IMPACT OF UNBALANCEDNESS AND HETEROSCEDASTICITY ON CLASSIC PARAMETRIC SIGNIFICANCE TESTS OF TWO-WAY FIXED-EFFECTS ANOVA TESTS

by

LYSON CHAKA

46287582

Submitted in accordance with the requirements for the degree of

MASTER OF SCIENCE

in the subject

STATISTICS

at the

University of South Africa

(UNISA)

Supervisor: Ms S. Muchengetwa

Co-Supervisor: Dr. E Rapoo

November 2016

Acknowledgments

I would like to extend my heartfelt gratitude and appreciation to my supervisor, Ms S. Muchengetwa, who assisted me to carry out and mould this thesis.

I would not have done justice to myself if I fail to appreciate the guidance and support that my coordinator, Dr Rapoo, tirelessly offered me throughout my studies.

I also feel indebted to my wife, Beauty, my family and friends who encouraged and stood by me to the finishing point.

Over and above, glory and honour be to God Almighty, for his mercies, goodness and wisdom that he bestows upon us all the time.

Declaration by Student

I declare that the submitted work has been completed by me, the undersigned and that I have not used any other than permitted reference sources or materials nor engaged in any plagiarism. All references and other sources used by me have been appropriately acknowledged in this work. I further declare that this work has not been submitted for the purpose of academic examination, either in its original or similar form, anywhere else.

Declared on the (date):.....

Signed:.....

Name: Lyson Chaka

Student Number: 46287582

Abstract

Classic parametric statistical tests, like the analysis of variance (ANOVA), are powerful tools used for comparing population means. These tests produce accurate results provided the data satisfies underlying assumptions such as homoscedasticity and balancedness, otherwise biased results are obtained. However, these assumptions are rarely satisfied in real-life. Alternative procedures must be explored. This thesis aims at investigating the impact of heteroscedasticity and unbalancedness on effect sizes in two-way fixed-effects ANOVA models. A real-life dataset, from which three different samples were simulated was used to investigate the changes in effect sizes under the influence of unequal variances and unbalancedness. The parametric bootstrap approach was proposed in case of unequal variances and non-normality. The results obtained indicated that heteroscedasticity significantly inflates effect sizes while unbalancedness has non-significant impact on effect sizes in two-way ANOVA models. However, the impact worsens when the data is both unbalanced and heteroscedastic.

Key words: Fixed-effects analysis of variance, unbalancedness, heteroscedasticity, homoscedasticity, effect size, eta-squared, traditional F-tests, robust tests, normality, outliers, Shapiro Wilk's tests.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background to the study

Analysis of variance (ANOVA) models have been useful and applicable tools in various disciplines other than statistics, especially for experimental design. Lewicki and Hill (2007) argued that ANOVA models have, in general, several advantages over other multivariate techniques in that they are robust and powerful tests in multivariate analysis. ANOVA models are a type of linear models appropriate when dealing with a metric or quantitative (usually continuous) response variable predicted by one or more explanatory factors that are measured on nominal or ordinal scale, and thus are qualitative in nature. With these two factors in consideration, there are basically two investigations that statistical analysts mainly focus on. Firstly, it is the **main effect**, which is the effect of one independent variable or factor on the response variable, averaging over the levels of the other independent variable(s). The second one is the **interaction effect**, which represents the combined effects of the two or more independent variables, called factors, in explaining the dependent measure.

An extension of ANOVA where two or more metric dependent variables are being influenced by one or more individual categorical variables gives rise to multivariate analysis of variance (MANOVA). Instead of performing multiple individual tests for each dependent variable, MANOVA makes it easier to conduct a single overall statistical test incorporating all the dependent variables involved. Sawyer (2009) postulated that, a MANOVA model aims at establishing how the response variable is influenced by the explanatory factors and/or combinations

of these, called factor interactions. The model also aims to investigate the differences in the means between and within factor levels. The interest is on how the altering of these factors could explain the variation in the combinations or interactions of the response variables at the same time. Similarly, MANOVA models have many advantages over several univariate ANOVA models used in isolation since one can collectively test a set of hypotheses of the differences in factor level means. This also considers the correlation between response variables and thus makes better use of the information in data. However, in some cases the investigators are only interested in the effect of one or more independent variables on a single response variable, thus the most appropriate technique to apply is ANOVA instead of MANOVA. The deciding factor is the number and nature of the dependent variables involved in the study. ANOVA is appropriate for one metric dependent variable, whereas MANOVA is best for two or more dependent variables.

In as much as the main focus of analysis of variance procedure is on investigating the main and interaction effects of the factors in question, it is not sufficient just to report that an effect is statistically significant. Brown (2008) argued that reporting the traditional ANOVA source table (with sum of squares due to the source (SS), degrees of freedom (df), mean sum of squares due to the source (MS), the F-statistic (F), and the probability of finding the observed results when the null hypothesis is true (p-value) and discussing the associated significance levels is not the end of the study, but it is just the beginning because we can learn much more by carefully plotting and considering the interaction effects and doing follow up analyses like planned or post-hoc comparisons, power and effect size analysis, and so forth. In support of this, Olejnik and Algina (2003) argued that researchers can improve the presentation of their research findings by supporting their statistical significance test with effect-size measure, which is a standardized index that is independent of sample size but seeks to quantify the magnitude of the difference between populations or the relationship between explanatory and response variables. To augment the significance tests, effect sizes are commonly used to provide important information on how strong the relationship between the variables involved is, if it ever exists (Lakens, 2013).

Nevertheless, ANOVA and MANOVA models have specific assumptions that must be satis-

fied if accurate analysis and results are to be achieved. These assumptions require the data to be normally distributed, homoscedastic (equal or homogeneous population variances) and completely balanced (having the same cell size in each factor combination), which is a rare situation to meet these assumptions in real-life data analysis. Furthermore, several studies which involve comparison of continuous responses variables among a variety of conditions that are discrete do apply analysis of variance (ANOVA), which is most appropriate only when the data conforms to a perfectly or completely balanced design (that is, when there are equal cell sizes). Normally, it is rare, due to various reasons, for a researcher to deal with analysis of data that is completely balanced. Recent study has shown that standard multivariate tests with balanced data, testing for factor effects, produce exact results. Other statisticians have discovered that, with the presence of unbalancedness in data, the tests can be biased, especially when heteroscedastic covariance matrices are involved.

Literature shows that many researchers have in the past tried to alleviate the problems of non-normality, heteroscedasticity and unbalancedness by applying data cleaning techniques like imputation of missing data, and transformation. However, transformations can not be a perfect solution to the problems of non-homogeneous population variances even if the data is somehow normally distributed. Zhang (2012) also supported this argument and tried to rectify the problem by applying the approximate Hotelling T^2 test to one-way MANOVA in the presence of heteroscedasticity. Generally, multivariate analysis of variance is one of the most popular techniques that is used especially when data involved is not normal and heteroscedastic. As noted by Erceg-Hurn and Mirosevich (2008), quite a number of recent powerful statistical techniques which are capable of rectifying the problems involved in assumption violations of classic parametric techniques are in place. However, it is unfortunate that most researchers do not apply these techniques, rather they attempt to apply the multivariate analysis of variance techniques without paying particular attention to the limitations of ANOVA when dealing with heteroscedastic and unbalanced data.

There is great need to conduct a thorough investigation on the effects of heteroscedasticity and unbalancedness on effect sizes in significance tests when dealing with ANOVA data. Kali-

nowski and Fidler (2010) argued that it is a common misconception that statistical significance indicates a large and/or important effect. The crucial message here is that the calculated probability (p-value) is a very limited piece of information, relating to false-positive (type I) error rates only. The same applies to statistical significance, which is merely a statement about the p-value relative to an arbitrary cut-off, so it too relates only to false-positive errors. There is much more to know about a set of empirical data. The best way to determine what went on in a study is to look at the effect size of the study, or consider any other measure that meaningfully summarises what went on in that study.

1.2 Motivation and Contribution

Statistical significance tests including Analysis of Variance (ANOVA), Multivariate Analysis of Variance (MANOVA) and Ordinary Least Squares (OLS) regression, that are widely applied in numerous disciplines have underlying assumptions, especially normality and homoscedasticity, which have to be satisfied. Thompson (2007) emphasized that, as researchers, we need to recognise that if we violate the assumptions of statistical methods, like the homogeneity of variance in ANOVA, we compromise not only our calculated probabilities (p-values) but also our effect estimates. In addition to that, Erceg-Hurn and Mirosevich (2008) supported the argument by pointing to the fact that when these assumptions underlying the parametric significance tests are sufficiently satisfied, the tests produce accurate results.

As a turnaround campaign, one of the most internationally respected statistician, Kirk (2003), painted a portrait of a possible future for a scientific world in which effect sizes play a pivotal role. Kirk (2003) based his argument on the view that the current practice of focusing exclusively on a dichotomous decision strategy of rejecting or failure to reject the null hypothesis test based on p-values is actually impeding scientific progress as well as distracting researchers from real goals. Contemporary research should focus on scientific hypotheses, what data tells us about the magnitude of effects, the practical significance of effects, and the steady accumulation of knowledge.

In line with this notion, Bakeman (2005) indicated that, given a complete explanation of the effect sizes and their applicability in various research designs, many more investigators would probably include them in their statistical reports. The magnitude of effect-size explains how strongly the explanatory variable(s) are related to the response variable. Eta squared (η^2), also equivalent to the usual correlation ratio (R^2) and Partial Eta squared ($\eta^2 partial$) are the basic effect-size measures among the list in ANOVA, some of which will be discussed in detail in the next chapter. Basically, eta squared (η^2) is defined as the ratio of the sum of squares of the effect of interest to the total sum of squares. On the other hand, Partial eta squared is statistically defined as the ratio of the sum of squares of whatever effect is of interest divided by the sum of squares of that effect and its associated error variance. Eta squared has its own disadvantages which the partial eta squared is able to take care of. In simple terms, eta squared and partial eta squared can be formulaically expressed as follows:

$$\eta^2 = \frac{SS_{effect}}{SS_{total}} \tag{1.1}$$

where:

SS_{effect} $\rightarrow$ represents the sum of squares of interest in ANOVA

SS_{Total} $\rightarrow$ represents the total sum of squares

$$\eta^2_{partial} = \frac{SS_{effect}}{SS_{effect} + SS_{error}} \tag{1.2}$$

where:

SS_{effect} $\rightarrow$ represents the sum of squares of interest in ANOVA

SS_{error} $\rightarrow$ represents the error sum of squares associated with the effect of interest.

In his report, Bakeman (2005) demonstrated the fact that in common data analytic approaches like the analysis of variance (ANOVA) with repeated measures, there is a lot of confusion in the choice of appropriate measure of effect size. He proposed the generalised eta squared (η^2_G), as defined by Olejnik and Algina (2003), as a preferred effect measure over eta squared or partial eta squared. The main argument for this effect measure was based on the fact that

it provides easy comparability across between-subjects and within subjects designs as well as being easy to compute from common statistical packages. Bakeman(2005) highlighted that the generalised eta squared is however the same as partial eta squared for factorial designs with between-subjects manipulated factors, even though in other designs the generalised eta squared is less than partial eta squared.

There has been some confusions and debates on which effect-size(s) are applicable in various designs. Most of these effect-size measures are the same but bearing different names. A recent study in mediation analysis, by Wen and Fan (2015), is one example of a disapproval of the recommendation that was previously done by Preacher and Kelly (2011), that the most appropriate mediation effect measure is the kappa squared (κ^2). As defined by Preacher and Kelly (2011), kappa squared is the ratio of the observed indirect effect relative to the maximum possible indirect effect for the given data conditions in the given model.

$$\kappa^2 = \frac{ab}{m(ab)}, \qquad 0 \leq \kappa^2 \leq 1 \tag{1.3}$$

where:

$ab \rightarrow$ represents the indirect mediation effect of predictor variable, X say, on response variable, Y

$m(ab) = m(a)m(b) \rightarrow$ represents the maximum possible value of the indirect mediation effect, and

$c \rightarrow$ represents the direct effect of X on Y

$m(a) \rightarrow$ represents the maximum possible value of a (given the values of b and c)in the mediation model

$m(b) \rightarrow$ represents the maximum possible value of b (given the values of a and c)in the mediation model.

According to Wen and Fan (2015), κ^2 is not an appropriate effect measure because it lacks the property of rank preservation, otherwise it is inversely affected by the mediation effect it represents. As a result, it gives paradoxical results in multiple mediation models. Wen and Fan

(2015) proposed that the traditional mediation effect size measure, P_M (the ratio of the indirect effect to the total effect), together with some other statistical information, should be preferred for basic mediation models. Recently, the effect size aspect has been one of the contentious issues in various statistical areas including the mediation analysis (Fritz, Taylor & MacKinnon, 2012).

With numerous estimates of effect sizes proposed to augment statistical significance tests in literature, these tools work appropriately under the conditions necessary for the design in question. Analysis of variance (ANOVA) is one of the areas in statistics where effect sizes implications are to be fully comprehended. As observed by Olejnik and Algina (2003), these effect sizes do not generalise beyond the limits of the research designs dealt with. This leads us to the fact that some effect measures are preferred over the others depending on such factors like the research design, Type I error probabilities and power of tests. No wonder, Kondo-Brown and Brown (2008) correctly preferred to use partial eta squared because their design was a MANOVA, which by definition involves non-independent or repeated measures.

Despite vast literature on simulation studies comparing type I error probabilities and powers of existing analysis of variance methods, there is need for thorough research on the effects of, and the remedies to the bias arising from such irregularities like unbalancedness and heteroscedasticity in ANOVA models. Many researchers have in the past tried to propose remedies to these individual multivariate diagnostics especially through monitoring and controlling the Type I error rate and power of the tests in one-way ANOVA models. The impact of unbalancedness and heteroscedasticity on effect size has not yet been fully comprehended, and as such, can never be underestimated especially in models such as the two-way fixed-effects ANOVA. It is based on these known and unknown problems of unbalancedness and heteroscedasticity that the motivation to bridge the gap by further investigating their impact on effect sizes on classic parametric two-way ANOVA tests, comparing four different models derived from the same real-life data, has triggered the execution of this thesis.

The research was designed in such a way that investigations on the impact of heteroscedasticity

and unbalancedness on effect size were conducted on a real-life dataset of two-way unbalanced ANOVA design with heterogeneous variances. Three samples were simulated from the original dataset, one that is balanced heteroscedastic, the other balanced homoscedastic, and the third being unbalanced homoscedastic. The primary concern was to establish how the magnitudes effect size are influenced by the presence of heteroscedasticity and unbalancedness in the datasets. The impact of heteroscedasticity and unbalancedness was interpreted based upon the changes in the Eta squared (η^2), Partial Eta squared ($\eta^2_{partial}$) and Omega squared (ω^2) effect size measures. The results of this research will act as a stepping stone and a bridging gap for further research in solutions to assumption violations and meaningful effects of statistical significance tests in multivariate analysis.

1.3 Objectives of the thesis

1.3.1 Objectives

This research tries to establish how the significance tests, effect sizes in particular, of a two-way fixed-effects ANOVA model can be affected if the essential analysis of variance assumptions of homoscedasticity and unbalancedness are violated. The aim is to investigate the impact of heteroscedasticity and unbalancedness on parametric statistical tests of two-way fixed-effects ANOVA model through observing the change in effect size measures. A comparison of the effect sizes was done on a balanced heteroscedastic two-way fixed-effects ANOVA model (simulated from the original dataset), against the unbalanced and heteroscedastic model from the original real-life dataset; the balanced homoscedastic model against unbalanced homoscedastic dataset model through testing the following hypotheses:

1.3.2 Hypotheses

(i) **Hypotheses testing under unbalancedness and heteroscedasticity**

Based upon the effect size benchmarks provided by Cohen (1988), the effect size will be deemed small when $0.01 \leq \eta^2_{partial} < 0.06$, medium when $0.06 \leq \eta^2_{partial} < 0.14$, and large

when $\eta^2_{partial} \geq 0.14$. The same guidelines will be used for Eta squared (η^2) and Omega squared (ω^2).

(1a) $H_{0(A)}$: There is no significant effect size of factor A

$H_{0(A)}$: $\eta^2 < 0.06$ (small effect size, Cohen (1988))

$H_{1(A)}$: Reject H_0 if $\eta^2 \geq 0.06$ (at least medium effect size, Cohen (1988))

(1b) $H_{0(B)}$: There is no significant difference in the measures of effect size on a balanced and unbalanced model

$H_{0(B)}$: $\eta^2_{partial}(balanced) = \eta^2_{partial}(unbalanced)$

$H_{1(B)}$: Reject H_0 if $\eta^2_{partial}(bal)$ - $\eta^2_{partial}(unbal) > 0.06$ (at least medium effect size, Cohen (1988))

(1c) $H_{0(C)}$: There is no significant difference in the measures of effect size on a homoscedastic and heteroscedastic model

$H_{0(C)}$: $\eta^2_{partial}(homosced) = \eta^2_{partial}(heterosced)$

$H_{1(C)}$: Reject H_0 if $\eta^2_{partial}(homosced)$ - $\eta^2_{partial}(heterosced) > 0.06$ (at least medium effect size)

(ii) **Hypotheses on testing Assumptions**

(2) H_0: The covariance matrices are the same (Homoscedasticity)

H_0: $\Sigma_1 = \Sigma_2 = ... = \Sigma_d$,

where d = 1,...,h

(3) H_0: The sample data was from a normally distributed population (Normality)

H_0: $X_h \sim N_p(\mu_h, \Sigma_h)$,

where X_h is h^{th} response variable group

Hypotheses 1(a) - 1(c) are the main hypotheses on which the thesis is based on, and will be dealt with in Chapter 4 and 5, whereas hypotheses (2) - (3), based on assumption tests, will be tested in Chapter 3.

1.4 Overview of Theories

A brief review of the theories and concepts involved in analysis of variance (ANOVA) is presented, with particular focus on the two-way fixed effects ANOVA. The concepts of heteroscedasticity and unbalancedness in ANOVA models are briefly explained.

1.4.1 Two-Way ANOVA

Loeza-Serrano and Donev (2014), define ANOVA as a commonly used traditional statistical technique for investigating how one or more qualitative predictor variables, called factors, affect a continuous dependent variable through considering the mean differences for the explanatory factor categories, known as factor levels. In order to estimate the values of the factor level means, given a certain combination of factor levels, outcomes called replicates must be observed. Equal number of replicates for each factor combination gives rise to balanced data. Usually, unbalanced data is generated when the number of replicates is not the same in factor levels.

The history of ANOVA models can be traced back to the time just after 1920 when Sir Ronald Fisher first used the technique to analyse agricultural and biological experiments. In support of that, Rutherford (2012), confirmed that the method is contained in several statistical packages and has been extensively applied in many other disciplines. However, according to Sahai and Khurshid (2005), ANOVA was primarily applied on balanced data until Frank Yates discovered that the technique can also be used for unbalanced data analysis in the 1930's. This has drawn attention to several researchers and statisticians who attempt to explore the means to address unbalanced data in ANOVA models (Sahai & Khurshid, 2005).

According to Olive (2010), fixed-effects models belong to a family of general linear models comprising of mixed and random effects models. In a two-way fixed effects ANOVA model, the response variable, Y say, is predicted by two categorical factors, D and H say. Sawyer (2009), argued that these two factors involved in fixed-effects models are assumed to be non-random, and their factor levels are not a random sample from a large population of levels. He further suggested that a fixed-effects model is particularly suitable if the main aim is to draw inferences precisely on the factor categories included in the data set of the model being studied.

A natural extension of a simple one-way ANOVA gives rise to a two-way ANOVA with interactions in that two independent variables (factors) effects are considered against one dependent (response) variable. The **interaction effect**, as defined by Loeza-Serrano and Donev (2014), is the means to compare the effect of a combination of two or more factors across their levels on the response variable. Following what several statisticians have proposed before, the two-way fixed-effects ANOVA model with interactions can be modeled as follows:

$$\mathbf{y_{ijh}} = \mu + \alpha_{\mathbf{i}} + \beta_{\mathbf{j}} + \alpha\beta_{\mathbf{ij}} + \epsilon_{\mathbf{ijh}} \tag{1.4}$$

in which y_{ijh} represents the response of the h^{th} replicate on the i^{th} level of factor A and j^{th} level of factor B, i = 1,2,...,k; and j = 1,2,...,m; h = 1,2,...,n. Analogously, μ is the overall mean, α_i and β_j are the main effects of the i^{th} and j^{th} levels of factors A and B respectively; and $\alpha\beta_{ij}$ is the interaction effect of the i^{th} and j^{th} levels of factors A and B. We assume that the ϵ_{ijh} are iid $\sim N(0;\delta_e^2)$, which implies that $E(y_{ijh}) = \mu + \alpha_i + \beta_j + \alpha\beta_{ij}$ and $Var(y_{ijh}) = \delta_e^2$.

The Null hypotheses of the two-way fixed-effects ANOVA can be expressed as given below:

$$\mathbf{H_{0(A)}} : \alpha_{\mathbf{1}} = \alpha_{\mathbf{2}} = ... = \alpha_{\mathbf{k}} = \mathbf{0} \tag{1.5}$$

$$\mathbf{H_{0(B)}} : \beta_{\mathbf{1}} = \beta_{\mathbf{2}} = ... = \beta_{\mathbf{m}} = \mathbf{0} \tag{1.6}$$

$$\mathbf{H_{0(AB)}} : \alpha\beta_{\mathbf{11}} = ... = \alpha\beta_{\mathbf{k1}} = ... = \alpha\beta_{\mathbf{1m}} = ... = \alpha\beta_{\mathbf{km}} = \mathbf{0} \tag{1.7}$$

Hypothesis 1.5 opines the non-existence of factor A main effect, and 1.6 opines the non-existence of factor B main effect, and 1.7 suggests the non-existence of the effect of interaction between factors A and B. According to Olive (2010), the test statistic for the three null hypotheses above follow the F-distribution, an analogy of the general linear hypothesis test statistic.

1.4.2 Unbalancedness

According to Olive (2010), a balanced design consists of all cells of the same size. Zetterberg (2013) further elaborated that data is balanced when there are equally many observations for each factor level combination. Thus unbalancedness can be defined as a situation whereby the cell sizes are not equal across the factor combinations being compared. As noted by Harrar and Bathke (2008), real life data is not balanced in most cases due to various reasons. Some of the reasons that give rise to unbalancedness in data include the research design, research instrument, or missing values due to other reasons beyond the control of the researcher. The case when data is unbalanced must be treated as a separate issue different from a special case for balanced data. The differences between balanced and unbalanced data are more than the similarities thereof.

There are three ways in which balanced data can be marred in two-way ANOVA models. The first situation is when the number of observations for the different factorial combinations are not equal. The second situation is when some cell values (factorial combinations) are completely missing due to several reasons. The third situation is when some response variables have not been measured on other experimental units, especially in multivariate data. Of the three situations, Xu et al (2013) argued that the most common form of unbalancedness in data is the first one, when the number of observations per cell (factor level combination) is not equal. The three types of imbalance are discussed in turn.

(i) Unequal cell values

Unequal number of observations in each cell is the most common type of imbalance sit-

uation in ANOVA data. This can happen due to various reasons. For an example, the unit of study, which can be an organism or human being, might cease to be part of study due to relocation, mortality or other reasons. In practice, ignoring the missing and only analysing the available units will result in biased and incomplete reflection of the effects of factors under study.

(ii) Some responses missing

This type of imbalance normally occurs when the researcher decides not to measure some of the response variables on some individual units. There are so many factors that may affect the collection of response measurements. This might be due to the difficulty to acquire the individual's response, or mortality of the subject during the course of study.

(iii) Some cells missing

This is a situation where by some cells (treatment combinations) are totally missing. This is the most extreme case of imbalance which needs special care in analysis of variance. It can be due to the fact that there was no information observed for other treatments. In the presence of missing cells, it is not appropriate to proceed with inferences using the traditional approaches of dealing with missing data that are presented later.

The traditional approach of imposing balance through deleting some of the observations randomly chosen from the cell with extra data before analysing the reduced dataset has been used by investigators of late (Hair et al., 2014). When the missing observations are few, an easier method is to fill in the gaps with estimates from the data, an approach called imputation (Hair et al., 2014). Contemporary technology has come with statistical packages such as SPSS, R and SAS, which have some methods for computing ANOVA sum of squares designated as Type I through Type IV sum of squares for hypothesis testing as well as the provisions to deal with missing data.

1.4.3 Heteroscedasticity

Heteroscedasticity in ANOVA can be simply defined as a situation where the error terms (also known as variances) are not equal between the groups of the predictor variables being compared. McDonald (2014) asserted that in ANOVA and other parametric tests, homoscedasticity assumption states that within-group standard deviations of the data set are all equal, otherwise they are heteroscedastic. It is the violation of this homoscedasticity assumption that is called heteroscedasticity, that is, when the error terms are unequal across the independent variable values. Moder (2007) argued that violation of this homoscedasticity assumption may cause an increase in Type I error rate which is not statistically desirable.

When the homoscedasticity assumption has been slightly violated, statistics has reliably proven that ANOVA and MANOVA model estimations are robust. Moreso, it has been discovered that the estimation in balanced ANOVA and MANOVA models is robust even with minor deviations from the homoscedasticity assumption. However, it is not always the case that covariances are equal for each factor combination as in balanced data. McDonald (2014) asserted that if the deviation is severe, remedies like data transformation and non-parametric tests might fail to rectify the heteroscedasticity problem.

According to Box (1949), Levene's test can be used to test the homogeneity of variance assumption in univariate ANOVA, whereas Box's M test is used for multivariate analysis. The Box's M test is a statistical procedure used to test for homogeneity of covariance matrices in multivariate analysis. It is basically meant to establish the existence or non-existence of homoscedasticity across the independent variables levels in a multivariate analysis of variance model, whereas Levene's test is applicable to univariate cases.

Very little efforts to alleviate the problems involved in two-way fixed-effects ANOVA models with unequal covariance matrices have been suggested in literature. The current situation implies that a lot of effort and ideas are still needed to unearth the limitations and problems associated with unbalancedness and heteroscedasticity when dealing with univariate and mul-

tivariate data.

1.5 Layout of the thesis

Chapter 1 : Introduction: This chapter dealt with the introduction aspects which include the background of the study, motivation and contribution that triggered the researcher to conduct such a study. A brief overview of the theories involved in multivariate statistical tests analysis was given, with particular focus on the theory of ANOVA, unbalancedness and heteroscedasticity in ANOVA data.

