UNIVERSITY OF CAPE COAST

COMPARING STUDENTS' PERFORMANCE IN EXAMINATIONS AT THE ST. JOSEPH SENIOR HIGH SECONDARY/TECHNICAL SCHOOL

EMMANUEL KWAME DEMORDZIE

2010

UNIVERSITY OF CAPE COAST

COMPARING STUDENTS' PERFORMANCE IN EXAMINATIONS AT THE ST. JOSEPH SENIOR HIGH SECONDARY/TECHNICAL SCHOOL

BY

EMMANUEL KWAME DEMORDZIE

Dissertation submitted to the Department of Mathematics and Statistics of the School of Physical Sciences, Faculty of Science, University of Cape Coast in partial fulfillment of the requirement for award of Master of Science degree in Statistics.

JULY 2010

DECLARATION

Candidate's Declaration

I hereby declare that this thesis is the result of my own original work and that no part of it has been presented for another degree in this university or elsewhere.

Candidate's Signature: …… Date:……..

Name: Emmanuel Kwame Demordzie

Supervisor's Declaration

I hereby declare that the preparation and presentation of the dissertation were supervised in accordance with the guidelines on supervision of thesis laid down by the University of Cape Coast.

Supervisor's Signature: ……Date:…….

Name: Francis Eyiah- Bediako

ABSTRACT

My motivation to embark on this project work was triggered by the academic performance of the students in St. Joseph Senior High Secondary/Technical. The main objectives of this study were to compare the performance of male and female students, use principal components to determine indices for ranking the students in descending order.

The Data used in the analysis were sourced from the students' of St. Joseph Senior High Secondary/Technical third term terminal examination results. Hotelling's T-squared test and Principal Component Analysis were the main statistical tools used in the study.

The test revealed that the males outperformed the females in the examination. The first principal component of the unrotated component matrix, was found to be the most suitable index that was used to determine the performance of students by ranking.

ACKNOWLEDGEMENTS

I am highly grateful to Mr. Francis Eyiah-Bediako a lecturer of the Department of Mathematics and Statistics of the University of Cape Coast for his critical supervision, comprehensive remarks, suggestions and criticisms to make this work complete.

Sincerely, I am much thankful to all the lecturers of the department especially those who were directly engaged in giving me guidance, counseling and training.

Finally, I am much indebted to my students of St. Joseph Senior High Secondary Technical in Bekwai of Ashanti, colleagues and friends who in diverse ways contributed to make this work successful.

DEDICATION

In memory of my parents.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Background of the study

Academic achievement is a crucial ingredient of learning during a course of study. It is directly related to students' performance. Invariably, teachers are confronted with the assessment of students' performance from time to time or from term to term. In higher education, the term "assessment" has taken on a rather broad dimension. It has been defined by Rowtree (1977) as "getting to know our students and the quality of their learning". Ramsden (1992) describes it as a way of teaching more effectively through understanding exactly what students know and do not know. Thus, assessment enables the teacher to understand the processes and outcomes of the student learning. It helps to determine what students actually achieve in their study. Such meaningful information on student learning can be useful for academic improvement. Assessment plays a key role in determining the quality of student learning.

The performance of students varies in every subject and the academic performance of students in school can never be the same. In effect, the obvious variation in academic performance of students is of interest to teachers, institutions, organizations and governments to determine and reward overall best students with the single purpose for motivation and recognition for outstanding academic performance. Many schools put in place an award schemes to reward outstanding performances of their students mostly, on Speech and Prize Giving

Days. Institutions give awards for many disciplines, especially for academic excellence among others. Universities all over the world also have various forms of award schemes, which are awarded students that distinguish themselves for academic excellence in various fields during graduation ceremonies. Therefore, many governments, for instance the Ghana Government instituted an award scheme known as the Presidential Awards. It is a yearly award given to outstanding students who excel in Basic Education Certificate Examination (BECE) and the May/June West African Senior Secondary Certificate Examination (WASSCE), selected from each of the regions in the country by the President on the eve of 6th March, Independence Day. Other institutions such as West African Examinations Council (WAEC) also initiated yearly awards scheme to reward outstanding candidates in the May/June WASSCE. International institutions like University of Cambridge International Examinations (CIE), the world's largest provider of international qualifications for 14 to 19 year old male and female students, also rewards its students for academic excellence.

This research work was highly motivated by the desire of the researcher to study the academic performance of students in St. Joseph Secondary /Technical school. It is one of the catholic institutions established by the Marist Brothers of the Catholic Church in 1991 in Ahwiren near Ashanti-Bekwai. It was started with thirty (30) students; however, the school currently has an enrolment of about 500 students.

The researcher was challenged by the average performance of the students in this school, and was motivated to determine whether there is any difference in

the performance of both male and female students. On the other hand, the researcher was interested in identifying the overall best student in one of the classes, using the end of the term examination scores. Many at times, we rank students in descending order of performance to determine the overall best student in classes. Simple methods like summing of individual scores in the various courses and arranging them in order of magnitude from biggest to lowest are applied manually to achieve the same goal. Nevertheless, we are highly motivated to use principal component method, so that the scores/marks are standardized such that a model is formulated to estimate indices for the students. The indices can then, be used to rank every member of the class. In effect, we are going to apply one of the multivariate methods (PCA) to analyze the data in this way and find out the outstanding student in the class. It would have been important to determine the overall best student by considering other areas of school life in its entirety. The ultimate goal of the research was to determine an index from the scores obtained by the students. Then use it to rank the students in order of performance to come out with the overall best students in the Senior High School (SHS) 2.

Objectives of the study

These are to:

(1) compare the performance of male and female students

(2) use principal components to determine the best index for ranking the students in descending order

(3) identify subjects that are influential in forming the principal components

(4) determine the order of performance of students using the index

Research Questions

The study seeks to answer the following questions:

(1) Is there any difference between the performance of male and female students?

(2) Is it possible to use the students' scores to determine an index for ranking the students according to performance?

(3) What subjects are influential in forming the principal components?

(4) What is the order of students by performance?

Literature Review

The desire of world leaders and stakeholders in many countries including Ghana to promote gender development and empowerment cannot be over emphasized. Besides, the widespread belief that males do well in academic fields than the females has been the concern of social researchers and the public. In view of these, many studies have been conducted to find out how the females are fairing in various fields of endeavour alongside their male counterparts. Similar and related studies by Felson (1991) stated that the widespread belief that males outperform females in Mathematics is apparently a myth. Besides, it was revealed in another study that states that, Gender differences in mathematics performance that favour males are usually attributed to gender socialization (Boswell 1980; Brush 1980; Linn and Peterson, 1986).

It also came to light and was reported that, basically, girls are taught that they have low aptitude for mathematics and that they will not need skills in advance mathematics as adults (Chipman and Thomas, 1985). Halpern (1986) concluded that, the finding that males outperformed females in tests of quantitative or mathematics ability is robust. She stated that the differences emerge reliably between 13 -16 years of age. Further, other researchers like Sells (1973) and Chipman and Thomas (1985) did similar studies. An article written by Agyei and Eyiah-Bediako (2008) which was published in the Journal for Gender and Behaviour, was on gender differences in Mathematics performance. In this study, it was found that there is no difference between Mathematics performance of male and female students of the Mathematics and Statistics Department of University of Cape Coast.

In another development, determination of overall best student is done in many institutions across the world depending on the motivation and the goal that is being pursued. In most schools in the United States of America (USA), Asia and Africa, the overall best student is considered according to his or her academic performance, behavior at home and other extra-curricular activities he or she participates at school. West Africa Examinations Council (WAEC) instituted an award scheme in 1984 to reward students for outstanding performance in West African Senior Secondary Examination Certificate (WASSCE). WAEC opens this award scheme for all students in senior secondary school for all the West African countries that are members of the council. They award only three students who emerge as the overall best candidate in the May/June WASSCE every year. In this

regard, several students across the member countries in West Africa have received awards since the inception of the scheme in 1984. In 2007 West African Senior Secondary Examination, Kwame Akoi, a medical student at Kwame Nkrumah University of Science and Technology (KNUST) was adjudged the overall best candidate (GNA, Friday, 5 December 2008).WAEC also moved a step further to award the schools for producing excellent students. There are also awards for the second and the third overall best students or candidates. The selection criteria as outlined by Patience Ayensu, Head of the National Office at WAEC in Accra explained that to be eligible to receive an award from WAEC, candidate must obtain a minimum of eight grade A1. These awards are in three categories namely, the Excellence Awards, Distinction Awards and the Merit Awards.

In line with the award scheme, an 18-year-old former student of King's High School, Satellite Town, Lagos, Master Maduka David Immanuel, has been adjudged the overall best candidate with a total score of 718.43. He recorded A1 in eight subjects and A2 in Biology. Remarkably, that was the first time since 1984 when the award was initiated by WAEC that only one candidate emerged the winner of the National Distinction Award for May/June 2008 WASSCE. This was because unlike previous awards that attracted the three candidates in the May/June WASSCE, Master Emmanuel was alone as other candidates did not meet the laid down criteria for the honour (Prince Education, Feb 16, 2010).

University of Cambridge International Examinations (CIE) is the world's largest provider of international qualifications for students of 14 to 19 years old. CIE is recognized internationally and provides courses, examinations and

qualifications to over 170 different countries. CIE examination results are expressed as grades and percentages and are internationally benchmarked. CIE also in its quest to motivate and give recognition to outstanding students initiated an award scheme in this direction. For this purpose, the CIE in association with the Knowledge and Human Development Authority (KHDA) in recent times awarded the Shaikh Maktoum Bin Mohammad Bin Rashid Al Maktoum Cambridge Outstanding Achiever Awards 2010 to two students of the Oxford School of Dubai, one of the leading British curriculum institutions in the United Arab Emirates (UAE). The awards which are endorsed and supported by HH Sheikh Maktoum Bin Mohammad Bin Rashid Al Maktoum, Chairman of the Dubai Technology and Media Free Zone were awarded in recognition of the students academic excellence in year 2009. The two students of The Oxford School, Tanvir Sajed and Syed Zeyd Abduraman at the AS Level were awarded in recognition of their outstanding performance in the Examinations held in 2009. Both Students had four As in their Science subjects, Biology, Chemistry, Physics and Mathematics.

Another prominent award is Sir John Monash Medal for Outstanding Achievement. This award is awarded to a student who has completed the academic requirements for the degree of Bachelor of Law in 2009 and is eligible to graduate and is adjudged to have an excellent academic record and to have demonstrated a significant commitment while at Monash to advancing the University's goal of social justice, human rights and a sustainable environment. The medal was first awarded in 2009, to a student who completed in 2008. Hugh

Evans was the first winner of the medal for the year 2008-09 (www.law.monash.edu/prize/sir-joh-monash).

Assessment of students is an integral part of education, teaching and learning. It plays a very important role in the academic performance of students. It enables the teacher to assess his or her methods of teaching, by taking stock of his interactions with the class, effective communication, efficient use of instructional time, handling of students challenges appropriately and reinforcement of skills taught among others. Students are able to identify their weaknesses and their strengths through assessment.

Data

The subjects taken in the examination by the students were Integrated Science, Core Mathematics, English Language, Economics, Elective Mathematics, Government, Geography and Frecnh and constituted the variables for the study. Where, Integrated Science, Core Mathematics and English Language and then Economics, Elective Mathematics, Government, Geography and Frecnh constituted the core and elective subjects respectively for the class. Principal component analysis technique is appropriate for analyzing the data, since the measurements obtained on these variables for each student in the class, constitute multivariate data. The PCA therefore, can provide an index for ranking the students. The data were obtained from third term terminal report of 2008/09 academic year, for 47 SHS General Arts 2A students' of St. Joseph

Secondary/Technical School. This was made up of the actual scores for 28 male and 19 female students.

