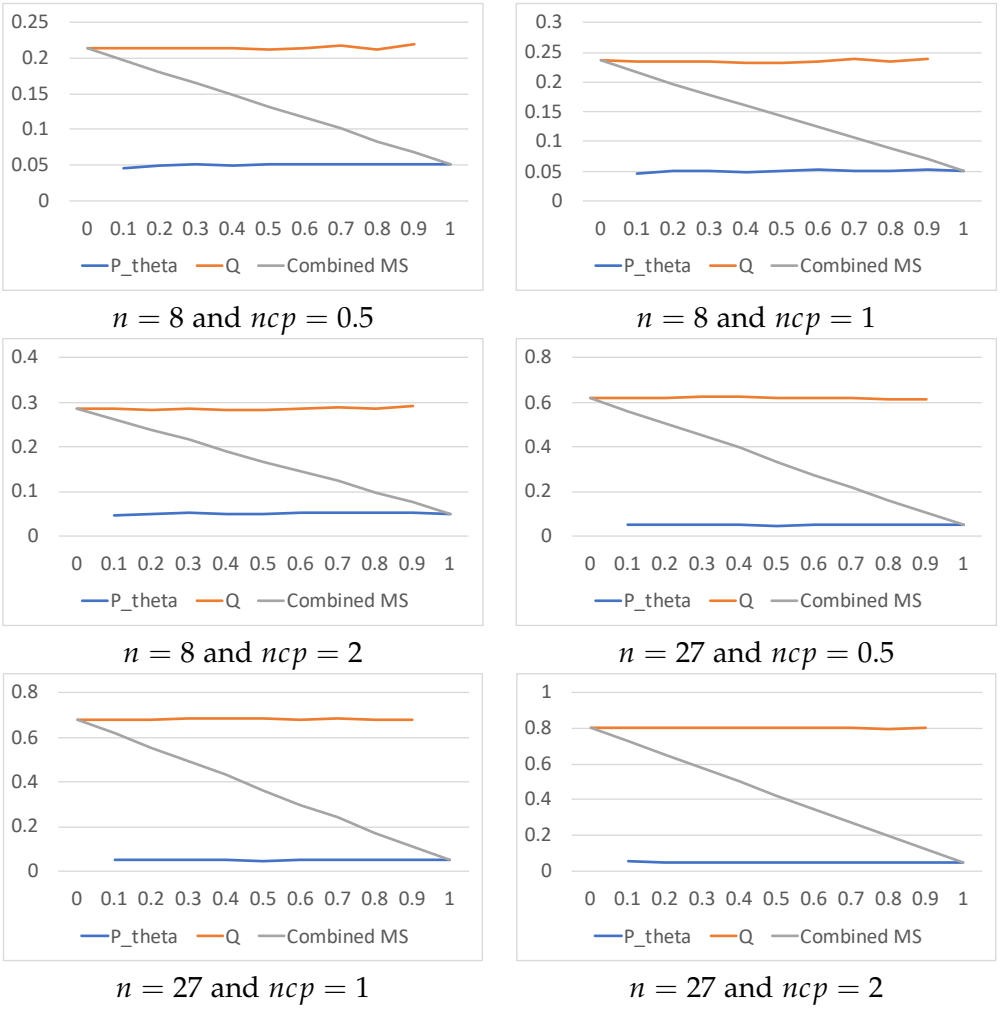


Figure 17: Rejection rates of the MS test at level 2.5%. The horizontal axis is λ, the vertical axis is the rejection rates. P_θ is the standard normal distribution, Q is the t_3 distribution, MC is the standard t-test and AU is the WMW test.

Figure 17 show the rejection rates when the MS test level is set at 2.5% or 0.025. This is because of the Bonferroni correction done due to the fact that we are testing two samples for the model assumption. Here, the blue line shows an almost horizontal line across the axis at the rejection rate is 0.05. This indicates that the Bonferroni correction is needed to control the level of the MS test at the nominal. Therefore it is recommended that the Bonferroni correction is used in the situation where the CP is carried out for a two-sample problem.

4.5 SIMULATIONS FOR VALUES OF δ_i AND τ

In this section we will look at some simulated values for δ_1, δ_2, δ_3, δ_4 and τ. The definitions of the δ_i (i = 1, 2, 3, 4) and τ has already been established in Chapter 3 but this section is mainly about looking at the simulated values of these variables.

Note that we are in the situation of the null hypothesis of the main test is violated, namely in the situation where the power is considered. The values of δ_i (i = 1, 2, 3, 4) and τ are taken from the simulations already done earlier in this chapter.

n, ncp	$\{\delta_1, \delta_2, \delta_3, \delta_4\}$	τ
8, 0.5	{0.01597, 0.01162, 0.02390, 0.00021}	0.00079
8, 1	{0.01953, 0.03465, 0.03869, -0.00002}	0.00283
8, 2	{0.00185, 0.06059, 0.00386, 0.00009}	0.00265
27, 0.5	{0.007508, 0.02850, 0.02758, -0.00018}	0.00935
27, 1	{0.00311, 0.05023, -0.00238, 0.00021}	0.00725
27, 2	{0, 0.00604, 0, 0}	0
125, 0.5	{0.00002, 0.05886, -0.00393, 0.00004}	0.00498
125, 1	{0, 0.00157, 0, 0}	0
125, 2	{0, 0.00014, 0, 0}	0

Table 13: δ_i (i = 1, 2, 3, 4) and τ values for $P_\theta \sim N(0,1)$, $Q \sim t_3$, MC is t-test, AU is WMW

n, ncp	$\{\delta_1, \delta_2, \delta_3, \delta_4\}$	τ
8, 0.5	{0.01597, 0.01200, 0.02390, 0.00020}	0.00012
8, 1	{0.01953, 0.03717, 0.03869, 0}	0.00186
8, 2	{0.00185, 0.07194, 0.00386, 0.00013}	0.00227
27, 0.5	{0.00751, 0.02984, 0.02758, -0.00016}	0.00937
27, 1	{0.00311, 0.05159, -0.00238, 0.00022}	0.00719
27, 2	{0, 0.00654, 0, 0}	0
125, 0.5	{0.00002, 0.05902, -0.00393, 0.00004}	0.00498
125, 1	{0, 0.00157, 0, 0}	0
125, 2	{0, 0.00014, 0, 0}	0

Table 14: δ_i (i = 1, 2, 3, 4) and τ values for $P_\theta \sim N(0,1)$, $Q \sim t_3$, MC is Welch's t-test, AU is WMW

n, ncp	$\{\delta_1, \delta_2, \delta_3, \delta_4\}$	τ
8, 0.5	{0.01330, 0.01059, 0.02989, -0.00002}	-0.00080
8, 1	{0.01865, 0.02754, 0.03773, 0.00024}	0.00015
8, 2	{-0.00184, 0.04907, 0.01682, 0.00028}	0.00218
27, 0.5	{0.00891, 0.02753, 0.02717, -0.00023}	0.00491
27, 1	{0.00177, 0.05278, 0.00727, -0.00011}	0.01128
27, 2	{0, 0.00560, -0.00026, 0.00001}	0.00010
125, 0.5	{-0.00006, 0.06716, -0.00485, 0.00016}	0.01302
125, 1	{0, 0.00191, 0.00004, 0}	0.00003
125, 2	{0, 0.00017, 0, 0}	0

Table 15: δ_i (i = 1, 2, 3, 4) and τ values for $P_{\theta,1} \sim N(0,1)$, $P_{\theta,2} \sim N(0,1.5)$, $Q_1 \sim t_3$, $Q_2 \sim t_4$, MC is Welch's t-test, AU is WMW

The values for δ_i (i = 1, 2, 3, 4) and τ given in Table 13 refers to the simulations done in Section 4.3.1.2. The values for δ_i (i = 1, 2, 3, 4) and τ given in Table 14 and Table 15 refers to the simulations done in Section 4.3.2.2. Table 15 differs from Table 14 only in

the sense that the ratio of variance of the two samples generated are not equal. Clearly, the definitions of δ_i (i = 1, 2, 3, 4) and τ given in the previous section are realistic. Note that τ = 0 when the difference between means (ncp) is large and the sample size is fairly large. The simulations were run M = 100, 000 times, the standard errors are very small, hence even a small value of τ is still statistically significant.

Assumption (V) provides for a $\delta > 0$ where $\delta = max\{|\delta_1|, |\delta_2|, |\delta_3|, |\delta_4|\}$. Also Lemma 2 will hold only if $\tau > 9\delta$. From the simulated values of δ_i (i = 1, 2, 3, 4) and τ there is not one situation where $\tau > 9\delta$ is confirmed. This could be due to the assumptions of Lemma 2 is still so strong that it does not apply in simulated simulations. The choice of a single δ to approximate all $\delta_1, \delta_2, \delta_3, \delta_4$ is not a very sharp approximation as some δ_is can be much smaller in absolute value than δ. $\tau > 9\delta$ is simple and of theoretical interest proving that weakening of the independence assumption is at least possible.

4.6 SIMULATIONS FOR VALUES OF THE CONDITIONAL PROBABILITIES

In this section we will look at some simulated values for $P_\theta(R_{AU}|R_{MS})$, $P_\theta(R_{MC}|R^c_{MS})$, $Q(R_{AU}|R_{MS})$ and $Q(R_{MC}|R^c_{MS})$ from Equation (3). Note again that we are in the situation of the null hypothesis of the main test is violated, namely in the situation where the power is considered. The values of $P_\theta(R_{AU}|R_{MS})$, $P_\theta(R_{MC}|R^c_{MS})$, $Q(R_{AU}|R_{MS})$ and $Q(R_{MC}|R^c_{MS})$ are taken from the simulations already done in the earlier sections.

λ	$P_\theta(R_{AU}\|R_{MS})$	$P_\theta(R_{MC}\|R^c_{MS})$	$Q(R_{AU}\|R_{MS})$	$Q(R_{MC}\|R^c_{MS})$
0	-	-	0.1567	0.1343
0.1	0.1506	0.1519	0.1512	0.1346
0.2	0.1669	0.1450	0.1526	0.1355
0.3	0.1694	0.1491	0.1535	0.1339
0.4	0.1702	0.1483	0.1519	0.1360
0.5	0.1730	0.1517	0.1581	0.1343
0.6	0.1814	0.1535	0.1535	0.1323
0.7	0.1797	0.1508	0.1527	0.1353
0.8	0.1731	0.1554	0.1545	0.1379
0.9	0.1832	0.1528	0.1539	0.1310
1	0.1716	0.1532	-	-

Table 16: Conditional probabilities for different values of λ when $n = 8$, $ncp = 0.5$ with $P_\theta \sim N(0,1)$, $Q \sim t_3$, MC is t-test and AU is WMW

The conditional power values presented in Table 16 are extracted from the simulations in Figure 2 which corresponds to Table 32 in the Appendix. The conditional power values presented in Table 17 are extracted from the simulations in Figure 5 which corresponds to Table 35 in the Appendix. To recapitulate the notations used, $P_\theta(R_{AU}|R_{MS})$ means the power of the AU test to reject the main null hypothesis given that the MS

λ	$P_\theta(R_{AU}\|R_{MS})$	$P_\theta(R_{MC}\|R_{MS}^c)$	$Q(R_{AU}\|R_{MS})$	$Q(R_{MC}\|R_{MS}^c)$
0	-	-	0.3918	0.3677
0.1	0.4544	0.4471	0.3922	0.3631
0.2	0.4361	0.4375	0.3945	0.3674
0.3	0.4413	0.4362	0.3907	0.3687
0.4	0.4441	0.4383	0.3897	0.3636
0.5	0.4490	0.4416	0.3900	0.3660
0.6	0.4488	0.4354	0.3915	0.3689
0.7	0.4487	0.4439	0.3901	0.3658
0.8	0.4378	0.4379	0.3929	0.3691
0.9	0.4418	0.4381	0.3845	0.3580
1	0.4396	0.4358	-	-

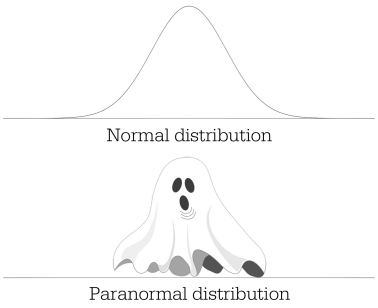
Table 17: Conditional probabilities for different values of λ when $n = 27$, $ncp = 0.5$ with $P_\theta \sim N(0,1)$, $Q \sim t_3$, MC is t-test and AU is WMW

test rejects the model assumption in the situation where the model assumption is not violated. $P_\theta(R_{MC}|R_{MS}^c)$ means the power of the MC test to reject the main null hypothesis given that the MS test does not reject the model assumption in the situation where the model assumption is not violated. $Q(R_{AU}|R_{MS})$ means the power of the AU test to reject the main null hypothesis given that the MS test rejects the model assumption in the situation where the model assumption is violated. $Q(R_{MC}|R_{MS}^c)$ means the power of the MC test to reject the main null hypothesis given that the MS test does not reject the model assumption in the situation where the model assumption is violated.

As λ changes, the powers of the conditional tests stay stable around the power values of the combined procedure given in Table 32 and Table 35 in the Appendix. This concurs with Equation (3) where the power of the combined procedure is a linear combination of the conditional probabilities weighted on λ and $(1-\lambda)$. One observation worth noting is that the values of $P_\theta(R_{AU}|R_{MS})$ are consistently larger than all the other conditional powers in both the selected situations presented here in this section.