Chapter 2 : Theory and Practice of heteroscedasticity and unbalancedness in ANOVA models: This chapter presents a detailed explanation of the theories applied in ANOVA, which include; the fixed-effects, random-effects and mixed-effects ANOVA models, their assumptions and hypotheses testing. The main focus will be on two-way fixed-effects ANOVA model with interactions. The impact of heteroscedasticity and unbalancedness on two-way fixed effects ANOVA will be reviewed. Lastly but not least, the methods of testing effect size in significance tests will be explained.

Chapter 3: Data Exploration: The original dataset is explored and cleaned for unnecessary information before it is used for analysis purposes. Missing values and outliers are checked and corrective measures undertaken. The three basic assumptions of ANOVA (normality, homoscedasticity and independence of observations) are tested, and remedies applied for any violation of these assumptions.

Chapter 4 : Materials and Methods: This chapter outlines the methodology and techniques used to analyse data. The Six-Stages model building proposed by Hair et al. (2014) will be used to present the thesis report. The main focus of this chapter will be on defining the materials used, research design, testing of ANOVA assumptions, estimating the ANOVA model, testing the model fit and validation of results.

Chapter 5 : Analysis and Discussion of Results: A detailed analysis and discussion of the research findings is dealt with in this chapter. Comparisons of different outputs from the statistical packages, SPSS and R, was used for data analysis and interpretation on both balanced unbalanced and heteroscedastic two-way ANOVA models. The comparisons were based on the methods used for significance testing and on the differences in effect sizes under the influence of heteroscedasticity and unbalancedness in two-way ANOVA.

Chapter 6 : Summary, Conclusions, Limitations of the Study and Areas of Further Research: The chapter presents a brief summary of the study; the conclusions derived from the analysis done; the limitations and constraints that affected the study; and finally, the suggested areas of further research that the research could not fully shed light on.

Chapter 2

Theory and Practice of Heteroscedasticity and Unbalancedness in ANOVA Models

2.1 Introduction

This chapter is a text-book type review of well known documented concepts about ANOVA. Among the concepts that will be reviewed are: the types of ANOVA models and their assumptions ranging from one-way to two-way fixed-effects ANOVA, random-effects and mixed-effects ANOVA models and their associated hypotheses tests; the statistical tools used to measure effect size; the origin of heteroscedasticity and unbalancedness in ANOVA data; the problems involved in the analysis of variance in the presence of heteroscedasticity and unbalancedness in research data; and the methods that have been used to stabilise heteroscedasticity and deal with unbalancedness in ANOVA data.

2.2 Theory of ANOVA

The development of analysis of variance techniques in analysing experimental data is attributed to Sir Ronald Fisher back in the 1920's. Although the ANOVA technique was initially applied to balanced data, it was later discovered that the method could be applied even on unbalanced data (Sahai & Khurshid, 2005). This discovery worked as an eye-opener to several researchers who started to explore other scenarios around ANOVA data like heteroscedasticity. As a result, the method is now available in several statistical packages and has been extensively applied in many disciplines.

Loeza-Serrano and Donev (2014), define ANOVA as a commonly used traditional statistical technique for investigating how one or more qualitative predictor variables, called *factors*, affect a continuous dependent variable through considering the mean differences for the explanatory factor categories, known as *factor levels*. In order to estimate the values of the factor level means, given a certain combination of factor levels, outcomes called *replicates* must be observed. Equal number of replicates for each factor combination balanced gives rise to balanced data. In most cases, the number of replicates varies over factor levels, giving rise to unbalanced data. Gaugler (2008) postulated that the balanced and unbalanced cases in multivariate analysis of variance and ANOVA models is one important concept that must be carefully considered when dealing with model estimation and hypothesis testing.

2.3 Models in ANOVA

There are two types of models that are used to describe the choice of levels of the independent variable in ANOVA, which have essential inferential interpretation drawn from that study: the factor levels can be deliberately chosen by the researcher, which is normally done; or they are randomly selected from some larger set of levels.

If the factor levels are deliberately chosen, based on the researcher's interest, and the levels are a set of all possible choices, then we have a **fixed-effect model** or fixed-factor model, also known as **ANOVA Model I**. The fact that the factor levels are not a random sample from some larger population implies that the inferences made from that model will only be generalisable to the levels involved. On the other hand, if the independent or factor levels were a random set selected from a larger list of levels, then we have a **random-effect model**, also known as the **ANOVA Model II** and the inference drawn from such model can be generalized to the whole population of levels from which the sample of levels was drawn. The third type of ANOVA model is the one which consists of a combination of fixed factor(s) and random factor(s), the **mixed-effects Model**, known as the **ANOVA Model III** type. These are discussed in detail in the next sections.

2.3.1 Fixed-effects ANOVA models

According to Olive (2010), fixed-effects models belong to a family of a large set of general linear models in which the levels of each factor are fixed and not a random sample from the population of levels. The interest of the experiment is the differences in response among these specific levels. Sawyer (2009), argued that a fixed-effects model is particularly suitable if the main aim is to draw inferences precisely on the factor categories included in the data set of the model being studied.

2.3.1.1 One-way fixed-effects ANOVA

One-way ANOVA has only one categorical independent variable or factor, which has two or more (theoretically any finite number) nominal levels called factor levels. Hence, the reason it is termed a single factor analysis of variance. We consider only one independent variable in one-way ANOVA, which divides the subjects under study into two or more levels or groups. However, the hypotheses formulated are based on the means of the groups of the single dependent variable involved, which is measured from the subjects under study, in quantitative and continuous nature.

In simple form, the one-way ANOVA model can be modeled as a means model, with a single response variable related to level means of a categorical independent factor. The one-way ANOVA means model can be expressed as follows:

$$\mathbf{y_{ih}} = \mu_\mathbf{i} + \epsilon_\mathbf{ih} \tag{2.1}$$

where y_{ih} is the response of the h^{th} replicate on the i^{th} level of factor A, i=1,...,k; and h = 1,2,...,n. μ_i is the mean of the i^{th} level of the independent factor , ϵ_{ih} is the random error. We assume that the ϵ_{ih} are iid $\sim$ N(0;σ_e^2), which implies that E(y_{ih}) = μ_i and Var(y_{ih}) = Var(ϵ_{ih}) = σ^2.

When considering the deviation of each factor level from the overall population mean, i.e let

$\alpha_i = \mu_i - \mu$ be the deviation, implying $\mu_i = \mu + \alpha_i$. Substituting the deviations in model (2.1), we formulate the following factor effects model:

$$\mathbf{y_{ih}} = \mu + \alpha_\mathbf{i} + \epsilon_\mathbf{ih} \tag{2.2}$$

where y_{ih} is the response of the h^{th} replicate on the i^{th} level of factor A, i=1,...,k; and h = 1,2,...,n. Analogously, μ is the overall mean, α_i is the main effects of the i^{th} levels of factors A. We assume that the ϵ_{ih} are iid $\sim$ N(0;σ_e^2), which implies that E(y_{ih}) = $\mu+\alpha_i$ and Var(y_{ih}) = σ_e^2.

This study focused on the two-way fixed-effects ANOVA design which the next section is going to talk about.

2.3.1.2 Two-way fixed-effects ANOVA

Two-way ANOVA with interactions is a natural extension of a simple one-way ANOVA in that the effects two independent variables (factors), either in isolation or in combination, influence the response or dependent variable. The means to compare the effect of one factor on the response variable across the levels of the second factor is through observing the interaction effect. In a two-way fixed-effects ANOVA model, the response variable, Y say, is predicted by two categorical factors, X_1 and X_2 say. Sawyer (2009), argued that these two factors involved in fixed-effects models are assumed to be non-random, and their factor levels are not a random sample from a large population of levels. Following what several statisticians have proposed before, the two-way fixed-effects ANOVA model with interactions can be modeled as follows:

$$\mathbf{y_{ijh}} = \mu + \alpha_\mathbf{i} + \beta_\mathbf{j} + \alpha\beta_\mathbf{ij} + \epsilon_\mathbf{ijh} \tag{2.3}$$

where y_{ijh} is the response of the h^{th} replicate on the i^{th} level of factor A and j^{th} level of factor B, i = 1,2,...,k; and j = 1,2,...,m; h = 1,2,...,n. Analogously, μ is the overall mean, α_i and β_j are the main effects of the i^{th} and j^{th} levels of factors A and B respectively; and $\alpha\beta_{ij}$ is the interaction effect of the i^{th} and j^{th} levels of factors A and B. We assume that the ϵ_{ijh} are iid $\sim$ N(0;σ_e^2), which implies that E(y_{ijh}) = $\mu + \alpha_i + \beta_j + \alpha\beta_{ij}$ and Var(y_{ijh}) = σ_e^2.

The hypotheses involved with the two-way fixed-effects ANOVA in 2.3.3 can be written as null hypotheses given below:

$$H_{0(A)} : \alpha_1 = \alpha_2 = ... = \alpha_k = 0 \qquad \textbf{(No factor A effect)}$$

$$H_{0(B)} : \beta_1 = \beta_2 = ... = \beta_m = 0 \qquad \textbf{(No factor B effect)}$$

$$\mathbf{H_{0(AB)}} : \alpha\beta_{\mathbf{11}} = ... = \alpha\beta_{\mathbf{k1}} = ... = \alpha\beta_{\mathbf{1m}} = ... = \alpha\beta_{\mathbf{km}} = \mathbf{0}, \qquad \textbf{(No interaction)} \quad (2.4)$$

Analogous to the general linear hypothesis, the test statistic for the three null hypotheses above follow the F-distribution.

2.3.1.3 Assumptions underlying the fixed-effects ANOVA Model I

ANOVA assumptions which this thesis focused on and tested include:

(i) Normality

This assumption of normality states that the dependent variable, from which the samples are drawn, is normally distributed in each of the groups. It is a theoretical requirement of the distribution of the populations from which the samples are drawn.

($\mathbf{H_0}$): There is no significant deviation from normality for each of the dependent variable's groups/levels.

($\mathbf{H_1}$): There is a significant deviation from normality.

The Shapiro-Wilk's test is a statistical test used to test whether the sample data was drawn from a normally distributed population or not. According to this test, the p-value of the test is compared against a specified level of significance usually denoted as α, and the following rejection criterion is used: Reject H_0 in favour of H_1 if $p < \alpha$, otherwise retain the null hypothesis.

(ii) Homoscedasticity

The assumption of homogeneity of variance requires the variances across the groups of the response variable to be equal. In conjunction with the normality assumption, the homogeneity assumption requires that the distributions in the populations are the same in all dimensions, that is, in means, shapes and variance. Otherwise, there is heteroscedasticity exists when the variances are unequal across the groups.

Levene's test was used to test the violation of the homogeneity of variance tests on both the balanced and unbalanced data sample. The Null hypothesis (H_0: There is equality of covariance matrices) was rejected at 5% significance level if p-value is less than α (5%).

(iii) Independence of observations

The independence of observations assumption requires that the error terms or residual effect ϵ_{ih} are independent from observation to observation. Furthermore, the ϵ_{ih} are randomly and normally distributed.

$\epsilon_{ih} \sim iN(0, \sigma^2)$; where $\mathrm{E}(\epsilon_{ih}) = 0$ and $\mathrm{Var}(\epsilon_{ih}) = \sigma^2$

Residual sequence plots can be used to check correlation of error terms, that is independence. However, in most cases, independence of observation is simply ensured by the nature of design (Hair et al., 2014).

(iv) Outliers

Outliers can be defined as anomalous values in the data set which tend to inflate the sample variance. This increase in sample variance has an inverse influence on the calculated F-Statistic of the ANOVA, hence decreasing the chances of rejecting H_0: Null hypothesis. The Normal Q-Q plots or the box-and-whisker plots can be used to detect the presence of outliers in the research data.

2.3.1.4 Hypothesis testing on one-way fixed-effects ANOVA

The general Null hypothesis tested in one-way ANOVA model is expressed as follows:

$$\mathbf{H_0} : \alpha_\mathbf{1} = \alpha_\mathbf{2} = ... = \alpha_\mathbf{k} = \mathbf{0} \tag{2.5}$$

The null hypothesis assumes that there is no effect of the independent factor on the response variable, against H_1: At least one $\alpha_i \neq 0$. The hypotheses of a one-way ANOVA can be tested by partitioning the total sum of squares into the following components:

$$\mathbf{SS_{TOTAL}} = \mathbf{SS_A} + \mathbf{SS_E} \tag{2.6}$$

where SS_{TOTAL} represents the total sum of squares, SS_A is the sum of squares of factor A and SS_E is the sum of squares of the error terms. These components are used to construct a one-way ANOVA table ($1 \leq i \leq k; 1 \leq h \leq n$) as follows:

Table 2.1: One-way fixed-effects ANOVA

Source of Variation	df	Sum of Squares	E(MS)	F
Factor A	k-1	$\sum_{i=1}^{k} \sum_{h=1}^{n} (\overline{Y}_{i.} - \overline{Y}..)^2$	$\frac{SS_A}{k-1}$	$\frac{MS_A}{MS_E}$
Error	n-k	$\sum_{i=1}^{k} \sum_{h=1}^{n} (Y_{ih} - \overline{Y}_{i.})^2$	$\frac{SS_E}{(n-k)}$	
Total	n-1	$\sum_{i=1}^{k} \sum_{h=1}^{n} \sum_{h=1}^{n} (Y_{ih} - \overline{Y}..)^2$		

There are three sources of variation in one-way ANOVA data. These are: the **factor**, which is the factor of interest in the study; the **error**, referring to unexplained random error; and the **total**, representing the total variation in the data that is associated with the grand mean when the factor of interest is ignored. The degrees of freedom (df) refers to the number of cell means that are free to vary when the grand mean is predetermined. The sum of squares involved are also determined according to the sources of variation, that is, quantifying the variability between the groups of interest (SS_{factor}), variability within the groups of interest (SS_{error}), and the total variability in the observed data (SS_{total}). The average of the sum of squares gives the mean squares (MS) for the factor and error, which are used to calculate the F-statistic as shown in Table 2.1 above.

2.3.1.5 Hypothesis testing on two-way fixed-effects ANOVA

Two-way fixed-effects ANOVA model, in relation to the general hypotheses (2.4) and the conditions for the basic ANOVA model, can assume as the following form:

$$H_{0(A)} : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$$
$$H_{0(B)} : \beta_1 = \beta_2 = ... = \beta_m = 0$$

$$\mathbf{H_{0(AB)}} : \alpha\beta_{\mathbf{11}} = ... = \alpha\beta_{\mathbf{k1}} = ... = \alpha\beta_{\mathbf{1m}} = ... = \alpha\beta_{\mathbf{km}} = \mathbf{0} \tag{2.7}$$

where $H_{0(A)}$ is the main effect hypothesis of the first factor, A: $H_{0(B)}$ is the main effect hypothesis of the other factor B, and $H_{0(AB)}$ is interaction effect hypothesis between the two factors. Analogous to the previous section, the three null hypotheses above use the F-statistics, with the total sum of squares partitioned as follows:

$$\mathbf{SS_{Total}} = \mathbf{SS_A} + \mathbf{SS_B} + \mathbf{SS_{AB}} + \mathbf{SS_{error}} \tag{2.8}$$

Due to the fact that SS_A, SS_B, and SS_{AB} are independent, the three hypotheses can be tested separately. The following ANOVA table outlines the breakdown of the sum of squares involved in a two-way fixed effects ANOVA model ($1 \leq i \leq k; 1 \leq j \leq m; 1 \leq h \leq n$).

Table 2.2: Two-way fixed-effects ANOVA

Source of Variation	df	Sum of Squares	E(MS)	F
Factor A	k-1	$\sum_{i=1}^{k}\sum_{h=1}^{n}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$\frac{SSA}{k-1}$	$\frac{MS_A}{MS_E}$
Factor B	m-1	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(\overline{Y}_{.j.} - \overline{Y}_{...})^2$	$\frac{SS_B}{m-1}$	$\frac{MS_B}{MS_E}$
A*B	(k-1)(m-1)	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(\overline{Y}_{ij.} - \overline{Y}i.. - \overline{Y}_{.j.} + \overline{Y}_{...})^2$	$\frac{SS_{AB}}{(k-1)(m-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	km(n-1)	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(Y_{ijh} - \overline{Y}ij.)^2$	$\frac{SS_E}{km(n-1)}$	
Total	kmn-1	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(Y_{ijh} - \overline{Y}...)^2$		

Of course there are many multi-way fixed-factor or effects (k-way fixed-effects)ANOVA models, but this thesis concentrated on two-way fixed-effects models only. The components of Table 2.2 above are analogous to Table 2.1 except that there are additional sources of variation, factor B and the interaction of the two factors.

2.3.2 Random-effects ANOVA models

So far only the fixed-factor or effects models whose factor levels are specifically determined by the researcher have been presented, in which the main interest is to compare the response effect for only those fixed factor levels. However, in most cases, researchers have to randomly select a sample of factor levels from the entire population of levels, and generalise the analysis results to the entire population of levels.

A random-effect ANOVA model is the one that has a response variable being influenced by one or more random factors which has many possible levels, and the main interest is to compare the differences in the response variable over the entire population levels. As such, inferences are drawn only from the random sample of levels included in the model.

2.3.2.1 One-way random-effects ANOVA

As stated before, the one-way random-effects ANOVA model consists of one independent factor whose levels are a random sample from the entire population of levels, where the interest is in the variability of the response variable over the entire population of the independent factor levels. The variance components model of a one-way random-effects model is given by:

$$\mathbf{Y_{ih}} = \alpha_{\mathbf{i}} + \epsilon_{\mathbf{ih}} \tag{2.9}$$

where $\alpha_i \sim idd\ \mathrm{N}(\mu, \sigma_A^2)$; $\epsilon_{ih} \sim N(0, \sigma^2)$; α_i and ϵ_{ih} are independent. $\mathrm{E}(\mathrm{Y}_{ih}) = \mu$; $\mathrm{Var}(\mathrm{Y}_{ih}) = \sigma_A^2 + \sigma^2$ implying $\mathrm{Y} \sim N(\mu, \sigma_A^2 + \sigma^2)$.

Analogously, letting $\mu_i = \mu + \alpha_i$ in 2.9, we can express the variance component model as a random-effects model as follows:

$$\mathbf{Y_{ih}} = \mu + \alpha_{\mathbf{i}} + \epsilon_{\mathbf{ih}} \tag{2.10}$$

where:

$\alpha_i \sim \mathrm{N}(0,\sigma_A^2)$ are normally distributed independent variables; i=1,...,k and

the $\epsilon_{ih} \sim \mathrm{N}(0,\sigma^2)$ are also iid.

The model can also be separated as:

$Y_{ih} \sim N(\mu, \sigma_A^2 + \sigma^2)$

$\mu \rightarrow$ is the overall mean,

$\alpha_i \rightarrow$ is the effect of the i^{th} random level of factor A, and $\alpha_i \sim$ *i.i.d* $N(0,\sigma_A^2)$

$\epsilon_{ih} \rightarrow$ is the error term.

2.3.2.2 Assumptions underlying the one-way random-effects ANOVA model

Before proceeding with any inferences in ANOVA, it is necessary to check the model assumptions first, otherwise biased conclusions may be obtained. The residual plot can be used to check the random-effects model assumptions since the least squares estimations are the same as those for fixed-effects models. The assumptions that need to be confirmed in random-effects ANOVA model are given below.

(i) Homogeneity of variances

The variance of the data in the groups should be homogeneous, that is, the error terms must be independently, identically and normally distributed,

$\epsilon_{ih} \sim N(0, \sigma^2)$.

Checking for this assumption is analogous to the fixed effects case.

(ii) Normality of random effects α_i

Random-effects ANOVA models are not robust to normality departure, hence, this assumption is important to check. Normal probability (Q-Q) plots can also be used to check for departure, noting that;

$\alpha_i \sim N(0, \sigma_\alpha^2)$ and $\text{Var}(Y_{ih}) = \sigma_\alpha^2 + \sigma^2$.

(iii) Independence of ϵ_{ih}'s from α_i's

This condition is very difficult to check, hence care must be taken when the design and implementation is being chosen. However, heteroscedasticity of the ϵ_{ih}'s can be an indication of violation of the independence assumption.

(iv) Independence of α_i's

The assumption is also difficult to check with the residuals, so care must be taken when the design is used.

2.3.2.3 Hypothesis testing on one-way random-effects ANOVA

The hypotheses that we are concerned with in one-way random-effects model are a bit different from the one-way fixed-effects model:

$$\mathbf{H_0} : \sigma_{\mathbf{A}}^{\mathbf{2}} = \mathbf{0} \qquad \textbf{against} \qquad \mathbf{H_1}; \sigma_{\mathbf{A}}^{\mathbf{2}} \neq \mathbf{0} \tag{2.11}$$

The purpose of these hypotheses is to check the effect of the random factor on the response variable as is the case with the fixed effects model. The composition and layout of the ANOVA table is almost the same as the fixed effects one. The only difference is that the expected mean squares, E(MS_A) now reflect randomness of α_i's, but E(MS_E) remains the same. That is, E(MS_A) $= \sigma_A^2 + \sigma^2$; and E(MS_E) $= \sigma^2$ as usual.

Table 2.3: One-way random-effects ANOVA

Source of Variation	df	Sum of Squares	E(MS)	F
Random Factor A	k-1	$SS_A = \sum_{i=1}^{k}\sum_{h=1}^{n}(\overline{Y}_{i.} - \overline{Y}_{..})^2$	$MS_A = \frac{\sigma^2 + kn\sigma_A^2}{k-1}$	$\frac{MS_A}{MS_E}$
Error	n-k	$SS_E = \sum_{i=1}^{k}\sum_{h=1}^{n}(Y_{ih} - \overline{Y}_{i.})^2$	$MS_E = \frac{\sigma^2}{(n-1)k}$	
Total	n-1	$SST = \sum_{i=1}^{k}\sum_{h=1}^{n}\sum_{h=1}^{n}(Y_{ih} - \overline{Y}_{..})^2$		

The sum of squares in Table 2.3 above are calculated in a similar way to fixed-effects model. Under H_0, it can be seen that $\frac{MS_A}{MS_E} \sim \text{F}_{(k-1),(n-1)k}$, the Fisher distribution with *k-1*, *(n-1)k* degrees of freedom.

2.3.2.4 Two-way random-effects ANOVA

Extending the one-way random-effects model by introducing an additional random factor with a subset of levels randomly selected from the entire population of levels as usual, we have a

two-way random-effects model. This model with both effects random has the following means and random effects model formats:

$$\mathbf{Y_{ijh}} = \mu_{\mathbf{ij}} + \epsilon_{\mathbf{ijh}} \tag{2.12}$$

Letting $\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$, we can decompose the two-way random-effects means model as follows:

$$\mathbf{Y_{ijh}} = \mu + \alpha_{\mathbf{i}} + \beta_{\mathbf{j}} + (\alpha\beta)_{\mathbf{ij}} + \epsilon_{\mathbf{ijh}} \tag{2.13}$$

where $1 \leq i \leq k$; $1 \leq j \leq m$; $1 \leq h \leq n$

$\mu \rightarrow$ is the overall mean,

$\alpha_i \rightarrow$ is the effect of the ith random level of factor A, and $\alpha_i \sim$ *i.i.d* N(0,σ_A^2)

$\beta_j \rightarrow$ is the effect of the jth random level of factor B, and $\beta_j \sim$ *i.i.d* N(0,σ_B^2)

$(\alpha\beta)_{ij} \rightarrow$ is the (i,j)th interaction effect of factors A and B, and $(\alpha\beta)_{ij} \sim$ *i.i.d* N(0,σ_{AB}^2)

2.3.2.5 Assumptions underlying the two-way random-effects ANOVA model

The assumptions that are associated with this model are:

(i) Homogeneity of error terms ϵ_{ijh}'s

The variance of the data in the groups should be homogeneous, that is, the error terms must be independently, identically and normally distributed

$\epsilon_{ijh} \sim$ *i.i.d* N(0,σ^2)

(ii) Normality of random effects

$\alpha_i \sim$ *i.i.d* N(0,σ_A^2)

$\beta_j \sim$ *i.i.d* N(0,σ_B^2)

$(\alpha\beta)_{ij} \sim$ *i.i.d* N(0,σ_{AB}^2)

As stated in the one-way case, the normal probability plots can be used to check these assumption.

(iii) Independence of ϵ_{ijh}'s from α_i's and β_j's

Care must be taken when the design and implementation is being chosen, otherwise heteroscedasticity of the ϵ_{ijh}'s can be an indication of violation of the independence assumption.

(iv) Independence of α_i's and β_j's

Also, this assumption is difficult to check with the residuals, so care must be taken when the design is used.

2.3.2.6 Hypothesis testing on two-way random-effects ANOVA

The hypotheses that we are concerned with in two-way random-effects model are an extension of the one-way random-effects model:

$$\mathbf{H}_0^A : \sigma_A^2 = 0$$
$$\mathbf{H}_0^B : \sigma_B^2 = 0$$
$$\mathbf{H_0^{AB}} : \boldsymbol{\sigma_{AB}^2 = 0} \tag{2.14}$$

The purpose of these hypotheses is to check the effects of the random factors and their interactions on the response variable. The major objective is to extend the test conclusion to the entire population of treatment levels. The composition and layout of the ANOVA table as follows:

Table 2.4: Two-way random-effects ANOVA

Source	df	Sum of Squares	E(MS)	F
Factor A	k-1	$\sum_{i=1}^{k}\sum_{h=1}^{n}(\overline{Y}_{i..}-\overline{Y}...)^2$	$\frac{\sigma^2+nm\sigma_A^2+n\sigma_{AB}^2}{k-1}$	$\frac{MS_A}{MS_{AB}}$
Factor B	m-1	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(\overline{Y}_{.j.}-\overline{Y}...)^2$	$\frac{\sigma^2+nk\sigma_B^2+n\sigma_{AB}^2}{m-1}$	$\frac{MS_B}{MS_{AB}}$
A*B	(k-1)(m-1)	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(\overline{Y}_{ij.}-\overline{Y}i..-\overline{Y}_{.j.}+\overline{Y}_{...})^2$	$\frac{\sigma^2+n\sigma_{AB}^2}{(k-1)(m-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	km(n-1)	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(Y_{ijh}-\overline{Y}ij.)^2$	$\frac{\sigma^2}{(n-1)km}$	
Total	kmn-1	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(Y_{ijh}-\overline{Y}...)^2$		

All the F tests in random models require appropriate test statistics that are determined by appropriate calculations of the expected mean squares. Correct denominators are supposed to be identified to perform the appropriate F tests as shown in Table 2.4 above. The mean

square of the random factors A (MS_A) and B (MS_B) both involve the interaction sum of squares (MS_{AB}) instead of the mean square error (MS_E) as in fixed model. Only the test H_0^{AB}: $\sigma^2_{AB} = 0$ in (2.14) uses the error sum of squares (MS_E) in the test statistic.