Outline of study

In this study, Chapter One dwells on the background, objectives, research questions, and the literature review of the study and then the variables used in the analysis. Chapter Two reviews the methods applied in the study such as Hotelling's T- squared and Principal Component Analysis. Chapter Three also dwells on preliminary analysis, which mainly outlines the descriptive statistics of the data. Chapter Tour dwells on further analysis of the data in which advance techniques reviewed in Chapter Two were employed. Finally, Chapter Five captured the summary, discussion and conclusions of the study.

CHAPTER TWO

REVIEW OF METHODS

Introduction

We are interested in analyzing whether there is significant difference between academic performance of both male and female students using Hotelling's T-squared test. For this purpose, the Hotelling's T-squared test is briefly reviewed. Besides, the concept of basic theory and methods of principal component analysis which is the main technique used in this research is also reviewed in this Chapter.

Hotelling's T- squared

Hotelling's T-squared is a statistic for testing the equality of vector means from two multivariate populations. It is an analogue of the univariate student's t-test. For a random sample of size $\boldsymbol{n_1}$drawn from population 1 and a sample of size $\boldsymbol{n_2}$ drawn from population 2, the observations on p- variables can be arranged as:

Population 1				**Population 2**			
$\boldsymbol{X_{11}}$	$\boldsymbol{X_{21}}$	$\cdots$	$\boldsymbol{X_{p1}}$	$\boldsymbol{X_{11}}$	$\boldsymbol{X_{21}}$	$\cdots$	$\boldsymbol{X_{p1}}$
$\boldsymbol{X_{12}}$	$\boldsymbol{X_{22}}$	$\cdots$	$\boldsymbol{X_{p2}}$	$\boldsymbol{X_{12}}$	$\boldsymbol{X_{22}}$	$\cdots$	$\boldsymbol{X_{p2}}$
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$		$\vdots$
$\boldsymbol{X_{1n_1}}$	$\boldsymbol{X_{2n_1}}$		$\boldsymbol{X_{pn_1}}$	$\boldsymbol{X_{1n_2}}$	$\boldsymbol{X_{2n_2}}$		$\boldsymbol{X_{pn_2}}$

(2.1)

where $\boldsymbol{X_{11}}$, $\boldsymbol{X_{12}}, \cdots, \boldsymbol{X_{1n_1}}$, $\boldsymbol{X_{21}}$, $\boldsymbol{X_{22}}, \cdots, \boldsymbol{X_{2n_1}}$, $\cdots$ are observations involved in variables 1, 2,$\cdots$,p respectively for population 1 and $\boldsymbol{X_{11}}$, $\boldsymbol{X_{12}}, \cdots,$

$\boldsymbol{X_{1n_1}}, \boldsymbol{X_{21}}, \boldsymbol{X_{22}}, \cdots, \boldsymbol{X_{2n_2}}, \cdots$ are observations in variables 1, 2,⋯ ,p respectively for population 2.

In comparison of vector means of the two populations using the Hotelling's T-squared, these assumptions must be followed; both populations are independent, multivariate normally distributed with means $\boldsymbol{\mu_1}$ and $\boldsymbol{\mu_2}$ and variance-covariance $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ respectively. There are two types of this test and each is used depending on first, when the variance-covariances are equal and the second is when the variance-covariances are not equal. However, for the purpose of this study, the researcher has chosen to use the test for the unequal variance-covariance matrix, since the population involved in the study has two variance-covariance that are not equal.

Thus, for unequal $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$ the test statistic is given as:

$$\boldsymbol{T^2} = [(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}}) - (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})]'[(\frac{1}{n_1}\boldsymbol{S_1} + \frac{1}{n_2}\boldsymbol{S_2})]^{-1}[(\bar{\boldsymbol{X}}_{\boldsymbol{1}} - \bar{\boldsymbol{X}}_{\boldsymbol{2}}) - (\boldsymbol{\mu_1} - \boldsymbol{\mu_2})]$$

(2.2)

This is distributed as $\boldsymbol{T^2} = \frac{(\boldsymbol{n-1})\boldsymbol{p}}{(\boldsymbol{n-p})}\ \boldsymbol{F_{p,\ n-p}}$ .The sample variance-covariance can be calculated using:

$$\boldsymbol{S_i} = \frac{\boldsymbol{1}}{(\boldsymbol{n_i-1})}\sum_{\boldsymbol{j=1}}^{\boldsymbol{n_i}}(\boldsymbol{X_{ij}} - \bar{\boldsymbol{X}}_{\boldsymbol{i}})(\boldsymbol{X_j} - \bar{\boldsymbol{X}}_{\boldsymbol{i}})'$$

(2.3)

Hypothesis Test

We can test the hypothesis,

$\boldsymbol{H_o}$: there is no difference between the vector means of the populations 1 and 2 (i.e. $\boldsymbol{\mu_1 - \mu_2 = 0}$) against the alternative hypothesis,

$\boldsymbol{H_1}$: There is a difference between the vector mean of the populations 1 and 2

(i.e. $\boldsymbol{\mu_1 - \mu_2} \neq 0$).

Principal Component Analysis

PCA is one of the techniques used widely with large multidimensional data sets. Its use allows reduction in the number of variables in multivariate data, whilst retaining as much as possible the variation present in the data set. It is concerned with explaining the variance-covariance structure through a few linear combinations of the original variables. The formation of the maximum number of new variables is equal to the number of correlated original variables; nevertheless, the new variables are uncorrelated among themselves. Hence, PCA is most useful if one simply wants to reduce relatively large number of variables into a smaller set of variables that captures the same or the original information (Sharma, 1996). The results of a PCA are usually discussed in terms of component scores and loadings (Shaw, 2003)

Objectives of Principal Component Analysis

The objective of PCA is to determine a new set of orthogonal axes such that:

1. The coordinates of the observations with respect to each of the axes give the values for the new variables. The new axes or variables are called principal

components and the values of the new variables are called principal component scores.

2. Each new variable is a linear combination of the original variables.
3. The first new variable (PC_1) accounts for the maximum variability or variance in the data.
4. The second new variable (PC_2) that is formed is such that its variance is the maximum amount of the remaining variance, which is orthogonal to the first principal component.
5. The p^{th}new variable is such that its variance is the maximum amount of the remaining variance that is orthogonal to p-1 variables.
6. The p new variables are uncorrelated.

Concept of Principal Component Analysis

Principal component analysis entails a mathematical procedure that leads to the transformation of $\boldsymbol{p}$-correlated variables into a set of $\boldsymbol{p}$-new orthogonal or uncorrelated variables. Each principal component is a weighted linear combination of the original variables.

Mathematically, Consider the original random vector $\boldsymbol{X'} = (\boldsymbol{X_1, X_2, X_3}, \ldots, \boldsymbol{X_p})'$ with the variance-covariance matrix Σ with eigenvalues $\boldsymbol{\lambda_1, \lambda_2, \lambda_3, \ldots \lambda_p} \geq \boldsymbol{0}$. If we let the variables $\boldsymbol{Y_1, Y_2, Y_3}, \ldots, \boldsymbol{Y_p}$ represent the linear combination of the original variables $\boldsymbol{X_1, X_2, X_3, \cdots, X_p}$

Then,

$$\boldsymbol{Y_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + \cdots + a_{1p}X_p}$$

$$Y_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + \cdots + a_{2p}X_p$$

$$\vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad (2.4)$$

$$Y_p = a_{p1}X_1 + a_{p2}X_2 + a_{p3}X_3 + \cdots + a_{pp}X_p$$

where a_{ij} is the weight of the j^{th} (j = 1,2,⋯p) variable for the i^{th}(i =1,2,⋯,p) principal component. In matrix notation, this can be written as $Y_i = a'X$.

The weights are estimated such that,

1) The first principal component accounts for the maximum variability of the p −variables of any linear combination of the data set; the second principal component that is formed is such that its variance is the maximum amount of the remaining variance that is orthogonal to the first principal component. It follows that each succeeding component accounts for as much variance that has not been accounted for by the preceding components.

   Generally, the p^{th} principal component accounts for the maximum variability that the first p-1 components do not accounted for.

2) The sum of squares of the weights is equal to one.

   Thus,

   $$a_{i1}^2 + a_{i2}^2 + \cdots + a_{ip}^2 = 1 \quad (i = 1, 2, \cdots, p) \text{ and} \qquad (2.5)$$

3) The sum of the products of the weights of the j^{th} variable and i^{th} principal component is zero:

   $$a_{i1}a_{j1} + a_{i2}a_{j2} + \cdots + a_{ip}a_{jp} = 0 \quad (\text{for all } i \neq j) \qquad (2.6)$$

These three conditions bring about maximization problem requiring an eigenanalysis of the variance-covariance structure. This is attainable by examining the eigenstructure of the covariance-matrix.

Analysis of the Eigenstructure of the Covariance-Matrix.

Let X be a *p*-component random vector where *p* is the number of variables. The covariance matrix, Σ, is given by $\boldsymbol{E}$(XX').

Let $\boldsymbol{\gamma}$= ($\boldsymbol{\gamma_1}$,$\boldsymbol{\gamma_2}$,⋯,$\boldsymbol{\gamma_p}$) be a vector of weights to form the linear combination of the original variables, and $\boldsymbol{y}$ =$\gamma'\boldsymbol{X}$ be the new variable, which is a linear combination of the original

The variance of the new variable is given by the $\boldsymbol{E}(\boldsymbol{\gamma}\,\boldsymbol{\gamma}')$ and is equal to $\boldsymbol{E}(\gamma' XX'\boldsymbol{\gamma})$ or $\boldsymbol{\gamma}'\Sigma\gamma$. The problem now reduces to determining the weight vector, γ, such that the variance, $\boldsymbol{\gamma}'\Sigma\gamma$ of the new variable is maximum over the class of linear combinations that can be formed subject to $\gamma'\gamma$=1

To maximize, the problem solution is obtained as follows:

Let $Z=\boldsymbol{\gamma}'\Sigma\boldsymbol{\gamma} - \lambda(\gamma'\boldsymbol{\gamma} - 1)$ (2.7)

where λ is the Lagrange's multiplier.

The p-component vector of the partial derivative is given by

$$\frac{\partial Z}{\partial \gamma} = 2\Sigma\boldsymbol{\gamma} - 2\lambda\,\boldsymbol{\gamma} \qquad (2.8)$$

setting the above vector of partial derivative to zero results in the final solution

$$(\Sigma - \lambda\,\mathrm{I})\,\gamma = 0 \qquad (2.9)$$

In order to ensure that the system of homogeneous equations have a nontrivial solution, the determinant of $(\Sigma - \lambda \mathrm{I})\ \gamma$ should be zero. Thus,

$$\left| \Sigma - \lambda \mathrm{I} \right| = 0 \tag{2.10}$$

Equation 2.7 is a polynomial in λ of order $\boldsymbol{p}$, and therefore has *p*- roots. Let $\lambda_{\mathbf{1}} \geq \lambda_{\mathbf{2}} \geq \ldots \geq \lambda_{\boldsymbol{p}}$ be the $\boldsymbol{p}$- roots. That is Equation. 2.7 results in $\boldsymbol{p}$ values for λ, and each value are called the eigenvalue or the root of the Σ matrix. Each value of λ is a result of set of weights or loadings given by the $\boldsymbol{p}$-component vector $\boldsymbol{\gamma}$ by solving the follow equations:

$$(\Sigma - \lambda_{\mathbf{1}} \mathrm{I})\gamma_{\mathbf{1}} = 0 \tag{2.11}$$

subject to

$$\boldsymbol{\gamma}'\boldsymbol{\gamma} = 1 \tag{2.12}$$

hence, the first eigenvector, $\boldsymbol{\gamma_1}$ corresponding to the first eigenvalue, $\boldsymbol{\gamma_1}$, is obtained by solving equations

$$(\Sigma - \lambda_{\mathbf{1}} \mathrm{I})\ \boldsymbol{\gamma_1} = 0 \tag{2.13}$$

subject to

$$\boldsymbol{\gamma_1'}\boldsymbol{\gamma_1} = 1 \tag{2.14}$$

Pre-multiplying Equation. 2.11 by $\boldsymbol{\gamma_1}$ gives

$$\boldsymbol{\gamma_1'}\ (\Sigma - \lambda_{\mathbf{1}} \mathrm{I})\ \gamma_{\mathbf{1}} = 0$$

$$\boldsymbol{\gamma_1'}\Sigma\ \gamma_{\mathbf{1}} = \lambda_{\mathbf{1}}\ \boldsymbol{\gamma_1'}\boldsymbol{\gamma_1}$$

$$\boldsymbol{\gamma_1'}\Sigma\boldsymbol{\gamma_1} = \lambda_{\mathbf{1}}, \quad \text{since } \boldsymbol{\gamma_1'}\boldsymbol{\gamma_1} = 1 \tag{2.15}$$

Hence, the left hand side of Equation 2.15 is the variance of the new variable, $\boldsymbol{y_1}$ and is equal to the eigenvalue $\lambda_{\mathbf{1}}$. The first principal component is, therefore, given by the eigenvector, $\gamma_{\mathbf{1}}$, corresponding to the largest eigenvalue, $\lambda_{\mathbf{1}}$.

