*IOSR Journal of Mathematics (IOSR-JM)*
*e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 12, Issue 6 Ver. VI (Nov. - Dec.2016), PP 104-108*
*www.iosrjournals.org*

# Application of Data Depth on Kruskal-Wallis Test Statistic

Okeke, Evelyn Nkiruka, and Okeke, Joseph Uchenna
*Department of Mathematics and Statistics, Federal University Wukari*

***Abstract:*** *Multivariate data depth functions are designed to provide a center- outward ordering (and thus a ranking) of points in* $R^p$ *. Three multivariate depth functions are studied in this article and used to test the equality of mean vectors. The depths of three different datasets were found and used as rank in Kruskal-Wallis non-parametric test to test the hypothesis of equality of means. The three data depth functions were compared based on the results of our analysis and we found Mahalanobis data depth to be the best among the three.*

***Keywords:*** *Spatial depth, Projection depth, Mahalanobis depth, Kruskal-Wallis test, Metric space, Norn and Mahalanobis distance*

## I. Introduction

Hotelling's t-squared test is the multivariate extension of Student's t-test used in multivariate analysis to test for the differences between the mean vectors of two different populations. Here we are interested in testing the hypothesis that the population mean vectors are equal against the general alternative that these mean vectors are not equal. This test is carried under the assumptions that the populations involved must be normally distributed, the samples will be independently drawn and the population covariance matrices will be equal. Most often, especially in real life situation it is not easy to come in contact with data that meet up with all these conditions and so the need for robust methods arises. Robust statistics develops methods that are less influenced by abnormal observations, often at the cost of higher computational complexity. Many robust methods, especially those based on ranks, are closely related to geometric or combinatorial problems. An often used criterion to judge the robustness of an estimator is its breakdown point. The breakdown point is the smallest fraction of data points that we need to replace in order to move the estimator of the contaminated data set arbitrarily far away.

In multivariate analysis some of the robust methods are discussed under non-parametric or semi-parametric methods. To test the equality of two or more multivariate populations under non-normality we wish to resort to the use of data depth. Multivariate depth statistics are particularly suited to analyze non-Gaussian or, more general, non-elliptical distribution in $R^p$. The notion of depth has been extended from data clouds, that is, empirical distributions, to general probability distribution on $R^p$, thus allowing for laws of large numbers and consistency results. It has also been extended from p-variate data to data in functional spaces (Mosler 2012) and also from finite to infinite dimensional spaces (see Chakraborty and Chaudhuri 2014). Associated with a given distribution $F$ on $R^p$, a depth function is designed to provide a $F$ -based center- outward ordering (and thus a ranking) of points $x$ in $R^p$. The depth of a data point is reversely related to its outlyingness, and the depth trimmed regions can be seen as multivariate set-valued quantiles. Higher depth corresponds to "centrality", low depth to "outlyingness". The "center" consists of the point(s) that globally maximize depth. Multimodality features of $F$ are ignored in depth function.

The depth function will be used in this article to reduce the mult-dimensional data set to lower dimension where we can apply a classical univariate non-parametric method and get a useful result. The idea of using a depth function to generate contours in higher dimensional space as analogues to rank and order statistics in univariate inference was introduced by Turkey (1975). Since our interest here is in testing the equality of two or more multivariate populations we intend to apply depth function on the data sets and then use Kruskal-Wallis test statistic to do the analysis. It is good to recall that Kruskal-Wallis test sometimes also called the "One way ANOVA on ranks" is a useful tool for testing the equality of k independent groups in non-parametric univariate statistics.

## II. Multivariate Depth

In this section different multivariate data depth functions and their properties will be studied in details. The depths functions are as follows:

**Spatial Depth**

Spatial depth (SPD) of a point $x$ can be defined as simply the difference between one and the norm of the mean of distances of $x$ from all the other points of X when divided by the norm of all such deviations.