2.3.3 Mixed-effects ANOVA model

Sometimes we have a combination of fixed effects factors and random effects factors in a single model. The simplest being a single fixed effect and a single random factor interacting to explain the differences in the response variable.

2.3.3.1 Two-way mixed-effects ANOVA

We now consider a two-way mixed-effects ANOVA model with factor A fixed and factor B random. The model is expressed thus:

$$\mathbf{Y_{ijh}} = \mu + \alpha_{\mathbf{i}} + \beta_{\mathbf{j}} + (\alpha\beta)_{\mathbf{ij}} + \epsilon_{\mathbf{ijh}} \tag{2.15}$$

where $1 \leq i \leq k$; $1 \leq j \leq m$; $1 \leq h \leq n$

$\mu \rightarrow$ is the overall mean,

$\alpha_i \rightarrow$ is the effect of the i^{th} level of fixed factor A, with $\sum_i \alpha_i = 0$

$\beta_j \rightarrow$ is the effect of the j^{th} random level of factor B, and $\beta_j \sim$ *i.i.d* $N(0,\sigma^2_B)$

$(\alpha\beta)_{ij} \rightarrow$ are the $(i,j)^{th}$ interaction effects of factors A and B.

$\epsilon_{ijh} \rightarrow$ is the error term. Since randomness is catching, the interaction between a random and fixed effect ends up random and has a distribution.

2.3.3.2 Assumptions underlying the mixed-effects ANOVA model

The assumptions of this model are that:

(i) Normality assumptions

$\epsilon_{ijh} \sim$ *i.i.d* $N(0,\sigma^2)$

$\beta_j \sim i.i.d$ N(0,σ_B^2)

$(\alpha\beta)_{ij} \sim i.i.d$ N(0,$\frac{k-1}{k}\sigma_{AB}^2$)

(ii) Independence assumption

$(\alpha\beta)_{ij}$ are independent of β_j

ϵ_{ijh} are independent of the β_j and $(\alpha\beta)_{ij}$

$(\alpha\beta)_{ij}$ not in the same column are independent, but those in the same column are dependent.

(iii) Sphericity across the levels of the fixed factor

With the additional restrictions that $\sum_i \alpha = 0$ for each j; and $\sum_i (\alpha\beta)_{ij} = 0$ $\forall$j; $(\alpha\beta)_{ij} \sim$ N(0,$\frac{k-1}{k}\sigma_{AB}^2$) it is important to note that testing the main effect of the fixed factor requires an additional assumption that all the pairwise differences of the fixed factor levels must be of homogeneous variance. This is called the assumption of sphericity across the fixed factor levels.

2.3.3.3 Hypothesis testing on two-way mixed-effects ANOVA

The hypotheses of interest on this model are given by:

$$\begin{aligned} \mathbf{H}_0^A : \alpha_i = 0 \qquad & \textbf{fixed} \quad \textbf{effect} \\ \mathbf{H}_0^B : \sigma_B^2 = 0 \qquad & \textbf{random} \quad \textbf{effect} \\ \mathbf{H_0^{AB}} : \boldsymbol{\sigma_{AB}^2 = 0} \qquad & \textbf{interaction effect} \end{aligned} \tag{2.16}$$

It can be noted that the random effect hypothesis (H_0^B) is tested by the error (F = $\frac{MS_B}{MS_E}$) whilst the fixed effect (H_0^A) is tested by the interaction (F = $\frac{MS_A}{MS_{AB}}$). More detail can be seen in the ANOVA table below.

Table 2.5: Two-way mixed-effects ANOVA

Source	df	Sum of Squares	E(MS)	F
Fixed A	k-1	$\sum_{i=1}^{k}\sum_{h=1}^{n}(\overline{Y}_{i..}-\overline{Y}...)^2$	$\frac{\sigma^2+\frac{(nm)}{(k-1)}\sum\alpha_i^2+n\sigma_{AB}^2}{k-1}$	$\frac{MS_A}{MS_{AB}}$
Random B	m-1	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(\overline{Y}_{.j.}-\overline{Y}...)^2$	$\frac{\sigma^2+nk\sigma_B^2}{m-1}$	$\frac{MS_B}{MS_{AB}}$
A*B	(k-1)(m-1)	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(\overline{Y}_{ij.}-\overline{Y}i..-\overline{Y}_{.j.}+\overline{Y}_{...})^2$	$\frac{\sigma^2+n\sigma_{AB}^2}{(k-1)(m-1)}$	$\frac{MS_{AB}}{MS_E}$
Error	km(n-1)	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(Y_{ijh}-\overline{Y}ij.)^2$	$\frac{\sigma^2}{(n-1)km}$	
Total	kmn-1	$\sum_{i=1}^{k}\sum_{j=1}^{m}\sum_{h=1}^{n}(Y_{ijh}-\overline{Y}...)^2$		

In this case, the F tests in the mixed model that involves the interaction of a fixed factor A and a random factor B would require testing the effects of both the fixed factor and the random factor by using the MS_{AB} as the error term. However, the interaction effect AB is the only one tested with the MS_E as the error term (Table 2.5).

2.4 Heteroscedasticity in ANOVA

In both standard ANOVA and MANOVA models, there is a basic assumption that the samples being dealt with are independent, normally distributed, and are homoscedastic over the levels of factor combinations. Violation of the homoscedasticity, called heteroscedasticity, occurs when the error terms are unequal across the independent variable values. A lot of negative effects and problems are associated with the violation of ANOVA assumptions, resulting in biased and optimistically inflated results. In order to have a strong grip on the concept, a brief review of the components of a heteroscedastic ANOVA model is outlined in the next section.

2.4.1 Heteroscedastic two-way ANOVA model

Analogous to the balanced ANOVA design, we consider a two-way ANOVA model with two fixed factors, factor A with levels i=1,...,k and factor B with levels j=1,...,m. Let Y_{ijh}, i,j and h as defined, be random variable whose observed sample values are y_{ijh}. Also let the sample mean ($\overline{Y}_{ij}$) and sample variance (S_{ij}^2) be defined as follows:

$\overline{Y}_{ij}=\sum_{h=1}^{n_{ij}}\frac{Y_{ijh}}{n_{ij}}$ and $\mathrm{S}_{ij}^2=\sum_{h=1}^{n_{ij}}\frac{(Y_{ijh}-\overline{Y}_{ij})^2}{n_{ij}}$

The two-way heteroscedastic ANOVA model is therefore given by:

$$\mathbf{Y_{ijh}} = \mu + \alpha_{\mathbf{i}} + \beta_{\mathbf{j}} + (\alpha\beta)_{\mathbf{ij}} + \epsilon_{\mathbf{ijh}}, \tag{2.17}$$

where μ is the overall mean; α_i is the effect of the ith level of factor A, β_j the jth level of factor B; whereas $(\alpha\beta)_{ij}$ represents the interaction effect of factors A_i and B_j, and $\epsilon_{ij} \sim N(0,\sigma^2_{ij})$

For coefficients μ, α_i, β_j and $(\alpha\beta)_{ij}$ to be uniquely defined, additional constraints, ω_i and ν_j, are needed.

Suppose ω_i and ν_j, $(1 \leq i \leq k\ ;\ 1 \leq j \leq m)$ are non-negative weights ($\sum_{i=1}^{k} \omega_i > 0$ and $\sum_{j=1}^{m} \nu_j > 0$), we apply the following constraints in model 2.17 above:
$\sum_{i=1}^{k} \omega\alpha_i = 0$; $\sum_{j=1}^{m} \nu_j\beta_j = 0$; $\sum_{i=1}^{k} \omega_i(\alpha\beta)_{ij} = 0$; $\sum_{j=1}^{m} \nu_j(\alpha\beta)_{ij} = 0$

where:

$\omega_i = u_{i.} = \sum_{m}^{j=1} u_{ij}$ and $\nu_j = u_{.j} = \sum_{i=1}^{k} u_{ij}$, with restrictions $\sum_i u_{i.}\alpha_i = 0$ and $\sum_j u_{.j}\beta_j = 0$, with $u_{ij} = \frac{n_{ij}}{\sigma^2_{ij}}$, and $1 \leq i \leq k$; $1 \leq j \leq m$.

2.4.2 Hypothesis testing on heteroscedastic two-way ANOVA

The hypotheses that we mainly focus on, defined for $1 \leq i \leq k$; $1 \leq j \leq m$, are:

$$\mathbf{H}_{0(A)} : \alpha_1 = \alpha_2 = ... = \alpha_k = 0$$
$$\mathbf{H}_{0(B)} : \beta_1 = \beta_2 = ... = \beta_m = 0$$

$$\mathbf{H_{0(AB)}} : (\alpha\beta)_{\mathbf{11}} = ... = (\alpha\beta)_{\mathbf{k1}} = ... = (\alpha\beta)_{\mathbf{1m}} = ... = (\alpha\beta)_{\mathbf{km}} = \mathbf{0} \tag{2.18}$$

$H_{0(A)}$ and $H_{0(B)}$ test the the presence of the main effects of factors A and B, respectively, and $H_{0(AB)}$ tests the presence of an interaction effect between factors A and B, against their usual alternative hypotheses.

For the standardised sum of squares due to factor A, factor B and the interaction sum of squares, and the related p-values for hypothesis testing, reference is made to Arnold (1981); Ananda and Weerahandi (1997); and Fujikoshi (1993). Nevertheless, as advised by Milliken and Johnson (1984), Type III Sum of Squares, readily available in statistical packages SAS and

SPSS, will be used when data in all treatment cells are available but with varying number of observations per cell.

2.4.3 Methods of dealing with heteroscedasticity in ANOVA

When testing the equality of two or more population means, classical F-tests are popularly used. However, due to the fact that these classical F-tests depend on normality, independence and homogeneity of variance assumptions, serious problems may arise especially when the homogeneity assumption is violated. Yiğit and Gokpinar (2010) attested that, proceeding with classical tests in the presence of heteroscedasticity may result in these classical F-tests failing to reject the null hypothesis even if there is enough evidence that an effect exists. This is especially problematic in small samples.

2.4.3.1 Data Transformation

Efforts to alleviate the problems of unequal variances involved in fixed-effects ANOVA models have been recorded in literature recently. When classical F-tests are used, the problems of unequal variances(heteroscedasticity) and non-normality can be addressed by data transformation techniques. According to Hair et al. (2014), heteroscedasticity can be a result of non-normality in one or more of the variables. As a result, correcting non-normality in these variables, through data transformation for example, can remedy the inequality of dispersion of variance. There are various data transformation procedures that can be used which were cited by Hair et al. (2014); Mosteller and Turkey (1977). Only three of the most common data transformations are discussed in this section.

(i) **The logarithmic transformation**

logarithmic transformation involves taking the logarithm of each observed value of the dependent variable. Any base can be used for the log, however, the most common are base-10 and base-e (known as the natural logarithm, where the constant $e = 2.7182818$). It does not matter which base is used because the bases are directly proportional to each other. Log transformations, as they are usually called, are suitable for the variables that are highly skewed. Since the logarithm of a negative number is undefined, a constant

must be added to the values to avoid loss of data

(ii) **The square root transformation**

Square root transformation involves taking the square root of each observed value. Since the square root of a negative number is not real, this technique is normally applied to count variables, such as the number of accidents, cases of theft, bacteria population, which assume only positive values. In case there are negative numbers in the dataset, there is need to disturb the dataset by adding a constant value to all values in order to uplift the negative values to above zero. However, it should be noted that square roots of numbers between 0 and 1 become bigger while square roots of numbers above 1 get smaller. Hence, this transformation is not suitable if the dataset contains a mixture of these.

(iii) **The inverse transformation**

Inverse transformation involves taking the inverse $(\frac{1}{y})$ of the observed variable (y). This kind of transformation changes very large numbers to very small numbers and vice-versa. It actually reverses the order of the data scores. Hence, prior to applying inverse transformation, there is need to reflect (multiply each variable by -1 to reverse the distribution) and then add a constant to uplift the values above 1, and then the inverse transformation will resemble the original data. This type of transformation is most powerful for positively skewed data since it compresses the right side of the distribution. In case of negatively skewed data, it is necessary to reflect, add a constant to uplift to above 1, transform, and then reflect again to restore to original order.

The three main transformations discussed above mitigate non-normality by compressing the data scores on the right of the distribution more than the left side. Basically, this process reduces the spacing between the data values, which is desirable to improve normality, but it has some negative connotations on interpreting the results. Hence, care must be taken when interpreting transformed data.

2.4.3.2 Approximations of test statistics

As another remedy to the problems of classical F-tests and exact procedures when testing the equality of means in the presence of unequal variances, some of the widely used alternatives are the test statistics approximations which surrogate the classical F-tests. These approximations include the Welch (1951) test; Schott-Smith (1971); Brown-Forsythe (1974) test; the Parametric Bootstrap test developed by Krishnamoorthy, Lu and Mathew (2007). According to Yiğit and Gokpinar (2010), these approximations of test statistics used to test the equality of means when population variances are unequal are based on the standardised between-group sum of squares and error sum of squares. A review of the standardised between-group sum of squares; error sum of squares; and the approximations of test statistics for comparing two or more population means in the presence of heteroscedasticity is presented as follows.

(i) The Standardised between-group sum of squares

Assume $Y_{j1}, ..., Y_{jn_j}$ is a random sample from $N(\mu_j, \sigma_j^2)$ j=1,...,k. The equality of means hypothesis concerned is given by:

$$\mathbf{H}_0 : \alpha_1 = \alpha_2 = ... = \alpha_k = \mathbf{0}; \text{ against}$$

$$\mathbf{H}_1\mathbf{: at\ least}\ \alpha_j \neq \mathbf{0}.j = 1, ..., k \tag{2.19}$$

The standardised between-group sum of squares when variances are not the same can be expressed as follows:

$$\hat{T}_A = \hat{T}(\sigma_1^2, ..., \sigma_k^2) = \sum_{j=1}^{k} \frac{n_j}{\sigma_j^2} \overline{Y}_j^2 - \frac{(\sum_{j=1}^{k} n_j \overline{Y}_j \sigma_j^2)^2}{(\sum_{j=1}^{k} n_j/\sigma_j^2)} \tag{2.20}$$

where $\overline{Y}_j = \sum_{h=1}^{n_j} Y_{jh}/n_j$ j = 1,...,k.

The standardised between-group error sums is then given by:

$$\hat{S}_{error} = \sum_{j=1}^{k} \frac{n_j S_j^2}{\sigma_j^2} \tag{2.21}$$

(ii) The Parametric Bootstrap Test (PB)

Parametric bootstrap (PB) can be defined as a process that is used to generate samples using sample statistics estimated from parametric models whose parameters have been replaced by estimates. This method is computer-intensive process that is used to generate samples called bootstrap samples based on the original dataset under study. The null hypothesis that the classical F-tests fail to reject in the presence of heteroscedasticity (variances not equal) can be analysed based on the bootstrap samples to achieve accurate results. Krishnamoorthy, Lu and Mathew (2007), have proposed a parametric bootstrap (PB) test for testing equality of means in one-way ANOVA in the presence of heteroscedasticity and unbalancedness.

As proposed by Krishnamoorthy, Lu and Mathew (2007), assuming δ_j^2 are unknown, a natural test statistic can be derived by replacing σ_j^2 by S_j^2 in the standardised between-group sum of squares in (2.20):

$$\hat{T}_A\ (\mathbf{S}_1^2,....,\mathbf{S}_k^2) = \sum_{j=1}^{k} \frac{n_j}{S_j^2} \overline{Y_j}^2 \text{ - } \frac{(\sum_{j=1}^{k} \frac{n_j \overline{Y_j}}{S_j^2})^2}{(\sum_{j=1}^{k} \frac{n_j}{S_j^2})} \tag{2.22}$$

where $\overline{Y}_j = \sum_{h=1}^{n_j} Y_{jh}/n_j$ and $\mathrm{S}_j^2 = \sum_{h=1}^{n_j} (Y_{jh} - \overline{Y}_j)^2/(n_j - 1)$, j = 1,...,k.

This test statistic in (2.22) is location invariant. Equating the common mean to zero, and letting $\overline{Y}_{B_j} \sim N(0, \frac{S_j^2}{n_j})$ and $\mathrm{S}_{B_j}^2 \sim S_j^2 \chi_{n_j-1}^2/(n_j - 1)$, j=1,...,k and replacing $\overline{Y}$, S_j^2 in (2.22) above by $\overline{Y}_{B_j}$ and $\mathrm{S}_{B_j}^2$ respectively, the parametric bootstrap pivot variable can be expressed as follows:

$$\mathbf{T}_{A_B} = \sum_{j=1}^{k} \frac{n_j}{S_{B_j}^2} \overline{Y}_{B_j}^2 - \frac{(\sum_{j=1}^{k} n_j \overline{Y}_{B_j}/S_{B_j}^2)^2}{(\sum_{j=1}^{k} n_j/S_{B_j}^2)} \tag{2.23}$$

It can be noted that the distribution of $\overline{Y}_{B_j}$ is $\mathrm{Z}_j \frac{S_j}{\sqrt{n_j}}$, where Z_j is a standard normal random variable. Without loss of generality, one can easily verify that the parametric

bootstrap (PB) pivot variable in (2.23) is distributed as

$$\widehat{T}_{A_B}(Z_j, \chi^2_{n_j-1}; S_j^2) = \sum_{j=1}^{k} \frac{Z_j^2(n_j-1)}{\chi^2_{n_j-1}} - \frac{[\frac{\sum_{j=1}^{k}\sqrt{n_j}Z_j(n_j-1}{S_j\chi^2_{n_j-1}}]^2}{\frac{\sum_{j=1}^{k} n_j(n_j-1)}{S_j^2\chi^2_{n_j-1}}} \tag{2.24}$$

For a given α, we reject H_0 in (2.19) if $P[\widehat{T}_{A_B}(Z_j, \chi^2_{n_j-1}; S_j^2) > \hat{T}_{A_0}] < \alpha$

where $\widehat{T}_{A_0}$ is the observed value of T_A in (2.22) above. However, for a fixed set of S_j (*j=1,...,k*), the probability above does not depend on any known parameters, hence it can only be approximated by Monte Carlo simulation as explained by the algorithm given next.

Algorithm 1

- Given the set n_i, $\overline{y}_i$ and S_i^2, for *i=1,...,k*, $\widehat{T}_{A_B}$ can be calculated and call it T_{A0}.
- For *j=1,...,m*, then generate Z_i and $\chi^2_{n_i-1}$ where $Z_i \sim N(0;1)$ for *i=1,...,k*
- Using $\widehat{T}_{A_B}$ in (2.24), compute $\widehat{T}_{A_B}(Z_i, \chi^2_{n_i-1}; S_i^2)$
- If $\widehat{T}_{A_B}(Z_i, \chi^2_{n_i-1}; S_i^2) > T_{A_0}$, set Q_j=1 and terminate the loop.
- $p = \frac{1}{m}\sum_{j=1}^{m} Q_j$ is the Monte Carlo estimate of the p-value α given above.

(iii) The Welch's Test (1951)

This test is a generalisation of the Behrens-Fisher problem, when testing equality of only two means is involved. Welch's (1951) test came as the first solution to solve the problems of unequal error variances when comparing means in one-way ANOVA. It is basically a form of one-way ANOVA that does not assume equal variances. The test is based on the Student's t distribution with degrees of freedom depending on both sample size and sample variances.

Considering equation (2.24), and letting $\omega_j = n_j/S_j^2$, j = 1,...,k, Welch (1951) derived a test statistic given by:

$$W_f = \frac{\hat{S}_a(S_1^2, ..., S_k^2)/(k-1)}{1 + \frac{2(k-2)}{k^2-1}\sum_{j=1}^{k}\frac{1}{n_j-1}(1-\frac{\omega_j}{\sum\omega_i})^2} \sim F_{(k-1,p)} \tag{2.25}$$

where $F_{(q,p)}$ is an F-distribution with (q,p) degrees of freedom, and $\hat{S}_a$ is given in (2.24) above, and

$p = [\frac{3}{k^2-1}\sum_{j=1}^{k}\frac{1}{n_j-1}(1-\frac{\omega_j}{\sum\omega_i})^2]^{-1}$

The test rejects H_0 given in (2.19) at a significance level α when the p-value $P(F_{k-1,p} > W) < \alpha$, for an observed ω of W.

The Welch test is built for non-homogeneous variances, but with the assumption of normality satisfied. It is considered a more robust and conservative statistic than other tests like the Student's t-tests in the presence of unequal population variances and unequal sample sizes in the sense that it can maintain the type I error rate close to nominal. This test achieves this robustness over the traditional F-test because it adjusts the denominator of the F ratio such that, despite the heterogeneity of the group variances, it has the same expectation as the numerator when the null hypothesis is true (see equation 2.25 above).

(iv) The Brown-Forsythe Test (1974)

This is also known as the Brown-Forsythe F-ratio. The test is appropriate when both the normality and homogeneous variance assumptions have not been satisfied. It also modifies the denominator of the traditional F-test in ANOVA, making it more robust than the classical F-test when ANOVA assumptions are violated.

Given the null hypothesis H_0 in (2.19), Brown and Forsythe (1974) proposed the test statistic:

$$\mathbf{B} = \Sigma_{j=1}^{k}\frac{n_j(\overline{Y}_j - \overline{Y})^2}{\Sigma_{j=1}^{k}(1-\frac{n_j}{n})\widehat{S}_j^2} \tag{2.26}$$

Under H_0, B has an F-distribution, $F_{k-1,p}$, where

$\mathbf{p} = \frac{[\Sigma_{j=1}^{k}(1-\frac{n_j}{n})S_j^2]^2}{\Sigma_{j=1}^{k}\frac{(1-\frac{n_j}{n})\widehat{S}_j^4}{n_j-1}}$

Under H_0 in (2.19), and given the value of α level, and an observed value B_s of B, the Brown-Forsythe rejects H_0 whenever the p-value is $P(F_{k-1,p} > B_s) < \alpha$.

The Brown-Forsythe F-test is a test of equal population variances that is robust based on the absolute differences within each group from the group median. Unlike the traditional F-test which is divided by the mean square error, the Brown-Forsythe test adjusts the mean square error (denominator) by using the observed variances of each group (equation 2.26). This gives it an edge over the classical F-test.

It is worthwhile to note that most of these robust ANOVA techniques (like Welch and Brown-Forsythe tests) are only available in one-way analysis of variance. However, if they are to be applied in two-way ANOVA, for an example in SPSS, the main factor effects can be tested by means of fixing one independent variable (factor) and assume it constant while the means of the other factor are compared using the usual one-way ANOVA process.

2.4.4 Case studies dealing with heteroscedasticity in ANOVA tests

Zhang (2015a) conducted a study in trying to use a parametric bootstrap approach (PB) to find solutions on one-way ANOVA in the presence of heteroscedasticity and unequal group sizes without using transformation of data technique. Based on the parametric bootstrap test proposed by Krishnamoorthy, Lu and Mathew (2007), Zhang (2015b) further extended the PB algorithm to a multiple comparison procedure (MCP) to test the equality of factor level means and to do pairwise comparisons of two-way ANOVA in the presence of unequal group sizes and heteroscedastic variances.

Simulation studies pertaining to this effect showed that, under heteroscedasticity assumption, the parametric bootstrap test proved to be one of the best technique for testing equality of factor level means. Furthermore, Zhang (2015a) proposed a parametric bootstrap test for multiple comparison in one-way ANOVA when error variances and group sizes vary. The research showed that a complete solution can be achieved when the proposed parametric bootstrap test

of multiple comparison is used together with the Krishnamoorthy, Lu and Mathew (2007) PB test. The simulation results achieved showed that the multiple comparison procedure and the Type I error of overall test were both close to nominal level.

Moreover, Zhang (2015b) had another study where a parametric bootstrap approach for simultaneous confidence intervals was proposed for all pairwise multiple comparisons in a two-way unbalanced design with unequal variances. Similarly, simulation results depicted that the Type I error of the multiple comparison test were close to the nominal level, even for small samples. The proposed method performed better than the Turkey-Kramer procedure under heteroscedastic variances and unequal group sizes.

In another study, Xu et al. (2015) proposed a parametric bootstrap (PB) test to compare it with the generalised F (GF) test for testing equal effects of factors of a two-way ANOVA model without interaction in the presence of heteroscedasticity. They used the Monte Carlo simulation to evaluate the powers of tests and the Type I error rates. In their study, it was discovered that, in the presence of heteroscedastic error variances and/ or as the number of factor levels increases, the classical F-test and the generalised F-test yield to serious Type I error properties. However, with the use of the parametric bootstrap (PB) test, the Type I error problems are kept under control. As a result of their research, Xu et al. (2015) concluded that, the parametric bootstrap (PB) test performs satisfactorily better than the generalized F (GF) test in two-way fixed effects models under heteroscedasticity, regardless of the number of factor levels involved, sample sizes or error variance values.

Xu et al. (2013), in their article, considered a two-way ANOVA model with unequal cell frequencies without the homoscedasticity assumption. They proposed a parametric bootstrap (PB) approach for testing main and interaction effects, and comparing it with the generalised F (GF) test. As usual, the Monte Carlo simulation was used to evaluate the The Type I error rates and powers of the tests. Their studies showed that the parametric bootstrap test performed satisfactorily better than the generalised F-test, even for small samples. As in the earlier studies reviewed, the results of their study indicated that the generalised F test portrayed poor

Type I error properties especially when the number of factor levels or treatment combinations increased.

Wang and Akritas (2011) developed an asymptotic theory for hypotheses testing in high-dimensional analysis of variance (HANOVA) in which the distributions are not specified at all. Most results in the literature have been restricted to observations of no more than two-way designs for continuous data. Wang and Akritas (2011) formulated a way that allowed the response variable to be either continuous, discrete or categorical. They developed an asymptotic theory to test the main and interaction effects of up to the third order in unbalanced designs with unequal error variances, arbitrary number of factors and unequal sample sizes, using two types of test statistics; one with χ^2 distribution to test low-dimensional parameters; and other with a limiting normal distribution for testing high-dimensional parameters.

