Original Research Article

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

ABSTRACT

This research work is focused 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 also 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 carried out to determine the relationship between the variables and the results showed that the relationship between the variables was little. Further methods were adopted using comparison of Rate between the means and total of the same variables to determine which continent was more affected by the bad health habits and it was figured 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 the same time (Krzanowski & Everitt, 1989). Almost all data collation processes result in multivariate data. Most multivariate data involves 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 important 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 differences and adjust outcome measures to control for 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 considered, it's termed univariate analysis and the following terms are used: Chi-square, T-test, Z-test, etc., however when two or more variables are involved, it's termed multivariate analysis, and the Hotelling T-square statistic can be used to test for a significant difference.

The test statistic T^2 is called "The Hotelling T-square" in honor of Harold Hotelling's pioneer in multivariate analysis (1931). T^2 is used when the members of response variables are two or more, although it can be used when there is only one response variable (Anderson, 1960).

Health habit from the medical dictionary is defined as a behavior that is 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 that of a woman. Good food provides vitamins, minerals, and nutrients that are important for human growth, well-being, and development.

Healthy habits are very important in human life because it helps to prevent certain bad health conditions such as heart diseases, stroke, cancer, high blood pressure, etc. This is especially important in postmenopausal women. Men are more likely to abuse smoking and alcohol, thereby becoming dependent on them. However, the impact of chronic alcohol use and smoking are greater on women than men due to their fragile nature. These complications include heart disease, breast cancer, death, etc. The hostile effects of obesity on women's health are extremely unbearable and inarguable.
As a result of this, an analysis is being carried out to determine the difference between these two continents to help curb the effects of these bad health habits, especially in women of childbearing age.

2. METHOD AND METHODOLOGY

The data used in this research as seen in the online appendix was obtained from World Bank (*www.canceratlas.cancer.org).*

The data to be analyzed is from the bad health habits between countries in these two continents, Africa and Europe. Africa shall be referred to as Sample 1 and Europe shall be referred to as Sample 2. 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:** The missing values in this data were handled using the method of overall mean imputation. 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.

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

b) Variance – Covariance Matrix (Covariance Matrix): 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,

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

c) Correlation Matrix: The correlation coefficient matrix of a covariance matrix is written as

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

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.*

d) **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-square statistic is

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

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

e) 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

f) **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:

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

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 matrix** is $S_1 = \begin{matrix} x_1 \\ x_2 \\ x_3 \end{matrix} \begin{matrix} \begin{pmatrix} 1.0901 & 0.3321 & -1.8055 \\ 0.3321 & 3.1666 & 1.7029 \\ -1.8055 & 1.7029 & 79.1563 \end{pmatrix} \end{matrix}$

$\qquad\qquad\qquad\quad x_1 \qquad x_2 \qquad x_3$

The **Correlation matrix** 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 $as \; \bar{X}_2 = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \bar{x}_3 \end{pmatrix} = \begin{pmatrix} 3.3581 \\ 15.1605 \\ 22.9395 \end{pmatrix}$

$\qquad\qquad\qquad\quad \bar{x}_1 \qquad \bar{x}_2 \qquad \bar{x}_3$

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

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

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

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

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

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

Since the $T^2_{cal} > T^2_{tab,}$ 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. Further investigation was made to determine the exact continent with more of the bad health habits using the method of rating the means and total sums of each same variables between the two samples by dividing the means and total of each same variable by the lowest mean value or total value, presented in table 1. It was seen using their mean values that the cancer rate for Africa was 1 while the cancer rate for Europe was 2, 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 smoking rate for Europe was 1, which showed that the continent with the higher rate was Europe. Also, the total of each variable was used whereby the cancer rate for Africa was 1 while that of Europe was 2, the smoking rate was 1 while that of Europe was 6 and also the obesity rate for Africa was 1 while that of Europe was 1 which 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.

In table 2, for sample one, the correlation between Cancer and Smoking was 0.1787, which shows that it is a positive slight correlation, and it implies that there is little relationship between the two variables. Also, the correlation between Cancer and Obesity was -0.1490, which shows that it's negatively slight correlation and it implies that there is little relationship between the two variables and then, the correlation between Smoking and Obesity was 0.1076, which also shows that it's positive slight correlation and it implies that there is little relationship between the two variables.

Furthermore, for sample two, the correlation between Cancer and Smoking was 0.2710, which shows that it's a positive fair correlation and it implies that there is little relationship between the

two variables. Also, the correlation between Cancer and Obesity was -0.1944, which shows that it's negatively slight correlation and it implies that there is little relationship between the two variables and then, the correlation between Smoking and Obesity was -0.0059, which also shows that it's negative slight correlation and it implies that there is little relationship between the two variables. Hence, the results imply that there is little or no relationship between the variables in both samples.

4. CONCLUSION

This research work investigated if there is a significant difference between the bad health habits in these two continents with the use of the Hotelling T-square method, which is applied in analyzing the data. The result gotten from the analysis shows that the bad health habits between these two continents are very different from each other since the test hypothesis used here was that of equality between two population means vectors and they were little or no correlation between the variables in the above table. It implies that both the results of the correlational analysis and the Hotelling T^2 are the same, indicating no significant association or difference. However, using the comparison of rate method the authors were able to figure out that Europe was the continent that was more affected by bad health habits.

The test statistics of the Hotelling T-square test indicate that there is a significant difference between the two continents in their bad health habits in women. These countries should publicize more on the danger of bad health habits and the damage it could cause to the human body especially the women in Europe.

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.