Spatial depth or $L_2$ of a point $x \in R^p$ where $R^p = \left\{ x = (x_1, ..., x_p) : x_i \in R, i = 1, ..., p \right\}$ with respect to the probability distribution of a random vector $\mathbf{x} \in R^p$ is defined as

$$D_s(x) = 1 - \left\| E\left\{ \frac{x - \mathbf{x}}{\|x - \mathbf{x}\|} \right\} \right\|$$

$$= 1 - \left\| \frac{1}{n} \sum_{i=1}^{n} \frac{x - \mathbf{x}}{\|x - \mathbf{x}\|} \right\| \qquad 1$$

This depth has been widely used for various statistical procedures including MANOVA, classification, clustering (see, e.g Okeke et. al 2015,Ghosh and Chaudhuri 2005, Jornsten 2004), construction of depth-based central and outlying regions and depth-based trimming (see Serfling 2006). This depth function according to Chakraborty and Chadhuri (2014) extends naturally to any Hilbert space **H** and it has many interesting properties like:

- It is invariant under the class of linear transformations, that is, If $T : \mathbf{H} \to \mathbf{H}$ where $T(\mathbf{x}) = cA\mathbf{x} + b$ for some $c > 0$, $b \in \mathbf{H}$ and an isometry A on $\mathbf{H}$.
- Spatial depth is a continuous function in $\mathbf{X}$ if the distribution of X is non-atomic.
- It has a unique maximum at the spatial median m of X and its maximum value is 1.
- If we consider the sequence $(m + n\mathbf{x})_{n \geq 0}$ for any fixed non-zero $\mathbf{X} \in \mathbf{H}$, it follows by a simple application of dominated convergence theorem that $D_s(m + nx) \to 0$ as $n \to \infty$ (see Mosler and Polyakova 2012, Liu 1990, and Zuo and Serfling 2000).
- The distribution of $D_s(x)$ is actually supported on the entire unit interval [0,1] for a large class of probability measures in a separable Hilbert space **H** including Gaussian probabilities.

One advantage of this depth is that its infinite dimensional extension does not suffer from degeneracy and thus can be conveniently used for analyzing infinite dimensional data. This depth function has attractive robustness and computational properties, however and serves as a basis for useful non-parametric multivariate descriptive measures.

**Projection Depth:**

Projection depth (PD) is the depth of datapoints defined using linear function (obtained through the projection of the original data) of a random element X. The PD of $x$ with respect to the distribution of X is defined as

$$PD(x) = \left[ 1 + \sup_{a \in \Phi} \frac{|a(x) - med(a(x))|}{MAD(a(x))} \right]^{-1} \qquad 2$$

where $a(x)$ is the projected data point, $med(y)$ is the median of the univariate random variable y, and $MAD(y) = med(|y - med(y)|)$ is the median of the absolute deviations from the median. The median of the projected dataponts were used in place of mean because mean has zero breakdown value (only one observation is needed to explode it when it is moved far away). Φ is the space of the random element X. It is worthy to note that Chakraborty and Chaudhuri (2014) showed that infinite extension of this depth has degenerate behavior in infinite dimensional space.

**Mahalanobis depths.**

Mahalanobis depth (MD) is obtained from little adjustment of Mahalanobis distance. Recall the mahalanobis distance $d^2 = (y_i - \bar{y})' S^{-1} (y_i - \bar{y})$, if $m_x$ is the vector that measures the location of X in a continuous and affine equivariant way and $C_x$ the matrix that measures the scatter of the distribution such that $C_{XA+c} = AC_X A'$ holds for any matrix A of full rank and any constant c. Then based on these parameters a simple depth fuction called the Mahalanobis depth is constructed as

$$M_D(x) = \left(1 + \|x - \mathrm{m_x}\|^2_{C_x}\right)^{-1} \qquad 3$$

$M_D(x)$ takes its unique maximum at the center $m_x$. Mahalanobis depth is continuous on $x$ and in the distribution of X. In particular, with $\mathrm{m_x} = \mathrm{E(X)}$ and $C_x = \Sigma_x$ the moment Mahalanobis distance is given as

$$M_{mD}(x) = [1 + (x - E(X))'\Sigma_x^{-1}(x - E(X))]^{-1} \qquad 4$$

The sample version is

$$MD(x_1, \dots, x_n;\ F_x) = [1 + (x - \bar{x})'S_x^{-1}(x - \bar{x})]^{-1} \qquad 5$$

where $\bar{x}$ is the mean vector and $S_x^{-1}$ is the empirical covariance matrix

