Appraising Obesity among Female Undergraduate Students of Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria: A Discriminant and Principal Component Analysis Approach

Chisimkwuo JOHN[1], Tal Mark POKALAS[2], and Kindness Chinma EGESIE[3]
[1,2,3]Department of Statistics, Michael Okpara University of Agriculture, Umudike, Nigeria.
Correspondence E-mail: john.chisimkwuo@mouau.edu.ng

ABSTRACT

The prevalence of obesity and its negative consequences is on the increase globally especially West Africa and Nigeria. This menace is fast increasing even among university students and if not properly checked it will have a far-reaching implication on the student's health and academic performance. Thus, this study measured the weight (Wt), Height (Ht), Waist Circumference, hip, body fat, systolic and diastolic blood pressure and pulse pressure of female students living in Block D of Michael Okpara University of Agriculture, Umudike, Abia State, Nigeria.

The variables were all measured using appropriate measuring instruments and 250 samples were collected. After validating the data for the necessary assumptions of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) approaches, these methods were employed for the data analysis. Employing the WHO criteria for classifying obesity into Normal weight (N), Obese (O) and Overweight (W) students, a comparison of group means shows that the obese group has a higher mean value for body mass index (BMI) than the other two groups. The prior percentage probabilities of an individual being in the non-obese, obese and overweight group is 67.8%, 8%, and 24.1% respectively. Indicatively, the PCA approach was able to reduce the dimension of the data with the first principal component (LD1) explaining 98.8% of the variation in the data while the second principal component (LD2) explains 1.2% of the variation. The model is used to predict obesity and was shown to possess 96.05% accuracy which implies that the error of mis-classification is 0.04%. It was concluded that there is only 8% chance of a female student being in the obese group as against 67.8% chance for the non-obese group.

Keywords: Obesity, Discriminant analysis, overweight, Body mass index, Principal component.

1. INTRODUCTION

In Nigeria, obesity is fast becoming one of the significant health concerns due to poverty and overdependence on the cheapest source of food irrespective of its nutritive value. It is an undisputable fact that obesity is more prevalent among males than their female counterpart (Aderonke *et al*., 2023; Oyechi and Okolo, 2008; WHO, 2021). This disparity cuts across higher education students as they constitute one of the groups that engage in unhealthy eating habits (Assaf *et al*., 2019; Peltzer *et al*., 2014). This is largely due to over dependence on fast food such as: carbonated drinks, snacks, processed food etc. in an attempt to meet up with submission deadlines for assignment and to optimize study times.

Many demographic studies have been conducted in the literature to ascertain the prevalence of obesity and overweight among undergraduate university students classifying the students into obese and non-obese (Peltzer *et al*., 2014; Onyechi and Okolo, 2008; Assaf *et al*., 2019; Patricia

and Pawa, 2022; Shaimaa *et al*., 2022; Rotich *et al*., 2023 and many more). However, none of the studies had been performed exclusively on classifying female students in Umudike hostels using a discriminant analysis approach.

Discriminant Analysis is a powerful statistical tool that is concerned with the problem of classification. This problem of classification arises when an investigator makes a number of measurements on an individual and wishes to classify the individual into one of the several population groups on the basis of these measurements (Morrison, 1967). It easily identifies the discriminant variables and allows for the development of predictive models, it addresses the limitations of traditional methods by simultaneously considering multiple variables and identifying the most influential factors in distinguishing between several groups. Thus, this study intends to discriminate and classify female students living in one of the three female hostels of Michael Okpara University of Agriculture into three groups: Non-obese, obese and overweight based on attributes such as: Height, weight, systolic blood pressure, waist circumference, BMI etc. using discriminant analysis and estimate the error of mis-classification and the percentage of students belonging to each of the three groups together with the corresponding prior probabilities of belonging to that group. In addition, this study will try to discover and select more relevant obesity variables using the principal component analysis approach. The prevalence of obesity among the female gender as reported by many studies is one of the motivations of this study.

