An-Najah National University

Faculty of Graduate Studies

Two-Sample Multivariate Test of Homogeneity Using Weighted Nearest Neighbours

By

Areej Ali Said Barakat

Supervised by

Dr. Mohammad Najib Ass'ad

This Thesis is Submitted in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computational Mathematics, Faculty of Graduate Studies, An- Najah National University, Nablus, Palestine.

2012

Two-Sample Multivariate Test of Homogeneity Using Weighted Nearest Neighbours

By
Areej Ali Said Barakat

This Thesis was defended successfully on30/01/2012 and approved by:

Committee Members		**Signatures**
1- Dr. Mohammad N. Ass'ad	(**Supervisor**)	……...
2- Dr. Saed F. Mallak	(**External Examiner**)	…..……
3- Dr. Samir Matar	(**Internal Examiner**)	……..

DEDICATION

First of all Thanks to Allah ,

Then,

Thanks to my supervisor

Dr. Mohammad Najib Ass'ad.

To my Father

Dr. Ali Said Barakat

For their indispensable help.

To my husband, mother, children, and my family.

Thanks to every one who helped me in my work.

الإقرار

أنا الموقع أدناه مقدم الرسالة التي تحمل العنوان :

Two-Sample Multivariate Test of Homogeneity Using Weighted Nearest Neighbours

أقر بأن ما اشتملت عليه هذه الرسالة إنما هي نتاج جهدي الخاص، باستثناء ما تمت الإشارة إليه حيثما ورد، وأن هذه الرسالة ككل، أو أي جزء منها لم يقدم لنيل أية درجة أو لقب علمي أو بحثي لدى أية مؤسسة تعليمية أو بحثية أخرى .

Declaration

The work provided in this thesis , unless otherwise referenced , is the researcher's own work , and has not been submitted elsewhere for any other degree or qualification.

اسم الطالب : **Student's Name :**

التوقيع: **Signature :**

التاريخ: **Date:**

List of Contents

List of Tables

List of Figures

Two-Sample Multivariate Test of Homogeneity Using Weighted Nearest Neighbours

By

Areej Ali Said Barakat

Supervised by

Dr. Mohammad Najib Ass'ad

Abstract

In this research we proposed a two-sample test related to that proposed by Schilling and Barakat tests , by taking into account all the points and their positions according to the nearest to the nearest technique starting from the point which represents the median . Schilling did not do that ,since he just applied his test by using the first nearest neighbour , two nearest neighbours, or first three nearest neighbours and we also took the nearest to the nearest point to the median which Barakat did not do since he took all the nearest neighbours to the starting point .

That's to say in our proposed test we will give more weight to the most nearest neighbour to the starting point which is the median than the remaining nearest neighbours and so on.

We proposed a test statistic and proved that its distribution is normal.

We construct a MATLAB computer program to compute our test statistic.

Chapter One

Literature Review

1. 1:Introduction

A branch of statistics is parametric and nonparametric studies, in our work we are talking about nonparametric studies which have been studied years ago.

Before we show our work we need to take a look to the history of nearest neighbour procedures which have been applied to many problems such as nonparametric classification, nonparametric regression, two sample tests, and goodness of fit.

The study of probability stems from the analysis of certain games of chance popular in the sixteenth and seventeenth centuries. It has since found application in most branches of science and engineering and this breadth and depth of applications makes it an interesting and important subject. (Ian, F. B. ,1979, page 1).

If a coin is tossed, we observe whether a head or tail is obtained and might describe this as the experimental of tossing a coin. There are essentially two types of experiments: deterministic and random. In a deterministic experiment, because of the physical situation, the observed result is not subject to chance. In other words, if we repeat a deterministic experiment under exactly the same conditions, we expect the same result. For example, if we have a length of straight wire and a ruler, measured in millimetres, an experiment might consist of asking an individual to

measure the length of the wire. If the experiment is repeated under identical conditions, we expect the same result since experimental error should be negligible and the experiment is essentially deterministic.

In a random experiment the outcome is always subject to chance. If the experiment is repeated, the outcome may be different as there is some random phenomena or chance mechanism at work affecting the outcome. Classical examples of such experiment occur in gambling casinos were games involving dice rolling, cards, roulette, coin tossing, and so forth are to be found. A characteristic of these games is that each time they are repeated the outcome has the opportunity of being different. Indeed the participants wager money often on the basis of their intuitive notions as to how likely the various outcomes are. (Ian, F. B., 1979, page 127-128)

1. 2 : History of Statistics

The Word statistics have been derived from Latin word “Status” or the Italian word “Statista”, the meaning of these words is “Political State” or a Government. Shakespeare used a word Statist in his drama Hamlet (1602). In the past, the statistics was used by rulers. The application of statistics was very limited but rulers and kings needed information about lands, agriculture, commerce, population of their states to assess their military potential, their wealth, taxation and other aspects of government.

Gottfried Achenwall used the word statistik at a German University in 1749 which means that political science of different countries. In 1771 W.

Hooper (Englishman) used the word statistics in his translation of Elements of Universal Erudition written by Baron B. F Bieford, in his book statistics has been defined as the science that teaches us what is the political arrangement of all the modern states of the known world. There is a big gap between the old statistics and the modern statistics, but old statistics also used as a part of the present statistics.

During the 18th century the English writers have used the word statistics in their works, so statistics has developed gradually during last few centuries. A lot of work has been done in the end of the nineteenth century.

At the beginning of the 20th century, William S. Gosset developed the methods for decision making based on small set of data. During the 20th century several statistician are active in developing new methods, theories and application of statistics. Now these days the availability of electronics computers is certainly a major factor in the modern development of statistics.

Statistics is a branch of mathematics that deals with the collection, organization, and analysis of numerical data and with such problems as experiment design and decision making.

The origin of the term statistics comes from the Italian word statista (meaning “statesman”), but the real term derived from statista was statistik which was firstly used by Gottfried Achenwall (1719-1772). He was a professor at Marlborough and Gottingen. But the introduction of the word

statistics was made by E.A.W. Zimmerman in England. However, before eighteenth century people were able to record and use some data.(Douglas, D., Jeffrey, C., 2010, page 2)

The popularity of statistics had started with Sir John Clair in his work Statistical Account of Scotland which includes the period of 1791-99. There are various techniques in statistics which can be applied in every branch of public and private enterprises. But statisticians generally divide it into two main parts: Descriptive Statistics and Inferential Statistics. Shortly, in descriptive statistics there is no generalization from sample to population. We can describe any data with tables, charts or graphs so that they do not refer any generalization for other data or population. On the other hand, in inferential statistics there is a generalization from sample to population. The generalization or conclusions on any data goes far beyond that data. So the generalization may not be true and valid, and statistician should specify how likely it is to be true, because it is based on estimation somehow. Inferential statistics could be also re-called as Statistical Inference. Statistical inference can be applied also in decision theory, which is a branch of statistics. Because there is a very close relationship between the two; decisions are made under the conditions of uncertainty. So statistical inference is very effective in decision making. (Douglas, D., Jeffrey, C., 2010, page 545)

Data are collections of any number of related observations. We can collect an information about the number of students in a university in any

country. We can divide them into the different categories such as nationality, gender, age groups, and etc.. A collection of data is called data set, and a single observation in the data is called a data point. People can gather data from past records, or by observation. Again people can use data on the past to make decisions about the future. So data plays very important role in decision making.