5

SIMULATIONS FOR THREE-STAGE MS TESTING PROCEDURE



In Chapter 4, a two stage procedure was simulated. This represents most of the simulation done in the literature that was reviewed. In this chapter, the three stage testing procedure is considered. Two model assumptions are checked before finally deciding on the model to be used. The main null hypotheses of the tests to choose from must be comparable, namely the main null hypotheses of all the tests considered must essentially test the same thing. For example, testing the equality of distributions of a sample or multiple samples.

From the literature, there is not much that has been done on the three stage MS testing procedure. One example is from Rasch, Kubinger and Moder (2011) which assessed the statistical properties of a three-stage procedure including testing for normality and for homogeneity of the variances. This is also recommended by Kim (2015). In most cases there are more than one model assumption, a two stage MS test might not be sufficient to check them. Spanos (2018) recommends that all model assumptions are listed and to test them all in some specific order. However, he does not precisely define a sequence on how to do this. In this chapter, we do not implement Spanos' recommendation completely, but we do take one step further in the direction of investigating an overall model checking process by defining a three stage procedure checking two model assumptions.

5.1 MAIN NULL HYPOTHESIS OF EQUAL DISTRIBUTIONS

The work in this section is similar to the work done by Rasch, Kubinger and Moder (2011) where two model assumptions are tested to specify the model to be used. The main null hypothesis is that two samples are equal. Two model assumptions are tested that is the assumption of equal variance and the assumption of normality. These two model assumptions will specify between three hypothesis tests which is the Wilcoxon-Mann-Whitney test (WMW), the Welch's t-test and the standard t-test. The Welch's t-test and the standard t-test does assume normality but the Welch's t-test does not assume equal variance while the standard t-test does assume equal variance. The WMW is a non-parametric test with lesser assumptions. The null hypothesis is that the two distributions are equal, so must be the variances. The alternative is that one distribution is stochastically larger than the other. Whereas this does not necessarily imply that the variances are equal, this does for example not hold for two normal distributions with different means and different variances.

The simulation starts by generating a multinomial Bernoulli vector (l_1, l_2, l_3) with probability $(\lambda_1, \lambda_2, \lambda_3)$ where $\sum_{i=1}^{3} \lambda_i = 1$. When $l_1 = 1$, $l_2 = 0$, $l_3 = 0$, two samples of the same size are generated from the t distribution with 3 degrees of freedom. When $l_1 = 0$, $l_2 = 1$, $l_3 = 0$, two samples of the same size are generated from the normal distribution with mean 0 and different variance, $\sigma_1 = 1$, $\sigma_2 = 1.5$. When $l_1 = 0$, $l_2 = 0$, $l_3 = 1$, two samples are generated from the standard normal distribution. The sample sizes considered are $n = 8, 27, 125$ representing a small sample size, a moderate sample size and a fairly large sample size respectively. The two samples are then tested with all three tests considered without any model checking. Next, the two samples are put through the combined procedure with model checking. If normality is rejected, the WMW test is used. If normality is not rejected, the two samples are tested for equal variance. If normality is not rejected and equal variance is rejected, the Welch's t-test is used and finally if normality and equal variance is not rejected, the standard t-test is used. This process is summarised in Figure 18. The simulations are repeated $M = 100,000$ times.

5.1.1 *Main null hypothesis is fulfilled*

Table 18 to Table 20 shows the rejection rates of the WMW test, Welch's t-test, the standard t-test and the combined, three stage Combined Procedure (CP). The values in the tables show that level of the tests and the CP are around the nominal level 5%. However, using the proportion test from Equation (12), quite a number of the values are

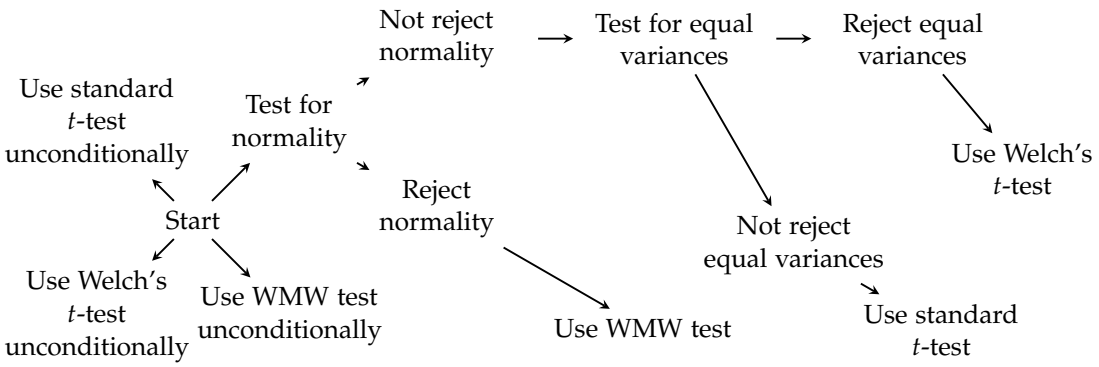


Figure 18: The decision tree for the unconditional testing and the three stage MS combined procedure for the main null hypothesis of equal distributions

statistically significantly different than 5%. The rejection rates where the null hypothesis that the levels are equal to 5% that are rejected by the proportion test are underlined in Table 18 to Table 20. The 95% confidence interval for the test is $(4.865, 5.135)$.

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	5.348 (0.071)	4.927 (0.068)	5.272 (0.071)	5.182 (0.070)
(0, 0, 1)	4.965 (0.069)	4.753 (0.067)	5.004 (0.069)	4.966 (0.069)
(1, 0, 0)	4.967 (0.069)	3.844 (0.061)	4.188 (0.063)	4.195 (0.063)
(0, 0.5, 0.5)	5.137 (0.070)	4.800 (0.068)	5.060 (0.069)	4.988 (0.069)
(0.5, 0.5, 0)	5.202 (0.070)	4.503 (0.066)	4.822 (0.068)	4.803 (0.068)
(0.5, 0, 0.5)	5.036 (0.069)	4.409 (0.065)	4.682 (0.067)	4.684 (0.067)
(1/3, 1/3, 1/3)	5.213 (0.070)	4.615 (0.066)	4.915 (0.068)	4.883 (0.068)

Table 18: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 8$

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	5.116 (0.070)	4.990 (0.069)	5.078 (0.069)	5.021 (0.069)
(0, 0, 1)	4.872 (0.068)	5.015 (0.069)	5.035 (0.069)	5.038 (0.069)
(1, 0, 0)	4.793 (0.068)	4.530 (0.066)	4.588 (0.066)	4.725 (0.067)
(0, 0.5, 0.5)	4.988 (0.069)	5.019 (0.069)	5.061 (0.069)	5.054 (0.069)
(0.5, 0.5, 0)	4.881 (0.068)	4.692 (0.067)	4.768 (0.067)	4.826 (0.068)
(0.5, 0, 0.5)	4.736 (0.067)	4.685 (0.067)	4.735 (0.067)	4.810 (0.068)
(1/3, 1/3, 1/3)	4.894 (0.068)	4.789 (0.068)	4.832 (0.068)	4.893 (0.068)

Table 19: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 27$

The CP procedure does seem to be the best at maintaining a stable level but the WMW and Welch's t-test are not that far behind. The Welch's t-test did have one instance where the level was at 3.844% in Table 18. When n = 8, the Welch's t-test constantly underperforms at level 5%. When n = 27, the WMW almost always have a smaller level than 5% but not statistically significantly so. Having said that, a smaller level is not a problem in itself, this means the error probability is lower. However, a low level can also

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	5.443 (0.072)	5.132 (0.070)	5.153 (0.070)	5.141 (0.070)
(0, 0, 1)	5.048 (0.069)	5.124 (0.070)	5.125 (0.070)	5.129 (0.070)
(1, 0, 0)	4.871 (0.068)	4.763 (0.067)	4.770 (0.067)	4.861 (0.068)
(0, 0.5, 0.5)	5.240 (0.070)	5.100 (0.070)	5.105 (0.070)	5.104 (0.070)
(0.5, 0.5, 0)	5.213 (0.070)	4.884 (0.068)	4.895 (0.068)	5.001 (0.069)
(0.5, 0, 0.5)	5.012 (0.069)	4.944 (0.069)	4.949 (0.069)	5.063 (0.069)
(1/3, 1/3, 1/3)	5.123 (0.070)	4.968 (0.069)	4.979 (0.069)	5.043 (0.069)

Table 20: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 125$

be caused by a power that is also low. A closer look at the power of the test is needed. Generally, the performances of the AU (WMW and Welch's t-test), MC (standard t-test) and CP are quite similar and no clear advantage can be seen looking at the levels alone.

5.1.2 *Main null hypothesis is violated*

Table 21 to Table 25 shows the power levels of the WMW test, Welch's t-test, the standard t-test and the three stage CP procedure for three different sample sizes and two ncps. When the sample size is small ($n = 8$), the WMW and the Welch's test perform slightly worse than the standard t-test and the CP procedure. In some situations, it is significantly worse, in others not significantly worse. Remember that because the simulations were done 100, 000 times, even a small difference is quite significant. The powers of the CP is quite close to that of the standard t-test. When $n = 27$, the powers of the WMW increases and in some cases, for example when $(\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = 0)$, the WMW has quite the advantage over the other three methods but this is due to the samples being generated in the model where WMW was chosen as the test.

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	11.350	10.942	11.591	11.405
(0, 0, 1)	14.774	14.837	15.303	15.209
(1, 0, 0)	14.074	11.485	12.263	12.373
(0, 0.5, 0.5)	13.166	13.046	13.646	13.497
(0.5, 0.5, 0)	12.606	11.206	11.903	11.842
(0.5, 0, 0.5)	14.363	13.155	13.801	13.784
(1/3, 1/3, 1/3)	13.086	12.476	13.161	13.086

Table 21: Powers of the rejection of a false main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 8$ and $ncp = 0.5$

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	29.910	29.996	31.350	30.900
(0, 0, 1)	44.183	45.188	46.094	45.912
(1, 0, 0)	40.070	34.215	35.965	36.146
(0, 0.5, 0.5)	36.929	37.578	35.965	38.367
(0.5, 0.5, 0)	34.685	31.722	33.299	33.173
(0.5, 0, 0.5)	41.963	39.415	40.750	40.766
(1/3, 1/3, 1/3)	38.067	36.418	37.819	37.666

Table 22: Powers of the rejection of a false main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 8$ and $ncp = 1$

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	27.716	28.946	29.218	29.035
(0, 0, 1)	41.206	43.546	43.602	43.592
(1, 0, 0)	38.427	33.678	33.920	35.896
(0, 0.5, 0.5)	34.315	36.180	36.341	36.257
(0.5, 0.5, 0)	33.169	31.418	31.673	32.570
(0.5, 0, 0.5)	39.922	38.774	38.919	39.911
(1/3, 1/3, 1/3)	35.856	35.292	35.494	36.116

Table 23: Powers of the rejection of a false main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 27$ and $ncp = 0.5$

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	77.694	80.498	80.704	80.554
(0, 0, 1)	93.659	94.659	94.997	94.997
(1, 0, 0)	90.926	84.078	84.282	84.423
(0, 0.5, 0.5)	85.831	87.721	87.844	87.753
(0.5, 0.5, 0)	84.474	82.283	82.476	83.980
(0.5, 0, 0.5)	92.288	89.466	89.561	91.142
(1/3, 1/3, 1/3)	87.577	86.457	86.621	87.626

Table 24: Powers of the rejection of a false main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 27$ and $ncp = 1$

$(\lambda_1, \lambda_2, \lambda_3)$	WMW	Welch's t-test	Standard t-test	CP
(0, 1, 0)	85.080	86.930	86.956	86.920
(0, 0, 1)	96.927	97.531	97.531	97.531
(1, 0, 0)	95.516	88.953	88.969	95.089
(0, 0.5, 0.5)	90.945	92.228	92.243	92.223
(0.5, 0.5, 0)	90.412	88.114	88.135	91.149
(0.5, 0, 0.5)	96.363	93.382	93.388	96.421
(1/3, 1/3, 1/3)	92.505	91.203	91.215	93.178

Table 25: Powers of the rejection of a false main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 125$ and $ncp = 0.5$

Generally, from the rejection rates in Table 18 to Table 25, it is difficult to choose a best method. However, a good balance between level and power across λs and sample sizes is the CP. The levels of the CP is well managed at around the nominal level 5% and the power is among the highest in most of the ratios of λs that was considered. Second to the CP, we would recommend using the WMW without any model checking.

5.2 MAIN NULL HYPOTHESIS OF REGRESSION SLOPE COEFFICIENT SIGNIFICANCE

The work in this section is similar to the setup presented in Mayo and Spanos (2004) where they looked at a probabilistic reduction approach to model checking in linear regression. Let's consider a simple linear regression model

$$y_t = \beta_0 + \beta_1 x_t + u_t,\ t = 1, 2, \ldots, n \tag{13}$$

where y_t is the dependent variable, x_t the independent variable, β_0 the intercept, β_1 the slope coefficient and u_t is the error component.

In Mayo and Spanos (2004), five model assumptions were considered namely Normality, Linearity, Homoskedasticity, Independence and t-homogeneity. According to Spanos (1999), t-homogeneity is short for time homogeneity, which is an assumption that the transition probabilities do not change over time. In this section, these five assumptions are reduced to two, namely Normality and Independence. The main null hypothesis that is of interest here is whether there is a significant relationship between the independent and dependent variables. This is achieved by testing the null hypothesis that the slope coefficient is equal to zero (β_1 = 0) as introduced in Scetion 2.3.5. The testing of the residuals for these two model assumptions will choose between three linear models which is either the AutoRegressive (AR(1)) model, a robust regression method or the standard linear regression model. The AR(1) model is used when the assumption of independence is rejected, as this could imply that the error terms are dependent or some autocorrelation is present. The robust regression method is used when the independence assumption is not rejected but the assumption of normality is rejected. Finally, the standard linear regression is used in the case where both model assumptions are not rejected. To the best of our knowledge, this kind of a three stage combined procedure has not been investigated, therefore the choice of model to use was purely imagined to be a good representation of a real world problem.