Let $\gamma_{\mathbf{2}}$ be the second $\boldsymbol{p}$-component vector of weights to form another linear combination. The next linear combination can be found such that the variability of $\gamma'_{\mathbf{2}}X$ is the maximum variance subject to $\gamma'_{\mathbf{1}}\ \gamma_{\mathbf{2}} = 0$ and $\gamma'_{\mathbf{2}}\ \gamma_{\mathbf{2}} = 1$. It can be illustrated that $\gamma_{\mathbf{2}}$ is the eigenvector of λ_2, the second largest eigenvalue of Σ. Similarly, it can be shown that the remaining principal components, $\gamma'_{\mathbf{3}}, \gamma'_{\mathbf{4}}, \ldots,$ $\gamma'_{\boldsymbol{p}}$, are the eigenvectors corresponding to eigenvalues, $\lambda_{\mathbf{3}}, \boldsymbol{\lambda_4}, \ldots, \lambda_{\boldsymbol{p}}$, of the matrix Σ. Thus, the problem of finding the weights reduces to finding the eigenstructure of the covariance matrix. The eigenvectors give the vectors of weights and the eigenvalues represent the variances of the new variables or the principal component scores.

Conditions under which Principal Component Analysis is Applicable

Correlation

With PCA, there is the need for critical examination of correlations, but, not necessarily the means for a set of variables. The acceptable correlation between two or more variables could be greater than or equal to 0.30 within variables of the same dimension. Correlation among variables from different dimensions should be close to zero, if dimensions are expected to be orthogonal

(uncorrelated) though some nonzero correlations are acceptable, particularly with dimensions that are expected to be oblique (correlated).

Although, the issue of collinearity is not as much a problem as with other methods, it still needs to be investigated. Variables within the same dimension are often seen as similar ways of expressing the correlation in the same dimension and thus can exhibit substantial correlation (for example, between (0.30-0.90). Nevertheless, if correlation exceeds 0.90, there could be problems associated with collinearity (i.e. instability of the weights or loadings). Should collinearity be suspected, then there is the need to consider collapsing the two variables involved into an average, summed composite, or even dropping one of the variables.

As regards most multivariate methods, it is vital to have access to a large data matrix with continuous variables. Unlike many methods, there will be less emphasis on meeting statistical assumptions, particularly when making descriptive summaries of the data. Certainly, interferences beyond a specific sample would be strengthened when meeting linear model assumptions (that is normality and linearity) in large and relevant sample, (Harlow, 2005). A number of assumptions such as normality, independence and linearity limit the applicability of PCA. Hence, the data set must meet these assumptions for PCA to be strictly applicable (Gorsuch, 1983).

Normality

Principal component analysis assumes that the underlying structure of the data is multivariate normal. Geometrically, a multivariate normal distribution

exists when the data cloud is hyper ellipsoidal with normality varying density around the centroid. Such a distribution exists when each of the original variables has a normal distribution about fixed value on all others. For this reason, we test for normality of each original variable to detect whether the data set is multivariate normal. Though multivariate normality implies univariate normality; the reverse is not always factual. Therefore, to detect multivariate normality appropriately, there is the need to test for the normality of each principal component. Besides, normally distributed principal component scores do not guarantee a multivariate normal distribution.

Independence

The independent random sample and the effect of outliers is also a vital condition that needs to be met. In PCA, it is assumed that the random observation vectors have been drawn independently from a *p*-dimensional multivariate normal population. To ensure independence, consideration needs to be given to it in the design of the study because there is no perfect means of determining whether a data set is independent. We can achieve this by constructing a univariate stem and leave, box and normal probability plots for each variable and checking for suspected outliers. Outliers in most cases exert unexpected pull on the direction of the component axes and therefore affect the efficacy of the ordination. However, it is extremely necessary to distinguish between extreme observations and outliers. Subsequently, we must take caution in discarding suspected points, so that there is no substantial loss of information.

Linearity

Lastly, PCA assumes that variables change linearly along underlying gradients and that linear relationship exists among variables such that the variables can be combined in a linear fashion to create principal components, (Johnson, 1981and Gorsuch, 1983).

Conveniently, statistical packages such as SPSS have provided for tests that are used to check whether a data set meets the above assumptions. The Kaiser-Mayer-Oklin (KMO), test and Bartlett's test of sphericity are both test of multivariate normality and sampling adequacy.

KMO tests is a measure of whether or not the distribution of values is adequate for conducting principal component analysis. It indicates levels of values with their respective interpretations or recommendations as follows; a measure greater than or equal to 0.90 is marvelous, a measure of 0.80+ is meritorious, a measure of 0.70+ is middling, a measure of 0.60+ is mediocre, a value less than or equal to 0.50 is miserable. The Bartlett test on the other hand measures the multivariate normality of the set of distributions. Besides, it tests for the linearity in the data set by checking whether the correlation matrix is an identity matrix. A value significantly less than 0.50 is an indication that the data set does not produce an identity matrix and is thus approximately multivariate normal, and is acceptable for PCA.

Principal Component as an Index

PCA is extensively used as dimensional reduction technique because it enables us to represent a *p*-dimensional data set in a lower dimensional space, where $\boldsymbol{m < p}$. It follows that the first few components may still be sufficient to represent most of the information in the original data.

Most principal components are interpretable, especially the first few ones. The first principal component, for instance, is in most cases the weighted sum of the original variables. Thus, given that this component accounts for a reasonably large proportion of the variability in the data, then it can be used as an index. To use the principal component as an index requires the determination of principal components scores and the factor loadings.

Analytically, the first principal component score for the first set of observations is the value obtained by substituting into the equation $Y_1 = \boldsymbol{a_{11}x_1} + \boldsymbol{a_{12}x_2} + \ldots + \boldsymbol{a_{1p}x_p}$, the estimated weights and values of the first observations on the *p* original variables.

Similarly, the first principal component scores for the second set of observations is the value obtained by substituting into the equation

$$Y_2 = \boldsymbol{a_{21}x_1} + \boldsymbol{a_{22}x_2} + \ldots + \boldsymbol{a_{2p}x_p},$$

the estimated weights and values of the first observation on the original variables and so on. The remaining principal components scores are similarly obtained.

Factor loadings measure the simple correlation between the original and the new variables. They give an indication of the extent to which the original variables are influential or important in forming new variables. The higher the

loading the more influential the variable is in forming the principal component score (or in this case the index) and vice versa. The loadings are given by

$$\boldsymbol{l_{ij}} = \frac{\boldsymbol{a_{ij}}}{\check{S}_{\boldsymbol{j}}}\sqrt{\lambda_{\boldsymbol{i}}},$$

where $\boldsymbol{l_{ij}}$ is the loading of the $\boldsymbol{j^{th}}$ variable for the $\boldsymbol{i^{th}}$ principal component, $\boldsymbol{a_{ij}}$ is the weight of the $\boldsymbol{j^{th}}$ variable for the $\boldsymbol{i^{th}}$ principal component, $\lambda_{\boldsymbol{i}}$ is the eigen value or variance of the $\boldsymbol{i^{th}}$ principal component and $\check{S}_{\boldsymbol{j}}$ is the standard deviation of the $\boldsymbol{j^{th}}$ variable. Statistical packages such as SPSS, SAS and Minitab are available for use to perform principal component analysis, including determination of scores and loadings.

Deciding on the Number of Principal Components to use

The percentage of variance in the variables that is accounted for by the components is useful index for assessing the variability of the components. Since the dimensions might not be expected to explain all the variation and covariation among the variables, it is reasonable that the dimensions account for at least 50%. We get an indication of the proportion of variation explained by the dimension by forming a ratio of an eigenvalue over the sum of all the eigenvalues.

The number of eigenvalues greater than 1.0 is often used as an upper bound estimate on the number of underlying components in PCA. Guttmann (1954) and Kaiser (1970) advocated suitable method that helps in deciding on the correct number of dimensions, but the true number may well be less than this. The rationale was that the variance of a single, standardized variable would be 1.0. If an underlying dimension were to be worth examining, then it needs to have at

least the same amount of variance as a single variable, nevertheless, ideally it should have much more variance.

Deciding on the number of PCs to be used depends largely on: the proportion of variance accounted for. The decision of enough variation explained by few PCs is subjective. Ideally, if $\frac{\sum_{i=1}^{k}\lambda_i}{\sum_{i=1}^{p}\lambda_i} \geq 80$ then it could be satisfactory. However, this may depend on the experimental requirement.

Another method of assessing the number of dimensions (PCs) is to examine the eigenvalues whenever they appear to be dropping off to a trivial and inconsequential size. In a scree graph, a plot of the number of components are on the x-axis and the values of the eigenvalues on the y-axis. The point at or before the elbow in a scree plot provides another estimation as to the number of underlying dimensions to use. Although this technique in most cases may lead to the inclusion of too many components compared with for example Kaiser Criterion, the scree graph or plot is often practical for data exploration.

Interpretation of Principal Component

PCA dwells much on the weight attached to specific variables. The loading or structure coefficient is the most interpretable component. We rotate component loadings in order to increase the interpretability, eliminate cases of bipolar factors, remove negative loadings and make the sizes of the negative loadings negligible of the dimensions (PCs). Several rotation procedures are available but the most common ones are Varimax and Quartimax. The major objective of Varimax rotation is to have a factor structure in which each variable

loads highly on one and only one factor. This means that a given variable should have high loadings on one factor and near zero loadings on the other factors. Varimax rotation destroys or suppresses the general factor and should be used when the presence of general factor is suspected

In addition, the major objective of the Quartimax rotation technique is to obtain a pattern of loading such that: (1) all the variables have a high loading on one factor. (2) Each variable should have a high loading on one other factor and near zero loadings on the remaining factors. Even though most computers use the varimax orthogonal rotation as a default option, yet it is essential to consider an oblique rotation, if we expect the dimensions to be related. In either case, we usually strive to rotate the weights so that each dimension has several variables that load highly with the remaining variable loading close to zero. This kind of partten is referred to as "Simple Structure" (Thurstone, 1985).

The loadings range between -1 to +1 irrespective of the structure used. It reveals how correlated a variable is with an underlying dimension (component). In PCAs, the same criterion is used as with other methods that rely on loadings; variables with loadings of 0.30 or greater are interpreted as having a meaningful impact on overall dimension. However, this is subjective to what goal one seeks to achieve. In trying to describe the nature of each dimension, it is worth noting the kind of variables that highly load on the component.