Most of the times it is not possible to gather data for the population. So what statisticians do is to gather the data from a sample. They use this information to make inferences about the population that the sample represents. A population is a whole, where as a sample is only a fraction of the population.

The word statistics has three different meanings (sense) which are discussed below:

(1) Plural Sense **(2)** Singular Sense **(3)** Plural of the word "Statistic"

(1) Plural Sense:

In plural sense, the word statistics refer to numerical facts and figures collected in a systematic manner with a definite purpose in any field of study. In this sense, statistics are also aggregates of facts which are expressed in numerical form. For example, Statistics on industrial production, statistics or population growth of a country in different years etc.

(2) Singular Sense:

In singular sense, it refers to the science comprising methods which are used in collection, analysis, interpretation and presentation of numerical data. These methods are used to draw conclusion about the population parameter.

For Example :

If we want to have a study about the distribution of weights of students in a certain college. First of all, we will collect the information on the weights which may be obtained from the records of the college or we may collect from the students directly. The large number of weight figures will confuse the mind. In this situation we may arrange the weights in groups such as: "50 Kg to 60 Kg" "60 Kg to 70 Kg" and so on and find the number of students fall in each group. This step is called a presentation of data. We may still go further and compute the averages and some other measures which may give us complete description of the original data.

(3) Plural of Word " Statistic " :

The word statistics is used as the plural of the word "Statistic" which refers to a numerical quantity like mean, median, variance etc…, calculated from sample value.

For Example : If we select 15 student from a class of 80 students, measure their heights and find the average height. This average would be a statistic.

1. 3 :Nearest Neighbours

“ When I see a bird that walks like a duck and swims like a duck and quacks like a duck, I call that bird a duck. ”

The famous words by James Whitcomb Riley quoted above actually capture the intuitive essence of nearest neighbor classification ; if the nature of an object is unknown to us, we assume it's of the same kind as objects with similar features. (Christian Alfons, Eirik Fredäng, Per Lind 2009).

Nearest neighbour classification is a technique for dividing datasets into different classes. The categorization of a new object is determined by the labels of the most similar already existing objects. To optimize the technique for a specific dataset, a different number of nearest neighbors to evaluate can be chosen.

The technique can be improved by weighting the influence of the nearest neighbours based on their distances.

Using a distance function other than regular Euclidean distance, the method can be extended to operate on datasets with symbolic attributes.

One of the strengths of the technique is that it is intuitive to understand and easy to implement, and compared to other object classification techniques it performs well, although it can become quite computationally demanding when operating on large datasets.

When gathering data through a data mining process, it is often useful to be able to categorize the sampled entities by arranging them into predefined groups. In order to accomplish satisfactory classifications, well-chosen classification models are often used. (Christian Alfons, Eirik Fredäng, Per Lind 2009).

Classification is used to categorize an object as belonging to a certain group based on the similarity in attributes with instances of the group.

When trying to determine the classification of an object in a set of data, a simple technique commonly used is the nearest neighbour classifier method.

As with other classifiers, the data is split into a training set and a test set in nearest neighbour classification. Based on the training items, that have already been correctly classified in the training set, the algorithm predicts which group the test data belongs to the nearest neighbour classifier method.

In nearest neighbour classification, a classification model isn't built prior to the labelling of the test items. Instead, existing values are evaluated

during the classification process. This kind of approach is known as a lazy learner.(Christian Alfons, Eirik Fredäng, Per Lind, 2009).

A k-nearest neighbour (k-NN) classification algorithm was described, and its advantages and disadvantages were presented and discussed. Two methods for comparing symbolic attributes – overlapping and a value difference metric – are compared.

Other classification techniques, such as rule-based classifiers and Bayesian classifiers, provide alternative approaches to solve the same types of problems. Naïve Bayesian classifiers assume independence between different attributes, allowing classification based on attribute probability, whereas rule-based classifiers make classification conclusions based on fulfilment of predefined conditions. (Christian Alfons, Eirik Fredäng, Per Lind,2009).

Nearest neighbour (NN) methods include at least six different groups of statistical methods. All have in common the idea that some aspect of the similarity between a point and its NN can be used to make useful inferences. In some cases, the similarity is the distance between the point and its NN; in others, the appropriate similarity is based on other identifying characteristics of the points. NN methods for spatial point processes and field experiments which are commonly used in biology and environmetrics were discussed in details and very briefly discuss of NN designs for field experiments, in which each pair of treatments occurs

as neighbours equally frequently, but NN estimates of probability density functions and NN methods for discrimination or classification or NN linkage (i.e. simple linkage) in hierarchical clustering where not discussed in the same paper. Although these last three methods have been applied to environmetric data, they are much more general. (Philip M. Dixon, 2001).

Spatial point process data describe the locations of 'interesting' events and some information about each event. Some examples include locations of tree trunks, locations of bird nests, locations of pottery shards, and locations of cancer cases. Focusing on the most common case where the location is recorded in two dimensions (x, y). Similar techniques can be used for three-dimensional data (e.g. locations of galaxies in space) or one-dimensional data (e.g. nesting sites along a coastline or along a riverbank). Usually, the locations of all events in a defined area are observed (completely mapped data), but occasionally only a subset of locations is observed (sparsely sampled data). Univariate point process data include only the locations of the events; marked point process data include additional information about the event at each location. For example, the species may be were recorded for each tree, some cultural identification were recorded for each pottery shard, and nest success or nest failure were recorded for each bird nest. (Philip M. Dixon, 2001)

Location or marked location data can be used to answer many different sorts of questions. The scientific context for a question depends on the area of application, but the questions can be grouped into general categories.

One very common category of question concerns the spatial pattern of the observations. Are the locations spatially clustered ? Do they tend to be regularly distributed, or are they random(i.e. a realization of a homogeneous Poisson process)? A second common set of questions concerns the relationships between different types of events in a marked point process. Do two different species of tree tend to occur together ? Are locations of cancer cases more clustered than a random subset of a control group ? A third set of questions deals with the density (number of events per unit area). What is the average density of trees in an area ? What does a map of density look like ? Methods to answer each of these types of question were discussed in the work of nearest neighbour methods.

Theoretical treatments and Applications of NN methods for spatial point patterns can be found in many articles and books. (Philip M. Dixon, 2001).

The motion planning problem consists of finding a valid path for a robot (movable object) from a start configuration to a goal configuration without colliding with any obstacle. Probabilistic road map (PRM) methods use randomization to construct a graph (road map) of collision-free paths that attempts to capture the connectivity of the configuration space.