2. METHODOLOGY

2.1 Sources and method of data collection

Between April 2022 and October 2023, we distributed a one-page questionnaire among the 250 study participants who were female students living in Goodluck Jonathan Hostel (Block D) of MOUAU and conducted a separate meeting session for each student at the university clinic in order to conduct the anthropometric measurements with the aid of trained nurse in accordance with the WHO standards. The data for this study is a primary data collected based on convenient sampling from the study area. A total of 250 female students aged 18- 29 participated in the study. The data collected include: weight (Wt), Height (Ht), Waist Circumference, hip, body fat, systolic and diastolic blood pressure and pulse pressure were all measured using standard techniques via the help of the Sphygmomanometer, weighing scale, calibrated scale, flexible and rigid tape, skin fold caliper, and questionnaires.

The students were classified as Normal weight students (N), Obese students (O) and Overweight students (W). The body mass index was computed using the WHO criteria. To conduct the anthropometric measurements the participants were asked to take off caps/ hats, wrist watch, tie, veil, hand bags, shoes and any outer garment that might influence the accuracy of measurements. Participants were directed to put on light attires and ensure their pockets are empty. The body mass index was estimated by dividing the student's weight by height. The body mass index was define using the WHO standards classification as obesity (Greater than 30kg/m^2) with overweight (25- to <30kg/m^2). The data collected from the study was analyzed using the R- software.

2.2 Method of data analysis

The data for this study will be analyzed using the Discriminant analysis as follows. Some basic assumptions necessary for the application of discriminant analysis can be carried out, it is assumed that such identifiable groups must have its underlying distribution from a multivariate normal distribution and Holgerson (2006) illustrated a graphical approach to multivariate normality. Also, the T^2-test is used to ascertain the test of difference between two different group mean vectors to be compared, while the MANOVA approach is used for cases with more than two group mean vectors. In addition, the Box-M test was used to test for the equality of the covariance matrices. These approaches are enumerated in Johnson and Wichern (2007)

2.3 Principal Component Analysis

Principal Component Analysis (PCA) dates back to the early 90's when Pearson (1901) and Hoteling (1947) arrived at the same result using different approaches. Their approaches view the PCA in two perspectives, namely, as a dimension reduction technique and as an approximation technique. As a dimension reduction technique, PCA reduces the dimensionality of a large data matrix, described by several inter-correlated variables, while retaining as much as possible, the variation present in the original data matrix. Abdi and Williams (2010) defined PCA as a multivariate technique that uses an orthogonal transformation to convert a data matrix of possibly correlated variables into a new data matrix of uncorrelated (orthogonal) variables called principal components. These principal components are obtained as linear combinations of the original variables.

Johnson and Wichern (2007) showed that the first principal component (PC) is chosen to have the largest possible variance. In their formulation, each of the succeeding components are orthogonal to the previous components and has the largest possible variance. Let the data matrix $\mathbf{X}$ has n samples and p variables, i.e., $(n \times p)$. A summary of obtaining the principal components are summarized in the following algorithm:

Algorithm 1: *The PCA Pseudocode*

1. **Require:** $\mathbf{X}: n \times p$ matrix
2. **Evaluate:** the centralize $\mathbf{X_0}$ by subtracting the mean $\bar{\mathbf{x}} = \frac{1}{n}\mathbf{X'1}$ from $\mathbf{X}$
3. **Calculate:** $\mathbf{X}_0 = \mathbf{U\Sigma V'}$ the SVD of $\mathbf{X}$
4. **Get:** $\mathbf{V}$ (the right robust bootstrapped singular vectors)
5. **Obtain:** the k dimensional principal components: $\mathbf{Z} = \mathbf{XV}_{[k]}$
6. **Return:** $\mathbf{Z}$, the principal components (scores)