The autoregressive model is widely used in areas of statistics such as econometrics and signal processing as a representation of a type of random process specifically time-varying process. Simply put, the autoregressive model specifies that the output variable

depends linearly on its own previous values. The AR(1) means the first-order autoregressive model of a time series Y_t which is defined as follows;

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t, \tag{14}$$

where t is time and u_t is noise at time t.

In robust statistics, robust regression is a form of regression analysis designed to overcome some limitations of traditional parametric and non-parametric methods. Robust regression provides an alternative to least squares regression that works with less restrictive assumptions. Specifically, it provides much better regression coefficient estimates when outliers are present in the data. Outliers violate the assumption of normally distributed residuals in least squares regression. They tend to distort the least squares coefficients by having more influence than they deserve. The robust method that was chosen for this section is by Koller and Stahel (2011) and Koller and Stahel (2017). This is an MM-estimator that is recommended in the R documentation of the `lmrob` function. An example of a standard linear regression method was already discussed in Section 2.3.5.

The simulation starts by generating a multinomial Bernoulli vector (l_1, l_2, l_3) with probability $(\lambda_1, \lambda_2, \lambda_3)$ where $\sum_{i=1}^{3} \lambda_i = 1$. When $l_1 = 1$, $l_2 = 0$, $l_3 = 0$, a residuals vector is generated from an AR(1) process. This is done by first generating a sample from a standard normal distribution as residuals. Autocorrelation is then added to the residuals using the function `filter` in R. The recursive linear filter was chosen to introduce autocorrelation to the error terms generated. Let the filter coefficient be f_1 and the residuals generated from the standard normal distribution is ε. Note that there is an implied coefficient 1 at lag 0 in the recursive filter which is

$$\varepsilon_{filter}[i] = \varepsilon[i] + f_1 \times \varepsilon[i-1].$$

We chose $f_1 = 0.05$.

When $l_1 = 0$, $l_2 = 1$, $l_3 = 0$, a residuals vector is generated from a t distribution with 3 degrees of freedom. We choose this distribution because of the heavy tails. When $l_1 = 0$, $l_2 = 0$, $l_3 = 1$, a residuals vector are generated from the standard normal distribution. The sample sizes considered are $n = 16, 27, 125$ representing a small sample size, a moderate sample size and a fairly large sample size respectively. According to White (2019), the minimum time needed to achieve a reasonable power is 15.91. Therefore, the sample size selection was modified and the smallest sample size that was considered is 16.

A linear model is built using the residuals that were generated. The three methods were then fitted without any MS testing and the significance of the slope coefficient was tested using the standard t-test. For the combined procedure, the residuals are first tested for autocorrelation using the Durbin-Watson (DW) test. The DW test is a widely used test for independence in a statistical regression analysis. In particular, the error term in the standard linear regression model is extended to allow for the possibility that the errors are correlated with their past. Referring to Equation (13), the term u_t now becomes $u_t = \rho u_{t-1} + \epsilon_i$. The DW test assesses whether or not $\rho = 0$, namely

$$H_0 : \rho = 0 \qquad \text{against} \qquad H_1 : \rho \neq 0.$$

The DW test statistic (Durbin and Watson (1971)) is given as;

$$DW = \frac{\sum_{t=2}^{T}(u_t - u_{t-1})^2}{\sum_{t=1}^{T} u_t^2}.$$

The DW statistic will always have a value between 0 and 4. A value of 2.0 means that there is no autocorrelation detected in the sample. If autocorrelation or dependence is not rejected, the AR(1) time series model is used and the slope coefficient is tested for significance using the t-test. If autocorrelation or dependence is rejected, the residuals are tested for normality. If both autocorrelation and normality is rejected, the robust regression model is used and the slope coefficient is tested for significance using the t-test. If autocorrelation is rejected and normality is not rejected, the the standard linear regression model is used and the slope coefficient is tested for significance using the t-test. This process is summarised in Figure 19. The level for both MS tests and also the main tests are set at 0.05. The simulations are repeated M = 10, 000 times when n = 16, 27 and repeated M = 1, 000 times when n = 125 due to the limitation of the processing power and time and computer memory.

5.2.1 *Main null hypothesis is fulfilled*

Table 26 to Table 28 shows the rejection rates of the AR(1) model, robust regression, the standard linear regression and the three stage Combined Procedure (CP). The values in the tables show that level of the tests and the CP are around the nominal level 5% except when using the AR(1) model.

For Table 26 and 27 the 95% confidence interval for the proportion test (Equation (12)) is (4.57, 5.43) and for Table 28 the 95% confidence interval for the proportion test is (3.7, 6.3). Values outside of this interval are considered rejected for the null hypothesis

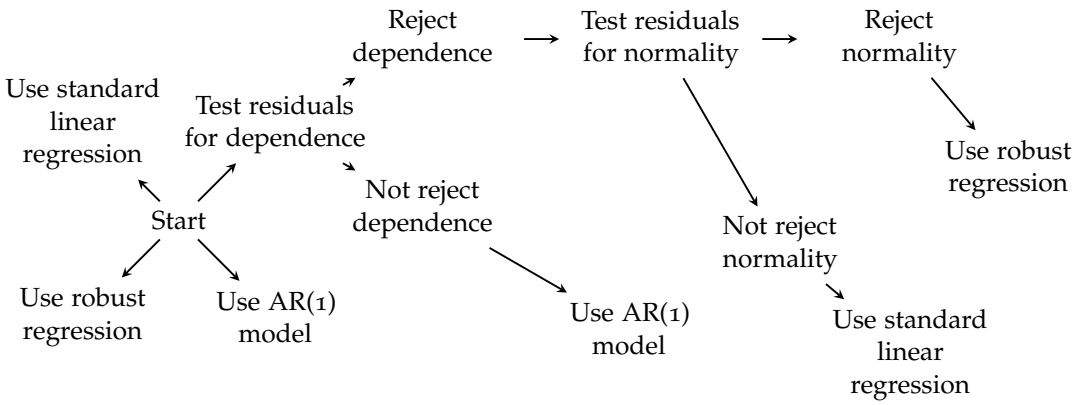


Figure 19: The decision tree for the unconditional testing and the three stage MS combined procedure for the main null hypothesis of regression coefficient significance

that the level is equal to 5% and are underlined. The respective standard errors of the rejection rates are presented in the parentheses. The CP does seem to be the best at maintaining a stable level when the sample sizes are 16 and 27 but when the sample size is 125, the level can drop considerably especially when $\lambda_2 = \lambda_3 = 0.5$. Note that the standard errors for the rejection rates in Table 28 are quite large due to the relatively small number of iterations, so these rejection rates values can fluctuate quite significantly.

$(\lambda_1, \lambda_2, \lambda_3)$	AR(1)	Robust regression	Linear regression	CP
(0, 1, 0)	7.47 (0.263)	5.07 (0.219)	4.77 (0.213)	4.66 (0.211)
(0, 0, 1)	6.99 (0.255)	5.23 (0.223)	4.94 (0.217)	4.86 (0.215)
(1, 0, 0)	7.22 (0.259)	6.05 (0.238)	5.89 (0.235)	5.71 (0.232)
(0, 0.5, 0.5)	6.91 (0.254)	4.97 (0.217)	4.69 (0.211)	4.71 (0.212)
(0.5, 0.5, 0)	7.08 (0.256)	5.35 (0.225)	5.59 (0.230)	5.28 (0.224)
(0.5, 0, 0.5)	7.30 (0.260)	5.35 (0.225)	5.34 (0.225)	5.06 (0.219)
(1/3, 1/3, 1/3)	7.23 (0.259)	5.64 (0.231)	5.55 (0.229)	5.27 (0.223)

Table 26: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 16$

$(\lambda_1, \lambda_2, \lambda_3)$	AR(1)	Robust regression	Linear regression	CP
(0, 1, 0)	6.19 (0.241)	4.64 (0.210)	4.89 (0.216)	4.18 (0.200)
(0, 0, 1)	6.32 (0.243)	5.14 (0.221)	5.02 (0.218)	4.70 (0.212)
(1, 0, 0)	6.64 (0.249)	6.11 (0.240)	6.07 (0.239)	5.59 (0.230)
(0, 0.5, 0.5)	6.29 (0.243)	5.27 (0.223)	5.02 (0.218)	4.73 (0.212)
(0.5, 0.5, 0)	6.11 (0.240)	5.65 (0.231)	5.33 (0.225)	4.73 (0.212)
(0.5, 0, 0.5)	6.88 (0.252)	5.65 (0.231)	5.78 (0.233)	5.35 (0.225)
(1/3, 1/3, 1/3)	6.29 (0.243)	5.70 (0.232)	5.57 (0.229)	5.37 (0.225)

Table 27: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 27$

$(\lambda_1, \lambda_2, \lambda_3)$	AR(1)	Robust regression	Linear regression	CP
(0, 1, 0)	5.6 (0.727)	6.0 (0.751)	5.5 (0.721)	3.6 (0.589)
(0, 0, 1)	4.0 (0.620)	4.4 (0.649)	4.5 (0.656)	3.7 (0.600)
(1, 0, 0)	4.0 (0.620)	6.1 (0.757)	5.6 (0.727)	4.3 (0.641)
(0, 0.5, 0.5)	4.3 (0.641)	4.9 (0.683)	4.4 (0.649)	2.9 (0.531)
(0.5, 0.5, 0)	5.2 (0.702)	5.6 (0.780)	6.0 (0.751)	4.2 (0.634)
(0.5, 0, 0.5)	5.7 (0.733)	6.6 (0.785)	5.9 (0.745)	5.5 (0.721)
(1/3, 1/3, 1/3)	6.5 (0.780)	6.8 (0.796)	6.9 (0.801)	5.3 (0.708)

Table 28: Rejection rates of the main null hypothesis (%) and standard errors (in parentheses)(%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 125$

5.2.2 *Main null hypothesis is violated*

The simulation in the previous section is repeated but now the main null hypothesis is violated namely $\beta_1 \neq 0$. We chose $\beta_1 = 0.1$. Table 29 to Table 31 show the powers of the three unconditional tests and the combined procedure when the main null hypothesis is violated. The powers of the CP are not particularly high, only in one situation the power of the CP is higher than other unconditional methods. When n = 27, the unconditional AR(1) has the lowest power across all λ values considered. When the levels and powers are considered together, it is difficult to choose one best method. Nevertheless, there is no evidence to conclude that the CP is much worse than the unconditional tests. Therefore, the CP can be favoured to be used in this situation as the Type I error is quite well controlled in small to moderate sample sizes.

$(\lambda_1, \lambda_2, \lambda_3)$	AR(1)	Robust regression	Linear regression	CP
(0, 1, 0)	24.90	28.08	23.81	27.06
(0, 0, 1)	40.70	38.45	40.88	39.50
(1, 0, 0)	38.86	39.19	41.51	39.61
(0, 0.5, 0.5)	32.36	32.43	31.61	32.87
(0.5, 0.5, 0)	30.70	34.26	32.91	33.39
(0.5, 0, 0.5)	39.20	38.18	40.63	39.13
(1/3, 1/3, 1/3)	34.56	34.87	35.11	35.09

Table 29: Powers of the rejection of the main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 16$

$(\lambda_1, \lambda_2, \lambda_3)$	AR(1)	Robust regression	Linear regression	CP
(0, 1, 0)	68.76	84.27	70.24	80.02
(0, 0, 1)	95.36	96.75	97.46	96.36
(1, 0, 0)	93.61	96.12	96.87	95.06
(0, 0.5, 0.5)	81.86	90.19	84.16	88.01
(0.5, 0.5, 0)	81.36	90.18	83.82	87.89
(0.5, 0, 0.5)	94.68	96.16	97.02	95.75
(1/3, 1/3, 1/3)	86.15	92.35	88.61	90.58

Table 30: Powers of the rejection of the main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 27$

$(\lambda_1, \lambda_2, \lambda_3)$	AR(1)	Robust regression	Linear regression	CP
(0, 1, 0)	1	1	1	1
(0, 0, 1)	1	1	1	1
(1, 0, 0)	1	1	1	1
(0, 0.5, 0.5)	1	1	1	1
(0.5, 0.5, 0)	1	1	1	1
(0.5, 0, 0.5)	1	1	1	1
(1/3, 1/3, 1/3)	1	1	1	1

Table 31: Powers of the rejection of the main null hypothesis (%) for various λ_1, λ_2 and λ_3 in the three stage procedure for $n = 125$

This chapter illustrates some examples of a three stage Combined Procedure with two model checking stages. This is to show that there are many more combinations of tests and model checking that can still be explored and researched. The model that was chosen as an example is the AR(1) model. One could argue that it is not realistic that AR(1) is the only possible deviation from dependence. The reason it was chosen was because it was the simplest dependence model and so it was a good start.

Nonetheless, the simulations show that the CP controls the Type I error well and has good power when the main null hypothesis is violated. The same conclusions in Chapter 4 can be made i.e. the CP can be useful in certain situations in a three stage CP, particularly when the sample size is small to moderate.

There are of course many more combinations of λ to consider which could potentially inform a three dimensional representation of the powers.

6

SUMMARY AND CONCLUDING REMARKS

> ‘‘The quiet statisticians have changed our world; not by discovering new facts or technical developments, but by changing the ways that we reason, experiment and form our opinions about it.’’ Ian Hacking

This thesis presented a novel methodology to study the performance of a combined procedure where a final main test is chosen by carrying out one or several misspecification tests. To come up with this methodology, an intensive but by no means exhaustive literature survey in Chapter 2 was carried out to understand how researchers carry out statistical hypothesis tests in practice, specifically how they deal with model assumptions. The literature shows that many researchers, particularly non-statisticians, do not fully understand the implications of making sure that model assumptions are fulfilled before using any statistical inference tests.