Random sampling strategies are the methods used to select collision-free robot configurations. Unlike uniform samplers, they also yield configurations in the narrow passages which is very important in crowded situations. The configurations generated by random sampling are to be

connected to the nearby configurations. Some road maps contain thousands of configurations, which can lead to substantial computation time for determining the nearby configurations. Therefore, an approach that efficiently finds nearest neighbours can dramatically improve the performance of path planners. The brute force method of nearest neighbour search has a complexity of $O(n^2)$. This increases the running time and hence the cost. The cost of nearest-neighbour calls is one of the bottlenecks in the performance of sampling-based motion planning algorithms. Therefore, it is crucial to develop efficient techniques for nearest-neighbour searching, so analysing the existing k-closest nearest neighbour search methods, and comparing it with the brute force method and evaluates the pros and cons of each of the existing methods were done in Nearest Neighbour Search Method. (Surbhi Chaudhry, Xiabing Xu, Dr. Nancy Amato, 2008).

1. 4 :Two Sample Tests

One sample hypothesis testing is obtained from a single population. Practical applications more frequently involve comparing the means of two or more populations.

For example:

- To compare the mean response of individuals on an experimental drug treatment to those taking a placebo.

- To compare birds living near a toxic waste site with birds living in a pristine area.

One of the most popular statistical testing procedures is the two sample t-test used for comparing the means of two populations.

There was a study of several tests for the equality of two unknown distribution. Two are based on empirical distribution functions, three others on nonparametric probability density estimates, and the last ones on differences between sample moments. They suggested controlling the size of such tests(under nonparametric assumptions) by using permutational versions of the tests jointly with the method of Monte Carlo tests properly adjusted to deal with discrete distributions. They also proposed a combined test procedure, whose level is again perfectly controlled through the Monte Carlo test technique and has better power properties than the individual tests that are combined. Finally, in a simulation experiment, they showed that the technique suggested provides perfect control of test size and that the new tests proposed can yield sizeable power improvements. (Jean-Marie DUFOUR, Abdeljelil FARHAT, 2001).

The classical tests of homogeneity, such as the two sample Kolmogorov-Smirnov test (Smirnove,1939), do not have a natural extension to comparing two multivariate populations. G. J. Sz´ekely and N. K. Bakirov have proposed a new test based on Euclidean distance between sample elements. Distance correlation is a new measure of dependence between

random vectors. Distance covariance and distance correlation are analogous to product-moment covariance and correlation, but unlike the classical definition of correlation, distance correlation is zero only if the random vectors are independent. The empirical distance dependence measures are based on certain Euclidean distances between sample elements rather than sample moments, yet have a compact representation analogous to the classical covariance and correlation. Asymptotic properties and applications in testing independence were discussed. Implementation of the test and Monte Carlo results were also presented. (Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov,2007)

This test can be applied to test homogeneity of any two multivariate populations with finite second moments, and the test is rotation invariant and consistent. Discussing the theoretical background of the proposed test, and showing that a practical implementation is possible via nonparametric bootstrap for the composite hypothesis of equal distributions when both distributions are unspecified. Empirical results in the univariate case suggest that this test compares favourably with existing tests. Also presenting empirical results for multivariate distributions, all of this was made in A Test of Homogeneity For Two Multivariate Populations. (Maria L. Rizzo, 2002).

Fortunato Pesarin and Luigi Salmaso provided solutions for univariate and multivariate testing problems with ordered categorical variables by working within the nonparametric combination of dependent permutation

tests. Two applications and Monte Carlo simulations for power comparisons of NPC solutions to most competitors from the literature were shown. (Fortunato Pesarin and Luigi Salmaso, 2006).

A new class of simple tests is proposed for the general multivariate two-sample problem based on the (possibly weighted) proportion of all k nearest neighbour comparisons in which observations and their neighbours belong to the same sample. Large values of the test statistics give evidence against the hypothesis H_0 of equality of the two underlying distributions. Asymptotic null distributions are explicitly determined and shown to involve certain nearest neighbour interaction probabilities. Simple infinite-dimensional approximations are supplied. The unweighted version yields a distribution-free test that is consistent against all alternatives; optimally weighted statistics are also obtained and asymptotic efficiencies are calculated. Each of the tests considered is easily adapted to a permutation procedure that conditions on the pooled sample. Power performance for finite sample sizes is assessed in simulations. (Schilling, M. F., 1986) This was the study of Schilling which we are depending on it in our work.

Given independent multivariate random samples $X_1, X_2,....$, and $Y_1, Y_2,.....$, from distributions F and G, a test is desired for H_0: F = G against general alternatives. Consider the $k \bullet (n_1+n_2)$ possible ways of choosing one observation from the combined samples and then one of its k nearest

neighbours, and let S_k be the proportion of these choices in which the point and neighbour are in the same sample.

SCHILING proposed S_k as a test statistic, but did not indicate how to determine k. (Schilling,M.F,1986).

BARAKAT, QUADE, and SALAMA proposed a test statistic, which is equivalent to a sum of N Wilkoxon rank sums (Barakat, A. S., Salama, I.A, and Quade, D., 1996). The limiting distribution of the test has not been found yet. They suggest as a test statistic T_m = Sum h (m, j) ، where h (m, j) = I{ j^{th} nearest neighbour of the median m is a y}. The limiting distribution of T_m is normal. A simulation with multivariate normal data suggests that there test is generally more powerful than Schilling's test using k = 1, 2 or 3. (Barakat, A. S., 2003).

A new classification method for enhancing the performance of K-Nearest Neighbour is proposed which uses robust neighbours in training data. This new classification method is called Modified K-Nearest Neighbour, MKNN. Inspired the traditional KNN algorithm, the main idea is classifying the test samples according to their neighbour tags.

This method is a kind of weighted KNN so that these weights are determined using a different procedure. The procedure computes the fraction of the same labelled neighbours to the total number of neighbours. The proposed method is evaluated on five different data sets. Experiments show the excellent improvement in accuracy in comparison with KNN

method. (Hamid Parvin, Hosein Alizadeh and Behrouz Minaei-Bidgoli, 2008).

1. 5 : Goodness of Fit

In a goodness-of-fit problem, the statistician wishes to know if a sample of n random variables has a certain specified distribution function. Weiss considered a multivariate goodness-of-fit test which can be formed by constructing a d-dimensional sphere with center at X_i for each point X_i, i = 1, 2,. .., n, where X_1, X_2,. .., X_n is a random sample in R^d with unknown density f(x). The hypothesis to be tested is that f(x) = g (x), where g(x) is a given continuous function. The volume of each sphere is taken to be l/[ng(x)]. The test compares the proportion of spheres containing exactly one point of the (n-l) points $X_{l,}$ X_2,. .., $X_{i-\,1,}$ X_{i+1},...,X_n to the proportion e^{-1} which is expected under the null hypothesis. (Weiss L., 1958).