As regards other methods that focus on weights, the sign appended to the loading provide information about the nature of the relationship. A positive value indicates that a variable is very similar to the underlying dimension,

whereas a negative loading connotes the higher score on the dimension on which the variables loads. There are several guidelines in evaluating the variables, thus, those with loading greater than or equal to 0.30 would be retained as marker variables for a dimension. Also, variables loading less than 0.30 on all dimensions could be ignored.

This would certainly mean that the variables do not have enough in common with other variables. In fact, variables with loadings greater than or equal to 0.30 on more than one dimension would be classified as complex variables. Since it would not be clear as to which dimension the variable might be describing, the complex variable would be discarded.

CHAPTER THREE

PRELIMINARY ANALYSIS

Introduction

This chapter and the next present the analysis of the data. This chapter dwells on the preliminary analysis, which mainly outline the descriptive statistics of the data. The second part dwels on further analysis of the data in which advanced techniques are applied to determine the overall best student and to find out if differences exist between the performance of male and female students.

Descriptive Statistics

Table 3.1 shows the Age Distribution of the Students who took part in the examination.

Table 3.1: Age frequency distribution of students

Ages (x)	Frequency(f)	Percent
17	7	14.89
18	14	29.79
19	18	38.30
20	5	10.63
21	3	4.26
22	1	3.13
Total	47	100

There are 47 students who took part in the examination, which is made up of 28 males and 19 females. This represented 59.6% for the males and 40.4% for the females.

From Table 3.1, it can be seen that the minimum and the maximum ages of the students are 17 and 22 respectively. The range of their ages is 5 indicating the age difference between the oldest and the youngest student. The mean age of the students is 19 years. The modal age of the distribution is 19. This indicates that students of 19 years of age constitute the majority in the class with the highest percentage of about 38. Those of 18 years of age were the second highest constituting about 30%. The highest age 22, recorded just about 2% of the distribution.

Table 3.2 shows the overall mean score and the mean scores for the males and the females by subjects respectively.

Table 3.2: The Mean Scores for Males and Females in the Examination

Subject	Male ($\bar{X}_1$)	Female ($\bar{X}_2$)
English Language	46.857	51.895
Core Mathematics	59.857	60.105
Integrated Science	56.750	57.000
ICT	49.893	53.579
French	45.571	42.579
Geography	51.321	40.947
Government	60.964	52.842
Economics	51.250	46.053
Mathematics	52.179	51.053
Grand Mean	$\bar{X}_m$ = 52.738	$\bar{X}_f$ = 50.673

From Table 3.2, it can be observed that the females out-performed the males in English Language with the mean score of about 52 to that of males of about 47. Interestingly, the mean score for the females in Core Mathematics, Integrated Science and Information and Communication Technology (ICT) are all slightly higher than that of the males. It can be seen that the females performed better than their male counterparts in this subjects. However, comparing the mean scores of French, Geography, Government, Economics and Elective Mathematics for the males and that of the females, it can be seen that the males did better in these subjects than the females. When the overall mean score of about 53 is compared to that of the females of about 51. We can realize that the males performed better than the females in the examination.

These comparisons did not give us enough evidence to draw a valid conclusion that there is difference between the performances of both sexes. In view of these, it is necessary to go further to find whether there exist any significant difference between the performances of both males and females. In this case, the appropriate test to be used is the Hotelling's T-squared test, since there are two groups involved in the study. This will be done in Chapter Four.

Table 3.3 shows a brief descriptive statistics of the data set of scores obtained by students. It consists of the mean score, median score, mode, minimum score, maximum score, and standard deviation.

Subjects	Mean	Median	Mode	Min.	Max.	Std.Dev.
Englang	48.894	51	52	18	74	10.797
CMaths	59.957	61	55	38	82	9.374
Intsc	56.851	57	56	17	77	10.960
Ict	51.383	50	40	25	73	11.850
French	44.362	43	38	18	71	12.737
Geog	47.128	47	40	20	78	13.074
Govt	57.681	56	67	21	88	17.105
Econs	49.149	49	50	15	83	14.430
EMaths	51.723	52	48	28	73	9.675

Table 3.3: Descriptive Statistics of scores obtained in the examination

It can be observed from Table 3.3, that the mean score for English language is around 49. The mean scores for Core Mathematics, Integrated Science, ICT, French, Geography, Government, Economics and Elective Mathematics are about 60, 57, 51, 44, 47, 58, 49 and 52 respectively. Relatively, it can be observed that students did better in Core Mathematics, which has the

highest mean score. Economics was found to have the smallest minimum score whereas Core Mathematics has the highest minmum score with the smallest standard deviation as compared to other subjects. It can also be seen from the table that the mean score for French was the smallest with quite higher standard deviation indicating that students' performance was not too good. Government has the highest modal score of 67 with the minimum and maximum scores of 21 and 88 respectively.

Comparing English Language and French, it can be observed from the table that the mean score of English is relatively higher than that of French. The median score, which divided the whole set of scores for English into two halves is also relatively higher than the median for French. Similarly, the standard deviation of the two courses indicated that the students did better in English than they did in French. The performance of students in both Core Mathematics and Elective Mathematics indicated that students mean score of the former is higher than that of the latter. This means that the students did better in the Core Mathematics than in Elective Mathematics. The median, mode and the standard deviation of the two courses suggest that the performance of students in the Core Mathematics is relatively better than in Elective Mathematics.

Furthermore, if we consider the performance of the students in the courses offered under the sciences like Integrated Science, Core Mathematics, Information and Communication Technology (ICT) and Elective Mathematics, it can be realized that the mean scores for all of them were within the range 51-60. This indicated that the students did relatively well in the sciences as compared to

the Social sciences like Economics, Geography and Government which have their means within the range 47-58. So also is the mean scores of the Languages (English and French) were within the range 44-49. This is supported by the value of their standard deviations; this is because the Sciences recorded the highest range of the mean scores as compared to the Social Sciences and the Languages.

Distribution of Scores in Percentages (%) in each Subject

In this section, we discuss the data using histogram to assess whether the data set of each subject meet the condition of normality. Since principal component analysis assumes that, the underlying structure of the data is multivariate normal.

Figure 3.1, is a histogram of English Language scores in percentages (%) obtained by students in the examination. The y-axis shows the number of students possessing the range of the scores listed on the x-axis (frequency) whereas the x-axis shows the range of the scores or the marks in percentages.

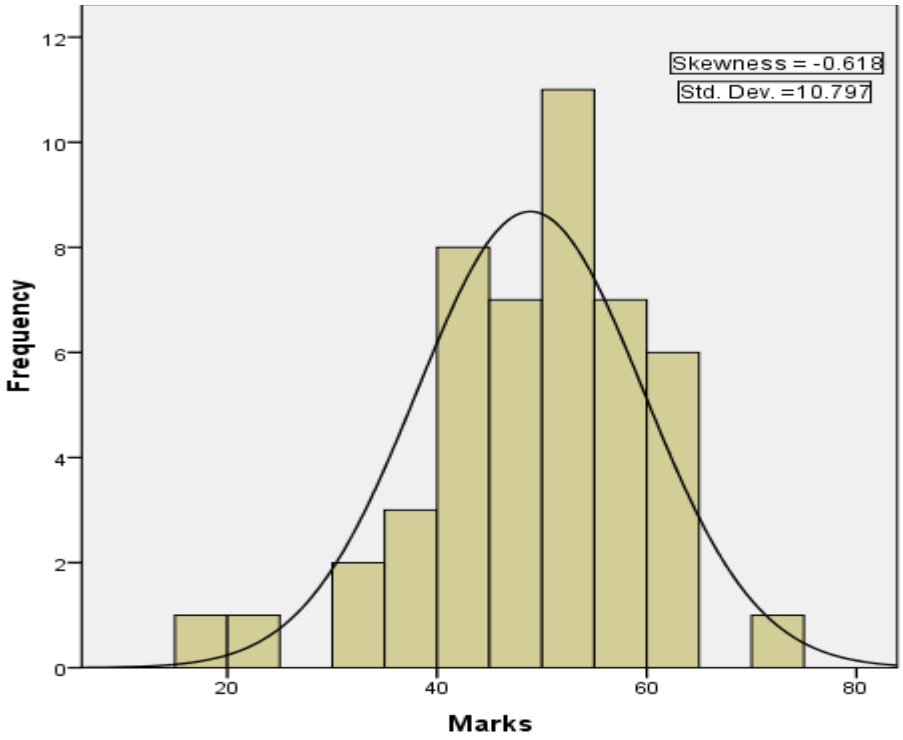

Figure3.1: Distribution of scores in English Language

From Fig. 3.1, it can be observed that there are two scores that appeared to be a bit away to the left and another one to the right. In this case the median mark can be estimated to be around 51%, indicating that relatively half of the students scored marks less than or equal to 51%. The graph also shows that the highest peak was around 50%. This indicates that more students have marks around 50% in the examination. The standard deviation of the distribution is around 11. This implies that the individual marks are relatively wide spread around the center (mean). From the graph, it can again be seen that the scores are not normally distributed. Therefore, the curve of the distribution is asymmetrical and negatively skewed. This can be supported by the coefficient of skewness (-0.618).

Figure 3.2, is a histogram of Core Mathematics scores in percentages (%) obtained by students in the examination.

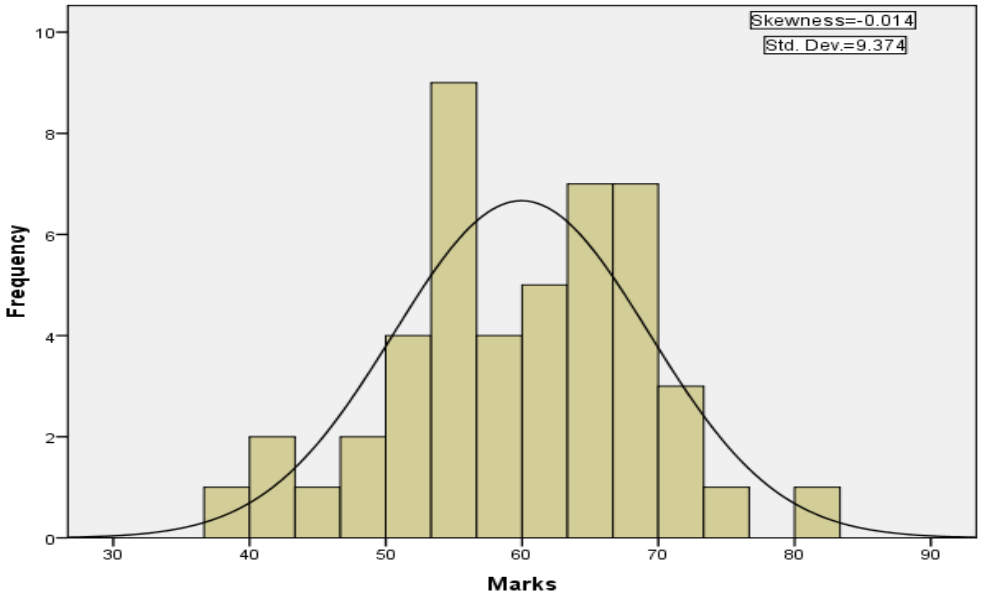

Figure 3.2: Distribution of scores in Core Mathematics

From Fig. 3.2, it can be observed that there was an observation or a score, which is quite large but not quite close to the majority of the scores in the distribution. The graph of the distribution shows that the highest peak is around 55%. This represents the modal mark for the distribution indicating that more students have scores around 55%. It can be seen from the graph that the median mark of the distribution is around 60%. This means that about half of the students have marks less than or equal to 60%. The distribution has a standard deviation of around 9. This indicates that the individual marks are relatively closely spread around the mean score of the data set. It can be deduced from the graph, that the marks are approximately normally distributed. However, the

coefficient of skewness (-0.014) indicated on the graph is quite close to zero. Therefore, the curve of the distribution is approximately normal.

Figure 3.3, is a histogram showing the distribution of the marks obtained in Integrated Science.