The misspecification paradox is introduced and discussed. Hennig (2007) first coined the term goodness-of-fit paradox, but we have decided to use misspecification paradox moving forward. A number of studies that discusses the combined procedure were surveyed to get a sense of a general recommendation, but we found that no agreement can be made whether MS testing should or should not be done. However, quite a number of authors do recommend that a test which has lesser assumptions should always be used without any model checking.

Chapter 3 presents a theory that shows that in certain conditions, the combined procedure has a higher power compared to the unconditional tests. We start by defining a few terms that is needed for the theory to be valid, which in the situation where the model assumption is not fulfilled, the power of the unrestricted model is higher than the power of the restricted model. Conversely, in the situation where the model assumption is fulfilled, the power of the restricted model is higher than the unrestricted model. Using the law of total probabilities as a starting point and the assumption that the MS test is independent of the main test, we were able to show that for a certain value of λ

(Equation (4)), the power of the combined procedure is indeed larger than the unconditional tests. The rather strict assumption of independence between the tests was then relaxed a little by adding some small measures of dependence, and the same conclusion still holds. As we have noted, the relationship that defines the dependence between the tests could potentially be more complex than a simple addition.

We also found that all the studies that simulate this combined procedure do it in a restrictive manner where all the samples are generated from a situation that either always violated the model assumption or always fulfills the model assumption. Therefore, a simulation setup was formulated in Chapter 4 where a random process is introduced in the beginning of the simulation to make the choice of either being in the situation where the model assumption is violated or fulfilled. Simulations were done and the Type I error rates were compared to the nominal level. In order to explore the power of the tests, a second set of simulations are carried out where one of each pair of samples has the mean shifted in varying degrees. A few combinations of distributions and hypothesis tests were considered. Looking at the levels of the tests and the combined procedure, there is not much evidence to support using the combined procedure over the unconditional tests. However, when inspecting the power plots, we can see that in the case where there is close to or equal chance of generating the samples from either a situation where the model assumptions are fulfilled or violated, the power of the combined procedure is larger than both the unconditional tests. This is not so apparent when the sample size is large. For this to be true there are some general requirements that have to be fulfilled, for example, the MS test must be at most weakly dependent or approximately independent to the main test. Secondly, in the situation where the model assumption is violated, the AU test must have a better power than the MC test and conversely, in the situation where the model assumption is fulfilled, the MC test must have a better power than the AU test. Finally, the MS test must have some use, namely it has a certain (possibly weak) ability to distinguish between the situation where the model assumption is fulfilled and violated.

Different levels of the MS test was also considered due to recommendations in some of the literature. It was found that the level of the MS test does not really affect the power of the CP. To choose a winner among the tests, both level and power must be examined. A level that is significantly larger than the nominal level can be a problem whereas a level that is lower than the nominal level is not a problem as this may indicate that the error probability is low. However, a level that is too low means that the power could be low as well which is also a problem. A power that is high should also further examined to check if the level is also too high. Hence, to choose a test is not an easy task

of balancing a level that is not too high whilst also requiring the power to be sufficiently high.

An example from the literature review (Rochon, Gondan and Kieser (2012)) shows that in terms of power, the two-sample t-test is better than the non-parametric WMW test if the underlying distributions are uniform. This clearly violates the normality assumption of the t-test (despite being asymptotically still correct), and will be picked up by many normality tests. Still it would be a bad decision to use the WMW test instead, even though its assumptions are fulfilled. An optimal combined procedure therefore should involve an MS test that picks up only those deviations from normality for which the WMW test (or whatever test is chosen as AU test) is actually helpful.

We then look at the level of the MS test conditional on the model and found that the level of the MS test is significantly higher than the nominal level, specifically twice of the nominal level. This leads us to recommend that the Bonferroni correction be used when a two-sample test problem is considered.

In some of the literature surveyed (Mayo and Spanos (2004), Rasch, Kubinger and Moder (2011)), a combined procedure with two or more model assumptions were considered. Therefore, in Chapter 5, a three stage combined procedure testing was studied. Two family of models were considered, equal distributions models and regression significance models. We first look at the main null hypothesis of equal distributions. The levels show slightly uncontrolled Type I error for the unconditional tests. The powers of the three stage combined procedure is larger than the unconditional tests especially when $(\lambda_1 = 1/3, \lambda_2 = 1/3, \lambda_3 = 1/3)$. This basically means that there is equal chance of getting two samples from either a t distribution with 3 degrees of freedom, normal distribution with mean 0 and unequal variances or standard normal distribution.

In all of the simulations that we considered, it can be argued that the combined procedure does indeed help in the choice of either using an MC or AU test. Certainly in some situations, it was shown that the CP has a better control of the level of the test and a good enough power. In some situations the CP has the best power of rejecting a false main null hypothesis. The CP also works well in cases where the sample size is small, $n = 8$. The CP is at the very least, almost as good as using a test without checking the model assumptions. Given a choice between not checking the model assumptions and checking them without significant loss of power and increased error probability, we feel that CP has the advantage here given the aforementioned requirements is fulfilled and also the Bonferroni correction is done in the case that a two-sample problem is considered.

Some suggestions for future work could be to ask, how much of a violation can cause problems for the level and power? For example, the variance of a t_2 distribution can-

not be calculated, therefore the Central Limit Theorem (CLT) does not hold. Numerous other violations of the model assumption has the potential to be further studied. The CLT works very well for a uniform distribution, so using something that assumes normality will still work quite well. Therefore, a good MS test for this reason would be one who detects those deviations from the normal that are really problematic, particularly heavy tails, non-existing variance or even if the variance exists but tails are heavy such as in the case of the t_3 distribution. Another example of the violation of the model assumption is independence of the samples which according to Cressie (1980) causes the biggest problems in the one-sample t-test. We are not aware of any literature examining of independence testing combined with the t-test.

There are also work such as Berk et al. (2013) and Hendry and Doornik (2014) that formally takes into account the effect of MS testing on subsequent tests which is called the post selection inference. Comparing post selection inference methods and CP could potentially be a good research area.

We believe that the focus of model checking is too much on the formal assumptions and not enough on deriving tests that can find the particular violations of model assumptions that are most problematic in terms of level and power. The development of MS tests that are better suited for this task and the investigation of the resulting combined procedures is a promising research area.

BIBLIOGRAPHY

[1] W. Albers, P. C. Boon and W. C. Kallenberg. 'Size and power of pretest procedures'. In: *Annals of Statistics* (2000), pp. 195–214.

[2] W. Albers, P. C. Boon and W. C. Kallenberg. 'Testing equality of two normal means using a variance pre-test.' In: *Statistics & Probability Letters* 38.3 (1998), pp. 221–227.

[3] W. Albers, P. C. Boon and W. C. Kallenberg. 'The asymptotic behavior of tests for normal means based on a variance pre-test'. In: *Journal of Statistical Planning And Inference* 88.1 (2000), pp. 195–214.

[4] L. A. Althouse, W. B. Ware and J. M. Ferron. 'Detecting departures from normality: A monte carlo simulation of a new omnibus test based on moments'. In: *Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, California* (1998).

[5] D. G. Altman. 'Poor-quality medical research: what can journals do?' In: *Jama* 287.21 (2002), pp. 2765–2767.

[6] B. C. Arnold. 'Hypothesis testing incorporating a preliminary test of significance'. In: *Journal of the American Statistical Association* 65.332 (1970), pp. 1590–1596.

[7] A. Azzalini. *The skew-normal and related families (Vol. 3)*. Cambridge University Press, 2013.

[8] T. A. Bancroft. 'Analysis and inference for incompletely specified models involving the use of preliminary test(s) of significance'. In: *Biometrics* 20.3 (1964), pp. 427–442.

[9] T. A. Bancroft. 'On biases in estimation due to the use of preliminary tests of significance'. In: *The Annals of Mathematical Statistics* 15.2 (1944), pp. 190–204.

[10] T. A. Bancroft and C. Han. 'Inference based on conditional specification: A note and a bibliography'. In: *International Statistical Review* 45.2 (1977), pp. 117–127.

[11] M. S. Bartlett. 'The effect of non-normality on the *t* distribution'. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 31 (1935), pp. 223–231.

[12] R. Berk et al. 'Valid post-selection inference'. In: *The Annals of Statistics* (2013), pp. 802–837.

[13] C. A. Boneau. 'The effects of violation of assumptions underlying the *t* test'. In: *Psychological Bulletin* 57 (1960), pp. 49–64.

[14] C. E. Bonferroni. 'Il calcolo delle assicurazioni su gruppi di teste'. In: *Studi in onore del professore salvatore ortu carboni* (1935), pp. 13–60.

[15] C. E. Bonferroni. 'Teoria statistica delle classi e calcolo delle probabilitá'. In: *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze* 8 (1936), pp. 3–62.

[16] G. E. P. Box. 'Robustness in the strategy of scientific model building'. In: *Robustness in Statistics* 1 (1979), pp. 201–236.

[17] G. E. P. Box and N. R. Draper. 'Empirical model-building and response surfaces'. In: New York: Wiley, 1987, p. 74.

[18] G. E. P. Box and D. A. Pierce. 'Distribution of residual autocorrelations in autoregressive-integrated moving average time series models'. In: *Journal of the American statistical Association* 65.332 (1970), pp. 1509–1526.

[19] M. B. Brown and A. B. Forsythe. 'Robust tests for the equality of variances'. In: *Journal of the American Statistical Association* 69.346 (1974), pp. 364–367.

[20] G. Casella and R. L. Berger. *Statistical Inference*. Vol. 2. Pacific Grove, CA: Duxbury, 2002.

[21] C. Chatfield. 'Model uncertainty, data mining and statistical inference (with discussion)'. In: *Journal of the Royal Statistical Society, Series B* 158.3 (1995), pp. 419–466.

[22] P. T. Choi. 'Statistics for the reader: What to ask before believing the results'. In: *Canadian Journal of Anesthesia/Journal Canadien d'anesthésie* 52 (2005), R1–R5.

[23] A. Cohen. 'To pool or not to pool in hypothesis testing'. In: *Journal of the American Statistical Association* 69.347 (1974), pp. 721–725.

[24] W. J. Conover, M. E. Johnson and M. M. Johnson. 'A comparative study of tests for homogeneity of variances, with applications to the outer continental shelf bidding data'. In: *Technometrics* 23.4 (1981), pp. 351–361.

[25] N. Cressie. 'Relaxing assumptions in the one-sample t-test'. In: *Australian Journal of Statistics* 22 (1980), pp. 143–153.

[26] H. E. Daniels. 'Rank correlation and population models'. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 12.2 (1950), pp. 171–191.

[27] P. L. Davies. *Data Analysis and Approximate Models*. Chapman & Hall/CRC, 2014.

[28] M. Delacre, D. Lakens and C. Leys. 'Why psychologists should by default use Welch's *t*-test instead of Student's *t*-test'. In: *International Review of Social Psychology* 30.1 (2017), pp. 92–101.

[29] D. Draper. 'Assessment and propagation of model uncertainty (with discussion)'. In: *Journal of the Royal Statistical Society, Series B* 57 (1995), pp. 45–97.

[30] J. Durbin and G. S. Watson. 'Testing for serial correlation in least squares regression. III'. In: *Biometrika* 58.1 (1971), pp. 1–19.

[31] R. G. Easterling. 'Goodness of fit and parameter estimation'. In: *Technometrics* 18.1 (1976), pp. 1–9.

[32] R. G. Easterling and H. E. Anderson. 'The effect of preliminary normality goodness of fit tests on subsequent inference'. In: *Journal of Statistical Computation and Simulation* 8.1 (1978), pp. 1–11.

[33] B. Efron. 'Estimation and accuracy after model selection'. In: *Journal of the American Statistical Association* 109.507 (2014), pp. 991–1007.

[34] P. J. Farrell and K. Rogers-Stewart. 'Comprehensive study of tests for normality and symmetry: extending the Spiegelhalter test'. In: *Journal of Statistical Computation and Simulation* 76.9 (2006), pp. 803–816.

[35] M. P. Fay and M. A. Proschan. 'Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules'. In: *Statistics Survey* 4 (2010), pp. 1–39.

[36] B. de Finetti. *Theory of Probability*. Wiley, 1974.

[37] F. M. Fisher. 'On the cost of approximate specification in simultaneous equation estimation'. In: *Econometrica: Journal of the Econometric Society* (1961), pp. 139–170.

[38] R. A. Fisher. 'Applications of "Student's" distribution'. In: *Metron* 5 (1925), pp. 90–104.

[39] R. A. Fisher. 'Moments and product moments of sampling distributions'. In: *Proceedings of the London Mathematical Society* 2.1 (1930), pp. 199–238.

[40] R. A. Fisher. 'On the Mathematical Foundation of Theoretical Statistics'. In: *Philosophical Transactions of the Royal Society of London A* 222 (1922), pp. 309–368.

[41] J. L. Gastwirth, Y. R. Gel and W. Miao. 'The impact of Levene's test of equality of variances on statistical theory and practice'. In: *Statistical Science* (2009), pp. 343–360.

[42] D. E. A. Giles and J. A. Giles. 'Pre-test estimation and testing in econometrics: Recent developments'. In: *Journal of Economic Surveys* 7.2 (1993), pp. 145–197.

[43] L. G. Godfrey. 'Misspecification tests and their uses in econometrics'. In: *Journal of Statistical Planning and Inference* 49.2 (1996), pp. 241–260.