Simulation results carried on the Arabidopsis Thaiana gene expression data show that the proposed test statistics performed well in both continuous and discrete HANOVA in terms of type I error accuracy, computing time and power. The ANOVA F-test was affected by unbalancedness and heteroscedasticity. The proposed test portrayed proved to be more powerful, producing reliable type I error rates as well as being computationally user-friendly when compared to the traditional HANOVA methods.

Gaugler and Akritas (2013) proposed a modification in the F-Statistic in testing the significance of the main random effects in two-factor random and mixed effects designs. Under the new test procedures that Gaugler and Akritas (2013) proposed, the symmetry assumption was not made, that is, the interaction term was not assumed independent from the main effect even though the two are uncorrelated in the random effects model. They based their asymptotic theory of deriving adjusted F-statistics based on the Neyman-Scott framework taking the notion that the number of factor levels in both factors can be large whereas the sizes of the groups can remain constant. As such, the test statistics can be derived by considering the difference of suitably defined mean squares (MSB-MSE* for the mixed effects and MSB-MSAB for the random effects, say) instead of the usual ratio, MSB/MSE.

Using these newly proposed test statistics under fully nonparametric models, the simulations done proved beyond doubt that these proposed statistics performed sufficiently well in situations where classical F statistic seems to violate the underlying assumptions, especially balancedness, symmetry and homoscedasticity.

On a different occasion, Zhang (2012) proposed a simple and accurate approximate degrees of freedom (ADF) test to address the problem of heteroscedastic two-way ANOVA. This attempt came as a means of amending the bias of blindly employing classical F-tests especially when the ANOVA model is heteroscedastic. In the study, Zhang (2012) noted that simulations reflected that ADF test produces good results in different cell sizes whereas the classical F-tests perform badly in the presence of heteroscedasticity.

All in all, recent study shows that in the presence of heteroscedasticity, F-tests suffer from lack of power, resulting in serious biased conclusions. In an empirical study conducted by Moder (2007) on ANOVA problems, the assumption of equal variance seemed to be more problematic in ANOVA models with wider ratios of standard deviations.

2.5 Unbalancedness in ANOVA Data

In simple terms, a balanced ANOVA design is one that consists of cells or factor combinations of the same size. Having more observations in some factor combinations gives rise to more information on the effect of those factor level combinations than for other cells with fewer observations. Consequently, the factor levels can not necessarily be independent, hence the tests and estimates of the effects are eventually not independent too. This lack of balance distorts the potential of an experiment to achieve the intended accurate results. Several methods to alleviate this problem have been proposed in literature,(Xu et al., 2013) however, choosing the most appropriate method is usually not an obvious task.

Literature in unbalanced data design was first realised in the mid 1930's. Gaugler (2008) argued

that, from a theoretical point of view, designing linear models from unbalanced data, and finding suitable ways of deriving inferences from them is still not fully comprehended. Following the same argument, Larson (2008), emphasised that the lack of balance in one-way or two-way ANOVA analysis may cause serious problems if the investigator did not choose an appropriate statistical package to handle the calculations.

2.5.1 Unbalanced two-way ANOVA model

Expressing the unbalanced ANOVA model design as a linear model has been a subject of debate over a long time since its introduction in the 1930's. Analogous to the balanced design model, the two-way unbalanced fixed effects ANOVA model can thus be expressed as follows:

$$\mathbf{Y_{ijh}} = \mu + \alpha_{\mathbf{i}} + \beta_{\mathbf{j}} + (\alpha\beta)_{\mathbf{ij}} + \epsilon_{\mathbf{ijh}} \tag{2.27}$$

where model assumptions and notation are the same as those in the balanced model design except that h = 1,...,n_{ij} , where n_{ij} represents the $(i,j)_{th}$ replicate of factor A_i and B_j; 1 $\leq i \leq k$; 1 $\leq j \leq m$.

2.5.1.1 Testing hypotheses in unbalanced two-way fixed-effects ANOVA model

Harrar and Bathke (2008) proposed that, unlike in balanced data, special precautions must be taken when dealing with unbalanced data. It is worthwhile to note that due to the fact that the independence of the sum of squares of both interaction and main effects in unbalanced data is affected, testing these effects in the two-way fixed-effects ANOVA model calls for an approach different from balanced data model. After realising the inexactness of F-tests in unbalanced data, Zhang (2012), proposed that there is need for modifying the procedures of determining the degrees of freedom and the effect of sum of squares. To this effect, Zetterberg (2013) supported the idea that the four methods of partitioning sum of squares for factors in ANOVA models, Type I, Type II, Type III and Type IV, be implemented. Type III sum of squares is adjusted for all other effects in the ANOVA models, hence, this is going to be used for unbalanced data with varied cell values.

Analogous to the balanced two-way fixed-effects ANOVA model, we can state the hypotheses for unbalanced data as follows:

$$\mathbf{H}_{0(A)} : \alpha_1 = \alpha_2 = ... = \alpha_k = \mathbf{0}$$

$$\mathbf{H}_{0(B)} : \beta_1 = \beta_2 = ... = \beta_m = \mathbf{0}$$

$$\mathbf{H}_{0(AB)} : \alpha\beta_{11} = ... = \alpha\beta_{k1} = ... = \alpha\beta_{1m} = ... = \alpha\beta_{km} = \mathbf{0} \tag{2.28}$$

$H_{0(A)}$ and $H_{0(B)}$ test the the presence of the main effects of factors A and B, respectively, and $H_{0(AB)}$ tests the presence of an interaction effect between factors A and B against their usual alternative hypotheses.

2.5.2 Methods of imposing balance on unbalanced data

(i) Deleting observations

Applying the traditional approach to amend the problems of unbalancedness in statistical data, investigators in the past used to impose balance through deleting observations randomly chosen from the cells with extra data before analysing the reduced data-set (Hair et al., 2014). This approach may be statistically attractive, but the problem is that it reduces accuracy of the model estimates and the power of the hypothesis tests. Hence, it is not recommended because it leads to loss of essential information depending on the eliminated observations.

(ii) Imputation

Another alternative is to impute (fill in estimated values from the data) especially when the missing observations are few, and use standard ANOVA for balanced data (Hair et al., 2014). However, it might look nice that the imbalance has been treated, parameters correctly estimated, but the significance tests produced are flawed. A more powerful and robust imputation method to deal with missing data is multiple imputation, which involves creating several different copies of imputed datasets and appropriately combining the results from each dataset. This method has an edge over other imputation methods in that it takes into consideration the variability in results between the imputed datasets,

at the same time showing the uncertainty associated with the missing values. Recent statistical packages, like SAS, SPSS and R, come with various provisions to deal with imbalance in ANOVA data, and a number of methods for computing ANOVA sum of squares and testing hypotheses designated as Type I through Type IV sum of squares.

When dealing with missing data, Milliken and Johnson (1984) advised that a great deal of thought is needed, statisticians should therefore avoid the practice of simply run a computer program on the data and then select number to include in the report.

2.5.3 Dealing with unbalancedness in ANOVA tests

The term "unbalanced" is not easy to define precisely. As highlighted earlier on, there are three situations that we consider when defining unbalancedness in ANOVA data. Whichever the case might be, the key issue is that, unbalanced data affect the grand mean and the effect mean, which are the basis of group means comparisons and ultimately the factor effects to be detected. Complexities arise when some of the factors under study are considered random, whereas with fixed effects, the challenge can be solved by making use of appropriate Sum of Squares (designated as Type I through Type IV).

Dealing with unbalanced data in ANOVA often presents various problems. However, some of these challenges can be mitigated by comprehensive understanding of the methods and assumptions involved. In as much as various methods of imposing missing data in unbalanced data are available for use, negative impacts, like loss of essential information (when some values are eliminated), or producing biased significance tests (when imputation is used) are a cause of concern. Statistical computational methods specifically designed for unbalanced data in are preferred (Milliken & Johnson, 1984).

Some of the recent efforts to curb the problems of unbalanced data include Zetterberg (2013) study, who used two numerical examples in order to establish the advantages of newly modified tests against standard tests in multivariate analysis of variance (MANOVA). The research

results indicated that the deviations between the tests were significantly smaller on balanced data, whereas significant discrepancies could be seen on unbalanced data, in favour of modified tests.

Prior to Zetterberg's study, Zhang and Xiao (2012) had previously proposed some kinds of adjustments on the standard test statistics with the aim of accommodating unbalancedness and heteroscedasticity of covariance matrices in unbalanced MANOVA data. In their study, they discovered that MANOVA model was robust when balanced data is involved especially in the presence of a slight violation to the homoscedasticity assumption. However, Zhang and Xiao (2012) further concluded that bias in standard tests grows with the severity of heteroscedasticity. The aim of their thesis was to use adjustments of the standard multivariate tests to protect against this bias. Zhang and Xiao (2012) subsequently proposed modifying the Wilks' Λ, Hotelling-Lawley Trace and Pillai's Trace as a means to improve unbiased results in the presence of unbalancedness and heteroscedasticity.

Even though we have some ways of treating unbalanced data using available Statistical computation methods, these packages come with their shortcomings. As a result, researchers must be very cautious especially when using F-tests in any kind of unbalanced data. The F-tests are just approximations which are severely affected by the degree of imbalance in the data and type of factors. This is a potential study gap that has to be filled in future.

2.6 Impact of heteroscedasticity and unbalancedness in ANOVA

Dealing with unequal error variances in ANOVA tests has been a serious challenge that has been overlooked (Krishnamoorthy, Lu & Matthew, 2007). As noted by Krishnamoorthy, Lu and Matthew (2007), one of the main problems caused by heteroscedastic error variances is the increase in the type I error rate in both one-way fixed effects and one-way random effects models especially when testing the significance of the main and interaction effects. After comparing various approaches to control the Type I error rate, the mentioned authors proposed the use of

the parametric bootstrap approach, instead of the Welch test, generalised F-test or the James (1951) second order test. Their tests results showed that the parametric bootstrap approach could tame the type I error rate satisfactorily well, closely above the nominal level 0.05 in the presence of heteroscedasticity regardless of the values of error variances, sample sizes or the number of means compared.

Olejnik and Algina (2003) argued that the estimation of effect sizes depends largely on the type of research design involved. Based on previous researches done on this matter, unbalanced designs have little impact on effect size estimations, whereas with the combination of an unbalanced design and heterogeneous variances, the effect sizes tend to be overestimated. Empirical studies by several authors have shown that the standard errors of effect size measures like the eta squared and omega squared tend to grow large in small sample sizes in the presence of heteroscedasticity. The magnitude of the effect size is generally reduced by population heterogeneity (Olejnik & Algina, 2003).

Moreso, a study conducted by Wang and Akritas (2006) in nonparametric tests investigating the effect of unbalancedness and heteroscedasticity on the main effects of the models involved, established that the p-values (calculated probability of getting the observed results when the null hypothesis is true) increase as a result of the disturbance in the variances. This was also evidenced by the disturbance in the classical F-tests due to the prevalence of unequal variances. Increase in p-values leads to rejection of the null hypothesis when in fact it is true, that is Type I error. The impact was worse when the model is both unbalanced and heteroscedastic.

Kesselman et al. (2008) postulated that, in the ANOVA context, the presence of heteroscedasticity, coupled with skewness and/or outliers, can lead to devastating depressed Type I error rates, decreased power to detect effects and inappropriate probability coverage for confidence intervals (CIs). On another note, when unequal variances couples with unequal sample sizes, the performance of classical F-tests is heavily compromised. Possible solutions have been proposed in recent researches that, in situations like this, it is advisable to adopt the non-pooled test statistics with trimmed means, like the Welch (1951); Brown and Forsythe (1974), and

others, since they do not pool across heterogeneous sources of variations.

Most of the past and recent researches in this area focused on the impact of non-homogeneous error variances on power analysis and comparison of tests in an attempt to control the type I error rate in various models, especially the one-way ANOVA. Other studies concentrated on the behaviour of different procedures in the presence of heteroscedasticity, non-normality and unbalanced group sizes. The current study will be useful to future researchers as it tries to further unearth the behavior of two-way fixed effects ANOVA models, with particular interest in changes in effect sizes, under the influence of non-homogeneous error variances and unbalancedness.

2.7 Calculating Effect Size

In addition to hypotheses testing and statistical significance tests, researchers are interested in testing and estimating effect sizes. Effect size is a term that generally refers to a family of numerical indices that quantify the magnitude of a treatment effect. Erceg-Hurn and Mirosevich (2008) defined an effect size as a measure that gives information about the magnitude of an effect, which determines whether the effect is of practical significance or not. Furthermore, Nandy (2012) defined effect size as simply a way of quantifying the size of the difference between two groups. It measures the strength of the relationship. Basically, there are two ways effect sizes are measured, resulting to two classes of effect sizes: the standardised mean difference effect sizes measured between two means (examples include Cohen's d, Hedge's g, Glass's delta); and the proportion of variance effect sizes measured as the correlation between the explanatory variable classification and the scores of the response variable (e.g Eta Squared (η^2), partial Eta Squared ($\eta^2_{partial}$), Epsilon Squared (ϵ), Omega Squared (ω^2), intra-class correlation (r_i). Each of these two classes of effect size has a number of univariate and multivariate types.

Effect sizes in analysis of variance measure the magnitude of association between the grouping variable (factor) and the dependent variable through the main and interaction effects. The commonly used effect-size measures in ANOVA include Eta Squared (η^2), partial Eta Squared

($\eta^2_{partial}$), Omega Squared (ω^2) and the intra-class correlation (r_i). There are various effect sizes suitable for different designs and experiments, readers are referred to Nandy (2012), Keselman et al. (2008); Algina, Keselman and Penfield (2006), for more detailed information.

However, it has been discovered that most of the effect size measures used in Statistics are robust to violation of normality and homoscedasticity assumptions. According to Erceg-Hurn and Mirosevich (2008), if parametric assumptions are violated, it is not statistically wise to report standard effect sizes (nor confidence intervals) because the degree of confidence would be biased. The two authors further argued that it is unfortunate that most of the commonly used standard parametric effect sizes, like the Cohen's d and η^2, are estimated under these restrictive assumptions. The next section reviews the effect sizes that are commonly used for standardised mean differences and proportion of variance between groups.

2.7.1 The Standardised Mean Difference

This is the most popular effect size measure suitable for interpreting the magnitude of a treatment, a contrast between any give two treatment groups, or any other numerical comparison. The most common contrast is the mean difference, $(\mu_i - \mu_j)$, where mean μ_i and mean μ_j are not the same. The mean difference effect measure depends on the scale of measurement used to compute the means of the variable of interest. This scale-dependent problem can be overcome by standardising the mean difference. Several standardised mean differences have been proposed in literature, from univariate standardised mean differences (when the contrast is for one response variable) to multivariate standardised mean differences (contrast applied on several response variables).

A simple univariate **standardised mean difference** is given by:

$$\delta = \frac{\mu_{\mathbf{1}} - \mu_{\mathbf{2}}}{\sigma} \tag{2.29}$$

The Standardised Mean Difference is simply the difference between the two means (μ_1 - μ_2) from the groups being compared, divided by the standard deviation (σ) of the population from which the two groups were sampled. Being standardised means that the effect-size measure can

be used to compare effect-sizes across different tests measured on different variables, or even measured in different scales of measurement. If the population variance (standard deviation) is not known, the standard deviation can be estimated in a number of ways giving rise to different effect-size indices discussed below. One of the simplest way to estimate δ is by replacing μ_1 and μ_2 with the means of group A and B respectively, σ estimated by sample standard deviation (SD).

2.7.2 The Cohen's (1965) *d*

The Cohen's d effect size estimator assumes equality of variances, otherwise it is biased. It standardizes the effect-size of the difference between two means, such that the difference between the two means is "d" standard deviation (Cohen, 1965). Cohen's d, a variant of the Standardised Mean Difference, is given by:

$$d = \frac{(\overline{\mathbf{X}}_\mathbf{1} - \overline{\mathbf{X}}_\mathbf{2})}{\mathbf{S}_\mathbf{pooled}} \tag{2.30}$$

where $\mathrm{S}_{pooled} = \sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2}} \approx \sqrt{\frac{(SD_1^2+SD_2^2)}{2}}$ is referred to as the standardiser.

Cohen's d effect size is suitable for two sample independent groups with homogeneous or approximately equal variances. This justifies the assumption that the two sample standard deviations estimate the same population standard deviation. Hence, a common standard deviation is pooled from the two standard deviations. Otherwise, with different sample standard deviations, pooling a common standard deviation to estimate Cohen's d is inappropriate.

A guideline on the interpretation of Cohen's d is as summarised below by Cohen (1992):

Table 2.6: Guideline on Cohen's d

Effect Size	d-Standardized mean difference	Percentage of variance explained
Small	0.20	1%
Moderate	0.50	10%
Large	0.80	25%

According to Cohen (1992), this guideline is not a set of hard-and-fast rule, it should just be used as a guideline. It is advisable to use these benchmarks based on the meaningful context and after assessing all the contributing factors that may affect the interpretation of the study

in question. An effect size $d \leq 0.20$ may be considered small, $d = 0.50$ is moderate, while $d \geq 0.80$ is deemed large.

2.7.3 Glass's Δ

In a situation where the two groups standard deviations are very different (heterogeneous), the Cohen's pooled standard deviation does not apply. The appropriate procedure of estimating the pooled standard deviation when the group variances are heterogeneous was proposed by Glass, McGaw and Smith (1981). One of the populations' standard deviations (control group) can be used as the standardizer to estimate a non-pooled standardizer for the effect size as follows:

$$\boldsymbol{\Delta} = \frac{\hat{\mu}_{\mathbf{1}} - \hat{\mu}_{control}}{\hat{\sigma}_{control}} \tag{2.31}$$

The control group standard deviation is inserted as the standardizer of the difference between the treatment group mean (μ_1) and the control group mean ($\mu_{control}$) to estimate the Glass Δ.

Alternatively, in the presence of unequal variances, Kulinska and Staudte (2006), proposed a pooled standardizer of a weighted sum of group variances and modified the estimator. The estimator that they came up with depends on the sample size:

$$d = \frac{\hat{\mu_{\mathbf{1}}} - \hat{\mu_{\mathbf{2}}}}{\sqrt{\frac{\mathbf{n_1}\hat{\sigma}_{\mathbf{1}}^{\mathbf{2}} + \mathbf{n_2}\hat{\sigma}_{\mathbf{2}}^{\mathbf{2}}}{\mathbf{N}}}} \tag{2.32}$$

The effect size estimator works in the same way as the other standardized mean difference estimators. The last estimator in this section looks at the the situation when the two groups are of different sample sizes (unbalanced)

2.7.4 The Hedges' *g*

This is another standardised mean difference type of effect-size measure that is suitable for two groups of different sample sizes. The estimator is expressed as follows (Hedges, 1981):

$$\textbf{Hedges}' \; g = \frac{\hat{\mu_1} - \hat{\mu_2}}{\sqrt{\frac{(\mathbf{n_1}-\mathbf{1})\mathbf{SD_1^2}+(\mathbf{n_2}-\mathbf{1})\mathbf{SD_2^2}}{\mathbf{n_1}+\mathbf{n_2}-\mathbf{2}}}} \tag{2.33}$$

In this case, each group standard deviation is weighted by its sample size (n_i). There are so many modifications and adjustments made to these standardised mean difference effect-size measures. This study focused only on the effect-size indices used in analysis of variance, and these are discussed in the next section.

2.7.5 Eta Square (η^2)

Effect-size indices in ANOVA measure the degree of association and the magnitude of the effect of the independent factor to the dependent variable. Normally, this measure of association is squared in order to relate to the proportion of variance in the response variable attributed to the grouping factor. Eta Squared is one common measure of this type. From the ANOVA table, Eta Square can be obtained by the ratio of the sum of squares as expressed below:

$$\eta^{\mathbf{2}} = \frac{\mathbf{SS_{Treatment}}}{\mathbf{SS_{Total}}} \tag{2.34}$$

where $0 \leq \eta^2 \leq 1$

The interpretation of η^2 is just similar to the linear regression R^2. It (η^2 x 100%) measures the proportion of the shared variance between the dependent variable and the independent categorical factor(s). It is worthwhile to note that Eta Squared estimates the association for sample. According to Nandy (2012), eta squared η^2 is biased and tends to over-estimate the population variance, and however, decreases as the sample size increases.

2.7.6 Partial Eta Square ($\eta^2_{partial}$)

Analogous to Eta squared, the partial Eta squared measures the degree of association in the sample. It is an adjustment to Eta squared by replacing the total sum of squares in the denominator with the combined treatment and error sum of squares.

$$\eta^{\mathbf{2}}_{\mathbf{partial}} = \frac{\mathbf{SS_{Treatment}}}{\mathbf{SS_{Treatment}} + \mathbf{SS_{Error}}} \tag{2.35}$$

where $0 \leq \eta_p^2 \leq 1$

It (η^2 x 100%) measures the proportion of the total variability in the dependent variable explained by the categorical factor(s). The interpretation is similar to η^2, and this index estimates the magnitude of the association in the sample too.

2.7.7 Epsilon Squared (ϵ^2)

Since Eta squared inflates the population strength and is only best for measuring the effect size of a particular sample, two adjustments to this effect were suggested. The first one is the Epsilon squared estimator which is expressed as follows:

$$\epsilon^{\mathbf{2}} = \frac{\mathbf{SS_{Treats}} - \mathbf{df_{Treats}} * \mathbf{MS_{Error}}}{\mathbf{SS_{Total}}} \tag{2.36}$$

The adjustment is made on the numerator of Eta squared where the mean square error is subtracted from the treatment or effect sum of squares. Epsilon squared can assume a negative value, which is typically equated to zero effect.

2.7.8 Omega Squared (ω^2)

A further adjustment to Epsilon Squared is the Omega Squared (ω^2). The adjustment was needed to address the problems of Eta squared which overestimates the population strength in the association. Hence, by adding the mean square error to the total sum of squares on the denominator of the Epsilon squared, the Omega Squared (ω^2) is approximated.

$$\omega^{\mathbf{2}} = \frac{\mathbf{SS_{Treats}} - \mathbf{df_{Treats}} * \mathbf{MS_{Error}}}{\mathbf{SS_{Total}} + \mathbf{MS_{Error}}} \tag{2.37}$$

where $0 \leq \omega_p^2 \leq 1$

Omega ω^2 estimates the population variance whilst Eta η^2 measures the sample variance, hence Eta is always greater than Omega ($\omega^2 < \eta^2$). Both Omega and Epsilon Squared can be negative values, which are treated as zero values.

When dealing with fixed factors in ANOVA, the partial Eta squared, Epsilon squared and the Omega squared are the best measures of effect size (Olejnik & Algina, 2003). The following guideline on interpreting the partial Eta squared, epsilon squared and Omega Squared indices will be used for effect size:

Table 2.7: Guideline on Partial Eta Squared

Effect-size Value	**Magnitude of Effect Size**
$0.01 \leq \eta^2_{partial} < 0.06$	Small effect
$0.06 \leq \eta^2_{partial} < 0.14$	Medium effect
$\eta^2_{partial} \geq 0.14$	Large effect

These guidelines are based on Cohen (1988) benchmarks for interpreting Eta squared effect size, as such the partial Eta Squared ($\eta^2_{partial}$), Epsilon squared (ϵ^2) and Omega Squared (ω^2) will be similarly used as measures of effect size in this study. Statistical packages, SPSS, will be used to generate most of these effect sizes estimators.

2.8 Post Hoc Tests

The term "post-hoc" is derived from the Latin terminology which means "after this". Post hoc tests are statistical tests that are carried out after an analysis of variance test to establish where the differences lie between groups. They are run when there is an overall significant group mean in the significance tests. There are several post hoc tests to choose from, however, each test has its own strengths and weaknesses over the other.

Generally, post hoc tests are based on the error termed **familywise error** (FWE). According to Iker (2013), familywise error can be defined as the probability that any one of the group comparisons or significance tests is a Type I error. It is worthwhile to note that as the number of tests conducted increases, the Type I error (probability that one or more tests are significant by mere chance) increases too. Hence, the familywise error is also called the cumulative or alpha inflation Type I error. The familywise error (fwe) is thus defined as follows:

$$\alpha_{\mathbf{fwe}} \leq \mathbf{1} - (\mathbf{1} - \alpha_{\mathbf{ec}})^{\mathbf{c}} \tag{2.38}$$

where:

α_{fwe} is the familywise error rate

α_{ec} is the normal alpha rate for each individual significance test (0.05, say)

c is the number of groups or comparisons in question.

Given the set of multiple tests or comparisons to be performed, the α_{fwe} estimates the true alpha level in each test. Normally, there is a discrepancy between this α_{fwe} and the normal alpha rate (0.05, say). In order to address this problem, especially when multiple comparisons or test are performed, a number of solutions and corrections have been developed in Statistics. Two of these approaches that will be used in this study are the Bonferronni and Games-Howell test discussed below.

2.8.1 Bonferroni

The Bonferonni procedure is a multiple-comparison post hoc test that is used when performing many independent or dependent statistical tests simultaneously. In this situation, carrying out simultaneous tests leads to the increase in the Type I error, which increases with each single test run. Bonferroni post hoc test is designed to address this problem. It is simply a newly calculated familywise alpha error rate that is built to keep the familywise alpha value at 5% or at any other stipulated level (10%, say). It is calculated as follows:

$$\alpha_{\mathbf{B}} = \frac{\alpha_{\mathbf{fwe}}}{\mathbf{C}} \tag{2.39}$$

where:

α_B is the Bonferroni alpha value to be used

α_{fwe} is the family error rate (given in 2.39)

C is the number of comparisons done in the tests

This test is the most widely used post hoc test due to its simplicity and flexibility nature. It can be used in other statistical tests other than the post hoc tests, for example in correlations. However, a lot of modifications to Bonferonni test and other familywise error corrections have been proposed in literature.

