Original Research Article

A STATISTICAL ANALYSIS ON THE EFFECT OF BAD HEALTH HABITS IN TWO CONTINENTS (AFRICA AND EUROPE)

ABSTRACT

This research work focuses on determining the difference between the health habits of countries in Africa and Europe, especially in females. It is crucial because it could help enlighten women on the dangers of bad health habits. Multivariate Hotelling T- square test is adopted to determine the significant difference between the two continents, Africa and Europe, having Cancer deaths caused by alcohol consumption, smoking prevalence, and Obesity prevalence as the variables and the correlation between the variables. The result showed that there is indeed a significant difference between the bad health habits in the two continents in the females. Correlation analysis was also carried out to determine the relationship between the variables, and the results showed that the relationship between the variables was little. More methods were adopted using comparison of Rate between the means, and the sum of the same variables, to determine which continent was more affected by the bad health habits, and it was figured out to be Europe.

Keywords: Multivariate Hotelling T- square test, bad health habit, Obesity, Correlation.

1. INTRODUCTION

Multivariate analysis is a principle that is based on the use of multivariate statistics. It involves observation and analysis of more than one statistical outcome variable at a time (Krzanowski & Everitt, 1989). Almost all data collation processes result in multivariate data. Most multivariate data involve analysis, estimation, construction of confidence interval sets, and hypothesis testing for means, variance, covariance, correlation coefficient, and related, more complex population characteristics (Bradley & Jackson, 1961).

Multivariate statistical analysis is crucial in social research because researchers in this field are often unable to use randomized laboratory experiments that their counterparts in medicine and natural science often use. Multivariate techniques try to statistically account for these disparities and adjust outcome measures to the portion that can be attributed to the differences (Shabbir, 1998).

Test hypotheses are of different types, depending on the nature of the data and the number of independent variables. When just one variable is involved, it is called univariate analysis, and the following terms are applied: Chi-square, T-test, Z-test, etc., however when two or more variables are involved, it is called multivariate analysis, and the Hotelling T-square statistic is applied to test for its significant difference.

The test statistic T-squared is called “The Hotelling T-squared” in honour of Harold Hotelling’s pioneer in multivariate analysis (1931). Hotelling T-squared statistic is used as a generalization

of students' t-distribution. T-squared is applied when the members of response variables are two or more, although it is also applied when there is only one response variable (Anderson, 1960).

Health habits from the medical dictionary are defined as behavioural attributes that are beneficial to one's physical or mental health, often linked to a level of discipline and self-control. Examples of good habits are Regular exercise, consumption of alcohol in moderation, balanced diet, etc. Examples of bad habits are Smoking, drug abuse, gambling, sexual promiscuity, poor sleep hygiene, etc.

Healthy habits are the best way to avoid diseases, prolong your life span and live congruously. A nourishing diet is a keystone to a healthy lifestyle. Excluding weight loss and maintenance, eating a balanced diet is crucial to every human life, especially in that of a woman. Good food provides vitamins, minerals, and nutrients that are crucial for human growth, well-being, and development.

Healthy habits are crucial in human life because it helps to prevent certain health conditions such as heart diseases, stroke, cancer, high blood pressure, e.t.c. It is very crucial for postmenopausal women. Today, versed changes have occurred in the life of every individual. Malnutrition, unhealthy diet, smoking, alcohol consumption, drug abuse, stress, e.t.c., are the presentations of the unhealthy lifestyle seen as the dominant form of lifestyle (Dariush, 2015). Men are more likely to abuse smoking and alcohol, thereby becoming dependent on them. However, the effects of chronic alcohol and smoking are much more impactful on women than men due to their fragile nature. These complications include heart disease, breast cancer, death, e.t.c. The hostile effects of obesity on women's health are extremely unbearable and inarguable. Obesity negatively impacts the health of women in so many ways. It affects both contraception and fertility as well. (Lash & Armstrong, 2009). Thereby increasing the risk of neonatal mortality, malformation, higher risk of multiple cancers, e.t.c. Obesity is becoming more prevalent and has wide-ranging effects on a variety of women's health. Clinicians should sensitize all women on the effects of obesity and the importance of controlling weight to prevent negative outcomes (Teresa et al, 2011).