## III. Kruskal –Wallis Test

Kruskal-Wallis test (sometimes also called the "One way ANOVA on ranks") was developed by Kruskal and Wallis (1952) as a useful tool for testing the equality of k independent populations. Consider k independent samples $Y_{11},...Y_{1n_1}, Y_{21},...,Y_{2n_2},...,Y_{k1},...,Y_{kn_k}$ of sizes $n_1, n_2,..., n_k$ drawn from k independent continuous (not necessarily normal) populations. We wish to test the hypothesis that these populations are identically distributed against its alternative that the populations are not identically distributed. To carry out the test we calculate the sum of ranks $R_1,..., R_k$ of the samples $Y_i$ s (that is $Y_{11},...Y_{1n_1}, Y_{21},...,Y_{2n_2},...,Y_{k1},...,Y_{kn_k}$) in an ordered arrangement of the k samples and use the Kruskal-Wallis test statistic H to carry out the test

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_i^{\,2}}{n_i} - 3(n+1) \qquad 6$$

where $n = \sum_{i=1}^{k} n_i$ is the sum of the samples sizes. The null hypothesis is rejected if $H \geq H_{1-\alpha}$ where $H_{1-\alpha}$ is $\alpha$ exact critical value of the H statistic.

## IV. Basic Mathematical Technique

In this section we are to describe some basic mathematical techniques that are used by various data depths.

- Metric space

We shall recall that from the knowledge of topology that the following notions hold:
For any finite space $p$

$$R^p = \left\{x = (x_1,..., x_p) : x_i \in R, i = 1,..., p\right\}$$

is a metric space with the metric $d(x, y)$defined by

$$d(x, y) = \left(\sum_{i=1}^{p}(x_i - y_i)^2\right)^{\frac{1}{2}},\ x, y \in R^p \qquad 7$$

If $l_\infty$ is the set of all bounded sequence $x = \{x_k\}$ of real or complex number, it is a metric space with the natural metric

$$d(x, y) = \sup_k |x_k - y_k| \qquad 8$$

Let $X = R^p$ and $d_\infty : X * X *... \in R$ be defined by

$$d_\infty\left(\bar{x}, \bar{y}\right) = \max\left\{|x_1 - y_1|, |x_2 - y_2|,..., |x_p - y_p|\right\} \qquad 9$$

$$= \max_{1\leq i\leq p}\left\{|x_i - y_i|\right\}$$

where $\bar{x} = (x_1,..., x_p)$ and $\bar{y} = (y_1,..., y_p)$ are arbitrary element of X and Y, then $d_\infty$ is a metric on $R^p$

- Norm

Given the $d$-dimensional vector

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$$

A general vector norm $|x|$or$\|x\|$ is a non-negative norm defined such that

I. $|x| > 0$ when $\neq 0$ and $|x| = 0$ if $= 0$
II. $|k| = |k|\,|x|$ for any scalar k.
III. $|x + \boldsymbol{y}| < |x| + |\boldsymbol{y}|$

The norm $p$ of a vector $\mathbf{x}$ can be written as $\|x\|^p$. This can be calculated thus

$$\|x\|^p = \sqrt[p]{x_1^p + x_2^p + \cdots + x_d^p} = \left(x_1^p + x_2^p + \cdots + x_d^p\right)^{\frac{1}{p}} \qquad 10$$

for $d$ number of variables. When $p = 2$ we have this to be

$$\|x\|^2 = \sqrt[2]{(x_1^2 + x_2^2 + \cdots + x_d^2)} \qquad 11$$

- Mahalanobis Distance

To obtain a useful distance measure in a multivariate setting, we must consider not only the variances of the variables but also their covariances or correlations. The simple Euclidean distance between y and $\bar{y}$, $(y_i - \bar{y})(y_i - \bar{y})$ is not useful because there is no adjustment for the variance or the covariance. For a statistical distance, we standardize by inserting inverse of the covariance matrix: $d^2 = (y_i - \bar{y})' S^{-1} (y_i - \bar{y})$ 12

The (squared) distance $d^2$ between two vectors were first proposed by Mahalanobis (1936) and are often referred to as Mahalanobis distances. If a random variable has a larger variance than another, it receives relatively less weight in a Mahalanobis distance. Similarly, two highly correlated variables do not contribute as much as two variables that are less correlated. In essence, then, the use of the inverse of the covariance matrix in a Mahalanobis distance (involving random variables) has the effect of (1) standardizing all variables to the same variance and (2) eliminating correlations.