A suitable inferential study is also obtainable in most exploratory techniques like the PCA and its related methods. Since the PCA approaches are majorly used as a dimension reduction technique, focus will be channeled to the number of PCs to select. Some strategies to determine the number of principal components are discussed by Jolliffe (2002). Keiser (1958) also discussed the Kaiser criterion, which proposed the method where the choice of the components is chosen by excluding the components for which the eigenvalues of the correlation matrix are less than one. This method excluded all the eigenvalues that are less than the average of all the eigenvalues. In this approach,

let p represent the total number of variables and λ_i be the i^{th} eigenvalue of $\mathbf{S_X}$, where $\mathbf{S_X}$ is the covariance matrix of $\mathbf{X_0}$, then any λ less than the average eigenvalue given by $\sum_{i=1}^{p}\frac{\lambda_i}{p}$ will be excluded. In another method, since Johnson and Wichern (2007) gave the total variation in the data matrix as $\sum_{i=1}^{p}\lambda_i = trace(\mathbf{S})$, Rossouw (2016) argued therefore that a possible approach is to retain the number of components that correspond to some percentage of the total information. Typically, components that forms up to 80% of the retained information are selected. Another method that merits consideration is the permutation test method discussed by Horn (1965) where a traditional method was used in choosing the number of components based on the scree test. Horn (1965) termed this approach as the "parallel approach" and opined that it is generally superior to the other methods in a comparative study on the selection methods. The scree test entails plotting the eigenvalues λ_i against i and visually identifying the point where the slope change. In the scree test selection criteria, Jolliffe (1986) suggested that the number of variables to be selected from the scree plot should correspond to the $min(p, n-1)$ eigenvalues of the covariance matrix. Finally, Hubert *et al.* (2005) considered the selection criteria that chooses k variables for which $\sum_{i=1}^{k}\hat{\lambda}_i \Big/ \sum_{j=1}^{r}\hat{\lambda}_j \approx 95\%$, where $\hat{\lambda}_i$ are the sorted eigenvalues and r is the rank of the covariance matrix.

2.4. Discriminant Function

Given samples $y_{11}, y_{12}, ..., y_{1n1}$ and $y_{21}, y_{22,}, ..., y_{2n2}$ from two separate population, where each vector y_{ij} consist of measurement on p variables. The discriminant function is the linear combination of these p variables that maximizes the distance between the two transformed group mean vectors. In Rencher (2002), the Fisher's Linear Discriminant function (FLDF) was presented as:

$$y = \left(\overline{X}_1 - \overline{X}_2\right)^T \boldsymbol{S}^{-1} X_1 = \hat{\boldsymbol{a}}^T X \qquad (1)$$

With the mean discriminant function given by $\overline{y}_1 = \hat{a}\overline{X}_1 = \left(\overline{X}_1 - \overline{X}_2\right)^T \boldsymbol{S}^{-1}\overline{X}_1$ and the critical value of the Fisher's linear discriminant function given by $\hat{y}_{critical} = \frac{\overline{y}_1 + \overline{y}_2}{2}$. The rule is to classify the students whose discriminant scores are greater than or equal to the critical value into Obese and those whose discriminant scores are less than the critical value into the Non-Obese.
This measures the performance that does not depend on the form of the parent populations and that can be calculated for any classification procedure. The Apparent Error Rate (APER) can be easily calculated from the confusion matrix, which shows actual versus predicted group membership. For n_1 observations from G_1 and n_2 observations from G_2, the confusion matrix has the form; obese
rate (APER) layout

	Predicted membership		
Actual membership	G_1	G_2	
G_1	n_{1c}	$n_{1m} = n_1 - n_{1c}$	n_1
G_2	$n_{2m} = n_2 - n_{2c}$	n_{2c}	n_2