In the goodness of fit testing problem one is given

$$\left\{x_i\right\}_1^N$$

each of which is a data set of N measured observations presumed to be randomly drawn independently from some probability distribution with density p(x). The goal is to test the hypothesis that p(x) = p_0 (x), where p_0 (x) is some specified reference probability density. Ideally, the test should have power against all alternatives. That is as the sample size N becomes

arbitrarily large, $N \to \infty$, the test will reject the hypothesis for all distributions $p \neq p_0$ at any non zero significance α level.

A related problem is two-sample testing. Here one has two data sets: $\{x_i\}_1^N$ drawn from p(x) , and $\{z_i\}_1^M$ drawn from q(z). The goal is to test the hypothesis that p = q, again with power against all alternatives; as $N \to \infty$ and $M \to \infty$ the test will always reject when $p \neq q$. Two sample testing can be used to do goodness-of-fit testing. A random sample $\{z_i\}_1^M$ is drawn from the reference distribution $q = p_0$ and then a two sample test is performed on $\{x_i\}_1^N$ and $\{z_i\}_1^M$.

In univariate (one – dimensional) problems each observation x_i (and z_i) consists of only a single measurement. In this case there are a wide variety of useful and powerful goodness of fit and two – sample testing procedures. Some of these can be extended to two or perhaps three dimensions if the sample size is large enough. However, when each observation consists of many measured attributes $x_i = \{ x_{i1}, x_{i2}, \ldots, x_{in} \}$ (and $z_i = \{ z_{i1}, z_{i2}, \ldots, z_{in} \}$), for large n, these tests rapidly loose power because all finite samples are spares in high dimensional settings owing to the " curse – of – dimensionality ". (Jerome H. Friedman 2003).

A new approach to constructing nonparametric tests for the general two-sample problem were proposed. This approach not only generates traditional tests (including the two-sample Kolmogorov–Smirnov, Cramér–von Mises, and Anderson–Darling tests), but also produces new powerful

tests based on the likelihood ratio. Although conventional two-sample tests are sensitive to the difference in location, most of them lack power to detect changes in scale and shape. The new tests are location-, scale-, and shape-sensitive, so they are robust against variation in distribution. (Jin Zhang, 2006).

Chapter Two

Introduction and Definitions

2. 1 : Introduction

Interest in statistical procedures which depends on "nearest neighbour" has grown in recent years, and high speed computers increased the popularity of these procedures, and made the applications of non-parametric methods based on "nearest neighbours " techniques, because of their new theoretical developments.

Multivariate two-sample test is one of the procedures which depends on "nearest neighbours' was well studied by so many scientists. One of them is Schilling at 1986, he proposed a multivariate two sample test with a fixed number of nearest neighbour, k. In his research he did not take into account the position of the nearest neighbour. (Schilling, M. F.,1986).

Barakat also proposed a multivariate two-sample test by using the median as a starting point, and taking into account the position of the nearest neighbours to the median.(Barakat, A. S., 2003).

In this research we have proposed a two-sample test related to that proposed by Schilling and Barakat tests, by taking into account the position of the nearest neighbour to the starting point which Schilling did not do, and by taking the nearest to the nearest point using the median as a starting point.

That is to say in our proposed test we will give more weight to the most nearest neighbour to the starting point which is the median than the remaining nearest neighbours and so on for other points.

" Nearest neighbour methods " include at least six different groups of statistical methods. All have in common the idea that some aspect of the similarity between a point and its nearest neighbour can be used to make useful inference. In some cases, the similarity is the distance between the point and its nearest neighbour (which we will use in our research), in others, the appropriate similarity is based on other identifying characteristics of the point. (Philip, M. D., 2001).

2. 2: The Arithmetic Mean:

The arithmetic mean is simple average of a data set. We can calculate the average age in a class, average monthly expenditure of students in a university, average tourist number coming to a country each year, and etc.. The arithmetic mean for population is represented by the symbol of μ and for sample is $\overline{x}$. The formulas for μ and $\overline{x}$ are provided below:

Population :

$$\mu = \frac{\sum x}{N}$$

where N represents population size

Sample :

$$\bar{X} = \frac{\sum x}{n}$$

where n represents sample size

Example :

Table (2. 1) provides the ages of students in a class. In this case, we will assume that data represents a sample derived from the whole university.

Table (2. 1): Age of Students in a Class an Example to Solve the Mean

ID	Name	Age
1	Mohammed	25
2	Ahmed	24
3	Reem	23
4	Yazan	24
5	Remas	22
6	Ali	26
7	Samar	25
8	Mahmud	27
9	Sami	26
10	Salah	28

Now, let us calculate the arithmetic mean for this ungrouped data :

$$\bar{x} = \frac{\sum x}{n}$$

$$= \frac{25+24+23+24+22+26+25+27+26+28}{10} = \frac{250}{10} = 25$$

So the arithmetic mean of the ages in the class will be;

$$\bar{x} = 25$$

But what about if the data is grouped ! In a grouped data, we do not know the separate values of each observation. So we are only able to estimate the mean. But in ungrouped data, since we know all the observations in the data, whatever mean we find from the data will be the actual number. **(** Douglas, D., Jeffrey, C., 2010)

To calculate the arithmetic mean for a grouped data we use the following formula:

$$\bar{x} = \frac{\sum (f \times x)}{n}$$

where

• $\overline{x}$ = sample mean

• Σ = summation

• f = number of observations in each class

• x = midpoint of each class

• n = sample size

Today, we get the frequency distribution by using statistical packages and computers calculate the arithmetic mean from the original data. So the arithmetic mean for grouped data would be unnecessary in this situation.

The arithmetic mean is the best known and most frequently used measure of central tendency. One of the most important uses of the arithmetic mean is that we can easily make a comparison with different data. (Douglas, D., Jeffrey, C., 2010)

2. 3 : The Median

In its simplest meaning, median divides the distribution into two equal parts. It is a single number, which represent the most central item, or middle most item in the distribution or in the data. Half of the data lie below this number, and another half of the data lie above this number.

In order to calculate the median for ungrouped data, firstly, we arrange the data in ascending or descending order. If we have an odd number of data, then the median would be the most central item in the data. Let us consider the following simple data in table (2. 2) :

Table (2. 2) : Graduated students in each year an example to calculate the median for ungrouped odd number data

Year	**1991**	**1992**	**1993**	**1994**	**1995**
No. of Students	10	15	13	14	17

Firstly, let us arrange the data in ascending order :

10, 13, 14, 15, 17

In this case, the most central item for this odd-numbered data would be 14, which is the median of this data set at the same time.

Another way of finding the median is to use the following formula :

Median is the $\left(\frac{n+1}{2}\right)^{th}$ item in the data array and n represents number of items in the data. If we apply this formula for the above data ;

Median is the $\left(\frac{5+1}{2}=3\right)^{th}$ item in the data which corresponds to 14.

However, this formula is frequently used for even-numbered data, which takes the average of the two middle items in the data.