## V. Demonstration Using Real Data

In this section, we try to compare the performance of three multivariate depth functions in testing for the equality of mean vectors of two independent samples using Kruskal-Wallis univariate non-parametric method. When we have p-dimensional Gaussian random vector, this study is better carried out using Hotellings $T^2$. But since this test has some constraints and can produce misleading result when some of the assumptions are not met we decided to find the depth of the datasets and then applied Kruskal-Wallis to the transformed data to get result. Then here, we try to investigate to what extend those results reflected what it should be when the data is not transformed. First we shall consider a real life observation on a p-dimensional Guassian random vector $\mathbf{x} = (x_1, \ldots, x_p)$. The data is obtained from Methods of Multivariate Analysis, second edition by Rencher (2002) or http://www.amazon.com/methods-multivariate-Analysis...Rencher/dp/0470178965 and the data contains 2 types of coating for resistance to corrosion on 15 pieces of pipe. Two pipes, one with each type of coating were buried together and left for the same length of time at 15 different locations that provided a natural pairing of the observations. Corrosion for each type of coating was measured on $p = 2$ variable (maximum depth of pit in thousandths of an inch, and number of pits.) The Hotellings $T^2$ result has it that the two coatings differ in their effect on corrosion. The depths of the datasets were obtained using three multivariate different methods we discussed above. The depths obtained were used as ranks in Kruskal-Wallis non-parametric test. The three different methods have it that the two coatings do not differ in their effect on corrosion at different level of significant.

The second dataset is from a retrospective study of 823 women admitted for the treatment of female genital cancer at the Department of Obstetrics and Gynecology, University of Ibadan, Nigeria. The data is published in the Proceeding of Nigeria Statistical Association by Folorunso et. al (2015). The cancer types they studied include Cervix, Endometrium, Ovary, Placental, Vagina, and Vulva. The data was analyzed using cox regression model and the results as presented there have that all the age groups are not significant and that cancer type (ovary) was statistically significant factors that affect patient's survival probabilities, and that the factors levels are significant at 5% α level. This data is categorical, so we first analyzed it using Chi-squared and Akaike Information criteria on categorical data and discovered that the variables of classification are associated. It was also revealed by cluster analysis as shown in Fig. 1 that there exist two groups in the dataset. The groups were used to partition the data into two classes and the depths of the dataset computed using three different methods we have. The obtained depths were analyzed with Kruskal-Wallis test statistic and the results have it that the test is not significant, though Mahalanobis depth showed the test to be significant at $p = 0.25$ .

The last dataset is available at http:www.real-statistics.com/multivariate-statistics/hotellings-t-squ… and is on tropical disease characterized by fever, low blood pressure and body aches. A pharmaceutical company who is working on a new drug to treat this type of disease wanted to determine whether the drug is effective at reducing these three symptoms. In the data a random sample of 20 people treated with the new drug and 18 with a placebo were selected for the analysis and result has it that the test in not significant. The depths of this dataset were computed and analyzed and our results from the three depth functions also revealed also that the test is not significant.

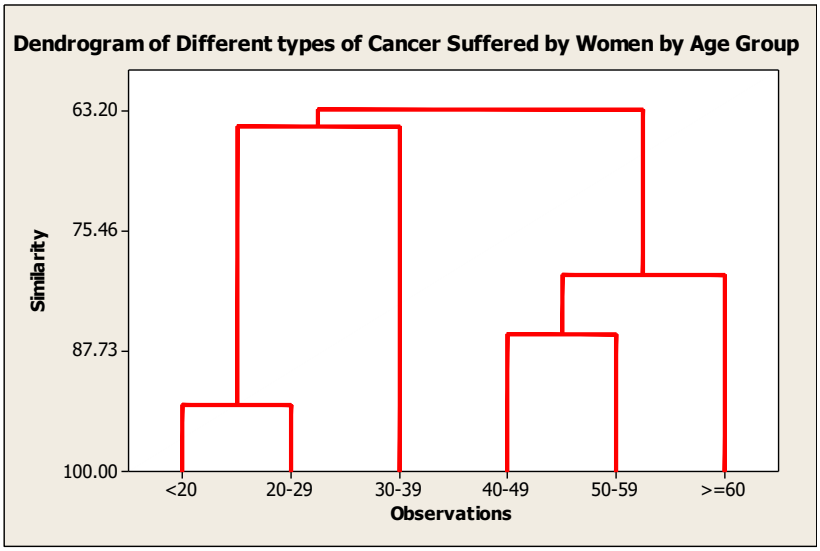


**Fig 1** Dendrogram of different types of cancer suffered by Nigerian women. The horizontal axis corresponds to different age group while vertical axis shows different similarity level.