Specifically, this work aims to test the significant difference between two populations using multivariate Hotelling T-squared statistics. This study, therefore, had some objectives. Firstly, it will attempt to determine the mean vector for each sample, the covariance matrix for each sample data set, and the correlation matrix. Secondly, it hopes to determine the significant difference between the population parameters using Hotelling's T-square test and the rate method to determine the continent more affected. This research helps to x-ray the differences in bad health habits between the countries in these two continents: Africa and Europe. This work will help women, especially those of childbearing age in these continents, to curb the effects of these dangerous health habits.

2. METHOD AND METHODOLOGY

The authors obtained the data used in this research work (as seen in the online appendix) from the World Bank (canceratlas.cancer.org/data).

From the available countries, the paper applies the inclusion and exclusion criteria to limit the data to the scope of this study. The authors choose to include variables relating to women's health and the continents with high data availability (95% of the data available). The authors exclude variables with no direct relationship with women's health. Also excluded were the continents with low availability of data. After applying the criteria, the authors chose Africa and

Europe since they met the inclusion criterion. The authors also included the health variables based on their effect on women's health, as seen in the existing literature.
The data to be analyzed is the bad health habits between countries in these two continents: Africa and Europe. Sample 1 refers to Africa, and Sample 2 refers to Europe. Each Sample contains three variables; cancer deaths caused by alcohol drinking, smoking prevalence, and obesity prevalence (all in women).

The methods used in this paper include:

a) **Missing Values:** Proper handling of missing values is crucial in every experiment (Ogoke & Nduka, 2012). There are many methods of handling missing values in data: deletion methods, prediction methods, imputation techniques, among others. The paper used imputation techniques consisting of many types: mean imputation, multiple imputations, random overall, among others. The study handled missing values in this data using the method of overall mean imputation, whereby it replaces the missing values with the mean of the individual variables involved. To determine the mean vector of a given set of variables $x_1, x_2, x_3, \dots, x_k$, the authors use the $\bar{x}_1 = \frac{\sum_{i=1}^{n_1} x_{i1}}{n_1}$, $\bar{x}_2 = \frac{\sum_{i=1}^{n_2} x_{i2}}{n_2}$, … , $x_k = \frac{\sum_{i=1}^{n_k} x_{ik}}{n_k}$, then expressed in terms of the vector $\bar{x}$

$$\text{i.e.} \quad \bar{X} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \vdots \\ \bar{x}_k \end{pmatrix}$$

The authors consider two of these vectors, i.e.

$$\bar{X}_1 = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \vdots \\ \bar{x}_k \end{pmatrix} \quad \text{And } \bar{X}_2 = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \\ \vdots \\ \bar{x}_k \end{pmatrix}$$

a. Variance – Covariance Matrix (Covariance Matrix): The covariance matrix is a measure of the relationship between two random variables. This matrix is essentially a measure of the variances of the variables between two variables. It indicates the linear relationship between variables. For the sample vector$(x_1, x_2, x_3, \dots, x_n)$, the variance-matrix is $= \frac{1}{n-1} \sum (x - \bar{x})' (x - \bar{x})$

$$\text{i.e. } S_x = \begin{pmatrix} S_1^2 & S_{12} & S_{13} & \cdots & S_{1k} \\ S_{21} & S_2^2 & S_{23} & \cdots & S_{2k} \\ S_{31} & S_{32} & S_3^2 & \cdots & S_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ S_{n1} & S_{n2} & S_{n3} & \dots & S_{nk}^2 \end{pmatrix}$$