In order to calculate the median of even-numbered data, we need to take the average of the two middle most items since we do not know the most

central item in the data set. So we should use the above formula to calculate the median. Now let us extend table (2. 3) to 1996 and try to calculate the median for the data. In this case, number of observations will be 6, (1991 – 1996). (Douglas, D., Jeffrey, C., 2010)

Table (2. 3) : Graduated students in each year example to calculate the median for ungrouped even number data

Year	**1991**	**1992**	**1993**	**1994**	**1995**	**1996**
No. of Students	10	15	13	14	17	21

Again we have to sort the data in ascending order ;

10 , 13 , 14 , 15 , 17 , 21

From the formula, median is $\left(\frac{6+1}{2}\right)=3.5$ item in the data which is

included between 14 and 15. And the average of 14 and 15 is 14.5. That is the median of this data set. So the median number of graduated students for the period of 1991- 1996 is 14.5.

For a grouped data, we have to find an estimated value for median that can fall into a class interval. Because we do not know all the observations in the

data, we are only given the frequency distribution with class intervals. The formula to calculate the median from the grouped data is given below:

$$\tilde{m} = L + \left(\frac{n + 1/2 - (F + 1)}{f} \right). \ w$$

where m̃ = the median assumed for the sample distribution.

• L = the lower limit of the class interval containing median

• F = the cumulative sum of the frequencies up to, but not including, median class

• f = the frequency of the class interval containing median

• w = the width of the class interval containing median

• n = total number of observations in the data

In case where we work with the population, m̃ would be replaced by Md and n by N. (Douglas, D., Jeffrey, C., 2010)

2. 4: Covariance:

The concept to which two random variables vary together (co – vary) can be measured by their covariance. Consider the two random variables x and y :

x_1, y_1
x_2, y_2
. .
. .
. .
x_n, y_n

For two random variables x and y having means E{ x } and E{ y }, the covariance is defined as :

$$Cov(x, y) = E\{[x - E(x)][y - E(y)]\}$$

The covariance calculation begins with pairs of x and y, takes their differences from their mean values and multiplies these differences together. For instance, if for x_1 and y_1 this product is positive, for that pair of data points the values of x and y have varied together in the same direction from there means. If the product is negative, they have varied in opposite directions. The larger the magnitude of the product, the stronger the strength of the relationship. (Classle, 2009).

The covariance is defined as the mean value of this product, calculated using each pair of data points x_i and y_i.

If the covariance is zero, then the cases in which the product was positive where offset by those in which it was negative, and there is no linear relationship between the two random variables.

Computationally, it is more efficient to use the following equivalent formula to calculate the covariance :

$$Cov(x, y) = E\{xy\} - E\{x\}E\{y\}$$

The value of the covariance is interpreted as follows :

- Positive covariance :

Indicates that higher than the average values of one variable tend to be paired with higher than average values of the other variable.

- Negative covariance :

Indicates that higher than average values of one variable tend to be paired with lower than average values of the other variable.

- Zero covariance :

If the two random variables are independent, the covariance will be zero. However, a covariance of zero does not necessarily mean that the variables are independent. A nonlinear relationship can exist that still would result in a covariance value of zero. (Classle, 2009).

Useful properties :

The variance of the sum of two random variables can be written as :

$$Var(x+y) = Var(x) + Var(y) + 2Cov(x,y)$$

When the random variables each are multiplied by constants a and b, the covariance can be written as follows :

$$Cov(ax, by) = ab\ Cov(x, y)$$

Limitations :

Because the number representing covariance depends on the units of the data, it is difficult to compare covariance among data sets having different scales. A value that might represent a strong linear relationship for one data set might represent a very weak one in another.(Classle, 2009) .

2. 5 : K-Nearest Neighbour Classification

Suppose we have a metric space, d, and suppose we have a sample x of observations ($x_1, x_2, .., x_n$) from any common population.

The nearest neighbour to an arbitrary point x in the metric is defined as any sample point x_i for which:

$$d(x, x_i) = \min_{1 \le j \le n} d(x, x_j)$$

A point x_i for which $d(x, x_i) > d(x, x_j)$ for at most (k-1), $j \neq i$, is called a k-nearest neighbour to x and we call it the k^{th} nearest neighbour to x If

$d(x, x_i) > d(x, x_j)$ for exactly (k-1), $j \neq i$, then we call it the k^{th} nearest neighbour to x.

To show a k- nearest neighbour, let us consider the task of classifying a new object (query point) among a number of known examples. This is shown in a figure below, which depicts the examples (instances) with the plus and minus signs and query point which is a bold circle. Our task is to estimate (classify) the outcome of the query point based on selected number of its nearest neighbours. In other words, we want to know whether the query point can be classified as a plus or minus sign.

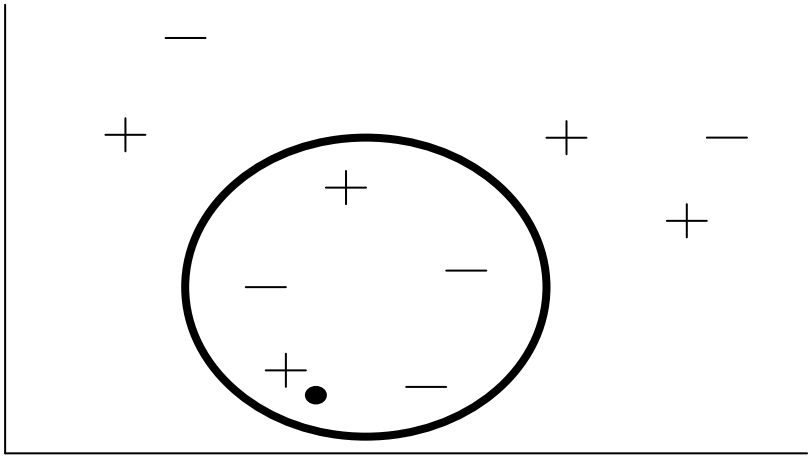


Figure (2. 1) : k – NN with signs

From figure (2.1) we can see:

1)One nearest neighbour outcome is plus.

2)Two nearest neighbour outcome is unknown.

3)Five nearest neighbour outcome is a minus.

(Hill, T., Lewicki, P., 2006).

To proceed, let us consider the outcome of KNN based on 1-nearest neighbour. It is clear that in this case KNN will predict the outcome of the query point with a plus sign (since the closest point carries a plus sign).