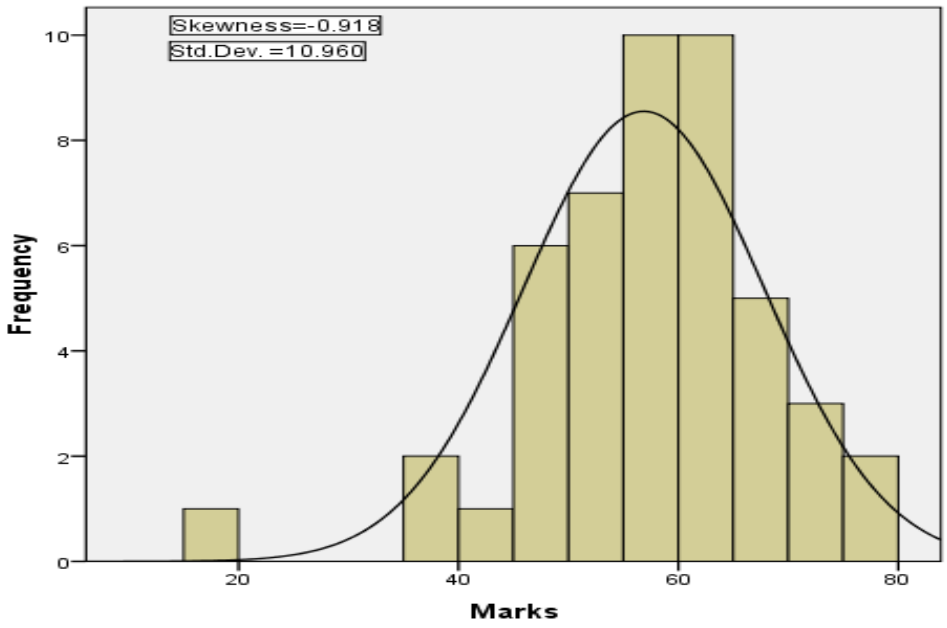

Figure 3.3: Distribution of scores in Integrated Science.

From Figure 3.3, it can be seen that there was an observation or a score which is at a distance from the rest of the scores. This extreme observation or score can be described as an outlier. The distribution of the scores shows that more students have scores between 55% and 60%. The standard deviation indicates that the individual marks are relatively wide spread around the mean of the data set. The curve of the distribution indicates that the scores are not normally distributed. Therefore, the distribution is negatively skewed. This can be supported by the value of coefficient of skewness (-0.918).

Figure 3.4, is a histogram depicting the distribution of the marks obtained by students in ICT.

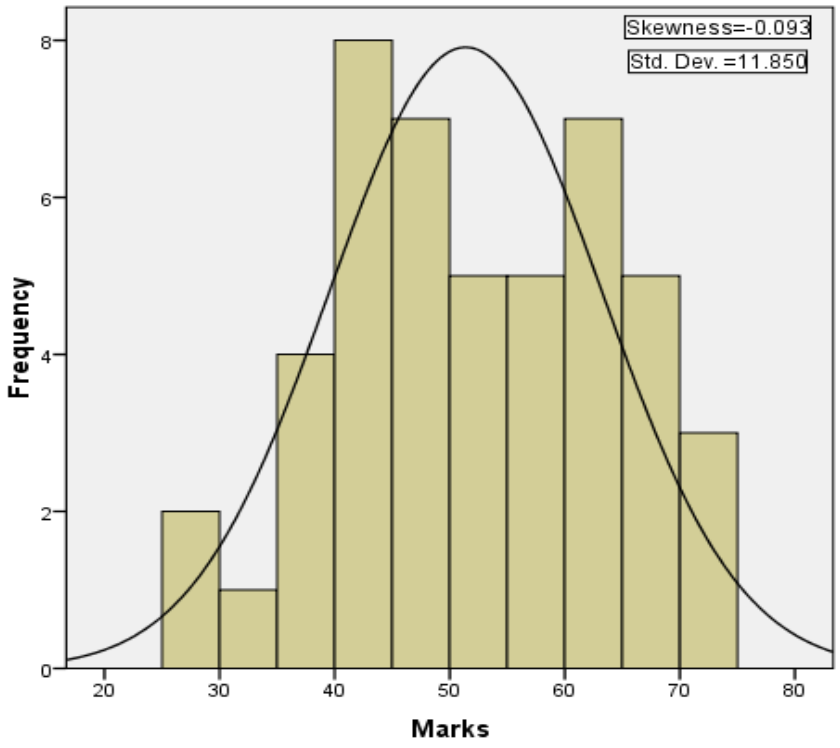

Figure 3.4: Distribution of scores in ICT

From Figure 3.4, it can be seen from the graph that the distribution shows the highest peak to be around 40%. The standard deviation of the distribution is about 13. This means that the individual scores are relatively wide spread around the mean of the data set. The curve of the distribution, indicates that the scores are not normally distributed. Therefore, the distribution is negatively skewed. This can be supported by the value of coefficient of skewness (-0.093).

Figure 3.5, is a histogram that depicts the distribution of the marks obtained by students in French.

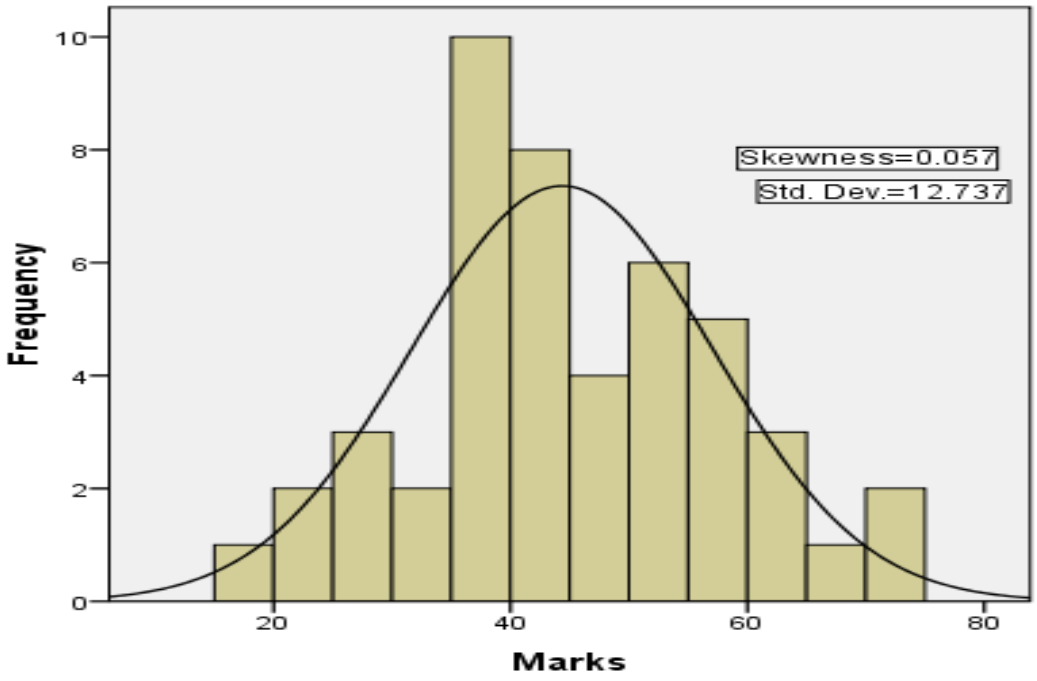

Figure 3.5: Distribution of scores in French

From Fig.3.5, it can be observed that the graph shows the highest peak of the distribution to be around 35%. This indicates that more students have marks around 35% representing the modal score in the examination. The standard deviation of the distribution of the scores is around 13. This shows that the scores are relatively wide spread around the center of the data set. A critical look at the curve of the distribution revealed that the scores are not normally distributed. Hence, the distribution is positively skewed. This can be supported by the value of coefficient of skewness (0.057).

Figure 3.6, is a histogram showing the distribution of marks obtained by students in

Geography.

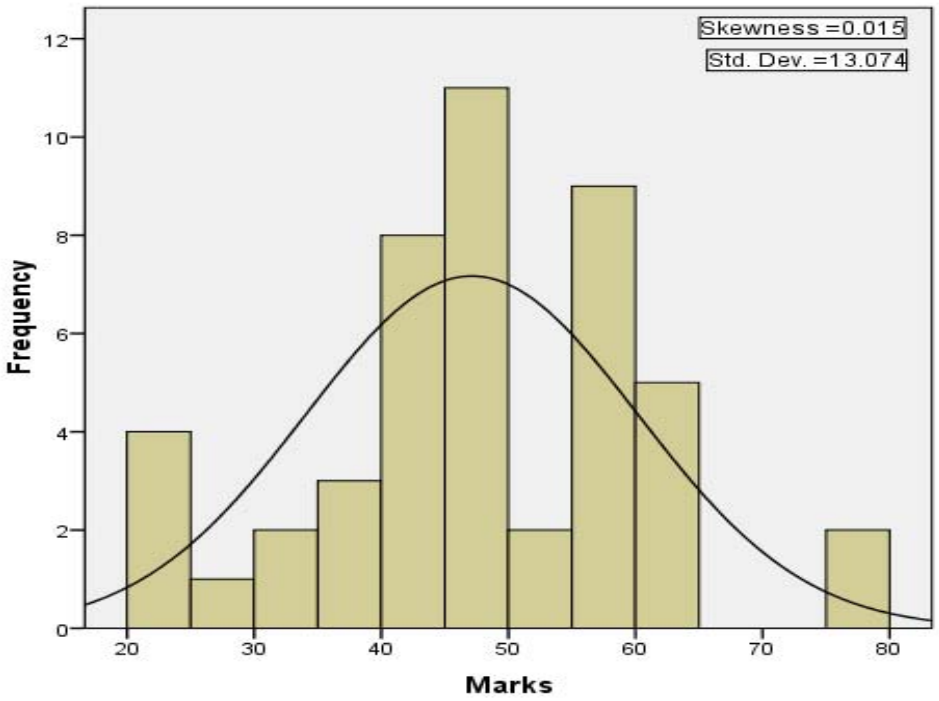

Figure 3.6: Distribution of scores in Geography

From Fig.3.6, it can be observed that one of the observations in the distribution is quite at a distance to the right side from the others. It appears to be the largest score, and unusual. The graph shows that the highest peak of the distribution is around 45%. This indicates that a lot more of the students have marks around 45%, which represents the modal mark. The distribution has a standard deviation to be around 13. This indicates that the scores were quite scattered. It can be realized from the graph that about 50% of the students have marks less than or equal to 47%. The curve of the distribution indicates that the

scores are not normally distributed. Therefore, the distribution is positively skewed. This can be supported by the coefficient of skewness (0.015).

Figure 3.7, is a histogram showing the distribution of marks obtained by students in Government.