Where,

$$S_i^2 = \frac{\sum x_i^2 - n\bar{x}_i^2}{n-1} \quad \text{And} \quad S_{ij} = \frac{\sum x_i x_j - n\bar{x}_i\bar{x}_j}{n-1}$$

b. Correlation Matrix: A correlation matrix is a table showing correlation coefficients between variables. It summarizes data and measures the strength of the relationship between pairs of variables in the table. The correlation coefficient matrix of a covariance matrix is written as

$$\ell_x = \begin{pmatrix} 1 & \ell_{12} & \ell_{13} & \cdots & \ell_{1p} \\ \ell_{12} & 1 & \ell_{23} & \cdots & \ell_{2p} \\ \ell_{13} & \ell_{32} & 1 & \cdots & \ell_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \ell_{1p} & \ell_{2p} & \ell_{3p} & \ldots & 1 \end{pmatrix}$$

Where,

$$\ell_{ij} = \frac{S_{ij}}{S_i S_j} \quad ; \quad ;$$

S_{ij} *is a covariance matrix and* S_i, S_j *are standard deviatons of variances.*

b) **Test of Hypothesis:** In the test of covariance significance between two population mean vectors, the hypothesis can be of the form:

$$H_0 : \hat{\mu}_1 = \hat{\mu}_2$$
$$H_1 : \hat{\mu}_1 \neq \hat{\mu}_2$$

The Hotelling T-squared statistic is used in testing the hypothesis. It is a generalization of student's t-statistic, and it is used in multivariate hypothesis testing. It measures distance from target using covariance matrix. It is written as:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(X_1 - X_2)^T S^{-1}(X_1 - X_2).$$

$$where S = \frac{1}{n_1 + n_2 - 2}\left[n_1 S_1 + n_2 S_2\right]$$ and $\bar{X}_1$ and $\bar{X}_2$ are the mean vectors of the two

samples, S_1 and S_2 are the sample variance-covariance matrix of the populations and the sample sizes are n_1 and n_2 and S^{-1} is the inverse of S.

The test statistic used is the F – distribution with P degree of freedom, where P is the number of variables. Since the sample sizes used are large.

c) Decision Rule: This is a rule that leads to the acceptance or rejection of the null hypothesis.

Reject $H_0 if T^2 \rangle \left(\frac{p(n_1 + n_2 - 2)}{n_1 + n_2 - p}\right) F_{(\alpha)p,(n_1+n_2-p)}$, otherwise accept H_0

d) **Rate:** Rate is the ratio between two related quantities in different units. The rate values used here were gotten by dividing the highest mean vector by the lowest mean vector of the same variables in the two continents. Another approach was to divide the highest total value by the lowest total value of the same variables in the two continents, that is:

$$R = \frac{n_1}{n_2};$$

where

n_1 is the higher value (mean vector or total of the same variable in the two samples)
n_2 is the lower value (mean vector or total of the same variable in the two samples)

In this case, the ratio gotten from the mean vectors and the total of the same variables used in the two samples.

3. RESULTS AND DISCUSSION

Let x_1 be Cancer deaths caused by alcohol consumption, x_2 be Smoking Prevalence, and x_3 be Obesity Prevalence.

Sample one

The **mean vector** of Sample one is given as $\overline{X_1} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} = \begin{pmatrix} 1.4377 \\ 1.9811 \\ 16.9717 \end{pmatrix}$

And the **Covariance matri**x is $S_1 = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{ccc} x_1 & x_2 & x_3 \end{array} \\ \begin{pmatrix} 1.0901 & 0.3321 & -1.8055 \\ 0.3321 & 3.1666 & 1.7029 \\ -1.8055 & 1.7029 & 79.1563 \end{pmatrix} \end{array}$

The **Correlation matri**x is $\ell_1 = \begin{bmatrix} 1 & 0.1787 & -0.1944 \\ 0.1787 & 1 & 0.1076 \\ -0.1944 & 0.1076 & 1 \end{bmatrix}$

Sample two

The **mean vector** of sample two is given $as\ \bar{X}_2 = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} = \begin{pmatrix} 3.3581 \\ 15.1605 \\ 22.9395 \end{pmatrix}$