Now let us increase the number of nearest neighbours to two, i.e., two nearest neighbours. This time KNN will not be able to classify the outcome of the query point since the second closest point is a minus, and so both the plus and the minus signs achieve the same score (i.e., win the same number of votes). For the next step, let us increase the number of nearest neighbours to five, (five nearest neighbours). This will define a nearest neighbour region, which is indicated by the circle shown in the figure above. Since there are two plus and three minus signs, in this circle KNN will assign a minus sign to the outcome of the query point. (Hill, T., Lewicki, P., 2006)

Another example of KNN classification :

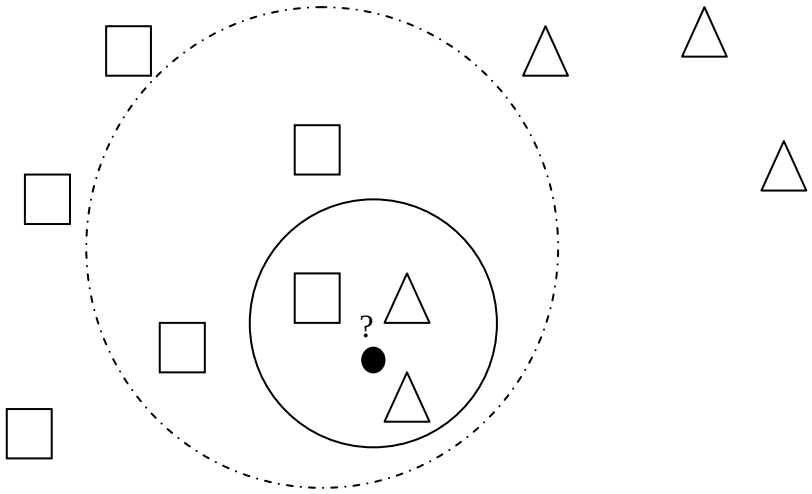


Figure (2. 2): k – NN for two classes

The test sample (bold circle) should be classified either to the first class of squares or to the second class of triangles.

If k = 3 it is classified to the second class because there are two triangle and only one square inside the inner circle. If k=5, it is classified to first class, (three squares vs. two triangles inside the outer circle).

(Hill, T., Lewicki, P., 2006).

Example of nearest neighbour Rule for k = 1 :

Two class problem : triangles and squares which we considered the triangles Θ_1 and squares as Θ_2. The bold circle represents the unknown sample x and as its nearest neighbour comes from class $\Theta_{1,}$, it is labelled as class Θ_1 .

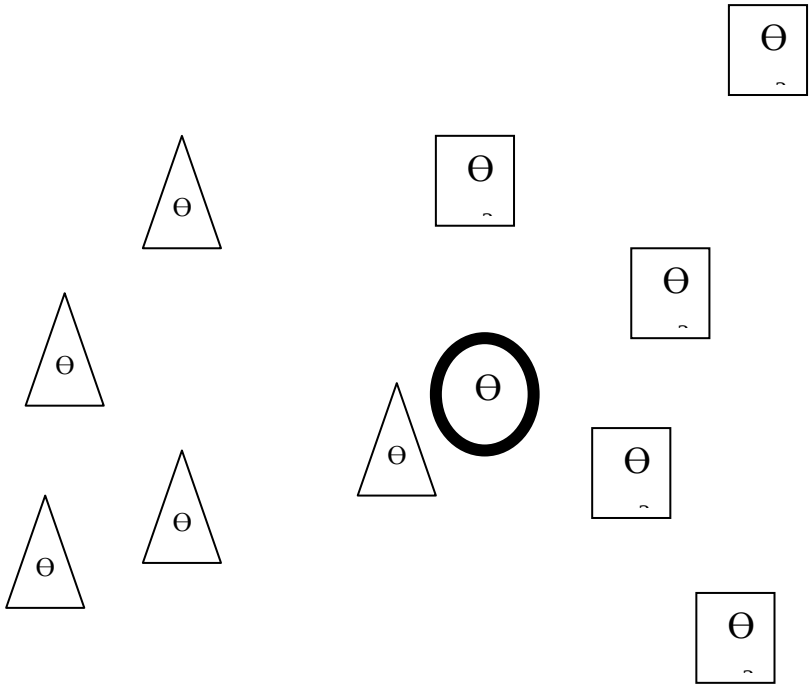


Figure (2. 3) : k- NN rule with k = 1

Example of K-nearest neighbour rule with k=3 :

There are two classes we considered the triangles Θ_1 and squares as Θ_2. The bold circle represents the unknown sample x and as two of its nearest neighbours comes from class Θ_2, and one comes from the class Θ_1 then the bold circle (unknown sample x) is labelled as class Θ_2.

(that's for the k th nearest neighbour for k = 3)

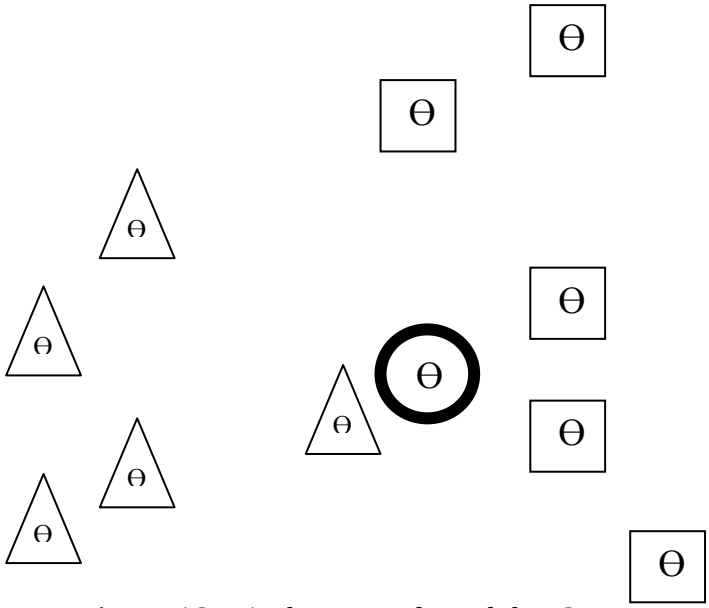


Figure (2. 4) : k – NN rule with k = 3

The number k should be :

1) large to minimize probability of misclassifying x.

2) small (with respect to number of samples) so that points are close enough to x to give an accurate estimate of the true class of x.

(Sargur N. Srihari, 2009, page 25)

K-nearest neighbours is part of supervised learning that has been used in many applications in the field of data mining, statistical pattern recognition, image processing and many others. Some successful applications are including recognition of handwriting, classifying the nearest segment for customer, etc.

The k-nearest neighbour algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-nearest neighbour is a supervised learning algorithm where the result of new instance query is classified based on majority of k-nearest neighbour category. The purpose of this algorithm is to classify a new object based on attributes and training samples. The classifiers do not use any model to fit and only based on memory.

Given a query point, we find k number of objects (or training points) closest to the query point. The classification is using majority vote among the classification of the k objects. Any ties can be neglected. K-nearest neighbour algorithm used neighbourhood classification as the prediction value of the new query instance. K is a positive integer, typically small. If k = 1, then the object is simply assigned to the class of its nearest neighbour.

K - nearest neighbour algorithm is very simple. It works based on minimum distance from the query instance to the training samples to determine the k-nearest neighbours. After we gather k-nearest neighbours, we take simple majority of these k - nearest neighbours to be the prediction of the query instance.

Euclidean distance between two instances

$$\sqrt{\sum_{r=1}^{n} (a_r(x_i) - a_r(x_j))^2}$$

such that an arbitrary instance is represented by

$$(a_1(x),a_2(x),a_3(x),...,a_n(x))$$

$a_i(x)$ denotes features. (Anjanita, D., 2010).

In pattern recognition, k – nearest neighbour is a type of instance – based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification.