Where n_{1c} = number of π_1 items correctly classified as G_1 items, n_{1m} = number of π_1 items misclassified as G_2 items, n_{2c} = number of π_2 items correctly classified as π_2 items, n_{2m} = number of π_2 items misclassified as π_1 items. The apparent error rate (APER) is the calculated as:

$$\text{APER} = \frac{n_{1m} + n_{2m}}{n_1 + n_2} \qquad (2)$$

RESULTS

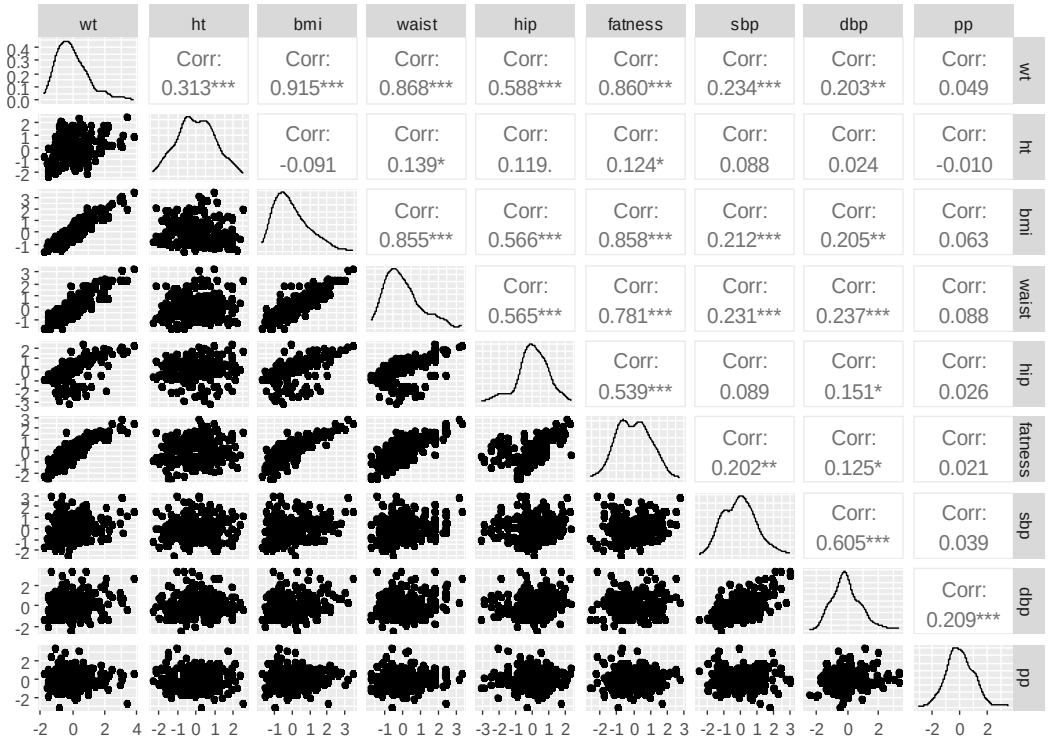

Figure 1: *Multivariate plot showing variable correlation*

The scree plot helps to check how well the principal component work on the data by determining the optimal number of factors and the amount of variation captured by each of the principal component. The elbow in the figure above appears to occur at the second principal component meaning that the first two principal component should be retained for the data analysis. In Figure 2, component 9 (PP – Pulse Pressure) is close to 0. These eliminates composition 9, and select other variables. This is supported by relative magnitude of the eigen value. Figure 3 shows the

display of the data in plane surface. It can be observed that there are variations in the sizes and level of variables.