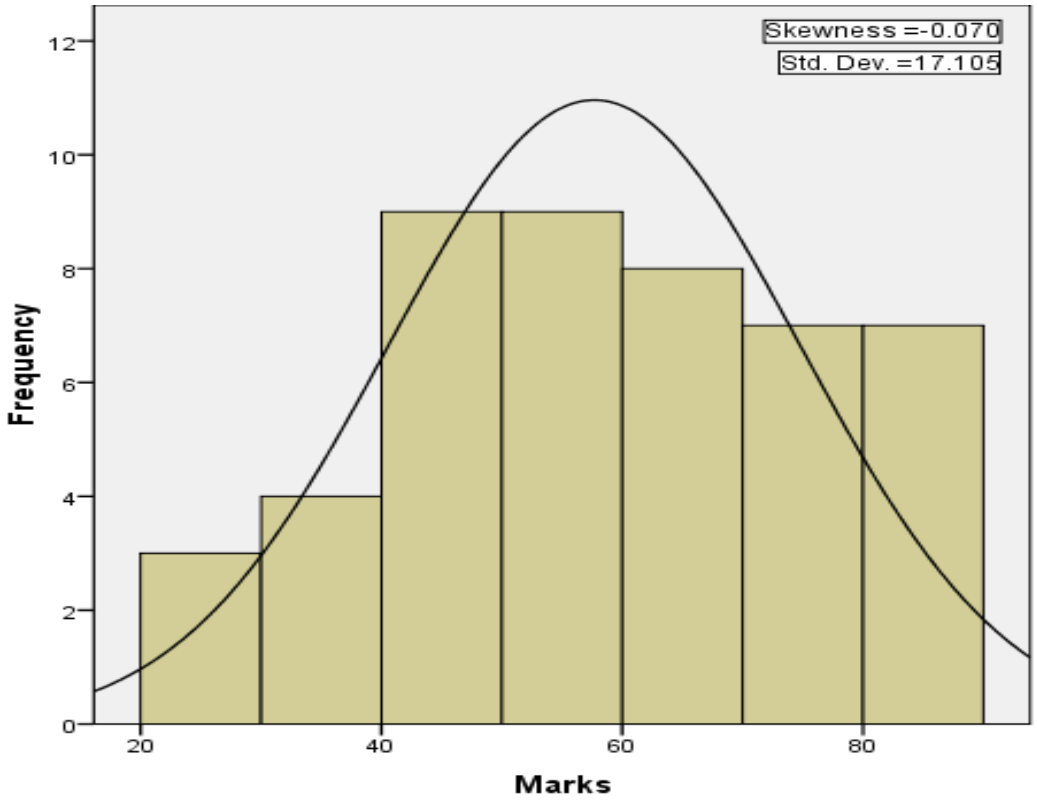

Figure 3.7: Distribution of scores in Government.

From Fig.3.7, it can be observed that equal number of students have marks between 40% and 60%. The standard deviation of the distribution is relatively high around 17. This indicates that the scores were quite scattered. From the graph, about 50% of the students have marks very close to 60%. The

curve of the distribution shows that the scores are not normally distributed. It is negatively skewed.

Figure 3.8, is a histogram, which shows the distribution of marks obtained by the students in Economics.

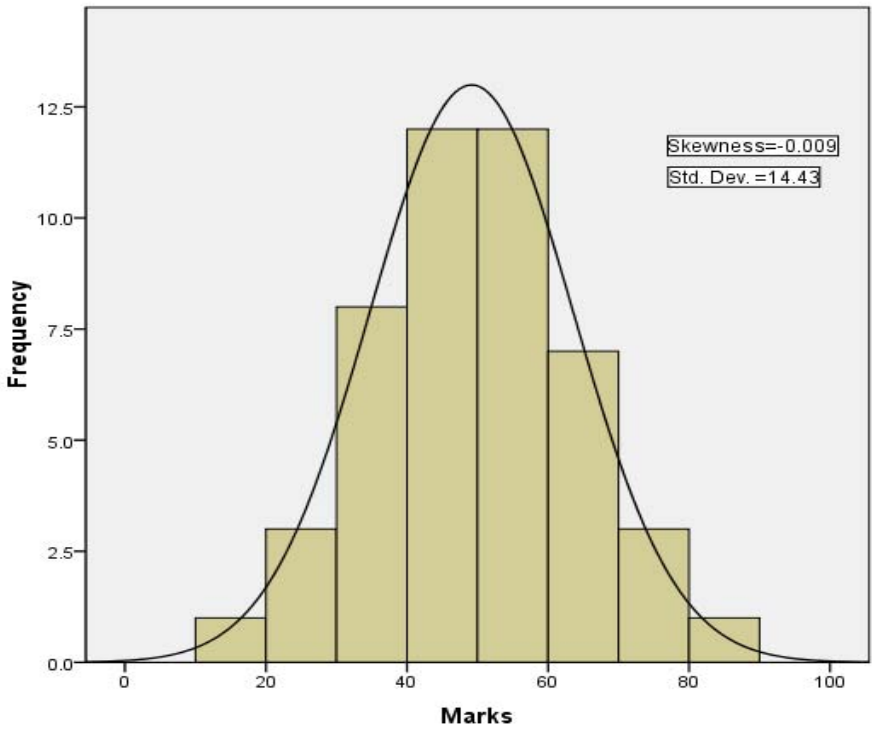

Figure 3.8: Distribution of scores in btained Economics.

From Fig. 3.8, it can be observed that there is equal number of students that have scores between 40% and 60%. It can be realized that the median mark of the distribution is around 60%. This implies that about 50% of the students have scores less than or equal to 60%. Further, the standard deviation of the distribution is relatively high around 14. This indicates that the scores were relatively quite scattered. The curve of the distribution is approximately normal.

This can be supported by the value of the coefficient of skewness, which is close to zero (-0.009).

Figure 3.9, is a histogram showing the distribution of marks obtained by the students in Elective Mathematics.

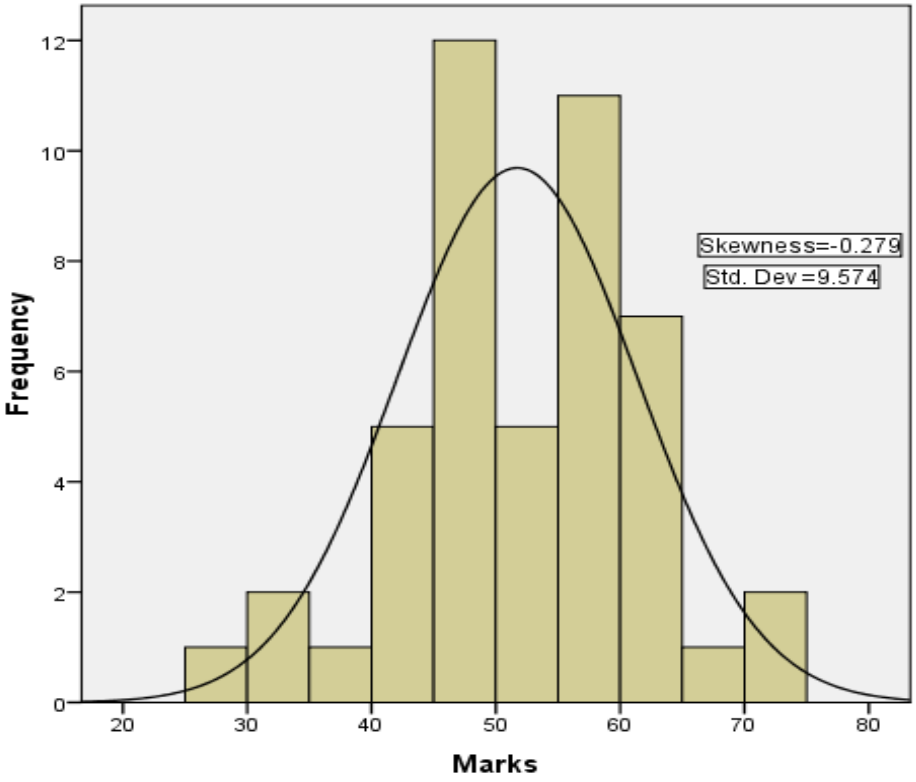

Figure 3.9: Distribution of scores in Elective Mathematics.

From Fig. 3.9, it can be seen that the highest peak of the distribution is around 48 %. This implies that more of the students have marks around 48%. The standard deviation of the distribution is around 10. This indicates that the individual scores are relatively closely scattered. Further, the curve of thedistribution indicates that the scores were not normally distributed. It is

negatively skewed. The value of the coefficient of skewness (-0.279) supports this.

Correlation Analysis

Table 3.4 is a correlation matrix showing the degree of association between a pair of courses. The courses involved are English Language, Core Mathematics, Integrated Science,

Information and Communication Technology (ICT), French, Geography, Government, Economics and Elective Mathematics. Marks obtained by students in the various courses are used in running the correlation coefficients.

Table 3.4: Correlation Matrix of marks obtained by Students

Var.	Englis	CMat	IntSc.	ICT	Frenc	Geo	Gov	Eco	EMat
Englis	1								
lCMat	0.44	1							
tIntSc.	0.63	0.38	1						
ICT	0.52	0.31	0.39	1					
dFrenc	0.53	0.17	0.63	0.45	1				
aGeog.	0.33	0.44	0.52	0.33	0.53	1			
nGovt.	0.52	0.40	0.55	0.50	0.53	0.61	1		
Econs.	0.59	0.46	0.64	0.50	0.62	0.27	0.73	1	
hEMat	0.42	0.89	0.47	0.42	0.35	0.50	0.48	0.53	1

It can be observed from Table 3.4, that there is a strong relationship between Core Mathematics and Elective Mathematics since the two subjects have a high correlation coefficient of 0.888. This implies that a very good performance in one can relatively reflect very good performance in the other. The correlation coefficient for Economics and Government is 0.734, which is relatively high. It indicates that there is a strong relationship between them. The coefficient for French and Elective Mathematics is 0.174, which is the lowest. It indicates that there is weak relationship between the two. It can be observed from the table that the association between the science courses like Integrated Science, Elective Mathematics, ICT and Core Mathematics is relatively weak since majority of their correlation coefficients are below 0.5. However, there is a strong association between the social science courses like Economics, Geography and Government. Further, it can be estimated that about 60% of the correlation coefficients are around 0.5, which is quite significant and acceptable. This indicates that the necessary condition for the principal component analysis, a linear dependence, appears to have been adequately met.

Table 3.5 indicates the KMO and Bartlett's Test for multivariate normality and sampling adequacy.

Table 3.5: KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		0.801183005
Bartlett's Test of Sphericity	Approx. Chi-Square	259.9485964
	Df	36
	Sig.	9.94716E-36

From Table 3.5 we can observe that the KMO measures about 0.80, which indicates that the distribution of the values is adequate for conducting the PCA and can be labeled meritorious. This also indicates that the set of distributions is approximately multivariate normal and linear since the data set does not produce an identity matrix.

Eigen Analysis

Figure 3.10 is a scree plot indicating the number of principal components that can be used for the analysis. The vertical axis is represented by the eigenvalue and horizontal axis by the component number.

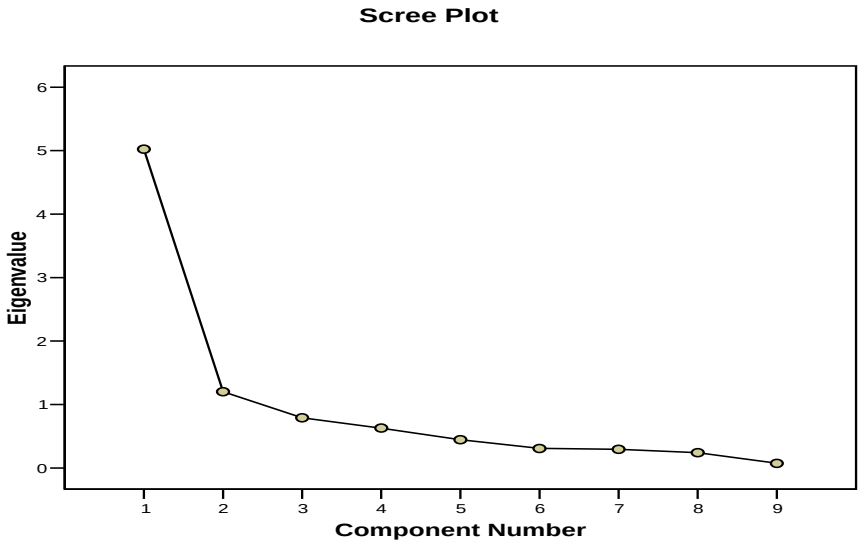

Figure 3.10: Plot of Principal Components against Eigenvalues of the scores.

It can be observed from Fig.3.10, that the elbow of the scree plot has indicated that two principal components are feasible or could be used in the

analysis. These two PCs are enough to explain the variation in the data. This can be supported by the percentage of the cumulative total variability of about 69%, which is explained by the two PCs. This is shown in Table 4.1 on next page.