[44] L. G. Godfrey. *Misspecification tests in econometrics. The Lagrange Multiplier principle and other applications*. Cambridge University Press, 1988.

[45] W. S. Gosset ("Student"). 'The probable error of a mean'. In: *Biometrika* 6 (1908), pp. 1–25.

[46] V. P. Gupta and V. K. Srivastava. 'Upper bound for the size of a test procedure using preliminary tests of significance'. In: *Journal of the Indian Statistical Association* 7 (1993), pp. 26–29.

[47] F. R. Hampel et al. *Robust Statistics*. Wiley, 1986.

[48] D. J. Hand. 'Wonderful examples, but let's not close our eyes'. In: *Statistical Science* 29.1 (2014), pp. 98–100.

[49] S. Hassan et al. 'Research design and statistical methods in Indian medical journals: a retrospective survey'. In: *PLoS One* 10.4 (2015), pp. 1–10.

[50] D. Hendry and J. Doornik. *Empirical Model Discovery and Theory Evaluation: Automatic Selection Methods in Econometrics*. MIT Press, 2014.

[51] C. Hennig. 'Falsification of propensity models by statistical tests and the goodness-of-fit paradox'. In: *Philosophia Mathematica* 15.2 (2007), pp. 166–192.

[52] C. Hennig. 'Mathematical models and reality: A constructivist perspective'. In: *Foundations of Science* 15.1 (2010), pp. 29–48.

[53] J. L. Hodges and E. L. Lehmann. 'The efficiency of some nonparametric competitors of the *t* test'. In: *Annals of Mathematical Statistic* 27 (1956), pp. 324–335.

[54] P. L. Hsu. 'Contributions to the theory of Student's *t*-test as applied to the problem of two samples'. In: *Statistical Research Memoirs* 2 (1938), pp. 1–24.

[55] T. C. Hsu and L. S. Feldt. 'The effect of limitations on the number of criterion score values on the significance level of the F-test'. In: *American Educational Research Journal* 6.4 (1969), pp. 515–527.

[56] B. G. Hwang et al. 'Investigating residentsfffdfffdfffd perceptions of green retrofit program in mature residential estates: The case of Singapore'. In: *Habitat International* 63 (2017), pp. 103–112.

[57] R. E. Kass et al. 'Ten simple rules for effective statistical practice'. In: *PLoS Computational Biology* 12.6 (2016), pp. 1–8.

[58] H. J. Keselman, C. J. Huberty et al. 'Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA analyses'. In: *Review of Educational Research* 68.3 (1998), pp. 350–386.

[59] H. J. Keselman, A. R. Othman and R. R. Wilcox. 'Preliminary testing for normality: Is this a good practice?' In: *Journal of Modern Applied Statistical Methods* 12.2 (2013), p. 2.

[60] S. Keskin. 'Comparison of several univariate normality tests regarding Type I error rate and power of the test in simulation based small samples'. In: *Journal of Applied Science Research* 2.5 (2006), pp. 296–300.

[61] T. K. Kim. 'T test as a parametric statistic'. In: *Korean Journal of Anesthesiology* 68.6 (2015), pp. 540–546.

[62] T. K. Kim and J. H. Park. 'More about the basic assumptions of t-test: normality and sample size'. In: *Korean Journal of Anesthesiology* 72.4 (2019), pp. 331–335.

[63] M. L. King and D. E. A. Giles. 'Autocorrelation pre-testing in the linear model: Estimation, testing and prediction'. In: *Journal of Econometrics* 25.1 (1984), pp. 35–48.

[64] D. Kokosińska et al. 'Heart rate variability, multifractal multiscale patterns and their assessment criteria'. In: *Physiological Measurement* 39.11 (2018), p. 114010.

[65] M. Koller and W. A. Stahel. 'Nonsingular subsampling for regression S estimators with categorical predictors'. In: *Computational Statistics* 32.2 (2017), pp. 631–646.

[66] M. Koller and W. A. Stahel. 'Sharpening wald-type inference in robust regression for small samples'. In: *Computational Statistics & Data Analysis* 55.8 (2011), pp. 2504–2515.

[67] J. D. Lee et al. 'Exact post-selection inference, with application to the lasso'. In: *Annals of Statistics* 44.3 (2016), pp. 907–927.

[68] H. Leeb and B. M. Pötscher. 'Model selection and inference: fact and fiction'. In: *Econometric Theory* 21.1 (2005), pp. 21–59.

[69] H. Leeb and B. M. Pötscher. 'On various confidence intervals Post-model-selection'. In: *Statistical Science* 30.2 (2015), pp. 216–227.

[70] E. L. Lehmann. 'Model specification: the views of Fisher and Neyman, and later developments'. In: *Statistical Science* 5.2 (1990), pp. 160–168.

[71] E. L. Lehmann. 'Robust tests for equality of variances'. In: *Contributions to probability and statistics: Essays in honor of Harold Hotelling* 2 (1960), pp. 278–292.

[72] J. A. Lewis. 'Statistical principles for clinical trials (ICH E9): an introductory note on an international guideline'. In: *Statistics in Medicine* 18.15 (1999), pp. 1903–1942.

[73] G. M. Ljung and G. E. P. Box. 'On a measure of lack of fit in time series models'. In: *Biometrika* 65.2 (1978), pp. 297–303.

[74] G. H. Lunney. 'Using analysis of variance with a dichotomous dependent variable: an empirical study 1'. In: *Journal of Educational Measurement* 7.4 (1970), pp. 263–269.

[75] H. B. Mann. 'Nonparametric tests against trend'. In: *Econometrica: Journal of the Econometric Society* (1945), pp. 245–259.

[76] C. A. Markowski and E. P. Markowski. 'Conditions for the effectiveness of a preliminary test of variance'. In: *The American Statistician* 44.4 (1990), pp. 322–326.

[77] D. G. Mayo. *Statistical Inference as Severe Testing*. Cambridge University Press, 2018.

[78] D. G. Mayo and A. Spanos. 'Methodology in practice: Statistical misspecification testing'. In: *Philosophy of Science* 71.5 (2004), pp. 1007–1025.

[79] M. Mendes and A. Pala. 'Type I error rate and power of three normality tests'. In: *Pakistan Journal of Information and Technology* 2.2 (1990), pp. 135–139.

[80] T. Micceri. 'The unicorn, the normal curve, and other improbable creatures'. In: *Psychological Bulletin* 105 (1989), pp. 156–166.

[81] F. C. Mills. *Statistical Methods: Applied to Economics and Business*. New York: Holt, 1924.

[82] K. Moder. 'Alternatives to *F*-test in one way *ANOVA* in case of heterogeneity of variances (a simulation study)'. In: *Psychological Testing and Assessment Modeling* 52 (2010), pp. 343–353.

[83] B. K. Moser and G. R. Stevens. 'Homogeneity of variance in the two-sample means test'. In: *The American Statistician* 46.1 (1992), pp. 19–21.

[84] B. K. Moser, G. R. Stevens and C. L. Watts. 'The two-sample *t* test versus Satterthwaite's approximate *F* test'. In: *Communications in Statistics-Theory and Methods* 18.11 (1989), pp. 3963–3975.

[85] H. R. Neave and C. W. J. Granger. 'A monte carlo study comparing various two sample tests for differences in means'. In: *Technometrics* 10 (1968), pp. 509–522.

[86] J. Neyman. *Lectures and Conferences on Mathematical Statistics and Probability*. 2nd. Washington: U.S. Department of Agriculture, 1952.

[87] J. Neyman and K. Iwaszkiewicz. 'Statistical Problems in Agricultural Experimentation'. In: *Supplement to the Journal of the Royal Statistical Society* (1935), pp. 107–180.

[88] J. Neyman and E. S. Pearson. 'On the Problem of the Most Efficient Tests of Statistical Hypotheses'. In: *Philosophical Transactions of the Royal Society of London A* 231 (1933), pp. 289–337.

[89] H. Nour-Eldein. 'Statistical methods and errors in family medicine articles between 2010 and 2014-Suez Canal University, Egypt: A cross-sectional study'. In: *Journal of Family Medicine and Primary Care* 5.1 (2016), pp. 24–33.

[90] K. Ohtani and T. Toyoda. 'Testing linear hypothesis on regression coefficients after a pre-test for disturbance variance'. In: *Economics Letters* 17.1-2 (1985), pp. 111–114.

[91] C. H. Olsen. 'Review of the use of statistics in infection and immunity'. In: *Infection and Immunity* 71 (2003), pp. 6689–6692.

[92] J. E. Overall, R. S. Atlas and J. M. Gibson. 'Tests that are robust against variance heterogeneity in k$\times$2 designs with unequal cell frequencies'. In: *Psychological Reports* 76.3 (1995), pp. 1011–1017.

[93] S. Park, E. Serpedin and K. Qaraqe. 'Gaussian assumption: The least favorable but the most useful [lecture notes]'. In: *IEEE Signal Processing Magazine* 30.3 (2013), pp. 183–186.

[94] H. O. Posten. 'Robustness of the two sample *t*-test'. In: *Robustness of statistical methods and nonparametric statistics*. Ed. by D. Rasch and M. L. Tiku. Netherlands: Springer, 1984, pp. 92–99.

[95] H. O. Posten. 'The robustness of the two sample *t*-test over the Pearson system'. In: *Journal of Statistical Computation and Simulation* 6 (1978), pp. 295–311.

[96] H. O. Posten, H. C. Yeh and D. B. Owen. 'Robustness of the two-sample *t*-test under violations of the homogeneity of variance assumptions'. In: *Communications in Statistics: Theory and Methods* 11 (1982), pp. 109–126.

[97] R. H. Randles and D. A. Wolfe. *Introduction to the Theory of Nonparametric Statistics*. New York: Wiley, 1979.

[98] C. V. Rao and K. P. Saxena. 'On approximation of power of a test procedure based on preliminary tests of significance'. In: *Communications in Statistics-Theory and Methods* 10.13 (1981), pp. 1305–1321.

[99] P. V. Rao. *Statistical research methods in the life sciences*. Pacific Grove, CA: Duxbury Press, 1998.

[100] D. Rasch and V. Guiard. 'The robustness of parametric statistical methods'. In: *Psychology Science* 46 (2004), pp. 175–208.

[101] D. Rasch, K. D. Kubinger and K. Moder. 'The two-sample *t* test: pre-testing its assumptions does not pay off'. In: *Statistical Papers* 52 (2011), pp. 219–231.

[102] N. M. Razali and Y. B. Wah. 'Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests'. In: *Journal of Statistical Modeling and Analytics* 2.1 (2011), pp. 21–33.

[103] J. Rochon, M. Gondan and M. Kieser. 'To test or not to test: Preliminary assessment of normality when comparing two independent samples'. In: *BMC Medical Research Methodology* 12.1 (2012), pp. 81–91.

[104] J. Rochon and M. Kieser. 'A closer look at the effect of preliminary goodness-of-fit testing for normality for the one-sample *t*-test'. In: *British Journal of Mathematical and Statistical Psychology* 64.3 (2011), pp. 410–426.

[105] J. P. Royston. 'An extension of Shapiro and Wilk's W test for normality to large samples'. In: *Applied Statistics* (1982), pp. 115–124.

[106] P. Royston. 'Approximating the Shapiro-Wilk W-Test for non-normality'. In: *Statistics and Computing* 2.3 (1992), pp. 117–119.

[107] P. Royston. 'Remark AS R94: A remark on algorithm AS 181: The W-test for normality'. In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 44.4 (1995), pp. 547–551.

[108] G. D. Ruxton. 'The unequal variance t-test is an underused alternative to Student's t-test and the Mann-Whitney U test'. In: *Behavioral Ecology* 17.4 (2006), pp. 688–690.

[109] A. M. E. Saleh and P. K. Sen. 'Asymptotic properties of tests of hypothesis following a preliminary test'. In: *Statistics and Risk Modeling* 1.4-5 (1983), pp. 455–478.

[110] F. E. Satterthwaite. 'An approximate distribution of estimates of variance components'. In: *Biometrics Bulletin* 2.6 (1946), pp. 110–114.

[111] S. S. Sawilowsky and R. C. Blair. 'A more realistic look at the robustness and type II error properties of the t test to departures from population normality'. In: *Psychological Bulletin* 111.2 (1992), pp. 352–360.

[112] H. Scheffé. 'Practical solutions of the Behrens-Fisher problem'. In: *Journal of the American Statistical Association* 65 (1970), pp. 1501–1508.

[113] V. Schoder, A. Himmelmann and K. P. Wilhelm. 'Preliminary testing for normality: some statistical aspects of a common concept'. In: *Clinical and Experimental Dermatology* 31.6 (2006), pp. 757–761.

[114] W. R. Schucany and H. K. T. Ng. 'Preliminary goodness-of-fit tests for normality do not validate the one-sample Student *t*'. In: *Communications in Statistics - Theory and Methods* 35 (2006), pp. 2275–2286.

[115] M. Shan, B. G. Hwang and K. S. N. Wong. 'A preliminary investigation of underground residential buildings: advantages, disadvantages, and critical risks'. In: *Tunnelling and Underground Space Technology* 70 (2017), pp. 19–29.

[116] S. S. Shapiro and M. B. Wilk. 'An analysis of variance test for normality (complete samples)'. In: *Biometrika* 52.3/4 (1965), pp. 591–611.

[117] A. Spanos. 'Akaike-type criteria and the reliability of inference: Model selection versus statistical model specification'. In: *Journal of Econometrics* 158 (2010), pp. 204–220.

[118] A. Spanos. 'Mis-specification testing in retrospect'. In: *Journal of Economic Surveys* 32.2 (2018), pp. 541–577.

[119] A. Spanos. *Probability theory and statistical inference: econometric modeling with observational data*. Cambridge University Press, 1999.