2.8.2 Games-Howell

This is a post hoc test used to detect the group differences when variances are not equal because it takes into account the unequal group sizes. When the variances are severely unequal, it leads to the increase in Type I error. Hence, Games-Howell is considered a better test especially in small samples (sample size per each cell is less than 5) and unequal variances. It is based on the modification of Welch's correction to the degree of freedom, *df*. Using the group sample sizes and their respective variances, Games-Howell test calculates *df* as follows (www.unt.edu):

$$df = \frac{(\frac{\mathbf{S_i^2}}{\mathbf{n_i}} + \frac{\mathbf{S_j^2}}{\mathbf{n_j}})}{\frac{(\frac{\mathbf{S_i^2}}{\mathbf{n_i}})^{\mathbf{2}}}{\mathbf{n_i - 1}} + \frac{(\frac{\mathbf{S_j^2}}{\mathbf{n_j}})^{\mathbf{2}}}{\mathbf{n_j - 1}}} \tag{2.40}$$

where:

i,j are the factor levels determining the $(i,j)^{th}$ group

S_i^2, S_j^2 are the sample variances of the i^{th} and j^{th} group respectively.

Games-Howell test is a post hoc test that is designed for the presence of unequal variances (heteroscedasticity) in ANOVA tests. It also takes care of unequal group sizes (unbalanced data) and small sample sizes (< 5, say). The combination of unequal variances unbalanced small samples leads to increased Type I error rate which can be taken care of by this test.

2.9 Conclusion

This chapter presented the theories involved in analysis of variance (ANOVA) process. First, it outlined the concept of ANOVA, the assumptions of ANOVA, and the types of ANOVA models. The difference between one-way and two-way fixed-effects and mixed-effects models; and hypothesis testing under unbalancedness and heteroscedasticity in ANOVA were elaborated. The concepts of heteroscedasticity, unbalancedness and the associated effects in analysis

of variance techniques were also discussed. Some of the methods that statisticians have been applying to solve the problems of unbalancedness, heteroscedasticity, and calculation of effect sizes in ANOVA were reviewed. Chapter 3 follows dealing with data exploration and remedies applied to assumption violations in the research data.

Chapter 3

Data Exploration

3.1 Introduction

In order to develop a high level of understanding of the data that will be used before analysis is done, it is prudent to explore the distributions and characteristics of the variables, and find out how a characteristic varies among the observations in the dataset. The primary concern is to have a chance to reduce the amount of unnecessary information in order to focus on the key aspects of the data. This will be done using exploration methods namely data visualisation techniques and summary statistics on the variables involved.

3.2 Data Description

An original dataset, *Grocery coupons.sav* dataset of 1404 observations, adopted from SPSS, was used to investigate the effects of heteroscedasticity and unbalancedness on effect sizes in two-way fixed-effects ANOVA design with interactions. The dataset is a survey that was carried out to investigate the spending patterns of customers coming to a certain shopping mall. This dataset was chosen because it qualifies to model a two-way ANOVA model with interaction due to the fact that it has a metric continuous dependent variable that is influenced by two independent variables, called factors, which are categorical in nature. Furthermore, based on the recommendations given by Hair et al. (2014), the sample size is large enough to cater for at least 20 observations per cell, the recommended minimum cell-size in multivariate analysis models. The dataset includes the following variables:

- The dependent variable *Amount spent* as the metric continuous response variable depend-

ing on two categorical factors.

- The independent variable: (Factor A) *Who shopping for*, with three factor levels: 1 = *Self*, 2 = *Self and spouse*, and 3 = *Self and family*.
- The independent variable: (Factor B) *Use coupons*, with four factor levels: 1 = *No*, 2 = *From newspaper*, 3 = *From mailings*, and 4 = *From both*.

The original *Grocery coupons.sav* dataset consisted of 1404 observations, with unequal cell sizes. Table 3.1 below shows number of observations (customers) in each factor combination or cell.

Table 3.1: Variables Types and Cell Sizes

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)				**Total**
	No	From Newspapers	From Mailings	From Both	
Self	140	148	120	76	484
Self & spouse	152	108	176	124	560
Self & family	112	72	104	72	360
Total	404	328	400	272	1404

The numerical values in Table 3.1 above represent the number of observations, in this case the number of customers, which fall under the factor level combinations of categorical factors A and B. The cell sizes for each factor combination are not the same, which suggests that an unbalanced two-way ANOVA model is proposed for analysis in this study.

3.3 Descriptive Statistics

The *amount spent* **means** and **standard deviations** in each factor level and combination has been summarised in Table 3.2 below. The dependent variable is *Amount spent*, and the two factors influencing the dependent variable are *Use coupons* and *Who shopping for*.

Table 3.2: Means (Standard deviations) Statistics

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*) No	From Newspapers	From Mailings	From Both	**Total**
Self	72.6525 (30.78268)	86.7428 (30.44535)	93.3618 (43.22774)	95.3165 (32.65347)	85.6544 (35.49042)
Self & spouse	97.3530 (42.54341)	87.1181 (42.75790)	97.9449 (40.4937)	96.8228 (47.93989)	95.4478 (43.29497)
Self & family	114.6829 (42.24326)	111.3415 (64.19930)	137.2683 (57.08706)	142.5358 (75.46532)	126.1099 (60.02387)
Total	93.5977 (42.16036)	92.2661 (44.86081)	106.7940 (49.48376)	108.5025 (57.00140)	99.9338 (48.54455)

A cursory look on the mean values in the table above, one can notice that most customers spend more when buying for self and family (mean = $ 126.1099), than for anything else using coupons from both (mean = $ 108.5025) mailings and newspapers. The same pattern can be noticed on standard deviations, however, the variations are not that significantly different across the factor combinations.

3.4 Distribution of the dependent variable

Two aspects are of vital interest when learning about the distribution of a numerical variable. These are location and spread. *Location* refers to the central tendency of values in relation to the central point, whereas *spread* refers to how dispersed or scattered the values are around the location. The histogram approach was used to check the distribution of the dependent variable *Amount spent*.

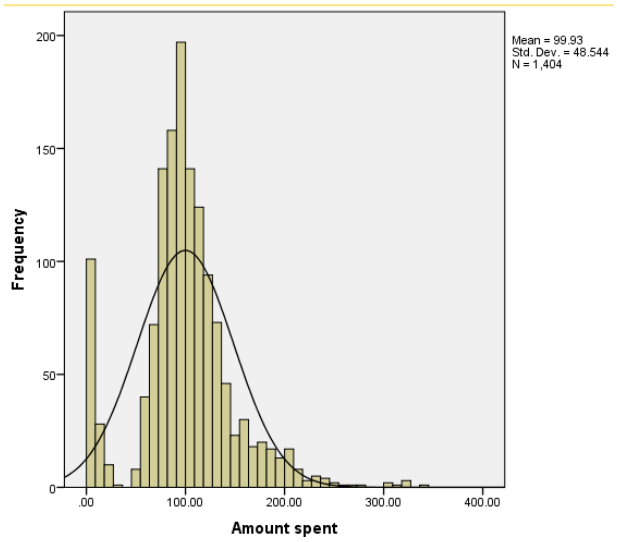

Figure 3.1: Histogram: Amount spent

The location indicated by the histogram (Figure 3.1) above is the mean amount spent (mean = 99.93). It can be noted that the distribution of the values around the mean is not well balanced. The histogram is slightly stretched to the right, giving it an elongated right tail. It is because more observations are accumulated on the left side of the location (99.93) than on the right side. Hence, this means that the data is right-skewed as displayed by a non-symmetric normal curve that has an elongated right (upper) tail. This problem will be dealt with in detail in the subsequent sections when assumptions are tested. The standard deviation of 48.544 implies a wide spread or deviation from the mean, of amounts spent in the dataset. The bigger the standard deviation, the more scattered the values are in relation to the mean of the dataset.

3.5 Data processing

Data processing, or preparing raw data for analysis, has to be considered before using the raw data in its original form for analysis. This involves thoroughly checking for missing values and possible outliers. Put in simple terms, an *outlier* is an observation that appears far away or diverges from the rest of the observations in the sample. There are two types of outliers: univariate and multivariate outlier. Univariate outliers are found in a single variable distribution, whereas multivariate ones are observed when distributions in multi-dimensional space, involving more than one variable. Causes of existence of outliers vary with the type of errors that can happen during data entry, measurement, experiment, data processing, or sampling if not

by natural means.

Outliers have several negative impact on the results of data analysis which include:

- increasing the error variance and decreasing statistical power,
- decreasing normality structure of the data set if the outliers are not randomly distributed,
- affecting the basic assumptions in ANOVA.

Missing values in the dataset may occur during data extraction or data collection. The *Grocery coupons.sav* dataset which was used in this research had no missing values. However, there is need to take into consideration the observations called *outliers*, which have suspicious characteristics that do not follow the overall patterns represented by the rest of the observations in the dataset. Outliers can negatively impact statistical tests if not taken care of before analysis is done. Box plots were used to identify possible outliers in the amount spent in *Grocery coupons.sav* dataset used. The box plots in Figure 3.2 below show the possible outliers.

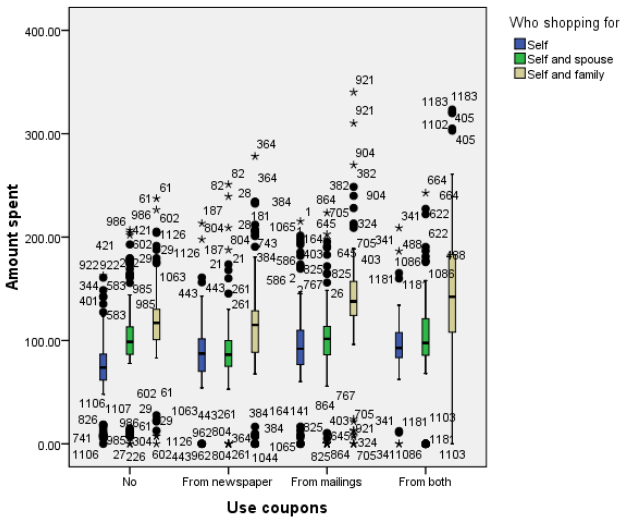

Figure 3.2: Amount spent: Box plots

Outliers are problematic and can impact on the group means and statistical tests. From Figure 3.2 above, the box plots indicate quite a number of possible outliers. Provided the exclusion of these outliers will not affect the recommended group size and overall sample size, total exclusion of extreme outliers from the dataset was called for since most of them were very much different from the rest of the data values.

After cleaning the extreme outliers, a reduced dataset consisting of 811 observations remained. Figure 3.3 below displays cleaned data without extreme outliers.

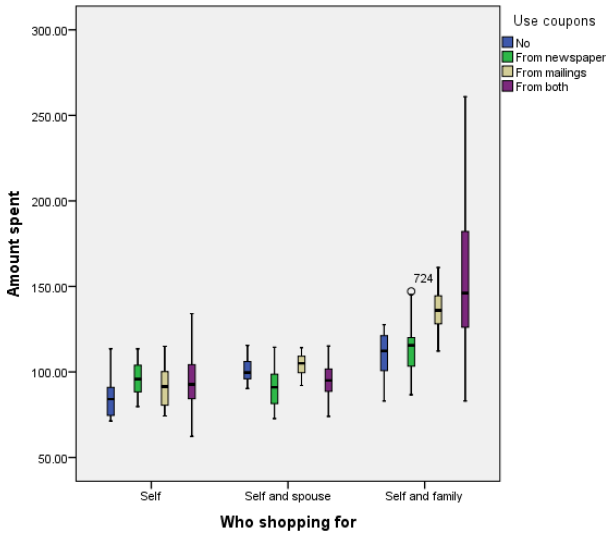

Figure 3.3: Amount spent: Box plots (Outlier-free)

Exclusion of outliers from the sample did not affect the recommended sample and cell size, minimum of 20 observations per cell, leaving the data free from the bias that these extreme values could inject into the analysis of significance tests (Hair et al., 2014).

3.6 Data Transformation

As noted before, the original data is skewed to the right. Data transformation techniques can be applied to reduce the influence of extreme values (skewness) that stretch the data away from the central location. Two commonly used transformations that address the problem of skewness are the **logarithm function** and the **square root** transformations. The P-P plots below show the results of these transformations done to evaluate the skewness of the transformed data.

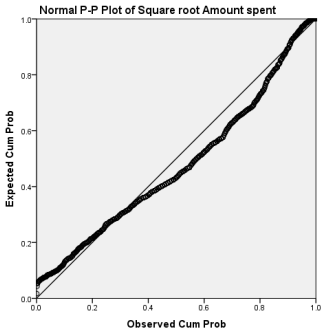

Figure 3.4: Square Root Transform

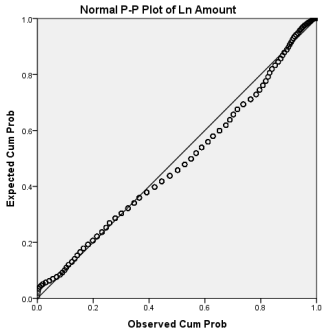

Figure 3.5: Natural Log Transform

In both transformations (Figure 3.4 and 3.5), the plotted points are fairly close to the diagonal normal line, but produce a slight upward bend to the left of the normal line, which indicates a slightly long tail to the right. This is a feature of presence of right-skewness in the data. However, when compared to the original dataset, the transformed data, especially the natural logarithm transformation, has significantly improved the skewness for the better. Since the natural logarithm transformation is a better remedy for skewness on this case, therefore the transformed data (*lnAmount*) will be used in subsequent analyses and significance tests.

3.7 Testing ANOVA Assumptions

Three basic analysis of variance model assumptions which must be tested include normality, homoscedasticity, and independence assumption. The transformed dataset will be tested for assumption violation.

3.7.1 Normality

The hypothesis being tested in this section is given by:

H_0: The sample data was from a normally distributed population (Normality)

H_1: The sample data was not from a normally distributed population

SPSS statistical package was used to generate the probability plots (p-p) plots, which are graphs that show how the observations behave in relation to the normal line. Normality assumption is violated if the plotted values deviate from the diagonal normal line of the p-p plot. To augment the p-p plot test, the histogram approach were generated in R for univariate normality test.

Normality assumption is violated if the normal curve on the histogram does not show a normal "bell shape". To supplement the graphical assessment of normality, the Shapiro-Wilk normality tests were conducted for each simulated sample data. In each case, the null hypothesis is retained when the p-value is greater than the stipulated alpha α level (0.05).

Based on the transformed *Grocery coupons.sav* original dataset, the following figures present the normality test conducted:

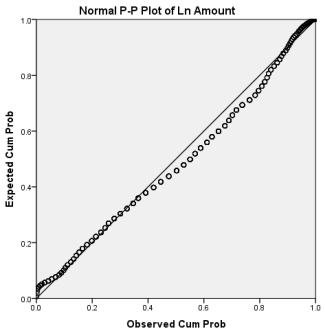

Figure 3.6: LnAmount p-p plot

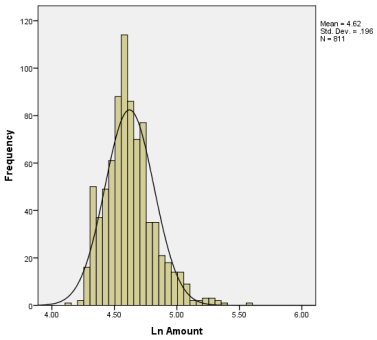

Figure 3.7: LnAmount histogram

Looking at the p-p plot (Figure 3.6), most of the data plots follow the normal line although there is slight deviation to the right of the line. Furthermore, the normal curve on the histogram (Figure 3.7) is not perfectly normal "bell-shape", an indication that there might be chances of normality violation. To augment the graphical normality assessment, Table 3.3 below gives the Shapiro-Wilk normality test statistics.

Table 3.3: Normality Tests for Original Data

	Shapiro-Wilk		
	Statistic	df	Sig.
Amount spent	0.9967	811	0.09498

The Shapiro-Wilk's normality tests gives non-significant p-value (p = 0.09498) which is slightly greater than the stipulated alpha (0.05). We fail to reject H_0 and conclude that the original data from the unbalanced and heteroscedastic model was from a normally distributed population. Hence, normality assumption was not violated.

3.7.2 Homoscedasticity

The null hypothesis under homogeneity of variance assumption states that the error variance of the dependent variable is equal across groups. The calculated probability of finding the result equal to, or more extreme than, what was actually observed when the null hypothesis is true is known as the *p-value*. It is based on this p-value that the significance of the test is determined. The probability of rejecting the null hypothesis in a statistical test when in fact it is true is known as the *significance level*, normally denoted by α, expressed as a percentage (5% or 10% say). Levene's test is used to check the violation of the homogeneity of variance assumption.

The Null hypothesis (H_0: There is homogeneity of error variance) is evaluated based on the rejection criterion: Reject H_0 at 5% significance level if p-value is less than α (5%). The results of the equality of error variances test are given below.

Table 3.4: Levene's Test of Equality of Error Variances[a]
Dependent Variable: lnAmount

F	df1	df2	Sig.
4.532	11	799	.000

The p-value < 0.001 from the Levene's test of equal error variances in Table 3.3 above, is clearly below alpha = 0,05. Hence, we reject H_0 at 5% significance level and conclude that there is no homogeneity in the error variances of the dependent variable (Amount spent) across the groups in the data set. The original data is heteroscedastic.

3.7.3 Independence of observations

As indicated by Hair et al. (2014), the independence of observations is ensured by the nature of design. The observations involved were customers visiting a shop, naturally independent of one another, hence there was no need to test the independence of observations assumption violation.

3.8 Simulation samples

The study involved comparison of four different ANOVA designs derived from the same original, unbalanced and heteroscedastic dataset. The other three models will be generated from the samples simulated from this original data. Each data sample has to be explored for ANOVA assumptions violation, and this is the purpose of the subsequent sections. Only normality and homoscedasticity assumptions were considered since the independence assumption was already met by the nature of design from the original data.

3.8.1 Balanced and heteroscedastic sample

A total of 100 samples of 864 observations each, with equal cell sizes, was randomly re-sampled from the original *Grocery coupons.sav* dataset generating a balanced design. In a similar way, the ANOVA assumptions were considered for this sample.

3.8.1.1 Normality

The probability plot and the histogram below displayed below provide a quick normality test for the simulated dataset.

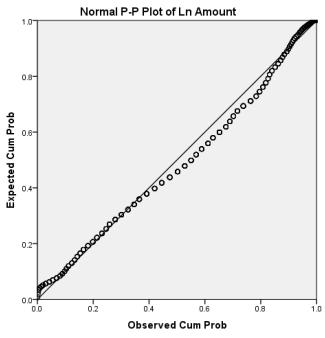

Figure 3.8: Bal-Heterosced P-p plot

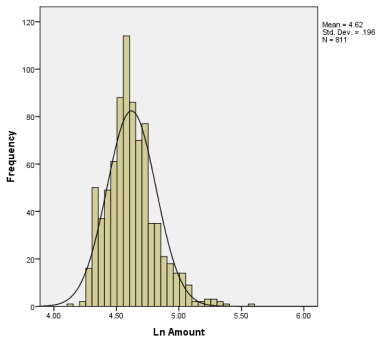

Figure 3.9: Normality histogram

Most of the data plots in Figure 3.8 above follow the normal line although there is slight deviation to the right of the line. Furthermore, the normal curve on the histogram, Figure 3.9, shows an almost normal "bell-shape". On the two figures, we have some indications that the balanced data sample might not be perfectly normal even though the departure is not so severe. For an informed insight, we consider the normality tests displayed in Table 3.5 below.

Table 3.5: Normality Tests for Balanced and Heteroscedastic Sample

	Shapiro-Wilk		
	Statistic	df	Sig.
Amount spent	0.9966	864	0.05917

The p-value from the Shapiro-Wilk's normality tests obtained is non-significant (p-value = 0.05917) and slightly more than the stipulated alpha (0.05). We fail to reject H_0 conclude that the simulated balanced heteroscedastic sample data was from a normally distributed population. Hence, normality assumption was fairly satisfied.

3.8.1.2 Homoscedasicity

Levene's Test of Equality of Error Variances for the balanced data sample are as summarized below:

Dependent Variable: Amount spent

F	df1	df2	Sig.
41.142	11	852	.000

Tests the null hypothesis that the error variance of the dependent variable is equal across groups

*a. Design: Intercept + shopfor + usecoup + shopfor * usecoup*

The p-value (Sig. = 0.000) from the Levene's test of equal error variances above is clearly less than alpha (0.05). Hence, we reject H_0 at 5% significance level and conclude that there is no homogeneity in the error variances of the dependent variable (Amount spent) across the groups of the two factors. The balanced sample data is heteroscedastic.

3.8.2 Balanced and homoscedastic sample

Similarly, 100 samples of 864 observations each were simulated from the *Grocery coupons.sav* data, customers who visited some shops to spend their money. The descriptive statistics from the balanced and heteroscedastic sample were used to simulate these homoscedastic samples with more or less the same distribution patterns as the original data except in the equality of variances. An overall mean (mean = 83.9) and average standard deviation (s.d = 15) estimated from the original dataset were used in the simulation process in R statistical package.

3.8.2.1 Normality

Considering the Q-Q plot in Figure 3.10 below, almost all the plots closely follow the normal line, which shows a normal pattern in the data plots. Hence, based on the graphical normality assessment, one can conclude that the assumption of normality for the residuals is met.

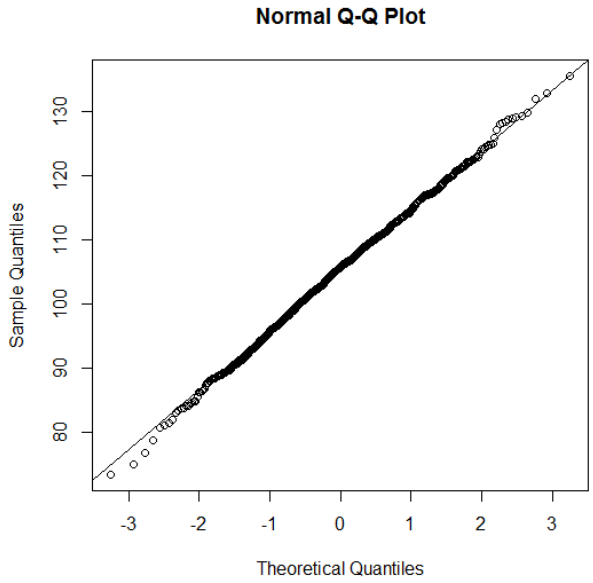

Figure 3.10: Balanced Homoscedastic Q-Q Plot

This is further supported by the Shapiro-Wilk's normality test given in Table 3.6 below.

Table 3.6: Normality Tests for Balanced and Homoscedastic Sample

	Shapiro-Wilk		
	Statistic	df	Sig.
Amount spent	0.9986	864	0.7372

The Shapiro-Wilk normality test on the balanced and homoscedastic data sample gave a non-significant p-value (p = 0.7372). The p-value is far greater than alpha (0.05). Hence, we retain H_0 and conclude that the dependent variable *Amount spent* in the balanced *Grocery coupons.sav* data is normal in the population.

3.8.2.2 Homogeneity

Violation of the homogeneity of variance on the balanced data sample was tested by Levene's test. The Null hypothesis (H_0: There is homogeneity of covariance matrices) and the rejection criterion: Reject H_0 at 5% significance level if p-value is less than α (5%), was used.

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Amount spent

F	df1	df2	sig.
1.003	11	852	.442

Tests the null hypothesis that the error variance of the dependent variable is equal across groups

*a. Design: Intercept + shopfor + usecoup + shopfor * usecoup*

The p-value = 0.442 from the Levene's test of equal error variances above is clearly greater than alpha = 0.05. Hence, we fail to reject H_0 at 5% significance level and conclude that indeed there is homogeneity in the error variances of the dependent variable (Amount spent) across the groups of the two factors. The balanced data is homoscedastic.

3.8.3 Unbalanced and homoscedastic sample

From the original *Grocery coupons.sav* dataset, 100 samples of 850 observations each, with unequal cell sizes, were drawn to build a balanced and homoscedastic model.

3.8.3.1 Normality

The Q-Q plots and Shapiro-Wilk's test were used to check the normality pattern of the simulated data. The graphical normality assessment shown by the Q-Q plot, Figure 3.11 below, gives an impression of a normally distributed data, plotted point closely follow the diagonal normal line.

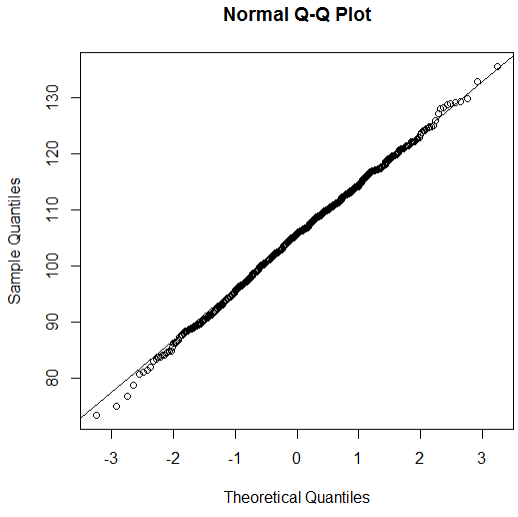

Figure 3.11: Unbalanced Homoscedastic Q-Q Plot

The Q-Q plot above portrays an almost normal pattern since most of the plots closely follow the diagonal normal line. To further clarify the assessment, the Shapiro-Wilk's normality test in Table 3.7 below confirms this pattern.

Table 3.7: Normality Tests for Unbalanced and Homoscedastic Sample

	Shapiro-Wilk		
	Statistic	df	Sig.
Amount spent	0.9983	850	0.5562

The p-value (p = 0.5562) from the Shapiro-Wilk's normality test is far greater than alpha = 0.05. Hence, we fail to reject H_0 and conclude that the dependent variable *Amount spent* in unbalanced *Grocery coupons.sav* sample is normal in the population.

3.8.3.2 Homoscedasticity

Levene's Test for the null hypothesis that the error variance of the dependent variable is equal across groups in the unbalanced data sample produced the following results:

Levene's Test of Equality of Error Variances[a]

Dependent Variable: Amount spent

F	df1	df2	sig.
1.071	11	838	.382

*a. Design: Intercept + shopfor + usecoup + shopfor * usecoup*

Similarly, the p-value = 0.382 from the Levene's test of equal error variances above is clearly more than alpha = 0.05. Hence, we fail to reject H_0 at 5% significance level and conclude that, again there is homogeneity in the error variances of the dependent variable (Amount spent) across the groups in unbalanced data. The unbalanced data is also homoscedastic.

3.9 Conclusion

This chapter presented the exploration and processing of the data that will be used to test the research hypotheses. The original dataset transformed and the three simulated samples constitute the basis for comparison of the four different two-way ANOVA models to be studied.