Table 3.6 shows the initial eigenvalues for the nine components corresponding to the number of subjects that are used in the study. It consists of the component number, total amount of variation explained by each of the components, the percentage of the variation explained by each component, and cumulative percentage of variation.

Table 3.6: Total variance explained by the principal components

	Initial Eigenvalues		
Component	Eigenvalue	% of Variance	Cumulative %
1	5.022	55.796	55.796
2	1.202	13.351	69.147
3	0.790	8.774	77.921
4	0.628	6.976	84.897
5	0.445	4.946	89.843
6	0.307	3.416	93.259
7	0.293	3.260	96.519
8	0.241	2.673	99.192
9	0.073	0.808	100.000

From Table 3.6, we can observe that first and the second principal components have eigenvalues respectively exceeding 1. The percentage of variance accounted for by the first component is 55.796, which is quite high. Thus, PC1 alone explains more than half of the total variation in the data. Further, the second component explained 13.351% of the total variation in the data. On this score, the two components explained quite a significant percentage of about 69 of the total variation. This revelation supported that of the scree plot in Figure 3.10**.**

CHAPTER FOUR

FURTHER ANALYSIS

Introduction

The previous chapter dealt with the exploratory analysis of the data. It was noted that the scores obtained by the students in the various courses vary from student to student. Similarly, it was found that there is variability in the performance of both males and females. This was revealed in the analysis of their mean scores. Nevertheless, these observations do not provide us with enough evidence to determine the differences in the performance between the male and female students in the class as well as the overall best student for an award. For these issues to be comprehensively addressed, we conducted T-squared test to find whether differences exist in the performance between both males and females. Subsequently, the data were subjected to further analysis using principal components in effects to determine the best index for identifying the best student.

As mentioned in Chapter Two, Hotelling's T-squared test statistics is used to find whether there is significant difference between vector means of two populations. The principal component analysis technique is used to determine an appropriate index that can be used to rank the students performance. The statistical software used in dealing with the Hotelling's T-squared is the MINITAB and that of the principal component was the SPSS.

Analysis of Data

The data is subjected to further analysis in order to draw the necessary inferences.

Test for Differences in Performance of Male and Female Students

We realized that the variance-covariance of the two samples are not equal, hence the Hotelling's T-squared test appears to be appropriate for unequal variances.

Testing whether or not the mean marks of the males are the same as that of the females is equivalent to conducting a Hotelling's test, therefore the hypotheses of interest are:

$\mathbf{H_o}$: There is no difference between the marks of males and females.

$\mathbf{H_1}$: There is difference between the marks of males and females.

We obtained the vector means for males and females as

$$\bar{X}_1 = \begin{pmatrix} 46.857 \\ 59.857 \\ 56.750 \\ 49.893 \\ 45.571 \\ 51.321 \\ 60.964 \\ 51.250 \\ 52.179 \end{pmatrix} \qquad \bar{X}_2 = \begin{pmatrix} 51.895 \\ 60.105 \\ 57.000 \\ 53.579 \\ 42.579 \\ 40.947 \\ 52.842 \\ 46.053 \\ 51.053 \end{pmatrix}$$

where $\bar{X}_1$ is the mean score for males and $\bar{X}_2$ is the mean

and the variance-covariance matrix for males and is

$$S_1 = \begin{pmatrix} 119.1 & 45.5 & 76.3 & 67.9 & 75.0 & 47.1 & 97.1 & 93.6 & 44.7 \\ 45.5 & 88.0 & 39.1 & 34.1 & 23.0 & 53.2 & 64.2 & 61.9 & 80.8 \\ 76.3 & 40.0 & 122.8 & 51.3 & 90.0 & 75.4 & 105.0 & 104.0 & 51.1 \\ 67.9 & 34.1 & 51.3 & 143.3 & 90.0 & 52.0 & 103.0 & 86.3 & 48.4 \\ 75.2 & 23.0 & 89.9 & 89.9 & 164.0 & 91.1 & 119.0 & 118.0 & 46.0 \\ 47.1 & 53.2 & 75.4 & 52.0 & 91.0 & 174.0 & 140.0 & 125.0 & 63.3 \\ 97.1 & 64.2 & 104.7 & 102.5 & 119.0 & 140.0 & 298.0 & 184.0 & 80.5 \\ 93.6 & 61.9 & 103.6 & 86.3 & 118.0 & 125.0 & 184.0 & 212.0 & 74.5 \\ 44.7 & 80.8 & 51.1 & 48.4 & 46.0 & 63.3 & 80.5 & 74.5 & 94.3 \end{pmatrix}$$

and then variance-covariance matrix for females is

$$S_2 = \begin{pmatrix} 84.3 & 26.5 & 44.5 & 55.5 & 76.3 & 59.0 & 60.6 & 71.2 & 37.2 \\ 26.5 & 79.0 & 33.9 & 46.1 & 21.9 & 54.8 & 51.5 & 60.7 & 75.4 \\ 44.5 & 33.9 & 88.2 & 54.3 & 73.7 & 54.8 & 82.1 & 68.8 & 46.1 \\ 55.5 & 46.1 & 54.3 & 194.0 & 84.9 & 79.1 & 135.0 & 134.9 & 76.3 \\ 76.3 & 21.9 & 73.7 & 84.9 & 171.2 & 82.0 & 119.0 & 115.1 & 53.1 \\ 59.0 & 54.8 & 54.8 & 79.1 & 82.0 & 116.9 & 109.0 & 84.8 & 57.8 \\ 60.6 & 51.5 & 82.1 & 135.0 & 18.8 & 09.3 & 41.0 & 111.4 & 81.8 \\ 71.2 & 60.7 & 68.8 & 135.0 & 115.1 & 84.8 & 111.0 & 164.1 & 8.3 \\ 37.2 & 75.4 & 46.1 & 76.3 & 53.1 & 57.8 & 81.8 & 83.3 & 92.9 \end{pmatrix}$$

With reference to Equation 2.2, we multiplied the vector means by the inverse of the sum product of the variance-covariances matrix and the sample sizes to obtain the $\boldsymbol{T^2}$.

The test statistics $\boldsymbol{T^2}$ yielded 21.4915 from the Minitab output, while the table value $\left[\boldsymbol{\chi_9^2}\ (0.05)\right]$ is 16.92

Since the $\boldsymbol{T^2}$ (21.4915) is more than $\boldsymbol{\chi_9^2}$ (0.05) (16.92), there is enough evidence against the null hypothesis at 5% significant level. We therefore reject the null hypothesis and conclude that, there is a significant difference between the mean scores obtained by males and females in the examination. This has confirmed, the result in Table 3.2 which indicated that the approximated overall mean score (53) for males is more than the overall mean score (51) for females. This suggested that males out-performed females.

Principal Component

The correlation matrix in Table 3.4 revealed the correlation coefficients between pairs of subjects. It can be observed from the table that a good number of coefficients are around the accepted value of 0.5. This suggests the presence of linear dependence, a necessary condition for using principal component analysis. From Appendix A1, we can see that the KMO value is approximately 0.8 indicating that the data is ideal for PCA. In addition, Bartlett's Test of sphericity approached statistical significance. Hence, it can be concluded that the data can be

analyzed using the principal components. On this score, two principal components were used in the analysis. This can be supported by the revelation from the scree plot of Figure 3.10 and the total variation explained by the first two principal components as revealed in Table 3.4 all in Chapter Three.

Interpretation of the Principal Components

The loadings of a variable on a principal component tell us how influential the variable is in the formation of the component.

Table 4.1, is a table that shows the value for the eigenvectors of the scores obtained in the examination. The variables are the subjects taken in the exam. There are two components having their individual eigenvectors, which correspond to each of the variables (subjects).

Table 4.1: Unrotated principal components matrix of eigenvectors

Variable	Component	
	1	3
English Language	0.742	-0.142
Core mathematics	0.659	0.720
Integrated Science	0.783	-0.205
I C T	0.646	-0.138
French	0.721	-0.458
Geography	0.735	-0.009
Government	0.801	-0.135
Economics	0.863	-0.139
Elect. Mathematics	0.747	0.596

From Table 4.1, it can be observed that all the variables loaded substantially quite high above 0.5 on the first principal component. Economics has the highest loading of about 0.86 followed by Government and Integrated Science with about 0.80 and 0.78 respectively. This indicates that Economics, Government and Integrated Science are influential in forming the first component. It is evident that all the subjects that do not require mathematical ability loaded quite significantly It is quite interesting to observe that Core Mathematics and Elective Mathematics have high loadings of 0.66 and 0.75 respectively on the first principal component with corresponding relatively high loadings on the second component.

This indicates that PC2 is mathematically skewed since only the Core and the Elective Mathematics loaded significantly on it. This has not indicated clear distinction as to which PC to described, since both are influential in forming the second principal component as well. It is therefore necessary to rotate the components to ensure that each variable loads high on one component and loads close to zero on the other component. The rest of the variables have relatively low and negative loadings on the PC2. French has negative loading of about -0.5 whereas Geography has a loading, which is near zero (-0.009).

Although, all the variables loaded substantially high on the first principal component, the second principal component was found to have about 75% of the variables having negative loadings. To this end, the component loadings were rotated in order to remove these negative loadings and make the sizes of the negative loadings negligible on the second principal component. Further, to

increase the interpretability of the principal components, we have also used the unrotated principal components and the rotated principal components to check if there is consistency in the indices with regard to ranking of the students by order of performance. Thereafter, the best index can be determined. Varimax and quartimax rotations were used. The varimax rotation (Table 4.2) was used specifically to remove the negative loadings and make their sizes negligible on PC2 to increase its interpretability, whereas the quartimax rotation (Table 4.3) was used in order to obtain a pattern such that all the variables have a high loadings on one factor or component.

Table 4.2, shows the varimax rotated component matrix of the variables (subjects) loadings. There are two components having eigenvectors corresponding to the individual subjects.

Table 4.2: Varimax Rotated principal component matrix of eigenvectors

	Component	
Variable	1	3
English Language	0.706	0.270
Core Mathematics	0.182	0.959
Int. Science	0.733	0.238
ICT	0.622	0.222
French	0.854	-0.010
Geography	0.630	0.380
Government	0.753	0.307
Economics	0.807	0.336
E. Mathematics	0.322	0.900

From Table 4.2, we can see that French, Economics and Government have high loadings of about 0.85, 0.80 and 0.80 respectively. This indicates that French, Economics and Government are influential in forming the first component. Considering that, the cut of point is 0.5 then this suggests that Core Mathematics (0.182) and Elective Mathematics (0.322) have relatively very low values of loading. However, on the second component Core Mathematics and Elective Mathematics are found to have quite high loadings of 0.96 and 0.90 respectively, indicating that they are influential in forming the component. French alone has the lowest and near zero loading of -0.01.

Table 4.3, shows the quartimax rotated component matrix of the variable (subject) loadings.

Table 4.3: Quartimax rotated principal component matrix of eigenvectors

Variables	Component	
	1	2
English Language	0.753	0.067
Core Mathematics	0.435	0.874
Integrated Science	0.809	0.019
I C T	0.659	0.045
French	0.820	-0.241
Geography	0.709	0.194
Government	0.808	0.091
Economics	0.868	0.105
Elect.Mathematics	0.554	0.779

On PC1 of Table 4.3, it can be seen that all the variables (subjects) have substantial loadings above 0.5 except Core Mathematics (0.435) on the first component. However, Economics, Government, French and Integrated Science loaded quite high with values 0.868, 0.808, 0.820 and 0.809 respectively. This suggested that they are influential in forming the first principal component. The subjects that do involve a little or no mathematical ability all have loadings that are close to zero on the second component. Core Mathematics and Elective Mathematics have substantial loadings of values 0.874 and 0.779 respectively. This suggested that they are influential in forming the second component. French subsequently was the only subject that loaded negatively on the second principal components with a value -0.241.