[120] K. Sridharan and S. Gowri. 'Reporting quality of statistics in Indian journals: Analysis of articles over a period of two years'. In: *Journal of Scientometric Research* 4.1 (2015), pp. 10–13.

[121] A. M. Strasak et al. 'The use of statistics in medical research: A comparison of The New England Journal of Medicine and Nature Medicine'. In: *American Statistician* 61.1 (2007), pp. 47–55.

[122] A. M. Strasak et al. 'The use of statistics in medical research: A comparison of Wiener Klinische Wochenschrift and Wiener Medizinische Wochenschrift'. In: *Austrian Journal of Statistics* 36.2 (2007), pp. 141–152.

[123] T. Toyoda and K. Ohtani. 'Testing equality between sets of coefficients after a preliminary test for equality of disturbance variances in two linear regressions'. In: *Journal of Econometrics* 31.1 (1986), pp. 67–80.

[124] S. B. Vardeman and M. D. Morris. 'Statistics and ethics: some advice for young statisticians'. In: *The American Statistician* 57.1 (2003), pp. 21–26.

[125] B. L. Welch. 'The generalisation of Student's problem when several different population variances are involved'. In: *Biometrika* 34 (1947), pp. 28–35.

[126] B. L. Welch. 'The significance of the difference between two means when the population variances are unequal'. In: *Biometrika* 29 (1938), pp. 350–362.

[127] E. R. White. 'Minimum time required to detect population trends: the need for long-term monitoring programs'. In: *BioScience* 69.1 (2019), pp. 40–46.

[128] W. Wiedermann and R. Alexandrowicz. 'A plea for more general tests than those for location only: Further considerations on Rasch & Guiard's 'The robustness of parametric statistical methods''. In: *Psychology Science* 49 (2007), pp. 2–12.

[129] R. R. Wilcox. 'How many discoveries have been lost by ignoring modern statistical methods?' In: *American Psychologist* 53.3 (1998), pp. 300–314.

[130] R. R. Wilcox, V. L. Charlin and K. L. Thompson. 'New Monte Carlo results on the robustness of the ANOVA F, W, and F* statistics'. In: *Communications in Statistics - Simulation and Computation* 15 (1986), pp. 933–943.

[131] S. Wu et al. 'Misuse of statistical methods in 10 leading Chinese medical journals in 1998 and 2008'. In: *The Scientific World Journal* 11 (2011), pp. 2106–2114.

[132] W. Wu et al. 'Design Assessment in Virtual and Mixed Reality Environments: Comparison of Novices and Experts'. In: *Journal of Construction Engineering and Management* 145.9 (2019), p. 04019049.

[133] D. W. Zimmerman. 'A note on preliminary tests of equality of variances'. In: *British Journal of Mathematical and Statistical Psychology* 57.1 (2004), pp. 173–181.

[134] D. W. Zimmerman. 'Two separate effects of variance heterogeneity on the validity and power of significance tests of location'. In: *Statistical Methodology* 3.4 (2006), pp. 351–374.

[135] D. W. Zimmerman and B. D. Zumbo. 'Rank transformations and the power of the Student t test and Welch t'test for non-normal populations with unequal variances'. In: *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 47.3 (1993), p. 523.

A

APPENDIX

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.13916	0.14082	0.14261	0.14328	0.14358	0.12263	0.14074
0.1	0.13919	0.1407	0.14222	0.1426	0.14252	0.12553	0.14011
0.2	0.13993	0.1418	0.14316	0.14346	0.1432	0.12741	0.14033
0.3	0.14111	0.14307	0.14427	0.14432	0.14441	0.13064	0.14066
0.4	0.14311	0.1445	0.14561	0.14578	0.14572	0.13347	0.1416
0.5	0.14615	0.14747	0.14856	0.14861	0.14819	0.13745	0.14301
0.6	0.14811	0.1491	0.14995	0.14993	0.14961	0.14086	0.1452
0.7	0.14869	0.14959	0.14979	0.14976	0.14965	0.14248	0.14454
0.8	0.15365	0.15424	0.15486	0.15457	0.15413	0.14912	0.14759
0.9	0.15353	0.15408	0.15427	0.15401	0.15331	0.15027	0.14736
1	0.15462	0.15506	0.1554	0.15516	0.15452	0.15303	0.14774

Table 32: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 8$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.39816	0.40249	0.40533	0.40614	0.40632	0.35965	0.4007
0.1	0.40219	0.40596	0.40816	0.40844	0.40827	0.36732	0.40301
0.2	0.40677	0.41003	0.4122	0.41257	0.41211	0.37562	0.40472
0.3	0.41531	0.41787	0.41938	0.41907	0.41884	0.38869	0.41084
0.4	0.42097	0.423	0.42411	0.42345	0.42304	0.39754	0.41343
0.5	0.43215	0.4339	0.4346	0.43383	0.43234	0.41214	0.42106
0.6	0.43475	0.43652	0.43706	0.43658	0.4358	0.41859	0.4247
0.7	0.44177	0.44297	0.44312	0.44199	0.44079	0.42904	0.4276
0.8	0.44955	0.45038	0.44927	0.44776	0.44606	0.44018	0.43213
0.9	0.45879	0.45911	0.45804	0.45648	0.45463	0.45332	0.43847
1	0.46244	0.46286	0.4622	0.46041	0.45818	0.46094	0.44183

Table 33: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 8$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.88208	0.886	0.88855	0.88911	0.88916	0.82358	0.88722
0.1	0.88862	0.89221	0.89494	0.89577	0.89585	0.83509	0.89289
0.2	0.89603	0.89922	0.90087	0.90137	0.90156	0.84852	0.89795
0.3	0.90585	0.90839	0.90992	0.91022	0.91046	0.86384	0.9073
0.4	0.91294	0.91527	0.91647	0.91659	0.91638	0.87694	0.91254
0.5	0.92186	0.92387	0.92476	0.92477	0.92435	0.89273	0.91936
0.6	0.92845	0.92957	0.93049	0.93016	0.92974	0.905	0.92486
0.7	0.93729	0.93785	0.93826	0.93762	0.93696	0.91904	0.93155
0.8	0.94431	0.94461	0.94459	0.94363	0.94328	0.93339	0.93721
0.9	0.95188	0.95188	0.95134	0.95062	0.94964	0.94582	0.94316
1	0.9605	0.96007	0.95902	0.95811	0.95698	0.96074	0.94932

Table 34: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 8$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.3829	0.38842	0.38497	0.38503	0.38513	0.3392	0.38427
0.1	0.38843	0.38962	0.3903	0.39029	0.39003	0.34989	0.38719
0.2	0.3953	0.3964	0.39691	0.39662	0.39611	0.36105	0.39149
0.3	0.39921	0.39981	0.39982	0.3994	0.3988	0.36964	0.39216
0.4	0.40372	0.40453	0.40411	0.40324	0.40248	0.37859	0.3944
0.5	0.41213	0.4125	0.41161	0.41044	0.40923	0.39088	0.39904
0.6	0.41573	0.41597	0.41504	0.41369	0.41243	0.39877	0.40092
0.7	0.42591	0.42594	0.42456	0.42324	0.42176	0.41301	0.40867
0.8	0.42807	0.42757	0.42623	0.42439	0.42281	0.41898	0.40815
0.9	0.43209	0.43233	0.4301	0.42794	0.42595	0.42775	0.40889
1	0.43664	0.43621	0.43452	0.43271	0.4305	0.43602	0.41206

Table 35: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 27$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.90594	0.9075	0.90873	0.90924	0.9094	0.84282	0.90926
0.1	0.90841	0.90969	0.91061	0.91091	0.91091	0.85162	0.91023
0.2	0.91403	0.91513	0.91558	0.91559	0.91549	0.86304	0.91422
0.3	0.91994	0.92073	0.92088	0.92067	0.92059	0.8752	0.91818
0.4	0.92233	0.92281	0.92278	0.92247	0.92198	0.88476	0.91901
0.5	0.92887	0.92923	0.92907	0.92867	0.92834	0.89701	0.92456
0.6	0.93058	0.93091	0.93069	0.9303	0.9295	0.90654	0.92532
0.7	0.93732	0.93709	0.93631	0.93553	0.93494	0.919	0.92999
0.8	0.94203	0.94173	0.94118	0.94035	0.9398	0.92962	0.93407
0.9	0.94532	0.94493	0.94361	0.94275	0.94166	0.93933	0.93463
1	0.94944	0.94864	0.94722	0.94593	0.94462	0.94997	0.93659

Table 36: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 27$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.99998	0.99997	0.99998	0.99998	0.99998	0.99389	0.99998
0.1	0.99997	0.99997	0.99998	0.99998	0.99998	0.99393	0.99998
0.2	0.99997	0.99997	0.99999	0.99999	0.99999	0.99472	0.99999
0.3	0.99999	0.99999	0.99999	0.99999	0.99999	0.99561	0.99999
0.4	0.99999	0.99999	0.99999	0.99999	0.99999	0.99614	0.99999
0.5	0.99999	0.99999	0.99999	0.99999	0.99999	0.99688	0.99999
0.6	0.99999	0.99999	0.99999	0.99999	0.99999	0.99774	0.99999
0.7	0.99999	0.99999	0.99999	0.99999	0.99999	0.99805	0.99999
0.8	0.99999	0.99999	0.99999	0.99999	0.99999	0.99853	0.99999
0.9	1	1	1	1	1	0.99948	1
1	1	1	1	1	1	1	1

Table 37: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 27$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.95512	0.95516	0.95516	0.95516	0.95516	0.88969	0.95516
0.1	0.95901	0.95898	0.95893	0.95877	0.95879	0.90096	0.95817
0.2	0.95912	0.95908	0.95895	0.95888	0.95879	0.90719	0.95826
0.3	0.96181	0.96168	0.9615	0.96135	0.96118	0.91646	0.96046
0.4	0.96452	0.96441	0.96414	0.96391	0.96385	0.92667	0.96244
0.5	0.96565	0.96555	0.96531	0.96487	0.96462	0.93397	0.9628
0.6	0.96819	0.96793	0.96766	0.96742	0.96719	0.94228	0.96511
0.7	0.96915	0.96893	0.96846	0.96814	0.96766	0.94936	0.9654
0.8	0.97189	0.97155	0.97106	0.97065	0.97025	0.95909	0.96748
0.9	0.97408	0.97369	0.97277	0.97237	0.97181	0.96804	0.96872
1	0.9749	0.97439	0.97366	0.97302	0.97243	0.97531	0.96927

Table 38: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 125$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	1	1	1	1	1	0.99843	1
0.1	1	1	1	1	1	0.99861	1
0.2	1	1	1	1	1	0.99878	1
0.3	1	1	1	1	1	0.99891	1
0.4	1	1	1	1	1	0.99896	1
0.5	1	1	1	1	1	0.99933	1
0.6	1	1	1	1	1	0.99938	1
0.7	1	1	1	1	1	0.99954	1
0.8	1	1	1	1	1	0.9998	1
0.9	1	1	1	1	1	0.99981	1
1	1	1	1	1	1	1	1

Table 39: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 125$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	1	1	1	1	1	0.99986	1
0.1	1	1	1	1	1	0.99983	1
0.2	1	1	1	1	1	0.99983	1
0.3	1	1	1	1	1	0.99982	1
0.4	1	1	1	1	1	0.99987	1
0.5	1	1	1	1	1	0.9999	1
0.6	1	1	1	1	1	0.99993	1
0.7	1	1	1	1	1	0.99994	1
0.8	1	1	1	1	1	1	1
0.9	1	1	1	1	1	0.99997	1
1	1	1	1	1	1	1	1

Table 40: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for $n = 125$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.1332	0.13558	0.1384	0.13975	0.14066	0.11485	0.14074
0.1	0.13381	0.13585	0.13827	0.1392	0.13976	0.11848	0.14011
0.2	0.13438	0.13686	0.13893	0.13989	0.14027	0.12035	0.14033
0.3	0.1353	0.13783	0.13988	0.14055	0.14115	0.12339	0.14066
0.4	0.13772	0.13965	0.14141	0.14211	0.14254	0.12688	0.1416
0.5	0.14044	0.14231	0.14428	0.14495	0.14503	0.1305	0.14301
0.6	0.14333	0.1448	0.14625	0.14668	0.14682	0.13501	0.1452
0.7	0.14365	0.14497	0.14591	0.14633	0.14657	0.1366	0.14454
0.8	0.14862	0.14959	0.15075	0.1509	0.15079	0.14353	0.14759
0.9	0.14839	0.14927	0.15004	0.15027	0.15	0.14462	0.14736
1	0.15015	0.15086	0.15164	0.15181	0.15154	0.14837	0.14774

Table 41: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 8$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.38646	0.39225	0.3969	0.3991	0.40024	0.34215	0.4007
0.1	0.39032	0.39547	0.39965	0.40134	0.40229	0.3502	0.40301
0.2	0.39513	0.39967	0.40363	0.40531	0.40595	0.35933	0.40472
0.3	0.40433	0.40812	0.41124	0.41222	0.41307	0.37352	0.41084
0.4	0.41025	0.41316	0.41596	0.41653	0.41703	0.38322	0.41343
0.5	0.42193	0.42465	0.42668	0.42714	0.42655	0.39884	0.42106
0.6	0.42492	0.42758	0.42934	0.43014	0.4302	0.40571	0.4247
0.7	0.43209	0.43402	0.43532	0.43511	0.43486	0.41692	0.4276
0.8	0.44062	0.44195	0.44196	0.44136	0.44047	0.42973	0.43213
0.9	0.44962	0.45036	0.45026	0.44961	0.44866	0.44311	0.43847
1	0.45394	0.45472	0.4547	0.45366	0.45229	0.45188	0.44183