Comparing the performance of the three different multivariate depth functions we studied using their different *p*-values, we have it from Table 1 that Mahalanobis depth function has best performance among the three, followed by Projection depth. Spatial depth performed poorly in all our datasets.

**Table1** Performance of three Different Multivariate Data Depth Functions According to their P-values

| Data set 1 | | Data set 2 (Cancer data) | | Data set 3 | |
|---|---|---|---|---|---|
| Data depth | p-value | Data depth | p-value | Data depth | p-value |
| SPD | 0.778 | SPD | 0.827 | SPD | 0.982 |
| PD | 0.724 | PD | 0.513 | PD | 0.640 |
| MD | 0.619 | MD | 0.275 | MD | 0.984 |

## VI. Comments

This results we obtained do not mean that spatial depth is worse than the other two methods. I supposed that different data sets may produce results that may be different from our own. It is also good to note that among the three methods, SPD is the most computational intensive. PD and MD can easily be done by any person that is good in using MATLAB and Microsoft office Excel. We could not cover infinite dimensional space in this study because we could not lay our hands on any of the algorithms that will handle that.

## References

[1]. Chakraborty, A. and Chaudhuri, P. (2014). On data depth in infinite dimensional spaces. *Annals of the Institute of Statistical Mathematics*, 66,303-324. DOI10.1007/s10463-013-0416-y

[2]. Folorunso, S. A., Chukwu, A. U. and Oluwasola, T. A. (2015). The cox regression model with application to predict admission lifetime of female genital cancer, Conference Proceedings of Nigerian Statistical Association Osun State Secretariat, Osogbo, 375-386.

[3]. Ghosh, A.K. and Chaudhuri, P. (2005). On maximum depth and related claaifier. *Scandinavian Journal of Statistics*, 32, 327-350.

[4]. Jornsten, R. (2004). Clustering and classificsation based on the $L_1$ data depth. *Journal of Multivariate Analysis*, 90, 67-89.

[5]. Kruskal, W.H. and Wallis, W.A. (1952).Use of rank in one criterion variance analysis, J*ournal of the American Statistical Association*, 47, 583-621.

[6]. Liu, R. Y. (1992). Data depth and multivariate rank tests. In $L_1$-statistical analysis and related methods (Neuch$\hat{a}$tel, 1992), pages 279-294. North-Holland, Amsterdam.

[7]. Mahalanobis, P. C.(1936). On the generalized distance in statistics. Proceeding of the National Institute of Science of India, pages 49-55.

[8]. Mosler, K. (2012). Depth statistics, http://arxiv.org/pdf/1207.4988

[9]. Mosler, K. and Polyakova, Y. (2012). General notions of depth for functional data. Discussion paper in Statistics and Econometrics 2/2012, University of Cologne (arXiv:1208.1981v1).

[10]. Okeke, E.N., Okeke, J.U., and Onyeagu, S.I.(2015). Multivariate rank discriminat classifier of small samples, *International Journal of Science: Basic and Applied Research (IJSBAR)*, Amman-Jordan. ISSN 2307 4531. 20(2):165-172.

[11]. Rencher, A.C. (2002). *Method of multivariate analysis*, 2nd Ed., John Wiley and Sons, Canada 280-281.

[12]. Serflinf, R. (2006). Depth function in nonparametric multivariate inference, computational geometry and application, DIMACS series in Discrete Mathematics and Theoretical Computer Science, 72, 1-16.

[13]. Tukey, J. (1975). Address to international congress of mathematics. Vancouv

[14]. Zuo, Y. and Serfling, R. (2000). General notion of statistical depth function. *Annals of Statistics*, 28(2):461-482.