And the **Covariance matrix** is $S_2 = \begin{array}{c} \\ \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{array} \begin{array}{c} \begin{array}{ccc} \bar{x}_1 & \bar{x}_2 & \bar{x}_3 \end{array} \\ \begin{pmatrix} 0.9873 & 1.6393 & -0.5784 \\ 1.6393 & 37.0605 & -0.1398 \\ -0.5784 & -0.1398 & 15.2681 \end{pmatrix} \end{array}$

The **Correlation matri**x is $\ell_2 = \begin{bmatrix} 1 & 0.2710 & -0.490 \\ 0.2710 & 1 & -0.0059 \\ -0.1490 & -0.0059 & 1 \end{bmatrix}$

Then the Hotelling's T-square is

$$T^2 = \frac{n_1 n_2}{n_1 + n_2}(X_1 - X_2)^T S^{-1}(X_1 - X_2).$$

$$where\ S = \frac{1}{n_1 + n_2 - 2}\left[n_1 S_1 + n_2 S_2\right]$$

$$T^2 = T_{cal}^2 = 88.862$$

Where $T_{tab}^2 = \left(\frac{p(n_1+n_2-2)}{n_1+n_2-p}\right)F_{(\alpha)p,(n_1+n_2-p)}$

Where $P = 3, (n_1 + n_2 - 2) = 94$ and $(n_1 + n_2 - P) = (53 + 43 - 3) = 93$

Hence $T_{tab}^2 = \left[\left(\frac{3(94)}{93}\right)\right] F_{(0.05)(3),(93)} = [(3.0323)]2.703 = 8.196$

Since the $T_{cal}^2 > T_{tab,}^2 H_0 \, is$ rejected.

The result above implies that there is indeed a significant difference in the bad health habits between the two continents.

Table 1: Rates between the Variables.

Variables	**Africa Mean**	**Europe Mean**	**Rate Africa**	**Rate Europe**
Cancer	1.4377	3.3581	1	2
Smoking	1.9811	15.1605	1	8
Obesity	16.9711	22.9395	1	1
Variables	**Africa Total**	**Europe Total**	**Rate Africa**	**Rate Europe**
Cancer	76.2	144.4	1	2
Smoking	105	651.9	1	6
Obesity	899.5	986.4	1	1

Table 2: Correlation between the Variables.

Samples–continents	Variables	Correlation coefficients	Remark
Sample 1: Africa	Cancer VS Smoking	0.1787	Positive Slight Correlation
	Cancer VS Obesity	-0.1944	Negative Slight Correlation
	Smoking VS Obesity	0.1076	Positive Slight Correlation
Sample2: Europe	Cancer VS Smoking	0.2710	Positive Fair Correlation
	Cancer VS Obesity	-0.1490	Negative Slight Correlation
	Smoking VS Obesity	-0.0059	Negative Slight correlation

Footnote: The correlation between these variables was classified using the characterizations reported by (Ogoke *et al.* (2013)). These characterizations range from 0.00 to 0.20 (Slight), 0.21 to 0.40 (Fair), 0.41 to 0.60 (Moderate), 0.61 to 0.80 (Substantial), 0.81 to 1.00 (Almost Perfect). Table 2 shows that there is little or no relationship between the variables in both samples

3.1. Discussion of Results

From the result, the T_{cal}^2 = 88.86 and the $T_{tab,}^2$= 8.2, which implies a difference between the bad health habits in Africa and Europe. To determine the continent with more negative health habits, the authors used Rating the Means and Sums method. Comparing the mean values of the dataset shows that cancer caused by alcohol consumption for Africa was 1, while cancer caused by alcohol consumption for Europe was 2. It also showed that the smoking rate for Africa was 1, while the smoking rate for Europe was 8, and the obesity rate for Africa was 1, while the obesity rate for Europe was 1, which showed that the continent with the higher rate was Europe. The authors also used the total of each variable. In this case, cancer caused by alcohol consumption for Africa was 1, while that of Europe was 2. The smoking rate was 1, while for Europe it was 6, and the obesity rate for Africa was 1, while Europe was 1. This result also showed that the higher rate was from Europe too. Hence, the continent most affected by these health habits is Europe since it has a higher rate value than Africa when compared.