Table 42: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 8$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.87468	0.87997	0.88398	0.88533	0.88614	0.80261	0.88722
0.1	0.88183	0.88683	0.89063	0.89236	0.89318	0.81708	0.89289
0.2	0.88938	0.89375	0.89683	0.89803	0.89884	0.83179	0.89795
0.3	0.90012	0.90365	0.90642	0.90732	0.908	0.84931	0.9073
0.4	0.90757	0.9108	0.91307	0.91369	0.91389	0.8634	0.91254
0.5	0.91731	0.92017	0.92195	0.92239	0.92232	0.88149	0.91936
0.6	0.92415	0.92592	0.92758	0.92775	0.9277	0.89546	0.92486
0.7	0.93365	0.93482	0.93576	0.93555	0.93516	0.91123	0.93155
0.8	0.94122	0.94192	0.94225	0.9417	0.94165	0.92727	0.93721
0.9	0.94929	0.94949	0.94931	0.94885	0.94814	0.9418	0.94316
1	0.95841	0.95808	0.9572	0.95649	0.95556	0.95849	0.94932

Table 43: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 8$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.38241	0.38409	0.38471	0.38483	0.38497	0.33678	0.38427
0.1	0.38783	0.38915	0.38995	0.39001	0.38982	0.34746	0.38719
0.2	0.39478	0.39596	0.39659	0.3964	0.39591	0.35938	0.39149
0.3	0.39863	0.39933	0.39952	0.39914	0.39859	0.36767	0.39216
0.4	0.40325	0.40409	0.40375	0.40294	0.4022	0.377	0.3944
0.5	0.4116	0.412	0.41117	0.41007	0.40888	0.38955	0.39904
0.6	0.4151	0.41542	0.41456	0.41328	0.41205	0.39726	0.40092
0.7	0.42519	0.42529	0.42399	0.42274	0.42132	0.41175	0.40867
0.8	0.42744	0.427	0.42574	0.42396	0.42239	0.41791	0.40815
0.9	0.43138	0.43164	0.42951	0.42743	0.4255	0.42681	0.40889
1	0.43614	0.43572	0.43405	0.4323	0.43013	0.43546	0.41206

Table 44: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 27$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.90575	0.90733	0.9086	0.90912	0.90931	0.84078	0.90926
0.1	0.90822	0.90953	0.91049	0.91082	0.91082	0.84957	0.91023
0.2	0.9138	0.91493	0.91544	0.91548	0.91539	0.8613	0.91422
0.3	0.9197	0.92055	0.92076	0.92058	0.9205	0.87357	0.91818
0.4	0.9221	0.92263	0.92263	0.92235	0.92186	0.88344	0.91901
0.5	0.92871	0.92911	0.92898	0.92859	0.92826	0.89587	0.92456
0.6	0.93041	0.93078	0.93057	0.93021	0.92944	0.90561	0.92532
0.7	0.9372	0.93697	0.93622	0.93545	0.93488	0.91857	0.92999
0.8	0.9418	0.94153	0.94102	0.9402	0.93967	0.9291	0.93407
0.9	0.94512	0.94474	0.94342	0.9426	0.94153	0.93899	0.93463
1	0.94923	0.94844	0.94703	0.94575	0.94447	0.94974	0.93659

Table 45: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 27$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.99998	0.99997	0.99998	0.99998	0.99998	0.99339	0.99998
0.1	0.99997	0.99997	0.99998	0.99998	0.99998	0.99358	0.99998
0.2	0.99997	0.99997	0.99999	0.99999	0.99999	0.99447	0.99999
0.3	0.99999	0.99999	0.99999	0.99999	0.99999	0.99531	0.99999
0.4	0.99999	0.99999	0.99999	0.99999	0.99999	0.99585	0.99999
0.5	0.99999	0.99999	0.99999	0.99999	0.99999	0.99679	0.99999
0.6	0.99999	0.99999	0.99999	0.99999	0.99999	0.99753	0.99999
0.7	0.99999	0.99999	0.99999	0.99999	0.99999	0.99794	0.99999
0.8	0.99999	0.99999	0.99999	0.99999	0.99999	0.99847	0.99999
0.9	1	1	1	1	1	0.99946	1
1	1	1	1	1	1	1	1

Table 46: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 27$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.95512	0.95516	0.95516	0.95516	0.95516	0.88953	0.95516
0.1	0.95901	0.95898	0.95893	0.95877	0.95862	0.90083	0.95817
0.2	0.95912	0.95908	0.95895	0.95888	0.95879	0.90704	0.95826
0.3	0.96181	0.96168	0.9615	0.96135	0.96118	0.91631	0.96046
0.4	0.96452	0.96441	0.96414	0.96391	0.96385	0.92656	0.96244
0.5	0.96565	0.96555	0.96531	0.96487	0.96462	0.93391	0.9628
0.6	0.96819	0.96793	0.96766	0.96742	0.96719	0.94222	0.96511
0.7	0.96915	0.96893	0.96846	0.96814	0.96766	0.94932	0.9654
0.8	0.97187	0.97153	0.97105	0.97064	0.97024	0.95905	0.96748
0.9	0.97408	0.97369	0.97277	0.97237	0.97181	0.968	0.96872
1	0.9749	0.97439	0.97366	0.97302	0.97243	0.97531	0.96927

Table 47: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 125$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.13067	0.13338	0.13541	0.13641	0.13715	0.11662	0.13791
0.1	0.12963	0.13212	0.13432	0.13545	0.13618	0.11693	0.13612
0.2	0.12784	0.13027	0.13232	0.13329	0.13378	0.11693	0.13341
0.3	0.1236	0.12573	0.12706	0.12793	0.12855	0.11374	0.12888
0.4	0.1238	0.12559	0.12708	0.12803	0.12855	0.11513	0.12885
0.5	0.1227	0.12424	0.12567	0.12672	0.12704	0.11433	0.126
0.6	0.11882	0.12035	0.12157	0.12217	0.12281	0.11195	0.12279
0.7	0.11909	0.12025	0.12174	0.12209	0.12231	0.11292	0.12169
0.8	0.11498	0.11647	0.1179	0.1181	0.11866	0.11035	0.11757
0.9	0.1129	0.11382	0.11516	0.11566	0.1158	0.10977	0.1149
1	0.11168	0.11277	0.11386	0.11415	0.11437	0.10942	0.1135

Table 48: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 8$, $ncp = 0.5$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.37918	0.384	0.38857	0.39065	0.39191	0.34609	0.39301
0.1	0.37307	0.37772	0.38184	0.38394	0.38468	0.34385	0.38463
0.2	0.36326	0.36711	0.37092	0.37282	0.37351	0.3377	0.37255
0.3	0.35317	0.35707	0.36036	0.36214	0.36302	0.32935	0.36252
0.4	0.3497	0.35315	0.35634	0.35735	0.35799	0.32905	0.3571
0.5	0.34233	0.34509	0.34826	0.3491	0.34942	0.32372	0.34696
0.6	0.32969	0.33257	0.33478	0.33548	0.33626	0.31571	0.33346
0.7	0.32697	0.32929	0.33115	0.33217	0.33215	0.31562	0.32749
0.8	0.31622	0.31834	0.32006	0.32035	0.32051	0.30867	0.31726
0.9	0.31042	0.31158	0.31308	0.31361	0.31388	0.30472	0.30997
1	0.30254	0.30383	0.30442	0.30462	0.30444	0.29996	0.2991

Table 49: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 8$, $ncp = 1$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.86862	0.87353	0.87739	0.87947	0.88014	0.82197	0.88086
0.1	0.86125	0.86546	0.86956	0.87114	0.87192	0.81848	0.87212
0.2	0.85613	0.85958	0.8628	0.86394	0.86455	0.81868	0.86358
0.3	0.85225	0.85573	0.85803	0.85902	0.85924	0.8183	0.85639
0.4	0.84696	0.84973	0.85161	0.85226	0.85234	0.81926	0.8481
0.5	0.84201	0.84425	0.84586	0.84625	0.84602	0.81816	0.84108
0.6	0.83815	0.83994	0.84084	0.8405	0.83992	0.81921	0.83436
0.7	0.8338	0.83508	0.83486	0.8347	0.83427	0.81974	0.82725
0.8	0.82506	0.82574	0.82607	0.82577	0.82509	0.81557	0.81695
0.9	0.8215	0.82172	0.82135	0.82051	0.81909	0.81654	0.81039
1	0.81476	0.81476	0.81374	0.81266	0.81119	0.81407	0.80165

Table 50: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 8$, $ncp = 2$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.37208	0.37323	0.37423	0.37453	0.37446	0.33202	0.37284
0.1	0.36321	0.36456	0.36561	0.36537	0.36565	0.32701	0.36326
0.2	0.35692	0.35805	0.35885	0.35884	0.35917	0.32536	0.35598
0.3	0.34941	0.3509	0.35161	0.35135	0.35105	0.32015	0.34552
0.4	0.34246	0.34343	0.34371	0.34375	0.34345	0.31847	0.33763
0.5	0.33384	0.33471	0.33534	0.3347	0.33438	0.31266	0.32778
0.6	0.32466	0.3254	0.32566	0.32483	0.32372	0.30811	0.31598
0.7	0.31863	0.31882	0.31885	0.31801	0.31696	0.3047	0.30885
0.8	0.30885	0.3096	0.30985	0.30916	0.30816	0.29919	0.29754
0.9	0.29923	0.29978	0.29921	0.29842	0.29744	0.29413	0.2863
1	0.29049	0.29069	0.29085	0.28991	0.28877	0.28946	0.27716

Table 51: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 27$, $ncp = 0.5$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.89438	0.89636	0.89815	0.89894	0.89924	0.83181	0.90008
0.1	0.88571	0.88756	0.88906	0.88953	0.88961	0.82823	0.88733
0.2	0.87903	0.88035	0.88094	0.88117	0.88092	0.82699	0.87737
0.3	0.86842	0.86926	0.86951	0.86925	0.86849	0.82491	0.8638
0.4	0.8608	0.86191	0.86164	0.86093	0.86035	0.82282	0.85276
0.5	0.85177	0.85186	0.8514	0.85009	0.8485	0.81998	0.83878
0.6	0.84066	0.8406	0.83982	0.83816	0.8366	0.81482	0.82665
0.7	0.83362	0.83334	0.83243	0.83041	0.82863	0.81403	0.81512
0.8	0.82511	0.82454	0.82274	0.82068	0.81869	0.8116	0.80458
0.9	0.8142	0.81345	0.81104	0.8088	0.80645	0.80649	0.79027
1	0.80632	0.80471	0.80183	0.79878	0.79575	0.80498	0.77694

Table 52: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 27$, $ncp = 1$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.99996	0.99996	0.99997	0.99997	0.99997	0.99431	0.99997
0.1	0.99998	0.99998	0.99998	0.99998	0.99997	0.99464	0.99995
0.2	0.99994	0.99993	0.99994	0.99994	0.99994	0.99548	0.99994
0.3	0.99991	0.99992	0.99992	0.9999	0.99989	0.99602	0.99986
0.4	0.99995	0.99996	0.99994	0.99994	0.99993	0.99676	0.99989
0.5	0.99994	0.99994	0.99994	0.99993	0.99991	0.99701	0.99986
0.6	0.99987	0.99985	0.99986	0.99987	0.99986	0.99754	0.99982
0.7	0.99989	0.99989	0.99985	0.99985	0.99985	0.9984	0.99978
0.8	0.99991	0.99991	0.99988	0.99987	0.99984	0.9988	0.99974
0.9	0.99989	0.99988	0.99987	0.99986	0.99985	0.9994	0.99979
1	0.99987	0.99986	0.99983	0.99981	0.99981	0.99989	0.99974

Table 53: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 27$, $ncp = 2$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.95129	0.95139	0.95146	0.95147	0.95146	0.88275	0.95147
0.1	0.94285	0.94282	0.94268	0.94249	0.94234	0.88117	0.94121
0.2	0.93296	0.93269	0.93226	0.93195	0.93165	0.87879	0.92946
0.3	0.92588	0.92542	0.92494	0.9245	0.92402	0.87856	0.92109
0.4	0.91924	0.91874	0.91791	0.91703	0.91616	0.8807	0.91247
0.5	0.90985	0.90929	0.90798	0.90717	0.90633	0.87692	0.90063
0.6	0.90203	0.9013	0.90023	0.89879	0.89764	0.87518	0.89149
0.7	0.89329	0.89244	0.89092	0.88952	0.88785	0.87353	0.88058
0.8	0.88443	0.88364	0.8819	0.88044	0.87905	0.87179	0.87033
0.9	0.87748	0.87637	0.8742	0.87224	0.87071	0.87179	0.86012
1	0.86828	0.8672	0.86466	0.86251	0.86093	0.8693	0.8508

Table 54: Powers of the combined procedure and the MC and AU tests where MC is the Welch's t-test, AU is the WMW test for $n = 125$, $ncp = 0.5$ and $\sigma_2/\sigma 1 = 1.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.14574	0.14599	0.14593	0.14513	0.14469	0.14414	0.13928
0.1	0.14680	0.14744	0.14755	0.14716	0.14577	0.14480	0.14006
0.2	0.14865	0.14947	0.14933	0.14837	0.14753	0.14694	0.14104
0.3	0.14842	0.14879	0.14878	0.14835	0.14736	0.14677	0.14114
0.4	0.14860	0.14949	0.14968	0.1493	0.14875	0.14692	0.14251
0.5	0.15214	0.15277	0.15263	0.15235	0.15116	0.15035	0.14434
0.6	0.15323	0.15386	0.15374	0.15333	0.15265	0.15165	0.14587
0.7	0.15344	0.15393	0.15381	0.15346	0.15290	0.15169	0.14591
0.8	0.15254	0.15283	0.15267	0.15221	0.15203	0.15041	0.14481
0.9	0.15453	0.15481	0.15498	0.15432	0.15321	0.15271	0.14578
1	0.15462	0.15506	0.15540	0.15516	0.154529	0.15303	0.14774