Figure 2: *The Scree plot of the data*

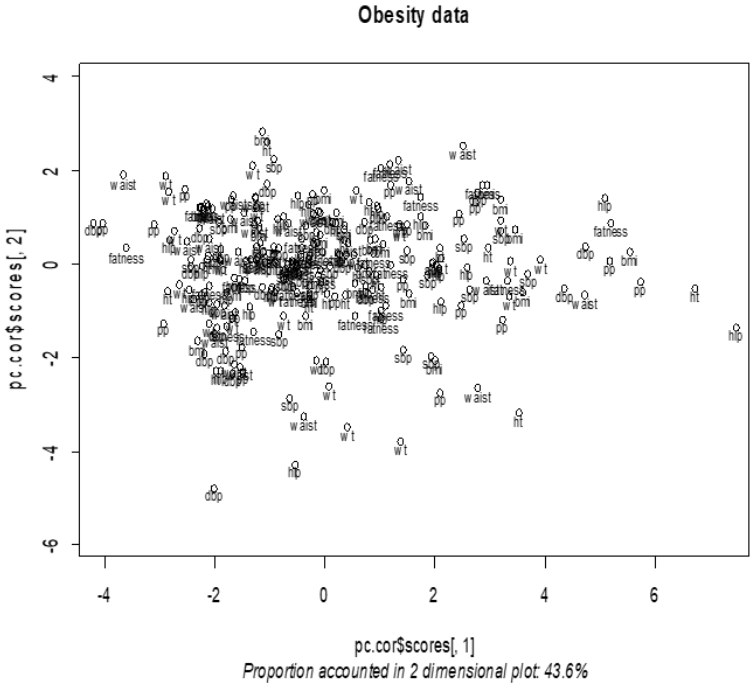

Figure 3: *The PC loading in 2D Configuration*

Table 2 *Multivariate Normality Test for Non-obese Students*

Test	Variable	Statistic	p value	Normality
Shapiro-Wilk	Wt	0.9933	0.6015	YES
Shapiro-Wilk	Ht	0.9883	0.158	YES
Shapiro-Wilk	BMI	0.9753	0.0033	No
Shapiro-Wilk	Waist	0.9673	4e-04	No
Shapiro-Wilk	Hip	0.9248	<0.001	No
Shapiro-Wilk	Fatness	0.9942	0.7267	YES
Shapiro-Wilk	Sbp	0.9776	0.0063	No
Shapiro-Wilk	Dbp	0.9582	<0.001	No

Table 3 *Multivariate Normality Test for Obese Students*

Test	Variable	Statistic	p value	Normality
Shapiro-Wilk	Wt	0.9425	0.2439	YES
Shapiro-Wilk	Ht	0.9563	0.4443	YES
Shapiro-Wilk	BMI	0.8763	0.0125	No
Shapiro-Wilk	Waist	0.9389	0.2072	Yes
Shapiro-Wilk	Hip	0.8988	0.0331	No
Shapiro-Wilk	Fatness	0.9540	0.4035	Yes
Shapiro-Wilk	Sbp	0.9691	0.7140	Yes
Shapiro-Wilk	Dbp	0.9537	0.3994	Yes

Table 4 *Multivariate Normality Test for Over-weight Students*

Test	Variable	Statistic	p- value	Normality
Shapiro-Wilk	Wt	0.9324	0.0046	No
Shapiro-Wilk	Ht	0.9707	0.2072	Yes
Shapiro-Wilk	BMI	0.9384	0.008	No
Shapiro-Wilk	Waist	0.9580	0.0559	Yes
Shapiro-Wilk	Hip	0.8043	<0.001	No
Shapiro-Wilk	Fatness	0.9884	0.8781	Yes
Shapiro-Wilk	Sbp	0.9697	0.1877	Yes
Shapiro-Wilk	Dbp	0.9374	0.0073	No

In Tables 2, 3, and 4, a normality test was conducted for all the three groups and it was observed that almost half of the variables for obese, non-obese and overweight are not normally distributed and this may be attributed to the presence of outliers. Outliers were detected and deleted using the robust squared Mahalonobis distance.

Table: 5: *Boxes' M Homogeneity of Covariance Matrices Test*

Variables	df	χ^2	P-value
Normal, Obese and Over weight	30800	538.69	1

Table: 6: *Boxes' M Homogeneity of Covariance Matrices Test*

	df	Pillai	Approx. F	num df	Den Df	Pr. (>F)
	22	0.84352	21.973	16	482	<2.2e-16***
Residual	247					
Signif. Codes	0	'***' 0.001	'**' 0.01'	*' 0.05	'.' 0.1	''1

Tables 5 and 6 suggests that there are significant difference between the variable means which indicate that the variance/ co-variance matrices are equal across all the three groups.