The next chapter outlines the methodology used to analyse the research data, outlining the first five stages of Hair et al. (2014) Six-Stage model building process.

Chapter 4

Materials and Methods

4.1 Introduction

Four models were considered to investigate the effects of unbalancedness and heteroscedasticity in two-way ANOVA models. These models are the balanced homoscedastic ANOVA model; unbalanced homoscedastic model; balanced heteroscedastic model; and unbalanced heteroscedastic model. Materials and methodologies used to present and analyse the data in each of these four models are outlined following the six-stage model building proposed by Hair et al. (2014). The six stages of model building are: Stage 1: The objectives of the study; Stage 2: The research design; Stage 3: Testing assumptions of the research design; Stage 4: Estimation of the two-way fixed-effects ANOVA models, assessing their overall fit; Stage 5: Validation of results; and Stage 6: Analysis and discussion of results. The focus of this chapter is on the first five stages.

4.2 Balanced and homoscedastic two-way ANOVA model

From the group means and standard deviations of the original dataset, *Grocery coupons.sav* dataset adopted from SPSS, 100 samples of 864 observations each were simulated in R and SPSS statistical packages. The individual group means used from the original dataset were as summarised in Table 4.1 below, and a common standard deviation of 15.0 was used to simulate 72 observations in each cell.

Table 4.1: Original data : Group Means

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)			
	No	From Newspapers	From Mailings	From Both
Self	83.9	95.7	91.1	95.6
Self & spouse	101.1	91.0	104.1	95.3
Self & family	110.2	113.7	136.1	150.4

The *self and family* level of the factor *shopfor* had higher group means than any other factorial combination. However, to achieve the desired homogeneity in error variances, a constant standard deviation of 15.0, estimated from the average group standard deviations of the original data, was used together with each group mean to simulate the required homogeneous sample.

4.2.1 Research Objectives and Design

In this model, the aim is to establish how the amount spent by customers on shopping is influenced by their reason for shopping (factor A) and the source of coupons used (factor B), when the model is balanced and has homogeneous error variances.

The research design was a total of 100 simulated balanced data samples of size 864 each, with homogeneous error variances, were used to investigate the effects of shopping options (*Who shopping for*) and the use of coupons (*Use coupons*) to the amount of money spent by customers (*Amount spent*). A two-way fixed-effects ANOVA design with interaction, having *Amount spent* as the metric continuous response variable depending on two categorical factors; factor A (*Who shopping for*) and factor B (*Use coupons*) is proposed.

4.2.2 Data Description and Sample Size

Each simulated sample of size 864, with equal cell sizes of 72 was generated from the original data descriptive statistics (group cell means and a homogeneous standard deviation of 15.0). Original data descriptive statistics were used in order to generate a sample that is closely related to the original dataset.

The data sample satisfied the recommended minimum cell size of 20 observations per cell (group) and an overall sample size above 250 to maintain a statistical power of 0.80 (Hair et al., 2014).

Table 4.2: Balanced & Homoscedastic Sample Cell Count

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)				**Total**
	No	From Newspapers	From Mailings	From Both	
Self	72	72	72	72	288
Self & spouse	72	72	72	72	288
Self & family	72	72	72	72	288
Total	216	216	216	216	864

Table 4.2 displays the simulated balanced data sample size, that is, the number of customers per cell. The recommended minimum cell size (at least 20 observations per cell) was achieved. Hence, the sample size was appropriate for analysis of variance.

4.2.3 Testing ANOVA Assumptions

This section was dealt with in Section 3.8.2 in detail. A recapitulation of the assumption tests is summarised below.

4.2.3.1 Normality Assumption

The Shapiro-Wilk normality tests done in R statistical package produced a test statistic W = 0.9986. With the p-value = 0.7372 greater than α (0.05), supported by the normal Q-Q plot in Figure 3.10, the balanced data sample was normal. Hence the assumption of normality for the residuals was not violated.

4.2.3.2 Homogeneity Assumption

Levene's test of equality of error variances resulted in a p-value = 0.442, clearly greater than alpha = 0.05. Hence, we failed to reject H_0 at 5% significance level and conclude that indeed there is homogeneity in the error variances of the dependent variable (*Amount spent*) across the groups of the two factors in the balanced sample data.

4.2.3.3 Independence of observations

This sample resembled an original dataset whose observations were independent from one another by the nature of the research design. Hence, the independence of observations assumption was satisfied.

4.2.4 Estimating the ANOVA model and assessing overall model fit

Since the sampled data satisfies all the ANOVA assumptions, classical F-tests were used to estimate the balanced ANOVA model. The **main** and **interaction** effects were calculated based on the type I sum of squares in which the F-tests and/or p-values as well as the effect sizes, were used to test the existence of group differences in the dependent variable.

The traditional F-tests were used to estimate the balanced homoscedastic model based on the simulated data sample. The ANOVA Table 4.3 below gives the estimated statistics for the model. The *shopfor* main effect, the *usecoup* main effect and the interaction effect, *(shopfor*usecoup)*, p-values are each clearly significant (p-value < 0.001) and less than α (0.05), which implies that theses effects are significantly contributing to the differences in amounts spent by customers. Furthermore, with the significant main and interaction effects in the model, the *Adjusted* $R^2 = 0{,}575$ (greater than 0.5000) shows a fairly good model fit. The fitted model explains about 58% (0.575 x 100) of the variability in the response variable, *Amount spent*, being attributed to the reason for shopping (*shopfor*) and the source of coupons they used (*usecoup*)

Table 4.3: Balanced & Homoscedastic ANOVA

Source	Type I Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected model	260952.807[a]	11	23722.982	106.943	.000	.580
Intercept	9654618.648	1	9654618.648	43523.121	.000	.981
shopfor	75416.045	2	37708.022	169.988	.000	.285
usecoup	35578.123	3	11859.374	53.462	.000	.158
shopfor*usecoup	1499958.639	6	24993.107	112.669	.000	.442
Error	188996.905	852	221.827			
Total	10552441.33	863				
Corrected error	449949.712	863				

a. R Squared = .580 (Adjusted R Squared = .575)

The partial eta squared ($\eta^2_{partial}$) effect-size statistics on the balanced homoscedastic model are displayed in the last column in the above ANOVA table. Considering the guidelines suggested by Cohen (1988), it can be noted that both the main effects and the interaction effect had large effect size ($\eta^2_{partial} > 0.14$), The greatest effect (0.442) is realized on the interaction between the reason for shopping (*Shopfor*) and the source of coupons (*Usecoup*). Though the effect sizes in this case are considered large, they are all below 50%.

Moreover, considering the profile plot for the same variables below, we can assess the interaction between the independent variables, *Who shopping for* and *Use coupons* as factors affecting *Amount spent* in Figure 4.1 below. Evidence of interaction between the factors *shopfor* and *usecoup* is clearly depicted by the non-parallel lines in the profile plot.

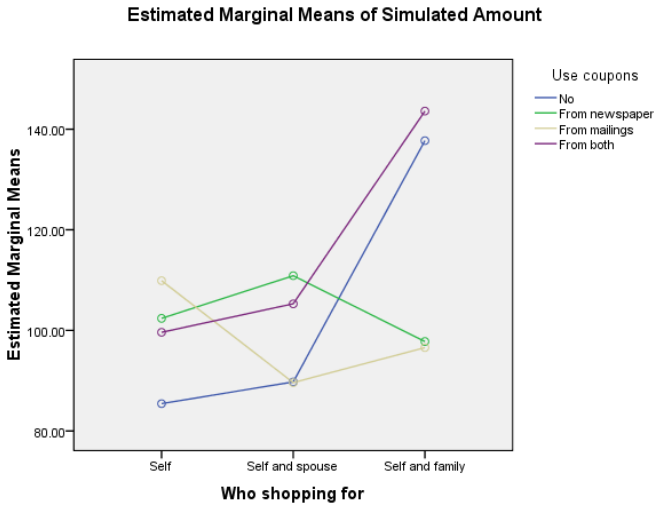

Figure 4.1: Profile Plot: Balanced Homoscedastic Model

Interaction is more evident in *self and spouse* level of *shopfor* factor than in the other two levels.

On the other hand, the *No* and *From both* levels of *usecoup* factor are almost parallel, indicating very little interaction where as the *From mailings* level does interact with the rest of the factor levels. One can safely conclude that, the differences in the amounts spent is explained by these two factors in isolation and in combination.

Post hoc tests were conducted for the balanced homoscedastic model fitted in order to establish the particular factor levels combinations that significantly contributed to the variations in the dependent variable means. Table 4.4 displays the Bonferroni post hoc tests for the two factors.

Table 4.4: Balanced & homoscedastic ANOVA : Post Hoc
Dependent variable: Amount spent

		Mean Difference			95% Confidence Interval	
(I)Who shop for	(J)Who shop for	(I-J)	Std. Error	Sig	Lower	Upper
Self	Self and spouse	.4517	1.24116	1.000	-2.5255	3.4289
	Self and family	-19.5893*	1.24116	.000	-22.5664	-16.6121
Self and spouse	Self	-.4517	1.2116	1.000	-3.4289	2.5255
	Self and family	-20.0410*	1.24116	.000	-23.1082	-17.0638
Self and family	Self	19.5893*	1.2116	.000	16.6121	22.5604
	Self and spouse	20.0410*	1.2116	.000	17.0638	23.0182
(I)Use coupons	(J)Use coupons	(I-J)	Std. Error	Sig.	Lower	Upper
No	From newspaper	.6204	1.43316	1.000	-3.1695	4.4103
	From mailings	5.6108*	1.43316	.001	1.8209	9.4007
	From both	-11.8652*	1.43316	.000	-15.6551	-8.0753
From newspaper	No	-.6204	1.43316	1.000	-4.4103	3.1695
	From mailings	4.9904*	1.43316	.003	1.2005	8.7803
	From both	-12.4856*	1.43316	.000	-16.2755	-8.6957
From mailings	No	-5.6108*	1.43316	.001	-9.4007	-1.8209
	From newspaper	-4.9904*	1.43316	.003	-8.7803	-1.2005
	From both	-17.4760*	1.43316	.000	-21.2659	-13.6861
From both	No	11.8652*	1.43316	.000	8.0753	15.6551
	From newspaper	12.4856*	1.43316	.000	8.6957	16.2755
	From mailings	17.4760*	1.43316	.000	13.6861	21.2659

**. The mean difference is significant at the .05 level*

The post hoc tests indicate that factor A (*Who shopping for*) had only one insignificant level combination, "*self - self and spouse*" (Sig. = 1.000), the rest of the factor levels and their combinations were significantly contributing to the differences in amount spent by customers. A similar case for the factor B (*Use coupons*) levels, only the level combination, "*No - From*

newspapers" had insignificant contribution to the dependent variable (Sig. = 1.000), whereas the rest of the factor level combinations were significant (Sig. < 0.05).

4.3 Unbalanced and heteroscedastic model

It is very rare to get a real-life dataset that satisfies all the ANOVA model assumptions. The original *Grocery coupons.sav* dataset adopted from SPSS, which was cleaned of all possible outliers in the previous chapter, now comprising 811 observations, was no exception. In line with the advice by Krishnamoorthy, Lu and Mathew (2007) when the ANOVA assumptions have been violated, the parametric bootstrap (PB) sample estimation approach is the best option in terms of controlling the type I error probability especially in the presence of unequal variances. Hence, this approach will be used to estimate the ANOVA model for the unbalanced and heteroscedastic dataset in this section. The first five stages of the model building process are presented as usual.

4.3.1 Research Objectives

The main aim of this study is to establish how the amount spent by customers on shopping is influenced by their reason for shopping (factor A) and the source of coupons used (fator B), under the influence of unbalancedness and heteroscedasticity.

4.3.2 Research Design

A real-life dataset, *Grocery coupons.sav*, adopted from SPSS, is used to investigate the effects of heteroscedasticity and unbalancedness on effect sizes of two-way fixed-effects ANOVA designs with interactions. In each case, a two-way ANOVA design with interaction, having *Amount spent* as the metric continuous response variable depending on two categorical factors; *Who shopping for (shopfor)* (with three factor levels: *self*, *self and spouse*, and *self and family*); and *Use coupons* (with four factor levels: *No*, *From newspaper*, *From mailings* and *From both*) is proposed.

4.3.3 Data Description and Sample Size

After cleaning the original *Grocery coupons.sav* dataset of all the possible outliers, the sample was reduced to a total of 811 observations with unequal cell sizes. Table 4.5 below is the two-way Anova design proposed, showing the varying standard deviations in each cell.

Table 4.5: Unbalanced & Heteroscedastic Sample Standard Deviations

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)				
	No	From Newspapers	From Mailings	From Both	**Total**
Self	10.2	9.6	11.5	16.2	13.0
Self & spouse	7.1	10.9	6.3	10.0	10.1
Self & family	12.1	13.8	12.8	38.0	27.1
Total	14.8	14.2	21.5	34.1	23.4

Variations in standard deviations, from as little as 6.3 to a maximum of 38.0, in the factorial combinations indicate unequal variances across the groups in the model, a justification that a heteroscedastic two-way ANOVA design should be proposed.

All the cell sizes in each factorial combination in Table 4.6 below were above 40. The dataset satisfied the recommended minimum cell size of 20 observations per cell (group) to maintain a statistical power of 0.80 (Hair et al., 2014).

Table 4.6: Unbalanced & Heteroscedastic Sample Cell Count

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)				
	No	From Newspapers	From Mailings	From Both	**Total**
Self	72	72	72	68	284
Self & spouse	72	72	72	72	288
Self & family	72	42	68	57	239
Total	216	186	212	197	811

Table 4.6 displays the original data sample size, that is, the number of customers per cell (factorial combination) who came for shopping with or without coupons. All the cell (group) sizes were adequate enough to meet the recommended minimum cell size (at least 20 observations

per cell). Hence, the sample was appropriate enough to warrant accurate analysis results. The next subsection looks at testing ANOVA assumptions.

4.3.4 Testing ANOVA Assumptions

ANOVA assumptions which were tested include normality, homoscedasticity, and independence assumption.

4.3.4.1 Normality

The hypothesis being tested in this section is given by:
H_0: The sample data was from a normally distributed population (Normality)

Based on the normality test done in Section 3.7, the p-p plots and the normal curve showed that the reduced original dataset was normally distributed, though not perfectly normal. It was established that the data had little skewness problems that could not be totally eradicated by data transformation. Hence, as a remedy to this problem, a parametric bootstrap approach will be used when estimating the ANOVA model for this dataset.

4.3.4.2 Homoscedasitcity

The p-value of ($p < 0.001$), from the Levene's test of equal error variances in Section 3.7.2, was clearly below alpha = 0,05. Hence, the homogeneity of the error variances assumption was not satisfied across the groups in the dependent variable (Amount spent). The original data is therefore heteroscedastic.

4.3.4.3 Independence of observations

The independence of observations is ensured by the nature of design (Hair et al., 2014). The observations involved were customers visiting a shop, naturally independent of one another, hence the independence of observations assumption was satisfied.

4.3.5 Estimating the ANOVA model and assessing overall model fit

Since the previous section shows that the dataset used was not perfectly normal and that it had some little skewness problems that could not be ironed out through data transformation, it was prudent to apply the parametric bootstrap approach (defined in Section 2.4.3) to approximate the test statistics in model estimation since it is robust to assumption violations. Basically, two types of hypotheses were tested: the **main** and **interaction** effects. Null hypotheses for the main effects and existence of interaction effects were rejected if the p-values for factor/or interaction factor exceeded the α level of significance (5%). Type III sum of squares was used for the unbalanced two-way ANOVA. In each case, the F-tests and/or p-values as well as the effect sizes, were used to test the existence of group differences in the dependent variables.

Overall model fitness was tested using the R^2 adjusted. The higher the R^2 adjusted (above 50% or 0,5000 ,say) the better the model fit. Alternatively, the F-test could also be used to check the overall model fitness.

The parametric bootstrap estimation, based on 100 samples, was used to approximate the heteroscedastic model based on the original unbalanced data. The ANOVA Table 4.7 below gives the estimated statistics of the model.

Table 4.7: Unbalanced & heteroscedastic ANOVA

Dependent variable: Amount spent

Source	**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**	**Partial Eta Squared**
Corrected model	268661.971[a]	11	24423.816	112.794	.000	.608
Intercept	88414226.278	1	8841426.278	46831.340	.000	.981
shopfor	182374.629	2	91187.315	421.120	.000	.513
usecoup	34393.629	3	11464.543	52.945	.000	.166
shopfor*usecoup	49141.645	6	8190.274	37.824	.000	.221
Error	173011.700	799	216.535			
Total	9289258.197	811				
Corrected error	441673.671	810				

a. R Squared = .608 (Adjusted R Squared = .603)

Considering the significance (p-value) column in Table 4.7 above, the *shopfor* main effect p-value (Sig = 0.000) is clearly less than α (0.05), hence we conclude that the main effect of

shopfor is significantly contributing to the differences in amounts spent by customers. Similarly, the *usecoup* main effect is also significant since the p-value (p-value < 0.001) is less than α (0.05) at 5% level of significance. The same applies to the interaction effect, *(shopfor*usecoup)*, which is clearly significant, p-value ($p < 0.001$) less α (0.05). More-so, the *Adjusted* $R^2 = 0.603$ (greater than 0.5000) shows a fairly good model fit. The fitted model explains 60% (0.603 x 100) of the variability in the response variable, *Amount spent*, being attributed to the reason for shopping (*shopfor*) and the source of coupons (*usecoup*) they used.

The effect sizes on the unbalanced heteroscedastic model were calculated using the partial eta squared ($\eta^2_{partial}$) statistics. The last column in the above ANOVA table gives the effect-size values of the calculated. Considering the guidelines suggested by Cohen (1988), it can be noted that both the main effects and the interaction effect had considerably large effect size ($\eta^2_{partial} > 0.14$). The greatest effect (0.513%) is attributed to reason for shopping (*shopfor*), whereas the source of coupons (*usecoup*) contributed less effect size (0.166), which resulted in a fairly high interaction effect size of 0.221.

Moreover, considering the profile plot for the same variables below, we can have a pictorial glimpse of the interaction between the independent variables, *Who shopping for* and *Use coupons* as factors affecting *Amount spent* in Figure 4.2 below.

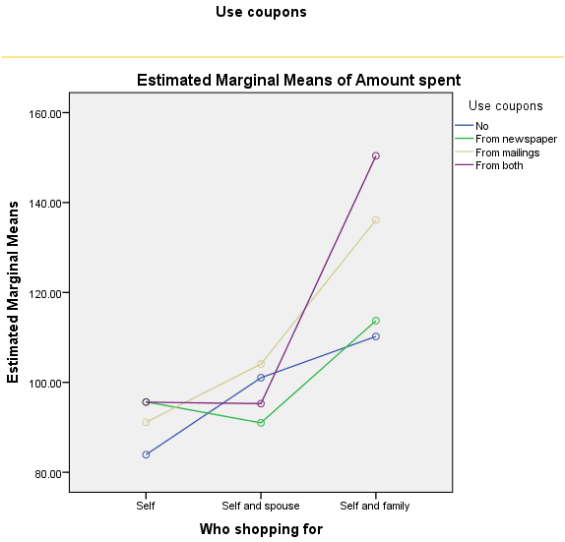

Figure 4.2: Profile Plot: Unbalanced Heteroscedastic Model

Evidence of interaction between the factors *shopfor* and *usecoup* is clearly depicted by the

non-parallel lines in the profile plot. Interaction of the two factors is more in *self & spouse* level of the factor *shopfor* than in any other level. One can safely conclude that, the differences in the amounts spent was influenced by these two factors in isolation and in combination.

Having a significant interaction effect as shown in the ANOVA Table 4.7, it was necessary to perform post hoc tests in order to establish where exactly the differences occurred between the groups of the amounts spent. Table 4.8 below displays the bootstrap Bonferroni post hoc tests for multiple comparisons for the two factors.

Table 4.8: Unbalanced & heteroscedastic Post Hoc

Dependent variable: Amount spent

		Mean Difference	Bootstrap[a]			
					BCa 95% CI	
(I)Who shop for	(J)Who shop for	(I-J)	Bias	S.E	Lower	Upper
Self	Self and spouse	-6.3297	-.0012	.9756	-8.1663	-4.4535
	Self and family	-36.2783	-.0870	1.8778	-39.9304	-32.8745
Self and spouse	Self	6.3297	.0012	.9756	4.3243	8.2253
	Self and family	-29.9485	-.0858	1.8037	-33.3516	-26.6864
Self and family	Self	36.2783	.0870	1.8778	32.6961	40.2191
	Self and spouse	29.9485	.0858	1.8037	26.4828	33.6967
(I)Use coupons	(J)Use coupons	(I-J)	Bias	S.E	Lower	Upper
No	From newspaper	.4513	-.0197	1.4537	-2.3658	3.3694
	From mailings	-11.5722	.0041	1.8301	-15.1920	-8.0712
	From both	-12.9482	-.1010	2.7052	-18.3750	-8.1598
From newspaper	No	-.4513	.0197	1.4537	-3.4686	2.4676
	From mailings	-12.0234	-.0237	1.7473	-15.2650	-8.3921
	From both	-13.3995	-.0814	2.7048	-19.4205	-8.3379
From mailings	No	11.5722	-.0041	1.8301	8.1320	15.0066
	From newspaper	12.0234	-.0237	1.7473	8.5296	15.1787
	From both	-1.3760	-.1051	2.9173	-7.0586	4.0376
From both	No	12.9482	.1010	2.7052	7.8951	18.7523
	From newspaper	13.3995	.0814	2.7048	8.3398	19.4200
	From mailings	1.3760	.1051	2.9173	-4.2759	7.2684

a. Unless otherwise noted, bootstraps are based on 1000 bootstrap samples

Reading from the bias-corrected and accelerated (BCa) 95% confidence intervals, the bootstrap post hoc tests indicate that all factor A (*Who shopping for*) levels had significant mean differences (BCa 95% Confidence Intervals do not cut through zero), hence these factor levels were significantly contributing to the differences in amount spent by customers. On the other

hand, factor B (*Use coupons*) had two level combinations, "*No - From newspapers*" and "*From both - From mailings*" which had insignificant contributions to the dependent variable (BCa 95% Confidence Intervals include zero), whereas the rest of the factor level combinations were significantly contributing to the differences among the amounts spent.

4.4 Unbalanced and homoscedastic model

Using the same approach as in the balanced model, a total of 100 samples of 850 observations each, with unequal cell sizes, were simulated from the original *Grocery coupons.sav* dataset using different group means and a common standard deviation from the original data descriptive statistics.

4.4.1 Research Objectives and Design

In this model, the aim is to investigate the effect of two independent categorical factors, *Who shopping for* and *Use coupons* on the numeric response variable *Amount spent* when the sample data has unequal cell sizes, but having homogeneous error variances.

The research was designed in such a way that 100 unbalanced simulated data samples of size 850 each, with homogeneous error variances, were used to investigate the effects of shopping patterns (*shopfor*) and the use of coupons (*usecoup*) to the amount of money spent by customers. A two-way fixed-effects ANOVA design with interaction, having *Amount spent* as the metric continuous response variable depending on two categorical factors, *Who shopping for* and *Use coupons*, is proposed.

4.4.2 Data Description and Sample Size

Each simulated sample of size 850 with unequal cell sizes was generated from the original data descriptive statistics (group cell means, homogeneous standard deviation of 15.0). Original data descriptive statistics were used in order to generate samples that are closely related to the

original dataset but of an unbalanced ANOVA design.

With a simulated sample of size 850 generated from the original data, having a minimum cell size of 68 across all the factor combinations, the dataset satisfied the recommended minimum cell size of 20 observations per cell (group) to maintain a statistical power of 0.80 (Hair et al., 2014).

Table 4.9: Unbalanced & Homoscedastic Sample Cell Count

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)				**Total**
	No	From Newspapers	From Mailings	From Both	
Self	68	70	72	72	282
Self & spouse	72	70	72	72	284
Self & family	70	72	72	70	284
Total	210	212	214	214	850

Table 4.9 above displays the cell sizes in the simulated unbalanced data sample. The recommended minimum cell size of at least 20 observations per cell was satisfied and thus, the sample is adequate for analysis of variance analysis.

4.4.3 Testing ANOVA Assumptions

The ANOVA assumptions for this model were tested in Section 3.8.3. Both the normality and homoscedasticity assumptions were satisfied. More-so, due to the nature of the research design, the independence of observations assumption was also ensured, hence no assumption violation was found.

4.4.4 Estimating the ANOVA model and assessing overall model fit

Since the sampled data satisfies all the ANOVA assumptions, classical F-tests were also used to estimate the unbalanced ANOVA model. The **main** and **interaction** effects were calculated based on the type III sum of squares due to unbalancedness in the dataset. The F-tests and/or p-values as well as the effect sizes, were used to test the existence of group differences in the dependent variables.

The traditional F-tests were used to estimate the unbalanced homoscedastic model based on the simulated data sample. The ANOVA Table 4.10 below gives the estimated statistics for the model.

Table 4.10: Unbalanced & Homoscedastic ANOVA
Dependent variable: Amount spent

Source	**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**	**Partial Eta Squared**
Corrected model	25114.810[a]	11	23192.255	104.105	.000	.577
Intercept	9496909.554	1	9496909.554	42629.721	.000	.981
shopfor	74881.889	2	37440.944	168.065	.000	.286
usecoup	35224.549	3	11741.516	52.705	.000	.159
shopfor*usecoup	148357.243	6	24726.207	110.991	.000	.443
Error	186686.894	838	222.777			
Total	9935506.316	850				
Corrected error	441801.704	849				

a. R Squared = .577 (Adjusted R Squared = .572)

The *shopfor* main effect, *usecoup* main effect and the interaction effect, *(shopfor*usecoup)* p-values are each clearly significant (p-value $< 0.001 < \alpha$ (0.05)), which implies that theses effects are significantly contributing to the differences in amounts spent by customers. Furthermore, with the significant main and interaction effects in the model, the *Adjusted* R^2 *= 0,572* (greater than 0.5000) shows a fairly good model fit. The fitted model explains about 57% (0.572 x 100) of the variability in the response variable, *Amount spent*, being attributed to the reason for shopping (*shopfor*) and the source of coupons they used (*usecoup*)

The partial eta squared ($\eta^2_{partial}$) effect-size statistics on the unbalanced homoscedastic model are displayed in the last column in the above ANOVA table. Considering the guidelines suggested by Cohen (1988), it can be noted that both the main effects and the interaction effect had large effect size ($\eta^2_{partial} > 0.14$). The greatest effect (0.443) is attributed to the interaction between the reason for shopping (*shopfor*) and the source of coupons (*usecoup*). Though the effect sizes in this case are considered large, they were less than 50%.