For sample one in table 2, the correlation between cancer caused by alcohol consumption and Smoking was 0.1787. This value shows a slightly positive correlation and little relationship between the two variables. Also, the correlation between cancer caused by alcohol consumption and Obesity was -0.1490. This value shows a slightly negative correlation and implies the two variables are a little related in the negative direction. The correlation between Smoking and Obesity was 0.1076. This value also shows a slightly positive correlation and implies a minor relationship between the two variables.

Furthermore, for sample two, the correlation between Cancer and Smoking was 0.2710, which shows a moderately positive correlation and implies a minor relationship between the two variables. Also, the correlation between Cancer and Obesity was -0.1944. This value indicates a

slightly negative correlation and means the two variables are a little related in the negative direction. Finally, the correlation between Smoking and Obesity was -0.0059. This value shows a slightly negative correlation and implies a minor relationship between the two variables. Hence, the results mean little or no relationship exists between the variables in both samples.

4. CONCLUSION

This research work investigated if there is a significant difference between the bad health habits in Africa and Europe using the Hotelling T-square method to analyze the data. The result obtained from the analysis shows that the bad health habits of these two continents are very different from each other. The reason for this is that the test hypothesis used was one of equality between two population mean vectors, and they were little or no correlation between the variables in the above table. However, using the comparison of rate method, the authors figured out that Europe was the continent that was more affected by these bad health habits.

The test statistics of the Hotelling T-square test indicate a significant difference between the two continents in terms of bad health habits in women. The authors believe the results make sense since the health habits of women in the two continents differ based on differences in culture, religious beliefs, and climate. Women in Africa are also historically not as independent or financially stable compared to European women. The paper recommends that countries, especially in Europe, publicize more on the danger of bad health habits and the damage they could cause to the human body.

REFERENCES

[**1**] T. W. Anderson, " Harold Hotelling's Research in Statistics.," *The American Statistician,* vol. 14, no. 3, pp. 17-21, 1960.

[2] R. A. a. J. J. E. .. Bradley, "Sequential X2 and T2-tests.," *The Annals of Mathematical Statistics.* , vol. 32, no. 4, pp. 1063-1077, 1961.

[3] J. .. a. E. .. Krzanowski, "Principles of multivariate analysis.," *Journal of the Royal Statistical Society Series A (Statistics in Society).* , vol. 152, no. 2, pp. 264-265, 1989.

[4] U. P. N. E. C. B. E. O. a. I. C. Ogoke, " A Comparative Study of Foot Measurements Using Receiver Operating Characteristics (ROC) Approach.," *Scientia Africana Journal of Pure and Applied Sciences.* , vol. 12, no. 1, pp. 76-88, 2013.

[5] J. Shabbir, "Understanding statistics: An introduction for the social sciences.," *Journal of Applied Statistics,* vol. 25, no. 5, p. 716, 1998.

[6] Ogoke U.P., and Nduka E.C, "Methods of Analyzing Missing Values in a Regression Model.," *Indian Journal of Science and Technology*. India. Vol. 5, no. 2,pp.1831-1833, 2012.

[7] Dariush D. F., "Impact of lifestyle on Health.," *Iranian Journal of Public Health,* vol. 44, no. 11, pp. 1442-1444, 2015.

[8] Lash M.M. and Armstrong A., "Impact of Obesity on Women's Health.," *Fertil Steril.*,vol. 91, no. 5, pp. 1712-6, *doi:10.1016/j.fertnstert.2008.02.141,PMID:18410940,* 2009.

[9] Kulie T., Slattengren A., Redmer J., Counts H., Eglash A., Schrager S., " Obesity and Women's health: an evidence-based review.,"*J Am Board Fam Med.* Vol. 24. no. 1, pp. 75-85, *doi: 10.3122/jabfm.2011.01.100076.PMID: 21209347,* 2011.