Table 55: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test where Q is the uniform distribution, $n = 8$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.43774	0.43623	0.43285	0.42827	0.42404	0.43792	0.40300
0.1	0.43814	0.43671	0.43313	0.42901	0.42528	0.43801	0.40541
0.2	0.4433	0.44187	0.43862	0.43482	0.43102	0.44323	0.41049
0.3	0.44101	0.44004	0.43722	0.43337	0.43006	0.4405	0.41115
0.4	0.44875	0.44783	0.44483	0.44164	0.43902	0.44833	0.41934
0.5	0.45102	0.45052	0.44852	0.44513	0.44182	0.45074	0.42307
0.6	0.45333	0.45254	0.45029	0.44734	0.44462	0.45241	0.42638
0.7	0.45649	0.45604	0.45442	0.45183	0.44940	0.45568	0.43057
0.8	0.45273	0.45267	0.45095	0.44904	0.44671	0.45178	0.4292
0.9	0.45772	0.45742	0.45575	0.45377	0.45221	0.45664	0.43614
1	0.46244	0.46286	0.4622	0.46041	0.45818	0.46094	0.44183

Table 56: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test where Q is the uniform distribution, $n = 8$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.9668	0.9613	0.95336	0.94779	0.9442	0.97175	0.93616
0.1	0.96635	0.96151	0.95346	0.94834	0.94479	0.97083	0.93769
0.2	0.96466	0.96032	0.95346	0.94894	0.94586	0.96856	0.93792
0.3	0.96458	0.96037	0.9541	0.94981	0.94665	0.96806	0.93868
0.4	0.96415	0.96061	0.95527	0.95172	0.94917	0.96735	0.94098
0.5	0.9629	0.95989	0.95484	0.95144	0.94923	0.96552	0.94175
0.6	0.96279	0.96079	0.9568	0.954	0.95174	0.96518	0.94426
0.7	0.96171	0.95991	0.95678	0.9543	0.95219	0.96344	0.94507
0.8	0.96078	0.95928	0.95681	0.95475	0.95327	0.96214	0.94606
0.9	0.96104	0.96014	0.95839	0.95691	0.95544	0.96189	0.94774
1	0.96050	0.96007	0.95902	0.95811	0.95698	0.96074	0.94932

Table 57: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test where Q is the uniform distribution, $n = 8$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.42623	0.41871	0.40852	0.40328	0.40022	0.43138	0.39503
0.1	0.42777	0.42086	0.41137	0.40634	0.40311	0.43267	0.39711
0.2	0.42717	0.42107	0.41239	0.4072	0.40419	0.43186	0.39682
0.3	0.42839	0.42251	0.41479	0.41014	0.40743	0.4319	0.39916
0.4	0.42895	0.42376	0.41691	0.41319	0.4108	0.43244	0.40086
0.5	0.43251	0.42846	0.42238	0.41857	0.41582	0.4348	0.40482
0.6	0.43505	0.43183	0.42617	0.42257	0.42026	0.43702	0.40709
0.7	0.43429	0.43163	0.42711	0.42474	0.42221	0.43548	0.40753
0.8	0.43629	0.43451	0.43113	0.42842	0.42654	0.43667	0.41067
0.9	0.43568	0.43464	0.43186	0.4296	0.4275	0.43573	0.4101
1	0.43664	0.43621	0.43452	0.43271	0.4305	0.43602	0.41206

Table 58: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test where Q is the uniform distribution, $n = 27$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.93435	0.92426	0.91607	0.91291	0.9114	0.95305	0.91015
0.1	0.93658	0.92783	0.92093	0.91777	0.91625	0.95304	0.9142
0.2	0.93821	0.93019	0.92295	0.92006	0.91865	0.95303	0.91622
0.3	0.93858	0.93103	0.9248	0.92248	0.9212	0.95191	0.91815
0.4	0.941	0.93464	0.92904	0.92675	0.9255	0.95308	0.92146
0.5	0.94206	0.93689	0.93231	0.93017	0.92903	0.95188	0.92454
0.6	0.94435	0.93991	0.93612	0.93432	0.93289	0.95217	0.92759
0.7	0.94581	0.94257	0.93901	0.93714	0.93606	0.95176	0.93065
0.8	0.94822	0.94559	0.94292	0.94132	0.94029	0.95222	0.93424
0.9	0.94808	0.9465	0.9445	0.94307	0.94183	0.95059	0.93519
1	0.94944	0.94864	0.94722	0.94593	0.94462	0.94997	0.93659

Table 59: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test where Q is the uniform distribution, $n = 27$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.96085	0.96085	0.96085	0.96085	0.96085	0.97597	0.96085
0.1	0.96284	0.96274	0.96268	0.96262	0.96251	0.9771	0.96217
0.2	0.96435	0.96432	0.96407	0.96402	0.96394	0.9771	0.96343
0.3	0.96509	0.96495	0.96461	0.96441	0.96429	0.97646	0.96322
0.4	0.96638	0.96622	0.96601	0.96589	0.96549	0.97582	0.96392
0.5	0.96887	0.9685	0.9682	0.96784	0.96752	0.97728	0.966
0.6	0.97024	0.97002	0.96957	0.96926	0.96896	0.97674	0.96668
0.7	0.97056	0.97027	0.96977	0.96937	0.96895	0.97531	0.96743
0.8	0.97267	0.97237	0.97189	0.97137	0.97098	0.976	0.96847
0.9	0.97405	0.97371	0.97319	0.97269	0.97223	0.97597	0.96951
1	0.9749	0.97439	0.97366	0.97302	0.97243	0.97531	0.96927

Table 60: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test where Q is the uniform distribution, $n = 125$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.1186	0.1207	0.12281	0.12315	0.12373	0.10444	0.12235
0.1	0.12464	0.12666	0.12842	0.12914	0.12904	0.11158	0.12646
0.2	0.13027	0.13214	0.13338	0.13387	0.13427	0.11812	0.13137
0.3	0.13507	0.13621	0.13724	0.13778	0.1378	0.12481	0.13378
0.4	0.13986	0.14115	0.14213	0.14224	0.1425	0.1303	0.13862
0.5	0.14548	0.14679	0.14751	0.14762	0.14729	0.13734	0.14287
0.6	0.15327	0.15443	0.15488	0.15465	0.1544	0.14566	0.14854
0.7	0.15648	0.15739	0.15812	0.15804	0.15779	0.15071	0.15154
0.8	0.16143	0.16216	0.16275	0.16266	0.1622	0.1574	0.15615
0.9	0.16452	0.16498	0.16513	0.16496	0.16446	0.16164	0.1579
1	0.17374	0.17393	0.17396	0.17335	0.17272	0.17214	0.16464

Table 61: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 8$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.32361	0.3283	0.33223	0.33411	0.33509	0.29209	0.33309
0.1	0.34375	0.34788	0.35157	0.35295	0.35404	0.31412	0.35134
0.2	0.36315	0.36685	0.36985	0.37112	0.37153	0.3368	0.36696
0.3	0.37913	0.38274	0.38562	0.38649	0.38621	0.35556	0.38046
0.4	0.39897	0.40147	0.40319	0.40337	0.40266	0.37891	0.39404
0.5	0.41852	0.42078	0.42184	0.42195	0.42102	0.40217	0.41125
0.6	0.43938	0.44112	0.4423	0.44223	0.44151	0.42534	0.43077
0.7	0.45928	0.46026	0.46043	0.45985	0.45901	0.44942	0.44567
0.8	0.47731	0.47801	0.47789	0.47717	0.47592	0.46953	0.46094
0.9	0.49407	0.49432	0.49331	0.49187	0.49057	0.48978	0.47423
1	0.51888	0.51877	0.51736	0.51543	0.51345	0.51773	0.49524

Table 62: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 8$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.7849	0.79005	0.79517	0.79706	0.7979	0.73915	0.79839
0.1	0.80021	0.80543	0.8096	0.81127	0.81205	0.75858	0.81147
0.2	0.82072	0.82546	0.82885	0.83047	0.83162	0.78398	0.83008
0.3	0.83944	0.84317	0.84643	0.8479	0.84833	0.80761	0.84629
0.4	0.85949	0.86271	0.86596	0.86709	0.86762	0.83253	0.86443
0.5	0.87795	0.88066	0.88258	0.88316	0.88347	0.85484	0.88069
0.6	0.89889	0.90074	0.9024	0.90302	0.90293	0.88115	0.89954
0.7	0.91896	0.92011	0.92166	0.92191	0.92171	0.90525	0.91852
0.8	0.93834	0.939	0.93961	0.93984	0.93957	0.92963	0.93592
0.9	0.95756	0.95772	0.95727	0.95685	0.95622	0.95303	0.95193
1	0.97832	0.97787	0.97719	0.97643	0.97562	0.9783	0.97089

Table 63: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 8$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.31384	0.31689	0.31995	0.32119	0.32165	0.2432	0.32263
0.1	0.32866	0.33196	0.33453	0.33523	0.3357	0.26383	0.33494
0.2	0.34823	0.35098	0.35344	0.35398	0.35425	0.29033	0.35077
0.3	0.36689	0.36954	0.37062	0.37054	0.37028	0.31543	0.36529
0.4	0.38306	0.38474	0.3855	0.38502	0.38443	0.33908	0.37751
0.5	0.40209	0.40348	0.40346	0.40289	0.40182	0.36674	0.39283
0.6	0.42147	0.42207	0.42176	0.42085	0.42006	0.39191	0.40828
0.7	0.43558	0.436	0.43577	0.43519	0.43364	0.41374	0.42153
0.8	0.45542	0.45589	0.45515	0.45349	0.4518	0.44124	0.43716
0.9	0.47216	0.47187	0.4711	0.46945	0.46714	0.46445	0.45008
1	0.48878	0.48819	0.48592	0.48358	0.4814	0.48839	0.46243

Table 64: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 27$ and $ncp = 0.5$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.81726	0.82315	0.82804	0.83041	0.83188	0.67122	0.8348
0.1	0.83038	0.8361	0.84083	0.84294	0.8442	0.69732	0.84617
0.2	0.84891	0.85375	0.85765	0.85912	0.86012	0.73121	0.86101
0.3	0.86143	0.86562	0.86951	0.87086	0.8716	0.75888	0.8719
0.4	0.87666	0.88065	0.88369	0.8847	0.88523	0.78826	0.8848
0.5	0.89171	0.89476	0.89745	0.89836	0.89894	0.81974	0.89754
0.6	0.91053	0.91273	0.91404	0.91448	0.91448	0.85237	0.91237
0.7	0.92363	0.92516	0.92598	0.92624	0.92625	0.87905	0.9239
0.8	0.94021	0.94097	0.94133	0.94118	0.94088	0.91196	0.93703
0.9	0.95453	0.95481	0.95451	0.95415	0.9537	0.94016	0.94936
1	0.97086	0.97047	0.96973	0.96884	0.96808	0.97141	0.96293

Table 65: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 27$ and $ncp = 1$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.99934	0.99948	0.99954	0.99956	0.99957	0.97592	0.99958
0.1	0.99951	0.99955	0.99961	0.99962	0.99964	0.97773	0.99965
0.2	0.99954	0.99959	0.99964	0.99969	0.99969	0.98118	0.99971
0.3	0.99971	0.99975	0.9998	0.99981	0.99981	0.9833	0.99982
0.4	0.99974	0.99978	0.99982	0.99983	0.99984	0.98507	0.99987
0.5	0.99971	0.99978	0.9998	0.99981	0.99982	0.98794	0.99983
0.6	0.99985	0.99987	0.99988	0.9999	0.9999	0.99022	0.99992
0.7	0.99978	0.99984	0.99984	0.99986	0.99988	0.99267	0.99988
0.8	0.99989	0.99995	0.99996	0.99997	0.99998	0.99526	0.99998
0.9	0.99993	0.99994	0.99994	0.99994	0.99995	0.99771	0.99996
1	1	1	1	1	1	1	1

Table 66: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 27$ and $ncp = 2$

λ	α_{MS}					MC	AU
	0.025	0.05	0.1	0.15	0.2		
0	0.91144	0.91156	0.91155	0.91156	0.91158	0.70112	0.9116
0.1	0.91742	0.9175	0.91749	0.91747	0.91742	0.72684	0.91723
0.2	0.92504	0.92496	0.92493	0.92484	0.92482	0.75734	0.92455
0.3	0.93249	0.93242	0.93225	0.93218	0.93207	0.78506	0.93163
0.4	0.9398	0.93975	0.93962	0.9396	0.93941	0.81451	0.93859
0.5	0.94806	0.94803	0.94788	0.94777	0.94772	0.84146	0.94682
0.6	0.95472	0.95457	0.95438	0.95415	0.95399	0.87084	0.95306
0.7	0.96573	0.96558	0.96525	0.9651	0.96477	0.90289	0.96357
0.8	0.97176	0.97152	0.97118	0.97084	0.97062	0.92968	0.96923
0.9	0.97977	0.97959	0.97921	0.97888	0.97859	0.95887	0.97709
1	0.98739	0.98718	0.98696	0.98654	0.98632	0.98774	0.98428

Table 67: Powers of the combined procedure and the MC and AU tests where MC is the standard t-test, AU is the WMW test for skewed distributions, $n = 125$ and $ncp = 0.5$

COLOPHON

This document was set in the Palatino typeface using LaTeX and BibTeX, composed with Overleaf (http://www.overleaf.com).