Table 7: *Prior probabilities of groups*

Non-obese	Obese	Overweight
0.67816092	0.08045977	0.24137931

The prior probabilities show the probability of an individual belonging to each group, given no additional information. In this study, the prior probabilities of being in the non-obese, obese and overweight group is 0.678, 0.080 and 0.241 respectively. This information can be useful for determining the expected distribution of individuals in each group, but such probabilities can change as more information is obtained.

Table 8: *Coefficients of linear discriminants:*

	LD1	**LD2**
Wt	-3.05392545	3.94153942
Ht	1.27952918	-1.75297457
BMI	5.31386149	-4.50775867
Waist	0.08105027	-0.06609472
Hip	0.06039902	-0.52654023
Fatness	-0.38489623	1.38789189
Sbp	0.11107695	0.40500949
Dbp	-0.03477627	-0.69483971
Proportion of Stress	LD1 = 0.988	LD2 = 0.012

The larger coefficients help identify key variables (Positive or Negative) that carry more weight with respect to the linear discriminant. In essence, body mass function (BMI) has a strong positive (5.3139) and negative (-4.5078) differences on the first and second linear principal component respectively. In Table 8, the first principal component LD1 explains 98.8% of the variation in the data while the second principal component LD2 explains 1.2% of the variation which means that LD1 contribute more to obesity than LD2. Predicting obesity using the model shows 96.05% accuracy which implies that the error of mis-classification is approximately 0.04%.

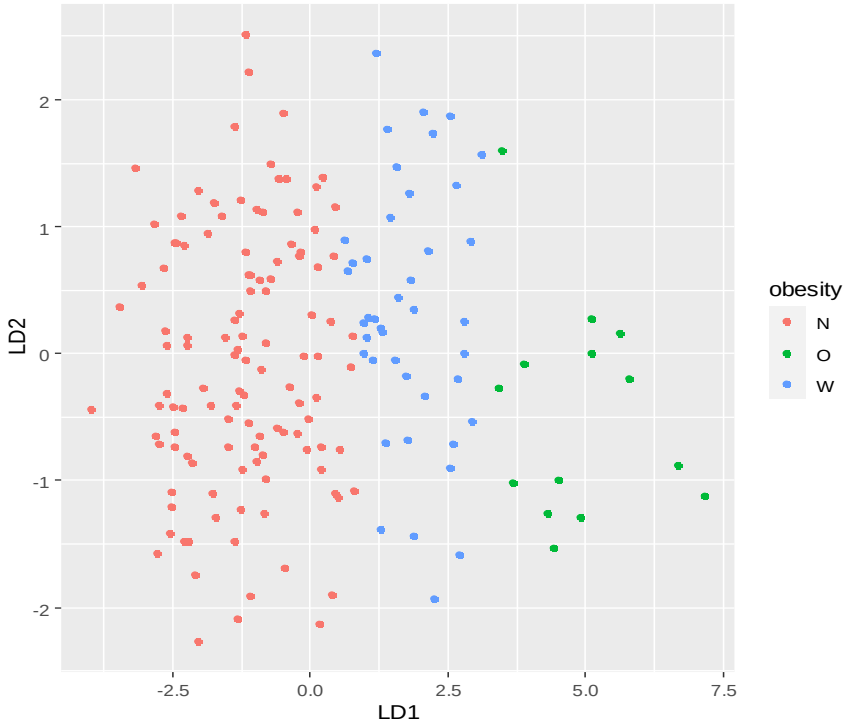

Figure 4: *Distribution of student's obesity Status*

From Figure 4, it is obvious that there is lesser distribution of students classified under the obese group than the other groups. However, the higher distribution of students is classified under the non-obese group.