Moreover, considering the profile plot for the same variables below, we can assess the interaction between the independent variables, *Who shopping for* and *Use coupons* as factors affecting

Amount spent in Figure 4.3 below.

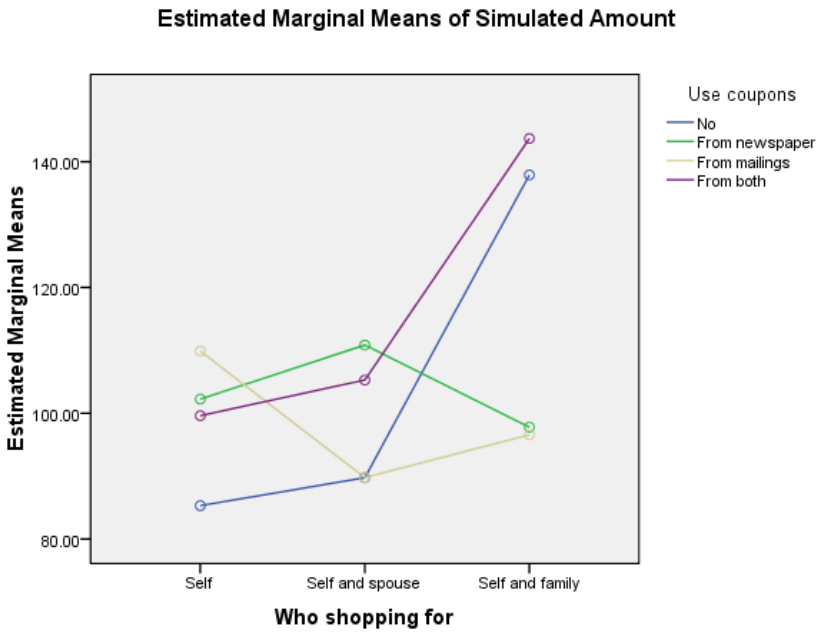

Figure 4.3: Profile Plot: Unbalanced Homoscedastic Model

Pictorial evidence of interaction between the factors *Shopfor* and *Usecoup* is clearly depicted by the non-parallel lines in the profile plot. There was little interaction between the *No* and *From both* levels of the *Usecoup* factor indicated by almost parallel line graphs in the profile plot (Figure 4.3), where as the *From mailings* and *From newspaper* levels exhibit interaction with the rest of the factor levels. On the other hand, the *Who shopping for* factor displays less interaction activity in the first two levels, *Self* and *Self and family*, while much association is seen in the third level *Self and family*. One can safely conclude that, the differences in the amounts spent is explained by these two factors in isolation and in combination.

Post hoc tests were conducted for the unbalanced homoscedastic model fitted to provide a further confirmation of the actual source of significant interaction between the factor levels. Table 4.11 displays the Bonferroni post hoc tests for the unbalanced homoscedastic model factors.

Table 4.11: Unbalanced & Homoscedastic ANOVA : Post Hoc

Dependent variable: Amount spent

(I)Who shop for	(J)Who shop for	Mean Difference (I-J)	S.E	Sig	95% Confidence Interval	
					Lower	Upper
Self	Self and spouse	.5436	1.25476	1.000	-2.4663	3.5535
	Self and family	-19.2411*	1.25476	.000	-22.2510	-16.2312
Self and spouse	Self	-.5436	1.25476	1.000	-3.5535	2.4663
	Self and family	-19.7847*	1.25254	.000	-22.7892	-16.7801
Self and family	Self	19.2411*	1.25476	.000	16.2312	22.2510
	Self and spouse	19.7841*	1.25254	.000	16.7801	22.7892
(I)Use coupons	(J)Use coupons	(I-J)	S.E	Sig	Lower	Upper
No	From newspaper	.7879	1.45316	1.000	-3.0551	4.6308
	From mailings	5.5304*	1.44978	.001	1.6964	9.3644
	From both	-11.5779*	1.44978	.000	-15.4119	-7.7439
From newspaper	No	-.7879	1.45316	1.000	-4.6308	3.0551
	From mailings	4.7425*	1.44632	.007	.9177	8.55674
	From both	-12.3658*	1.44632	.000	-16.1906	-8.5409
From mailings	No	-5.5304*	1.44978	.001	-9.3644	-1.6964
	From newspaper	-4.7425*	1.44632	.007	-8.5674	-.9177
	From both	-17.1083*	1.44292	.000	-20.9241	-13.2924
From both	No	11.5779*	1.44978	.000	7.7439	15.4119
	From newspaper	12.3658*	1.44632	.000	8.5409	16.1906
	From mailings	17.1083*	1.44292	.000	13.2924	20.9241

**. The mean difference is significant at the .05 level*

The post hoc tests confirm the profile plot graph results (Figure 4.3) that the *Who shopping for* factor had insignificant level combination in "*self - self and spouse*" levels (Sig. = 1.000), but the rest of the factor levels and their combinations were significantly contributing to the differences in amount spent by customers. Similar pattern can also be noted for the second factor (*Use coupons*) levels, only the level combination, "*No - From newspapers*" had insignificant contribution to the dependent variable (Sig. = 1.000), whereas the rest of the factor level combinations were significant (Sig. < 0.05). Generally, there is enough evidence of some interactions between the two factors in question and their respective factor levels.

4.5 Balanced and heteroscedastic model

A total of 100 balanced samples of 864 observations each, with equal cell sizes of 72 observations, were randomly sampled from the original *Grocery coupons.sav* dataset. This implies that the samples had unequal variances since it is just a subset of the original dataset. Similarly, based

on the argument by Krishnamoorthy, Lu and Mathew (2007) on heteroscedatic data samples, the parametric bootstrap approach was proposed to estimate the ANOVA model in order to control the type I error probability. The first five stages of the model building process are presented as usual.

4.5.1 Research Objectives and Design

The main aim of this study is to establish how the dependent variable *Amount spent* by customers on shopping is explained by the categorical factors, *Who shopping for* and *Use coupons*, under the influence of heteroscedasticity.

The research design was in such a way that 100 samples from the original dataset, *Grocery coupons.sav*, adopted from SPSS, were used to investigate the effects of heteroscedasticity on effect sizes of two-way fixed-effects ANOVA model with interactions. In this case, a two-way ANOVA design with interaction, having *Amount spent* as the metric continuous response variable depending on two categorical factors, *Who shopping for* and *Use coupons* is proposed.

4.5.2 Data Description and Sample Size

The 864 sampled observations were categorised into 4 by 3 factor levels of the explanatory variables. With equal cell sizes of 72 observations each, the sample had varying group standard deviations, indicating the presence of non-homogeneous group variances. Table 4.12 below shows the proposed balanced two-way ANOVA design, with varying group variances (standard deviations displayed in each cell).

Table 4.12: Balanced & Heteroscedastic Sample Standard Deviations

FACTOR A (*Who shopping for*)	**FACTOR B** (*Use Coupons*)				**Total**
	No	From Newspapers	From Mailings	From Both	
Self	10.21389	9.61191	11.54082	24.15467	15.92713
Self & spouse	7.09907	10.91701	6.30587	9.96651	10.10320
Self & family	12.10203	64.19930	14.67940	75.46532	52.14203
Total	14.79198	38.82084	21.41514	50.81324	34.98403

The largest variations in mean amount spent was realised in the *Self and spouse* level of *Who shopping for*. Varying standard deviations across the group cells is a justification that a heteroscedastic two-way ANOVA design should be used.

In each sample, the factorial combination had a cell size of 72 observations, the sample satisfied the recommended minimum cell size of 20 observations per cell (group) to maintain a statistical power of 0.80 (Hair et al., 2014). Hence the overall sample size was adequate to apply analysis of variance approach to test the factor effects.

4.5.3 Testing ANOVA Assumptions

ANOVA assumptions for this model were tested in Section 3.8.1. In brief, the assumption tests were as follows:

- Normality assumption was not perfectly met since the data exhibit little departure from normality due to skewness.
- Homogeneity of error variances assumption was violated. The Levene's equal variances test had a p-value = 0.000, which is less than 0.05 α significance level, resulting in the rejection of the null hypothesis for homogeneity of error variances.
- Independence of observations was satisfied due to the nature of research design.

With this scenario, the ANOVA assumptions are violated, which implies that classical F-tests would be inappropriate to apply (Erceg-Hurn & Mirosevich, 2008). However, robust parametric bootstrap approach can be used to estimate the model, since it can control the type I error rate.

4.5.4 Estimating the ANOVA model and assessing overall model fit

The fact that the basic ANOVA assumptions were violated, and that data transformation could not perfectly amend the problem as indicated in Chapter 3, the parametric bootstrap approach (defined in Section 2.4.3) will be used to approximate the test statistics in model estimation since it is robust to these assumption violations. The usual hypotheses of the **main** and **interaction** effects were tested, the Null hypotheses for the main effects and existence of interaction effects rejected if the p-values for factor/or interaction factor exceed the α level of significance (5%). Type I sum of squares were used for the balanced two-way ANOVA, with the F-tests and the effect sizes used to test the existence of group differences in the dependent variables.

The parametric bootstrap estimation, based on 1000 samples, was used to approximate the balanced heteroscedastic model based on the sampled data. The bootstrap ANOVA Table 4.13 below gives the estimated statistics of the model. The *shopfor* main effect p-value ($p < 0.001$) is clearly less than α (0.05), hence we conclude that the main effect of *shopfor* is significantly contributing to the differences in amounts spent by customers. Similarly, the *usecoup* main effect is also significant (p-value $< 0.001 < \alpha = 0.05$) at 5% level of significance. The same applies to the interaction effect, *(shopfor*usecoup)*, which is clearly significant (p-value $< 0.001 < \alpha = 0.05$). However, the *Adjusted* R^2 *= 0.224* (less than 0.5000) shows a poor model fit. The fitted model explains only 22.4% (0.224 x 100) of the variability in the response variable, *Amount spent*, being attributed to the reason for shopping (*shopfor*) and the source of coupons they used (*usecoup*).

Table 4.13: Balanced & heteroscedastic ANOVA

Dependent variable: Amount spent

Source	**Type I Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**	**Partial Eta Squared**
Corrected model	246772.346[a]	11	24423.850	23.613	.000	.234
Intercept	9496231.072	1	9496231.072	9995.564	.000	.921
shopfor	173817.489	2	86908.745	91.479	.000	.177
usecoup	31422.770	3	10474.257	11.025	.000	.037
shopfor*usecoup	41532.087	6	6922.014	7.286	.000	.049
Error	809437.914	852	950.045			
Total	10552441.33	864				
Corrected error	1056210.260	863				

a. R Squared = .234 (Adjusted R Squared = .224)

The effect sizes on the unbalanced heteroscedastic model were calculated using the partial eta squared ($\eta^2_{partial}$) statistics. The last column in the above ANOVA table gives the effect-size values of the calculated. Considering the guidelines suggested by Cohen (1988), it can be noted that only the *Who shopping for* effect had a fairly large effect size ($\eta^2_{partial} = 0.177 > 0.14$). On the other hand, the *Use coupons* contributed less 0.05 effect size (0.037), which resulted in a very low interaction effect size of 0.049 as well. Although both the main and interaction effects were all significant (Sig = 0.000), the strength or magnitude of their effects is very low (poor model fit) as indicated by a low Adjusted R^2 of 22.4%.

Furthermore, looking at the graphical presentation in form of a profile plot for the interaction of the variables in the model, we can assess the interaction between the independent variables, *Who shopping for* and *Use coupons* as factors explaining the *Amount spent* in Figure 4.4 below.

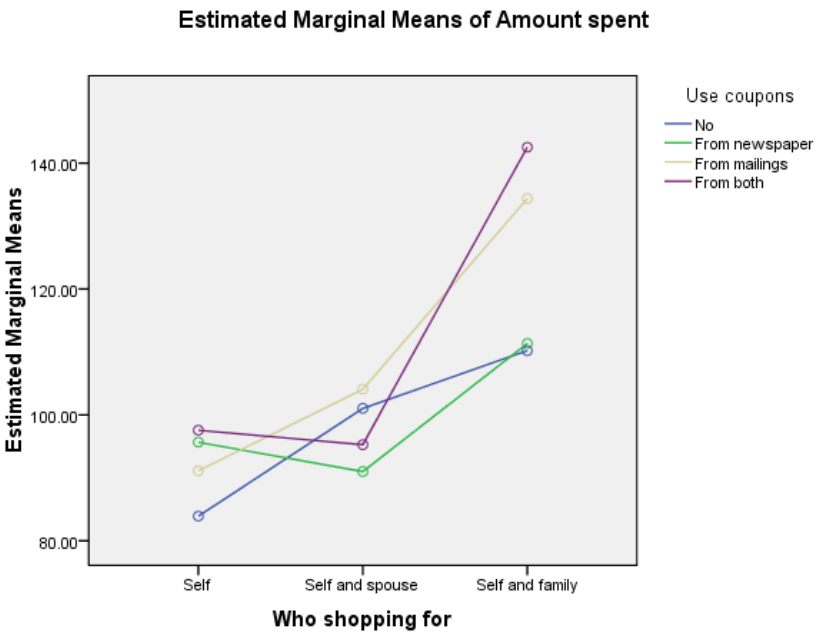

Figure 4.4: Profile Plot: Balanced Heteroscedastic Model

Evidence of interaction between the factor levels of *Who shopping for* is clearly displayed by crossing lines in the profile plot. Considerable interaction is vividly detected from the *Who shopping for* factor levels *self & spouse* and *self & family*, line graphs clearly non-parallel, whereas little interaction is displayed between the levels *self* and *self & spouse*, two pairs of lines almost parallel. On the other hand, the *Use coupons* factor levels seems to have non-significant interaction between the level combinations *From mailings - From both*, and *No - From newspapers*, since the respective pairs of plots are almost superimposed on one another. One can safely conclude that, the differences in the amounts spent was fairly influenced by these two factors in isolation and in combination.

Having a significant interaction effect as shown in Table 4.13, this leads us to the usual analysis of post hoc tests in order to establish the exact factor levels contributing to the differences in the group means of the amounts spent. Table 4.14 below displays the bootstrap Bonferroni post hoc tests for multiple comparisons for the two factors.

Table 4.14: Balanced & Heteroscedastic Post Hoc

Dependent variable: Amount spent

		Mean	Bootstrap[a]			
		Difference			BCa 95% CI	
(I)Who shop for	(J)Who shop for	(I-J)	S.E	Sig.	Lower	Upper
Self	Self and spouse	-5.7816	1.1279	.074	-8.0019	3.4824
	Self and family	-32.5594*	3.2488	.000	-39.0175	-26.3137
Self and spouse	Self	5.7816	1.1279	.074	-3.5013	7.9651
	Self and family	-26.7779*	3.2130	.000	-33.2729	-19.7124
Self and family	Self	32.5594*	3.2488	.000	26.0931	39.0807
	Self and spouse	26.7779*	3.2130	.000	20.6166	32.8515
(I)Use coupons	(J)Use coupons	(I-J)	S.E	Sig.	Lower	Upper
No	From newspaper	-.9478	2.8420	1.000	-7.4732	4.7628
	From mailings	-11.4725*	1.8101	.001	-15.1913	-7.7173
	From both	-13.4051*	3.6599	.000	-20.9670	-6.0973
From newspaper	No	.9478	2.8420	1.000	-4.0986	6.6362
	From mailings	-10.5247*	3.1074	.002	-16.2294	-4.6003
	From both	-12.4574*	4.4302	.000	-21.2139	-4.0733
From mailings	No	11.4725*	1.8101	.001	7.9412	14.9779
	From newspaper	10.5247*	3.1074	.002	3.3888	16.8415
	From both	-1.9327	3.7110	1.000	-9.1540	5.2537
From both	No	13.4051*	3.6599	.000	6.4423	20.6762
	From newspaper	12.4574*	4.4302	.000	4.0528	20.6762
	From mailings	1.9327	3.7110	1.000	-5.6144	9.8038

a. Unless otherwise noted, bootstraps are based on 1000 bootstrap samples

Reading from the bias-corrected and accelerated (BCa) 95% confidence intervals, the bootstrap post hoc tests indicate that in factor A (*Who shopping for*), the levels *Self* and *Self & spouse* had insignificant collective contribution (Sig.= 0.074 $>$ 0.05 α level) to the amount spent by customers because the BCa 95% Confidence Intervals cut through zero whereas the rest of the factor level combinations had significant interactions (Sig. = 0.000). On the other hand, factor B (*Use coupons*) had two level combinations, "*No - From newspapers*" and "*From both - From mailings*" which had insignificant contributions (Sig. = 1.000 $>$ 0.05 α level) to the dependent variable (BCa 95% Confidence Intervals include zero), whereas the rest of the factor level combinations were significantly contributing to the differences among the amounts spent.

4.6 Validation of Results

The bootstrap samples generated from the original dataset, based on 100 samples, for the heteroscedastic samples produced consistent results similar to the original sample. More-so, the

simulation samples (homoscedastic data samples simulated from the original dataset) also produced results consistent with the original data analysis results. Hence, based on the bootstrap and simulation sample results, we conclude that the results of this study had both internal and external validity.

4.7 Conclusion

This chapter presented the materials in a form of datasets used to test the research hypotheses, the methodology used to analyse the research data, outlining the first five stages of Hair et al (2014) Six-Stage model building process: the objectives of the research, research design, testing of ANOVA assumptions, estimation of the ANOVA models and assessing the model fit, as well as validation of results. The methods used to estimate the four ANOVA models, that is, the homoscedastic, heteroscedastic, balanced and unbalanced ANOVA model, and to assess their model fit were articulated. These methods spontaneously guaranteed the validity of the research results. The next Chapter 5, deals with the analysis and discussion of results in detail with regards to the main objective and hypotheses of the thesis.

Chapter 5

Analysis and Discussion of Results

5.1 Introduction

Based on the effect sizes estimated by Eta squared (η^2), Partial Eta squared ($\eta^2_{partial}$), and Omega squared (ω^2), the impact of heteroscedasticity and unbalancedness on the significance tests in each of the four models will be analysed. In each case, the changes in effect sizes will be evaluated and comparisons made for the three effect-size magnitudes. A detailed comparison on the effects of unbalancedness and heteroscedasticity on two-way fixed effects ANOVA models significance tests will be assessed based upon these changes in effect-sizes.

5.2 Effect-size analysis

A detailed comparative analysis of the impact of unbalancedness and heteroscedasticity was done through comparing the changes in Eta squared (η^2), Partial Eta squared ($\eta^2_{partial}$) and Omega squared (ω^2) effect sizes in the four models. Following the Cohen (1988) guidelines, the difference in effect size was considered significant if it exceeds 0.06 (6%), that is from moderate to large effect size. The effect sizes in each of the four models are summarised in Table 5.1 below.

Table 5.1: Effect sizes

Model	**Effect Size**		
1. Balanced Homoscedastic model	η^2	$\eta^2_{partial}$	ω^2
Factor A	0.007	0.285	0.007
Factor B	0.004	0.158	0.004
*Interaction A*B*	0.015	0.442	0.015
2. Unbalanced Homoscedastic model	η^2	$\eta^2_{partial}$	ω^2
Factor A	0.008	0.286	0.007
Factor B	0.004	0.159	0.003
*Interaction A*B*	0.015	0.443	0.015
3. Balanced Heteroscedastic model	η^2	$\eta^2_{partial}$	ω^2
Factor A	0.016	0.177	0.016
Factor B	0.003	0.037	0.003
*Interaction A*B*	0.004	0.049	0.004
4. Unbalanced Heteroscedastic model	η^2	$\eta^2_{partial}$	ω^2
Factor A	0.020	0.513	0.020
Factor B	0.004	0.166	0.004
*Interaction A*B*	0.005	0.221	0.005

A cursory look on the effect size summary Table 5.1 above, in all the four models the effect-size measures, eta squared (η^2) and omega squared (ω^2), estimated very small effect magnitudes (less than the hypothesised minimum 0.06 stated in hypothesis 1(a) $H_{0(A)}$). On the other hand, the partial eta squared ($\eta^2_{partial}$) estimated considerably greater effect sizes, ranging from small (less than 0.06) to large (more than 0.14). It can also be noted that the unbalanced heteroscedastic model had somehow exaggerated effect sizes amongst the four models. The ω^2 and η^2 estimated very low and almost equal effect-size magnitudes, far less than the $\eta^2_{partial}$ in each model. In support of Thompson (2007), the ω^2 effect sizes were always less than or equal to η^2, which indicated that ω^2 is more conservative than the η^2 and the $\eta^2_{partial}$, which many researchers argue that the later overestimates effect size. However, due to the fact that ω^2 and η^2 yielded insignificant effect-size magnitudes, the $\eta^2_{partial}$ is recommended for analysis purposes in this section since it had considerably significant effect measures that are above the hypothesised minimum (0.06).

5.2.1 Impact of unbalancedness on effect size

A summary of the effect-size differences of the balanced and unbalanced two-way fixed-effects ANOVA models is displayed in the last column of Table 5.2 below. Based on partial eta squared (η^2_p) effect-size measure and the hypothesis 1(b) $H_{0(B)}$ in the first chapter, the difference (d)

column in the table below summarises the discrepancies between the balanced and unbalanced models effect sizes.

Table 5.2: Effect changes due to unbalancedness

Homoscedastic models				
Effect Measure		**Balanced ANOVA**	**Unbalanced ANOVA**	**Difference *d***
Factor A:	η^2	0.007	0.008	0.001
	η_p^2	0.285	0.286	0.001
	ω^2	0.007	0.008	0.001
Factor B:	η^2	0.004	0.004	0.000
	η_p^2	0.158	0.159	0.001
	ω	0.003	0.003	0.000
A*B:	η^2	0.015	0.015	0.000
	η_p^2	0.442	0.443	0.001
	ω^2	0.015	0.015	0.000
Heteroscedastic models				
Factor A:	η^2	0.016	0.020	0.004
	η_p^2	0.177	0.513	0.336
	ω^2	0.020	0.016	0.004
Factor B:	η^2	0.003	0.004	0.001
	η_p^2	0.037	0.166	0.129
	ω^2	0.003	0.004	0.001
A*B:	η^2	0.004	0.005	0.001
	η_p^2	0.049	0.221	0.172
	ω^2	0.004	0.005	0.001

The first pair of homoscedastic models' effect-size differences were far below the hypothesised minimum effect magnitude of 0.06 especially in the main effects of both factors. Hence, based on the hypothesis 1(b), we fail to reject $H_{0(B)}$ and conclude that unbalancedness had little or no significant impact on main and interaction effect sizes in homoscedastic two-way fixed-effects ANOVA models.

However, considering the second pair of heteroscedastic models, significant partial eta squared (η_p^2) effect-size differences ($d > 0.06$) between the balanced and unbalanced models can be seen. Hence, we reject the hypothesis 1(c) $H_{0(C)}$ and conclude that, *ceteris paribus*, unbalancedness contributed to the differences (d) in main and interaction effect-size magnitudes in the heteroscedastic models.

Furthermore, a close comparison of the effect sizes from the balanced models against the unbalanced models counterparts indicates an increasing trend in the effect sizes as the model becomes unbalanced. This drives us to the conclusion that unbalancedness inflates the effect-size magnitudes in two-way fixed-effects ANOVA models.

5.2.2 Impact of heteroscedasticity on effect size

The effect of heteroscedasticity could be seen when comparing the balanced models against the unbalanced models effect size changes as shown in column labeled $\boldsymbol{d}$ in Table 5.3 below. Small to moderate effect size changes in the factor A and B main effects were depicted in balanced homoscedastic and heteroscedastic models. However, the interaction effect had a large effect difference in these models ($\boldsymbol{d} > 0.14$). Furthermore, medium to large effect changes ($\boldsymbol{d} > 0.06$) were seen when the model is unbalanced, with the largest effect change in the interaction effect realized by η_p^2.

Looking at the balanced models first in Table 5.3, it is clear that the homoscedastic model had generally higher partial eta squared (η_p^2) effect sizes ($\eta_p^2 > 0.14$) than the heteroscedastic counterparts, which suggests that the presence of heteroscedasticity reduced the effect-size magnitudes in the model. Based on hypothesis 1(c) $H_{0(C)}$ and considering the partial eta squared measure, the differences in the main and interaction effect sizes due to heteroscedasticity were significant (difference $\boldsymbol{d} > 0.06$) in the models involved. One can conclude that heteroscedasticity in balanced two-way fixed-effects ANOVA reduces the effect size of the model. However, the reduction depends on the severity of heteroscedasticity (Erceg-Hurn & Mirocevich, 2008).

Table 5.3: Effect sizes under heteroscedasticity

Balanced models				
Effect Measure		**Homoscedastic ANOVA**	**Heteroscedastic ANOVA**	**Difference *d***
Factor A:	η^2	0.007	0.016	0.009
	η_p^2	0.285	0.177	0.108
	ω^2	0.007	0.016	0.009
Factor B:	η^2	0.004	0.003	0.001
	η_p^2	0.158	0.037	0.121
	ω	0.003	0.003	0.000
A*B:	η^2	0.015	0.004	0.011
	η_p^2	0.442	0.049	0.393
	ω^2	0.015	0.004	0.011
Unbalanced models				
Factor A:	η^2	0.008	0.020	0.012
	η_p^2	0.286	0.513	0.227
	ω^2	0.007	0.020	0.013
Factor B:	η^2	0.004	0.004	0.000
	η_p^2	0.159	0.166	0.007
	ω^2	0.003	0.004	0.001
A*B:	η^2	0.015	0.005	0.010
	η_p^2	0.443	0.221	0.222
	ω^2	0.015	0.005	0.010

On the other hand, the pattern in the magnitude of effect sizes in unbalanced models was not consistent. The main effect-size magnitudes increased with heteroscedasticity, where as the interaction effect-size magnitude decreased. In conclusion, based on the results of this study, the presence of unequal variances yields inconsistent effect-size changes in unbalanced two-way fixed-effects ANOVA. Weird and over-estimated effect-size measures existed especially when the model is both unbalanced and heteroscedastic. However, there were significant differences in both the main and interaction effect sizes in these unbalanced models, leading to the rejection of hypothesis 1(c) $H_{0(C)}$ and a conclusion that heteroscedasticity does affect the main and interaction effect sizes in two-way fixed-effects ANOVA either positively or negatively.