The k – nearest neighbour algorithm is among the simplest of all machine learning algorithms: an object is classified by a majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours.

The same method can be used for regression, by simply assigning the property value for the object to be the average of the values of its k nearest neighbours. It can be useful to weight the contributions of the neighbours, so that nearer neighbours contribute more to the average than the more distance ones. (A common weighting scheme is to give each neighbour a weight of 1/d, where d is the distance to the neighbour. This scheme is a generalization of linear interpolation).

The neighbours are taken from a set of objects for which the correct classification (or, in the case of regression, the value of the property) is known. This can be thought of as the training set for the algorithm though

no explicit training step is required. The k – nearest neighbour algorithm is sensitive to the local structure of the data.

Nearest neighbour rules in effect compute the decision boundary in an implicit manner. It is also possible to compute the decision boundary itself explicitly. (Bremner, D., Demaine, E., Erickson, J., Iacono , J., Langerman, S., Morin, P., Toussaint, G., 2005).

The training examples in the algorithm are vectors in a multidimensional feature space, each with a class label. The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples.

In the classification phase, k is a user-defined constant, and an unlabelled vector (a query or test point) is classified by assigning the label which is most frequent among the k training samples nearest to that query point.

Usually Euclidean distance is used as the distance metric; however this is applicable to continuous variables. In cases such as text classification, another metric such as overlap metric (or Hamming distance) can be used, or any other ways for finding the distance can be used, and we are going to talk about them in the next section. Often, the classification accuracy of " k " - NN can be improved significantly if the distance metric is learned with specialized algorithms such as example Large Margin Nearest Neighbours or Neighbourhood components analysis.

A drawback to the basic " majority voting " classification is that the classes with the more frequent examples tend to dominate the prediction of the new vector, as they tend to come up in the k nearest neighbours when the neighbours are computed due to there large number, (Coomans, D., Massart, D.L.,1982). One way to overcome this problem is to weight the classification taking into account the distance from the test point to each of its k nearest neighbour.

K – nearest neighbour is a special case of a variable – bandwidth, kernel density " balloon " estimator with a uniform kernel.

(Terrell, D. G., Scott, D. W., 1992)

In Statistics, adaptive or " variable – bandwidth " kernel density estimation is a form of kernel density estimation in which the size of the kernels used in the estimate are varied depending upon either the location of the samples or the location of the test point. It is a particularly effective technique when the sample space is multi – dimensional.

Given a set of samples, $\{ \vec{X}_i \}$, if we want to estimate the density $P(\vec{x})$, at a test point, $\vec{x}$:

Then,

$$P(\vec{x}) \approx W/n$$

$$W = \sum_{i=1}^{n} w_i$$

$$w_i = K((\vec{x} - \vec{x}_i)/h)$$

Where n is the number of samples, K is the " kernel ", and h is its width. The kernel can be thought of as a simple linear filter.

Using a fixed filter width may mean that in regions of low density, all samples well fall in the tails of the filter with very low weighting, while regions of high density will find an excessive number of samples in the central region with weighting close to unity. To fix this problem, we vary the width of the kernel in different regions of the sample space.

There are two methods of doing this: balloon and point wise estimation. In a point wise estimator, the kernel width is varied depending on the location of the samples. (Terrell, D. G., Scott, D. W.,1992).

For multivariate estimators, the parameter, h, can be generalized to vary not just the size, but also the shape of the kernel.

A common method of varying the kernel width is to make it proportional to the density at the test point :

$$h = \frac{k}{\left[nP(\vec{x})\right]^{1/D}}$$

Where k is a constant and D is the number of dimensions while we can say W is a constant :

$$W = k^D (2\pi)^{D/2}$$

This produces a generalization of the k – nearest neighbour algorithm. That is, a uniform kernel function will return the k-NN technique. (Mills, Peter., 2011)

This method is particularly effective when applied to statistical classification. (Taylor, Charles,1997)

So an alternative nonparametric method which is called a k – nearest neighbour or k-NN is similar to kernel methods with a random and variable bandwidth, the idea is to base estimation on a fixed number of observations k which are closest to the desired point.

Nearest neighbour methods are more typically used for regression than for density estimation. (Bruce, E. H., 2009).

2. 6 : Distance Metric :

Since the k nearest neighbour depends on the distance between the points to find the nearest neighbour we are going to show some ways that the distance between two points in clustering input can be measured.

1) Manhattan (First Vector Norm) :

The formula for this distance between a point X = (x_1, x_2, …,x_n) and a point Y = (y_1, y_2, …,y_n) is

$$d = \sum_{i=1}^{n} \left| x_i - y_i \right|$$

Where n is the number of variables, and x_i, y_i are the values of the i^{th} variable at points X and Y respectively. (First vector norm).

The Manhattan distance function computes the distance that would be travelled to get from one data point to the other if a grid – like path is followed. (IOS., 2004).

2) Euclidean distance (Second Vector Norm):

The Euclidean distance function measures "as- the – crow – files " distance.

The formula for this distance between a point

X = (x_1, x_2, …,x_n) and another point Y = (y_1, y_2, …,y_n) is

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

The difference between Manhattan distance and Euclidean distance is showed in the following figures :