Table 9: *The Confusion matrix*

Actual Membership	**Predicted Membership**		**Total**
	To SLT =c (π_1)	**To STA=** s (π_2)	
From SLT= c (π_1)	68	32	100
From STA= s (π_2)	19	39	58

The apparent error rate is then given as $APER = \frac{n_{1\pi_2} + n_{2\pi_2}}{n_1 + n_2} = \frac{32+19}{158} = \frac{51}{158} = 0.32$

Observe that the probability of misclassification of SLT candidates $= \frac{n_{1\pi_2}}{n_1} = \frac{32}{100} = 0.32$ and the probability of misclassification of STA candidates= $\frac{n_{2\pi_2}}{n_1} = \frac{19}{100} = 0.33$

CONCLUSION

In this study we applied the principal component analysis to classify female students into groups based on their obesity status. The study found that the prevalence of obesity among female students of MOUAU is within a tolerable region but adequate intervention measures should be put in place to avoid further increase. Most of the students are non-obese and very few percentages of them are obese

REFERENCES

Abdi, H. and Williams, L. J. (2010), Principal Component Analysis, *John Wiley & Sons, Inc. WIREs Comp Stat,* 2, 433–459

Aderonke, M., Ifeoluwa, B., Kahinde, A., and Elizabeth, A. (2023). Overweight and Obesity are Prevalent among Female adults in selected areas in Ibadan, Oyo State, Nigeria. *Clinical Epidemiology and Global Health*, 22, 101314

Assaf, I., Brietch,F.T., and Faily,M. (2019). Students University Health Lifestyle Practice: quantitative analysis. *Health International Science Systems,*7,7-14.

Holgersson H. (2006). A graphical method for assessing multivariate normality. *Computational Statistics*, 21, 141-149.

Hoteling, H. (1947). *Multivariate Quality Control in Techniques of Statistical Analysis*, McGraw-Hill, New York, 111-184.

Horn, J. L. (1965). A Rationale and Test for the Number of Factors in Factor Analysis, *Psychometrika*, 30(2), 179-185.

Johnson, R. A. and Wichern D. W. (2007), *Applied Multivariate Statistical Analysis, 6th edition*, Pearson International Edition, Upper Saddle River, NJ

Jolliffe, I. T. (2002). *Principal Component Analysis (2nd Edition).* Springer, USA.

Kaiser, H. F. (1958). The Varimax Criterion for Analytic Rotation in Factor Analysis, *Psychometrika*, 23, 187–200.

Morrisson D.F. (1967). *Multivariate Statistical Methods*, McGraw-Hill Book Company, New York.

Onyechi, U.A and Okolo, A.C. (2008). Prevalence of obesity among Undergraduate Students living in a Hall of Residence, University of Nigeria Nsukka Campus. *Animal Research International*, 5(3), 456-467.

Patricia, P., and Pawa, P. (2022). Overweight and Obesity: a study among university students in Sarawak, Malaysia. *International Journal of Health promotion and Education*, DOI: 10.1080/14635240.20

Pearson, K. (1901). On Lines and Planes of Closest Fit to System of Points in Space, *Journal of Educational Philosophical Magazines*, 2, 559-572.

Peltzer, K, Pengpid, S and Samuel, T.A. (2014). Prevalence of Overweight and Obesity and its associated factors among University Students from 22 countries. *International Journal of Environmental Research and Public Health,*11, 7425-7441.

Rossouw, R.F. (2015*). Multivariate Statistical Process Evaluation and Monitoring for Complex Chemical Processes*, Ph.D Thesis, University of Stellenbosch.

Rotich, S., Kamau, J., Oketch, M., and Okube, O. (2023). Prevalence and predictors of Obesity among Undergraduates Students at a Private University, Nairobi, Kenya. *Open Journal of Endocrine and Metabolic Diseases,* 13(2), 23-38.

Shaimaa, M.H., Eaman, M.M. and Shimaa, A.E. (2022). Obesity/Overweight among University Students, Mania Egypt. *Minia Journal of medical Research.*

World Health Organization (2010). Fact sheet on Obesity. http://www.who.int/topics/obesity/en/, google scholar. World Health Organization (2021). Obesity and Overweight.