5.3 Robust versus Traditional F-tests

When the variances are not equal and the cell sizes are different, the F tests are not robust enough to produce accurate results (Yigit and Gokpinar, 2010). Literature suggests several subrogates to the traditional F-tests, like the Welch test and Brown-Forsythe F-test, which are good alternatives when dealing with one-way ANOVA in the presence of heteroscedasticity and

unbalancedness. In our case, the traditional F-test was compared against the robust parametric bootstrap approach. We consider first the magnitudes of effect size each method could detect in the presence of heteroscedasticity and unbalancedness. Table 5.4 below is a presentation of the traditional F-test ANOVA for the unbalanced and heteroscedastic transformed original dataset.

Table 5.4: Balanced & heteroscedastic Traditional F-test ANOVA

Dependent variable: Natural log Amount

Source	**Type III Sum of Squares**	**df**	**Mean Square**	**F**	**Sig**	**Partial Eta Squared**
Corrected model	19.327[a]	11	1.757	107.688	.000	.597
Intercept	17018.641	1	17018.641	1043092.872	.000	.999
shopfor	13.790	2	6.895	422.613	.000	.514
usecoup	2.112	3	.704	43.140	.000	.139
shopfor*usecoup	3.015	6	.503	30.802	.000	.188
Error	13.036	799	.016			
Total	17387.144	811				
Corrected error	32.363	810				

a. R Squared = .597 (Adjusted R Squared = .592)

The above table displays almost similar results pattern shown in the bootstrap ANOVA Table 4.7 for the same model discussed in the previous chapter. The *shopfor* main effect was clearly significant (Sig. = 0.000) since this p-value is less than the α (0.05). The same applies to the second factor (*usecoup*) and the interaction effect (*shopfor*usecoup*), which are both significantly contributing to the differences in amounts spent by customers. The three effect sizes for the main and interaction effects all exceeded the hypothesized threshold, 0.06. Furtermore, the Adjusted $R^2 = 0.592$ is greater than 0.5000 (50%), showing a fairly good model fit as well. Hence, based on hypothesis 1(a) $H_{0(A)}$, the traditional F-test managed to detect significant main and interaction effect sizes since all $\eta^2_{partial}$ were greater than 0.06

However, the main focus is on the differences, if any, between the performance of the traditional F-test against the robust test (parametric bootstrap (PB) method in this case) in the presence of both heteroscedasticity and unbalancedness. To illuminate the important differences, a comparison of the magnitudes of effect sizes based on partial eta squared ($\eta^2_{partial}$) for the two approaches is displayed in Table 5.5 below.

Table 5.5: Traditional versus Robust Effect Sizes Under Heteroscedasticity

Factor Effect	**Traditional (F-test)**	**Robust (PB)**	**Difference *d***
Factor A: η_p^2	0.514	0.513	0.001
Factor B: η_p^2	0.139	0.166	0.027
Interaction A*B: η_p^2	0.188	0.221	0.033
Adjusted R^2	0.592	0.603	0.011

Although the differences (d) in the last column of Table 5.5 above were less than the hypothesised threshold 0.06, it can be clearly noticed that the parametric bootstrap (PB) had generally higher partial eta squared effect sizes across all factors than the traditional F-test. This is further supported by a higher Adjusted R^2 value (0.011 more) than the traditional F-test one. Hence we can conclude, based on these test results, that the robust parametric bootstrap approach performs better than the traditional F-test approach in terms of measuring effect sizes in a two-way fixed-effects ANOVA under the influence of heteroscedasticity and unbalancedness. This echoes what Yigit and Gokpinar (2010) said, that the traditional F-tests are less robust in the presence of unequal variances.

However, there was no big difference to deliberate on post hoc tests since the two methods produced the same post hoc results in terms of significant and insignificant factor levels under the same conditions.

5.4 Conclusion

This chapter presented a detailed comparative analysis of the effects of unbalancedness and heteroscedasticity on effect sizes in the four different two-way ANOVA models. These two phenomena proved to have significant impact on the magnitudes of effect sizes in isolation as well as in combination, depending on their severity of course. A comparison of the traditional or classic F-tests against the robust parametric bootstrap approach when dealing with two-way fixed-effects ANOVA in the presence of heteroscedasticity and unbalancedness was presented. Chapter 6 follows summarising the findings of the thesis, highlighting the conclusions, limitations and areas of further study.

Chapter 6

Summary, Conclusions, Limitations of the Study and Areas of Further Research

6.1 Introduction

Analysis of variance (ANOVA) models are useful tools applicable in various disciplines when dealing with multivariate analysis. Important assumptions of ANOVA have to be satisfied in order to get statistically accurate analysis results. A brief summary of the study on effects of heteroscedasticity and unbalancedness on effect sizes in two-way fixed-effects ANOVA significance tests conducted and the conclusions deduced are outlined. The study had its own limitations which are exposed here. Lastly, areas of further research which the researcher could not shed light on will be indicated in this chapter.

6.2 Summary of the study

Analysis of variance techniques produce accurate results when the ANOVA assumptions are not violated. In real life situations, it is usually rare to have data that satisfies all the ANOVA assumptions, especially the normality and homoscedasticity assumptions. Once these assumptions are not met, especially the equality of variance and balanced cell sizes, then the ANOVA F-tests suffer in terms of accuracy (Yigit & Gokpinar, 2010).

The study aimed at assessing the effects of heteroscedasticity and unbalancedness in conducting

analysis of variance tests. Precisely, the dissertation tried to establish:

- the effects of heteroscedasticity on effect sizes in two-way fixed-effects ANOVA model significance tests.
- the effects of unbalancedness on effect sizes in two-way fixed-effects ANOVA model significance tests.
- the ways of dealing with heteroscedastic and unbalanced data in order to achieve more accurate ANOVA tests results.

Chapter 1 outlined the background of the problem of unbalancedness and heteroscedasticity. The major aim, objectives and hypotheses to be tested were presented in detail. In addition to that, the chapter concluded with a brief overview of the theories involved in analysis of variance.

Chapter 2 discussed the theory and practices of ANOVA, heteroscedasticity and unbalancedness. Some ways of dealing with the problems of heteroscedasticity and unbalancedness in ANOVA models were reviewed. There are several types of ANOVA models depending on the number of factors and the nature of factors involved. Section 2.4 deliberated on fixed-effects ANOVA models (ANOVA Model Type I), starting from one-way to multi-way ANOVA models. Section 2.5 dwelt on random-effects ANOVA models, whose factor levels are a random sample selected from the entire population of factors (Gaugler & Akritas, 2013). The study focused on two-way fixed effects ANOVA. A combination of fixed and random factors makes a mixed-effect ANOVA model. A detailed explanation on the assumption underlying each type of ANOVA model as well as the types of hypotheses associated was presented.

ANOVA models are affected by the violation of equality of variance assumption. In the past, researchers have proposed some methods to deal with the problems of heteroscedasticity in ANOVA models. Most of the suggested methods have been applied to one-way ANOVA models only. Some of the discussed methods include; data transformation techniques (Hair et al., 2010); the parametric bootstrap (PB) approach (Krishnamoorthy, Lu & Mathew, 2007); Schott-Smith test (Schott & Smith, 1971); the Brown-Forsythe test (Brown & Forsythe, 1974); Welch's test (Welch, 1951). The impact of unbalancedness and heteroscedasticity measuring tools sug-

gested were the effect-size measures (eta squared, partial eta squared and omega squared).

Chapter 3 was data exploration. The original dataset was explored and prepared for proper data analysis purposes. Missing values and outliers were checked and corrective measures applied. The three basic assumptions of ANOVA (normality, homoscedasticity and independence of observations) were tested, and remedies applied for any violation of these assumptions. A parametric bootstrap approach was proposed for analysis of variance in the presence of unequal variances in the data.

The ANOVA assumptions tested in R and SPSS were as follows:

- **Normality assumption**: Tested by The Shapiro Wilk's test and the Q-Q plots.
- **Homoscedasticity assumption**: Tested by Levene's test at 5% significance level
- **Independence of observation assumption**: Guaranteed by the nature of study

Box plots produced in R, were used to detect outliers in the data. The analysis data was cleaned of all problematic outliers as they appeared to contribute to biased results. Elimination of these extreme values stabilised the ANOVA model without jeopardising the required sample size.

Chapter 4 presented the materials and methods used to analyse data in line with the main objectives of the study. The study aimed at establishing the impact of unbalancedness and heteroscedasticity on significance tests and effect size of two-way fixed effects ANOVA with interactions. One secondary dataset, unbalanced and heteroscedastic, extracted from SPSS, was used to simulate three samples for comparison purposes: the balanced heteroscedastic, balanced homoscedastic and unbalanced homoscedastic datasets. Analysis was done in R and SPSS software packages. The *Grocery coupons.sav* dataset adopted from SPSS, was used to simulate the other three datasets for investigating the effects of heteroscedasticity and unbalancedness in two-way fixed-effects ANOVA design with interactions. A two-way ANOVA design with interaction, having *Amount spent* as the metric continuous response variable depending on two categorical factors; *Who shopping for (shopfor)* (with three factor levels: *self*, *self and*

spouse, and *self and family*); and *Use coupons* (with four factor levels: *No*, *From newspaper*, *From mailings* and *From both*) was proposed.

The analysis and discussion of research results of the two-way fixed effects ANOVA models were the main focus of this chapter. For the balanced two-way ANOVA model, Type I sum of squares was applied, whereas the Type III sum of squares was used for unbalanced two-way ANOVA model (Ecerg-Hurn & Mirosevic, 2008). The main and interaction effects significance were tested at 5% significance level. The overall model fit was assessed by interpreting the Adjusted R^2. The model was deemed fit if the Adjusted R^2 was high (above 0.5 or 50%).

Chapter 5 presented a comparative analysis and discussion of results. Effects of unbalananced-ness and heteroscedasticity on two-way fixed effects ANOVA models were noted from the differences and shifts in effect sizes, measured by the Eta squared (η^2), Partial Eta Squared ($\eta^2_{partial}$), and Omega squared (ω^2) statistics. A comparison of the performance of the traditional F-tests against the robust parametric bootstrap approach was also done. Cohen (1988)'s guidelines for small (effect < 0.06), medium ($0.06 \leq$ effect < 0.14) and large (effect > 0.14) effect size were used to interpret the effect calculated and to test the hypotheses involved as summarized below.

- **Significant effect size**: Tested using Cohen (1988) guideline on small, medium and large effect size.
- **Significant change in effect size**: Guided by Cohen (1988) benchmarks (effect-size difference considered significant when it exceeds 0.06)

6.3 Findings

- There was insignificant change in effect sizes ($\eta^2_{balanced} - \eta^2_{unbalanced} < 0{,}06$) due to unbalancedness between the models from samples with equal variances. Based on the results of this study, it can be concluded that unbalancedness has little or no significant impact on the effect size in two-way fixed effects ANOVA, especially when the homoscedasticity assumption is satisfied. Furthermore, the unbalanced models portrayed slightly greater

effect sizes than the balanced counterparts, leading to the conclusion that unbalancedness inflates the effect-size magnitudes in two-way fixed-effects ANOVA models.

★ Heteroscedasticity generally reduces the effect size of both balanced and unbalanced two-way fixed-effects ANOVA models. However, an inconsistent pattern in the magnitude of effect sizes was realised in the unbalanced models. The main effect-size magnitudes increased with heteroscedasticity, whereas the interaction effect-size magnitude were significantly reduced. In conclusion, the study revealed that the presence of unequal variances (heteroscedasticity) in unbalanced two-way fixed-effects ANOVA yields inconsistent and over-estimated effect-size changes.

★ Partial Eta Squared ($\eta^2_{partial}$) tends to over-estimate the effect size regardless of sample size, whereas Omega Squared (ω^2) and Eta Squared (η^2) are more conservative than Partial Eta Squared. It is recommended that researchers interested in determining effect sizes in ANOVA models may, but not solely, rely on the most popular partial eta squared.

★ Based on effect size estimation, the traditional F-tests were found to be less robust than the parametric bootstrap approach in the presence of unbalancedness and heteroscedasticity. The parametric bootstrap (PB) technique could detect more effect-size magnitudes than the traditional F-tests. Hence, it is advised to consider the robust parametric bootstrap (PB) approach when dealing with heteroscedastic and unbalanced two-way ANOVA models.

6.4 Limitations of the study

- There are many effect size measures proposed in literature. Focusing only on Eta squared (η^2), Partial Eta squared ($\eta^2_{partial}$) and Omega squared (ω^2) was a limitation to the research. More interesting features and behaviours in other effect size measures not included in the study could have been missed in the process.

- Traditional F-tests are not robust in the presence of heteroscedasticity and unbalancedness in ANOVA tests. Several robust tests have been proposed in literature to alleviate the problem. However, some of these robust tests, like the Brown-Forsythe and Welch's test, are only applicable to one-way ANOVA. Better results could be achieved if the other

robust tests that are applicable in two-way fixed-effects ANOVA with interactions are explored.

- One secondary data set and simulation samples were used in the study. The sets of factors that were used could be insufficient to reflect the accurate results and findings that can be generalised for other different datasets. Insufficient information can lead to biased conclusions. However, the dataset and the simulated samples were chosen for analysis purposes that are only in line with the major objectives of the study.
- There could be some shortfalls contributed by the methods applied in both estimations (simulation sample) and analysis methods, which affect the accuracy of results and yet they were not catered for in this thesis. These include the retention of some outliers in the analysis sample, and the elimination of the other extreme values. With the reduced sample, though the sample size was statistically acceptable and large, essential information can be lost.

6.5 Areas of further research

★ Many researchers have in the past tried to alleviate the problems of heteroscedasticity in one-way ANOVA and MANOVA. Little has been done to assess and address the issue in Analysis of Covariance (ANCOVA).

★ There is a wide gap in the analysis of the effects of heteroscedasticity and/or unbalancedness in random effects models. Very few researchers have attempted the area.

APPENDICES

APPENDIX A: R CODES

R Codes for Simulation samples

>sim1.amnt <- rnorm(72,mean=83.9,sd=15)

>sim1.amnt

>Usecoup<-C(rep("No",72),rep("Nespaper",72),rep("Mail",72),rep("Both",72))

>Usecoup

>Shopfor<-C(rep("Self",288),rep("SelfSpouce",288),rep("SelfFamily",288))

>Shopfor

R Codes for ANOVA

>Bal.homo<-data.frame(Amnt.spent,Shopfor,Usecoup)

>Bal.homo

>Bal.homo.anova<-aov(Amnt.spent Shopfor*Usecoup)

>Bal.homo.anova

R Codes for Normality tests

>qqnorm(Amnt)

>qqline(Amnt)

>shapiro.test(Amnt)

R Codes for Outlier detection

>boxplot(Amnt)

APPENDIX B: F CODES

Bootstrap

```
BOOTSTRAP
/SAMPLING METHOD=SIMPLE
/VARIABLES TARGET=amtspent INPUT=shopfor usecoup
/CRITERIA CILEVEL=95 CITYPE=BCA NSAMPLES=1000
/MISSING USERMISSING=EXCLUDE.


UNIANOVA amtspent BY shopfor usecoup
/METHOD=SSTYPE(1)
/INTERCEPT=INCLUDE
/SAVE=PRED
/EMMEANS=TABLES(OVERALL)
/EMMEANS=TABLES(shopfor)
/EMMEANS=TABLES(usecoup)
/EMMEANS=TABLES(shopfor*usecoup)
/PRINT=ETASQ HOMOGENEITY DESCRIPTIVE
/CRITERIA=ALPHA(.05)
/DESIGN=shopfor usecoup shopfor*usecoup.
```

APPENDIX C: EDITORIAL LETTER

REFERENCES

[1] Algina, J., Keselman, H.J. and Penfield, R.D. (2005). An alternative to Cohen's Standardized Mean difference interval in the two independent group case. *Psychological Methods*, 10: 317-328.

[2] Algina, J., Keselman, H.J., and Penfield, R.D. (2006). Confidence intervals for an effect size when variances are not equal. *Journal of Modern Applied Statistical Methods*, 5: 2-13.

[3] Ananda, M.A. and Weerahandi, S. (1997). Two-Way ANOVA with unequal cell frequencies and unequal variances. *Statistica Sinica*, 7:631-646.

[4] Arnold, S.F. (1981). *The Theory of Linear Models and Multivariate Analysis.* John Wiley: New York

[5] Bakeman, R. (2005) Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37 (3): 379-384

[6] Bathke, A. (2004) The ANOVA test can still be used in some balanced designs with unequal variances and nonnormal data. *Journal of Statistical Planning and Inference*, 126: 413-422

[7] Brown, J.D. (2008) Effect size and eta squared. *Shiken: JALT Testing and Evaluation SIG Newsletter.* 12 (2): 38-43

[8] Brown, M.B. and Forsythe, A.B. (1974) The small sample behavior of some statistics which test the equality of several means. *Technometrics*, 16: 129-132

[9] Box, E.G.P. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, 36: 317-346.

[10] Cohen, J. (1965). Some Statistical Issues in Psychological Research. (In B. B. Wolman (Ed.), *Handbook of Clinical Psychology*, (pp. 95 - 121). New York: McGraw-Hill).

[11] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, N.J: Erlbaum

[12] Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112 (1): 155-159.

[13] Erceg-Hurn, D.M. and Mirocevich, V.M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist*, 63: 591-601.

[14] Fritz, M. S., Taylor, A. B., and MacKinnon, D. P. (2012). Explanation of two anomalous results in statistical mediation analysis. *Multivariate Behavioral Research*, 47, 61-87

[15] Fujikoshi, Y. (1993). Two-Way ANOVA models with unbalanced data. *Discrete Maths*, 116:315-334.

[16] Gaugler, T. (2008). Nonparametric Models for Crossed Mixed effects Designs.
URL https:books.google.co.zw/books?isbn=1109016735

[17] Gaugler, T. and Akritas, M.G. (2013). Testing for Main Random Effects in Two-Way Random and Mixed Effects Models: Modifying the F Statistics. *Journal of Probability and Statistics*: Article ID 708540 *URL http://dx.doi.org/10.1155/2013/708540*

[18] Glass, G.V., McGraw, B. and Smith, M.L. (1981).*Meta-Analysis in social research.* New Park, CA: Sage

[19] Hair, J.F. Jr, Black, W.C., Babin, B. J. and Anderson, R. E., (2010). *Multivariate Data Analysis. A global perspective (7th ed.).*
Pearson Education, Inc.: USA

[20] Hair, J.F. Jr, Black, W.C., Babin, B. J. and Anderson, R. E., (2014). *Multivariate Data Analysis (7th ed.).*
Pearson Education, Inc.:Sussex, England

[21] Harrar, S.W. and Bathke, A.C. (2008). Nonparametric methods for unbalanced multivariate data and many factor levels. *Journal of Multivariate Analysis*, 99: 1635-1664.

[22] Hedges, L.V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, 6 (2): 107-128

[23] Iker, G. (2013). What method do you think is the best when you are carrying out an analysis of Variance between groups?
URL https://www.researchgate.net/post

[24] James, G.S. (1951). The comparison of several groups of observations when the ratios of population variances are unknown. *Biometrika*, 38: 324-329.

[25] Kalinowski, P. and Fidler, F. (2010). Interpreting significance: the difference between statistical significance, effect size, and practical importance. *Psychological Science*, 10 (1): 50-54

[26] Keselman,H.J., Algina, J., Lix,L.M., Wilcox, R.R. and Deering, K.N. (2008). A generally robust approach for testing hypotheses and setting confidence intervals for effect sizes. *Psychological Methods*, 13 (2):110-129

[27] Kirk, R.E. (2003). *The importance of effect magnitude.* In S.F. Davis (Ed.), Handbook of research methods in experimental psychology. Oxford, UK: Blackwell

[28] Kondo-Brown, K., and Brown, J. D. (Eds.) (2008). *Teaching Chinese, Japanese, and Korean heritage language students.* New York: Lawrence Erlbaum Associates.

[29] Krishnamoorthy, K., Lu, F. and Mathew, T. (2007). A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics and Data Analysis* 51:5731-5742.

[30] Krishnamoorthy, K. and Lu, F. (2010). A parametric bootstrap solution to the MANOVA under heteroscedasticity. *Journal of Statistics, Computing and Simulation* 80:(8), 873-887.

[31] Kulinska, E. and Staudte, R.G. (2006). Interval estimates of weighted effect sizes in the one-way heteroscedastic ANOVA. *British Journal of Mathematical and Statistical Psychology*, 59:97-111.

[32] Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative Science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4 (863):1-12

[33] Larson, M.G. (2008). Analysis of Variance. *Circulation*, 117:115-121

[34] Lewicki, P. and Hill, T. (2007). Statistics: Methods and applications. *URL https://www.statsoft.co.za*

[35] Loeza-Serrano, S. and Donev, A.N. (2014). Construction of Experimental Designs for

Estimating Variance Components. *Computational Statistics and Data Analysis*, 71:1168-1177

[36] Lungsrud, O. (2003). Anova for unbalanced data: Use type II instead of type III sum of squares. *Statistics and Computing*, 13:163-167.

[37] McDonald, J.H. (2014). Handbook of biological statistics.
URL http://www.biostathandbook.com/twowayanova.html

[38] Milliken, G.A. and Johnson, D.E. (1984). *Analysis of messy data. Volume 1: Designed experiments*. Van Nostrand Reinhold: New York, USA.

[39] Moder, K. (2007). How to keep the Type I Error Rate in Anova if Variances are heteroscedastic. *Ausrian Journal of Statistics*, 36 (3): 179-188

[40] Mosteller, F. and Turkey, J.W. (1977). *Data Analysis and Regression. Reading*. MA: Addison-Wesley

[41] Nandy, K. (2012). *Understanding and Quantifying Effect Sizes*. Department of Biostatistics, School of Public Health: university of California Los Angeles.

[42] National Institute of Standards and Technology (2015). Education and Training: Data Sets.
URL http://www.itl.nist.gov/div898/education/datasets.html

[43] Olejnik, S., and Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods*, 8, 434-447

[44] Olive, D.J. (2010). *Multiple Linear and 1D Regression*. Department of Mathematics: Southern Illinois University.

[45] Preacher, K. J., and Kelley, K. (2011). Effect size measures for mediation models: Quantitative strategies for communicating indirect effects. *Psychological Methods*, 16, 93-115

[46] Rutherford, A. (2012). ANOVA and ANCOVA: *A glm approach*. Wiley: New York

[47] Sahai, H. and Khurshid, A. (2005). A Biography on Variance Components: An Introduction and Update: 1984-2002. *Statistica Applicata*, 17 (2): 191-339

[48] Sawyer, S. (2009). Analysis of variance. The fundamental concepts. *The journal of Manual and Manipulative Therapy*, 17:E27-E38.

[49] Schott, J. (2007). Some high-dimensional tests for one-way MANOVA. *Journal of Multivariate Analysis*, 98: 1825-1839.

[50] Schott, A.J. and Smith, T.M.F. (1971). Interval estimates for linear combinations of means. *Applied Statistics*, 20:276-285

[51] Searle, S. (1987). *Linear models for unbalanced data.* Wiley: New York.

[52] Srivastava M.S. (2007). Multivariate theory for analysing high dimension data. *Journal of Japan Statistics Soc*, 37: 53-86.

[53] Srivastava, M.S. and Fujikoshi, Y. (2007). *Multivariate Analysis of Variance with fewer observations than the dimension.* Department of Statistics: University of Toronto

[54] Tian, L.L., Ma, C.X., and Vexler, A. (2009). A parametric bootstrap for comparing heteroscedastic regression models. *Communications in Statistics - Simulation and Computation*, 38:1026-1036.

[55] Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, 44, 423-432.

[56] Wang, L. and Akritas, M.G. (2006). Two-Way heteroscedastic ANOVA when the number of levels is large. *Statistica Sinica*, 16:1387-1408.

[57] Wang, H. and Akritas, M.G. (2011). Asymptotically distribution free tests in heteroscedastic unbalanced high dimensional ANOVA. *Statistica Sinica*, 21: 1341-1377

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

[59] Wen, Z. and Fan, X. (2015). Monotonicity of effect sizes:Questioning kappa-squared as Mediation Effect Size measure. *Psychological Methods*, 20 (2): 193-203

[60] Xu, L. and Wang, S.A. (2007a). A new generalised p-value for ANOVA under heteroscedasticity. *Statistics and Probability Letters*, 78: 963-969

[61] Xu, L. and Wang, S.A (2007b). A new generalised p-value and its upper bound for

ANOVA under unequal error variances. *Communications in Statistics Theory and Methods*, 37: 1002-1010

[62] Xu, L., Yang, F., Abula, A. and Qin, S. (2013). A parametric bootstrap approach for two-way ANOVA in presence of possible interactions with unequal variances. *Journal of Multivariate Analysis*, 115: 172-180

[63] Xu, L., Yang, F., Chen, R. and Yu, S. (2015). A parametric bootstrap test for two-way ANOVA model without interaction under heteroscedasticity. *Communications in Statistics - Simulation and Computation*, 44:(5), 1264-1272

[64] Yigit, E. and Gokpinar, F. (2010). A simulation study on tests for one-way ANOVA under the unequal variance assumption. *Communications de la Facult des sciences de l'Universit d'Ankara. Series A: Mathematics and Statistics*, 59:(2),15-34

[65] Zetterberg, P. (2013). *Effects of unbalancedness and heteroscedasticity on two-way MANOVA tests*. Department of Statistics: Stockholm University.

[66] Zhang, G. (2015a). A parametric bootstrap approach for One-way ANOVA under unequal variances with unbalanced data.*Communications in Statistics - Simulation and Computation*, 44:(4), 827-832

[67] Zhang, G. (2015b) Simultaneous confidence intervals for pairwise multiple comparisons in a two-way unbalanced design with unequal variances. *Journal of Statistical Computation and Simulation*, 85:(13), 2727-2735

[68] Zhang, J.T. (2012). An approximate degrees of freedom test for heteroscedastic two-way ANOVA. *Journal of Statistical Planning and Inference*, 142: 336-346

[69] Zhang, J.T. and Xiao, S. (2012). A note on the modified two-way manova tests. *Statistics and Probability Letters*, 82: 519-527.