Ranking of Students by Order of Performance Based on PCs (Index)

The ranking of the students was carried out using all the individual's first principal component of the unrotated principal components (Table 4.1), varimax rotated component (Table 4.2) and quartimax rotated components (Table 4.3). The unrotated principal component was found to have all the variables loading substantially quite high on its PC1 as compared to the other two but majority of the variable loaded negatively on PC2 except Core and Elective Mathematics. The quartimax rotated components matrix indicated PC1 having the variables with quite high loadings, whereas PC2 indicated very low loadings by some of the variables (subjects) except Core and Elective Mathematics. Similarly, the varimax

rotated PC has relatively high loadings by most of the variables on PC1 and quite low loadings on PC2 except Core and Elective Mathematics with quite high loadings.

The fitted models for the individual PCs were used to obtain the indexes by substituting the scores for each student in the nine subjects. The value of the index determine whether or not the performance of a student in relation to others is better or worse. That is, the higher the score for the index the better the performance. In effect, resultant scores from the indexes were compared with each other in order to assess the consistency of the PCs. It was found that all the principal components were very consistent when the students were ranked except that of varimax. On this basis, the PC1 of unrotated principal component was identified to be the best index and was used for the ranking to determine the outstanding overall best student in the examination.

It can be concluded that the overall best student that deserved to be given the ultimate award is the student with number 042, followed by 029 and 011 as second and third best runners up. The scores of the PCs used for the ranking can be found in Appendices B, C and D.

CHAPTER FIVE

SUMMARY, DISCUSSION AND CONCLUSIONS

This chapter presents a summary of the study, its findings and the main conclusion of the study.

Summary

The researcher was challenged by the average performance of the students in the school, and therefore gets inspired and motivated to find out whether males perform better than the female students do. Further, many people have the perception that males perform academically better than the females in school. In view of this, the researcher was highly motivated to compare the performance of the male and female students. Again, the researcher was interested in identifying the overall best student in the class chosen for the study. Many at times, students are ranked using a simple method by summing their individual scores in the various subjects and arrange them in descending order of magnitude to find the best student. However, the researcher was motivated to employ principal component method, to standardize the scores such that a model was formulated as an index for the ranking.

The study aims at, first to compare the performance of male and female students, second to use principal components to determine the best index for ranking the students in descending order. The best index obtained by PC1 was used for ranking the students in descending order of performance for the identification of the overall best student.

The study revealed that the males performed better than their female counterparts in the examination. The overall best student was revealed by the ranking to be a male followed by the second best who was also a male. However, a female emerged the third best.

Discussion

The study analyzed the scores obtained by the students in the various subjects in the third term examination at St. Joseph Senior High Secondary/Technical School. The relevant data were sourced from the students' terminal report. It covers only third term examination results for 2008/09 academic year for students in SHS 2A.

There were two sections of the analysis. The first section, which is the preliminary, looked at the variability in the scores obtained by the students for every subject in the examination. In addition, the normality of the individual subject scores was assessed and it was revealed that the scores of some subjects were approximately normally distributed. The results revealed that the pattern of the scores among the students was irregular. It varied from one subject to the other.

The subjects did satisfy the conditions discussed in Chapter Two for the factorability of the correlation matrix and therefore the objective of ranking the students was pursued. It was identified that the value of the various subjects on the principal components indicates or reflects the performance of the students.

Researchers have carried out several studies to find out how females are fairing in various fields of endeavour alongside their male counterparts. Similar and related studies discussed in Chapter One stated that the widespread belief that males out-perform females in mathematics is apparently a myth. Besides, it was revealed in another study that was stated in Chapter One that, Gender differences in mathematics performance that favour males are usually attributed to gender socialization. Further, it was revealed in another study which was indicated in the literature that, there was no difference between mathematics performance of male and female students. Similarly, when nine (9) subjects were considered in this study to find out if there is differences between male and female performance in the examination. It was revealed that males out-performed females.

Principal component analysis technique was used to further analyze the data of this study. Principal components were used to determine indices to rank the students in order of performance. It appears that the result of PCA coincides with the simple sum of scores in all subjects. However, this technique is preferred since it allowed or permitted the researcher to standardize the scores and examine the degree of influence of each subject in forming the respective principal components. It also enabled the researcher to determine the proportion of variability that was explained by the components and identify the subjects in

which students perform better or worse. The first principal component of the unrotated component matrix was identified to be the most suitable for fitting the index model that was used for the ranking of the students. Thus, the fitted model was given as:

$\boldsymbol{Y_1}$=0.742$\boldsymbol{x_1}$+0.659$\boldsymbol{x_2}$+0.783$\boldsymbol{x_3}$+0.646$\boldsymbol{x_4}$+0.721$\boldsymbol{x_5}$+0.735$\boldsymbol{x_6}$+0.801$\boldsymbol{x_7}$+0.863$\boldsymbol{x_8}$+ 0.747$\boldsymbol{x_9}$

(5.1)

where $\boldsymbol{x_1}$, $\boldsymbol{x_2}$, $\boldsymbol{x_3}$, $\boldsymbol{x_4}$, $\boldsymbol{x_5}$, $\boldsymbol{x_6}$, $\boldsymbol{x_7}$, $\boldsymbol{x_8}$, $\boldsymbol{x_9}$ represent the respective subjects. The outstanding overall best student in the examination was determined by using the PC1 of the unrotated component matrix. With this model it implies that we can determine the best student in the class under study from term to term to some extent ascertain its consistency.

Conclusions

The results obtain from this study revealed that, males outperformed females in the examination. The actual scores were used to determine an index for ranking the students in order of performance. PC1 of the unrotated component, was found to be the most suitable index, which has Economics, Gorvenment and Integrated Science being highly influential in forming it.

The rankings based on PC1 of unrotated component matrix and that of quartimax rotated component matrix appeared to be consistent, since each revealed that student with serial number 042 was the overall best, followed by student with serial number 029 and then student with serial number 011 as the

second and third runners up respectively. To conclude, it was identified that the overall best student was a male, the second best was also a male followed by the third best who was a female.

Suggestions for Further Research

This research was conducted for only one class of the whole form two. Therefore, it will be of interest to conduct this similar research on the three terms of the academic year. Another area of much interest would be to discover another robust technique to develop an index that could account for a greater percentage of the total variance in the sample used.

References

Agyei D.D. and Eyian-Bediako,F.(2008). Gender differences in Mathematics Performance, *Journal for Gender and Behaviour*. 1519-1529

Brush, L.R. (1980), *Encouraging Girls in Mathematics: The Problem and the Solution.* Cambridge, MA Abt. Books.

Cattel, R.B. (1966) *The scree test for the number of factors ,Multivariate Behavioural Research*. Addison-Wesley Publishing Company, Inc.

Cliff, N. (1987) *Analyzing Multivariate data*. New York: Harcourt, Brace Jovanovich.

Gorsuch, R. L.(1983). *Factor Analysis* (2nd edition). Hillsdale, New Jersey: Erlbaum.

Guttman, L.(1954). *Some Necessary Conditions for Common Factor Analysis* Psychometrika, 19, 149-160.

Harlow, L.L.(2005). *The essence of multivariate thinking*. Lawrence Erlbaum Associates, Mahwah Publishes. New Jersey, 199-206.

http://news.myjoyonline.com/education/2008 (accessed: July 1, 2010)

http://www.eyeofdubai.com (accessed: July 1, 2010)

Johnson, R.A. & Wichem, D. W. (1981).*Applied multivariate statistical analysis*.(5th ed) Englewood Cliffs, NJ: Prentice Hall.

Jolliff, I.T. (1986) *Principa*l *component analysis*. New York: Springer-Verlag.

http://www..law.monash.edu/prizes/sir-john-monash-medal.html (accessed: July 1, 2010)

Jonathon Shlens, *A Tutorial on Principal Component Analysis*, Institute for Nonlinear Science, University of California, San Diego

Kaiser, H.F. (1970). *A second generation Little Jiffy.* Psychometrika, 35, 401-415

Lucy, S. (1973). Maths Mystique: Fear of Figuring. *Time Magazine*, 171(3)

Rowtree, D. (1977) *Assessing Students*. London: Harper & Row

Sharma,S.(1996). *Applied Multivariate Techniques*. John Wiley & Sons,Inc. 58-84.

Thurstone, L.L. (1935). *The vectors of the mind*. Chicago: University of Chicago Press.

Ramsden, P. (1992) *Learning to Teach in Higher Education* London: Routledge

Written by Prince Education Feb 16, 2010 (accessed: July 1, 2010)

Appendix A1: Data set of scores obtained by students in the examination.

Engl	CMat	IntSc	French	Geog	Gov	Econs

Appendix B: Ranking of students based on Unrotated PC1 (index) scores

Ranking	Serial No.	Relative Indices
1	042	482.771
2	029	482.454
3	011	468.844
4	020	426.299
5	024	414.083
6	006	408.661
7	027	407.499
8	019	406.024
9	022	392.605
10	035	390.508
11	007	390.109
12	003	384.902
13	045	383.474
14	009	379.403
15	010	378.278
16	002	370.612
17	021	370.585

18	047	365.382
19	046	363.889
20	005	363.777
21	032	357.909
22	041	357.547
23	034	355.034
24	016	350.316
25	028	346.625
26	036	344.612
27	001	342.890
28	018	334.834
29	039	333.423
30	043	332.478
31	040	328.692
32	015	324.291
33	004	321.780
34	037	318.098
35	030	317.870
36	033	316.465
37	017	315.190
38	008	312.791
39	023	305.089
40	031	300.057

41	038	268.805
42	044	266.050
43	026	258.780
44	025	231.655
45	013	229.701
46	012	213.555
47	014	212.056

Appendix C: Ranking of students on Quartimax rotated PC1 (index) scores

Rank	Serial Number	Relative Indices
1	042	462.212
2	029	457.900
3	011	452.430
4	020	405.674
5	024	397.515
6	006	394.034
7	027	393.697
8	019	388.560
9	022	377.057
10	035	375.467
11	007	374.929
12	009	365.763
13	045	363.767
14	003	360.885
15	010	359.251
16	002	355.780
17	046	350.632
18	005	348.921

19	047	346.977
20	021	345.805
21	034	342.951
22	032	340.104
23	041	337.621
24	016	332.045
25	028	331.175
26	001	328.713
27	036	321.319
28	018	315.828
29	039	315.338
30	043	312.346
31	040	309.555
32	015	309.552
33	004	309.240
34	033	304.796
35	037	303.498
36	017	296.671
37	008	294.296
38	030	293.902
39	023	285.333
40	031	282.471
41	038	258.389

42	044	250.635
43	026	246.802
44	013	223.530
45	025	212.982
46	014	200.817
47	012	194.669

Appendix D: Ranking of students on Varimax rotated PC1 (index) scores

Ranking	Serial No	Relative Index
1	042	406.786
2	011	401.811
3	029	398.892
4	020	354.491
5	024	350.945
6	027	350.773
7	006	348.975
8	019	341.784
9	007	335.496
10	035	332.070
11	022	328.982
12	045	320.486
13	003	315.292
14	009	315.188
15	002	314.109
16	010	313.198
17	005	310.970
18	041	302.475
19	047	302.450
20	046	301.690
21	021	301.044

22	034	299.524
23	032	291.961
24	028	290.764
25	016	288.782
26	036	288.046
27	001	275.643
28	043	275.413
29	004	273.342
30	018	272.143
31	015	271.490
32	030	268.719
33	039	267.837
34	033	267.640
35	040	267.151
36	008	258.232
37	017	251.362
38	037	247.739
39	023	244.181
40	031	243.685
41	038	228.454
42	026	223.519
43	044	209.086
44	025	198.503

45	**013**	**180.350**
46	**012**	**172.992**
47	**014**	**162.760**