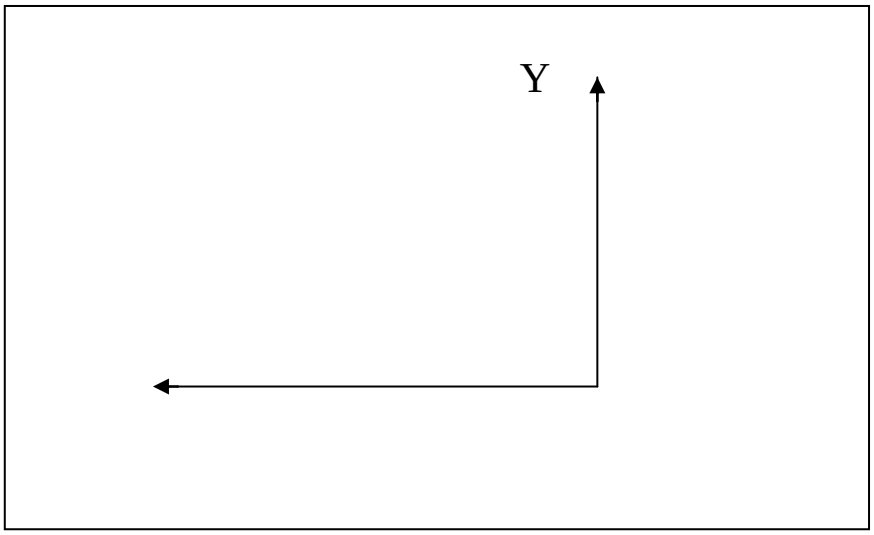


Figure (2. 5) : Manhattan Distance

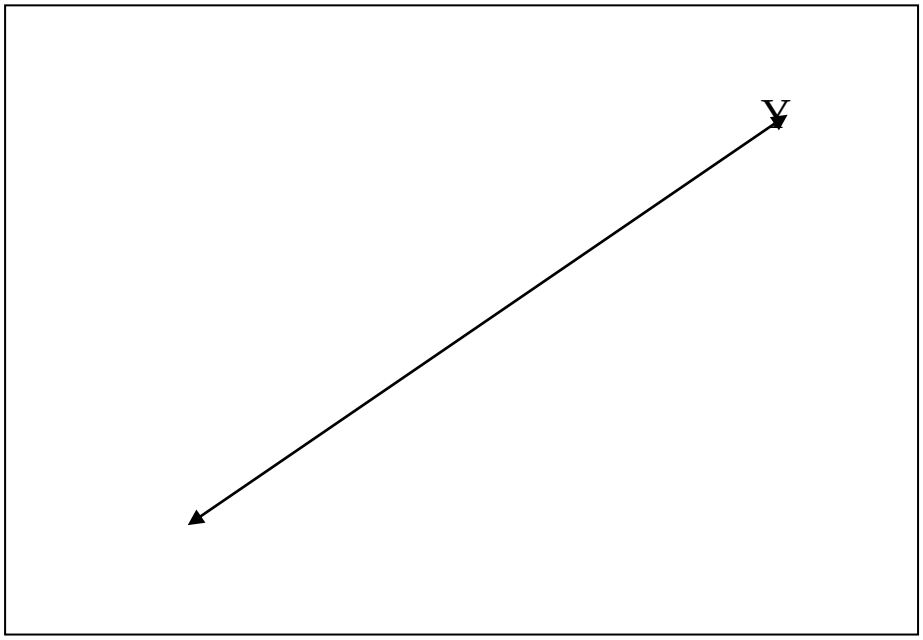


Figure (2. 6) : Euclidean Distance

3) Euclidean Squared Distance Metric :

It uses the same equation as the Euclidean distance metric, but does not take the square root. As a result, clustering with the Euclidean Squared

Distance Metric is faster than clustering with the regular Euclidean distance. (IOS., 2004)

4) Pearson Correlation :

It measures the similarity in shape between two profiles.

The formula for the Pearson Correlation distance is :

$$d = 1 - r$$

such that;

$$r = Z(X) \,.\, Z(Y)$$

is the dot product of the z – scores of the vectors x and y.

The z – score of x is constructed by subtracting from x its mean and dividing by its standard deviation. (IOS., 2004)

5) Pearson Squared :

This distance measures the similarity in shape between two profiles, but can also capture inverse relationships.

For example consider the following gene profiles :

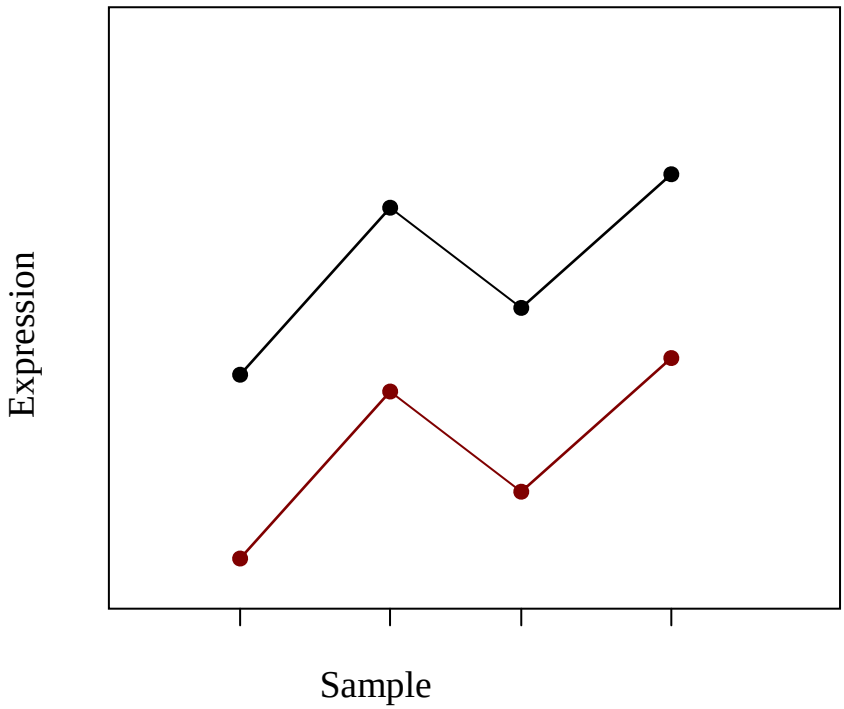


Figure (2. 7) : Pearson Squared measures the similarity in shape between two profiles.

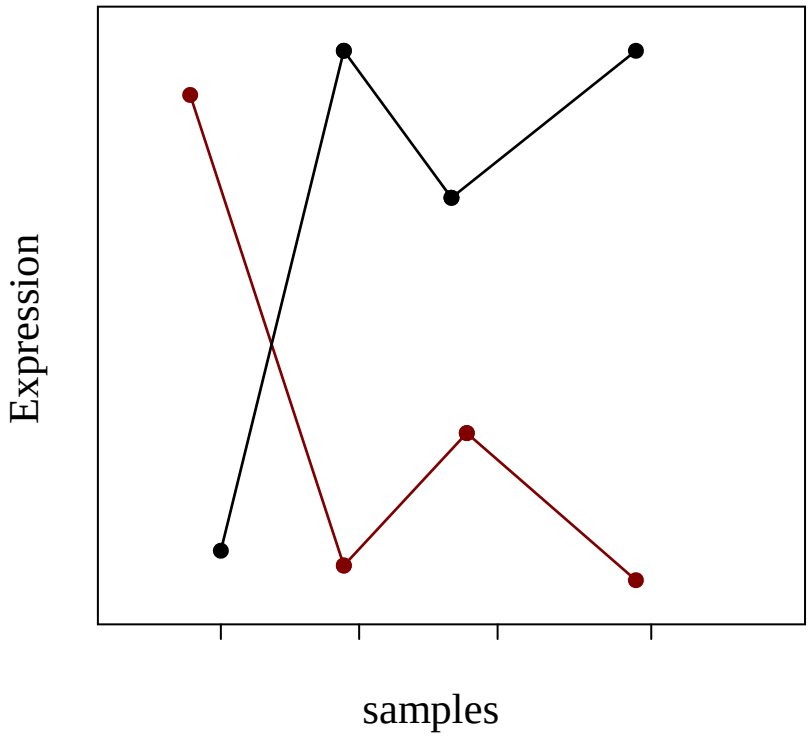


Figure (2. 8) : Pearson Squared distance for inverse relationship between two profiles.

In the figure (2. 7) , the black profile and the red profile have almost perfect Pearson Correlation despite the differences in basal expression level and scale. These genes would cluster together with either Pearson Correlation or Pearson Squared distance.

In the figure (2. 8), the black profile and the red profile are almost perfectly anti-correlated. These genes would be placed in remote clusters using Pearson Correlation, but would be put in the same cluster using Pearson Squared.

The formula for the Pearson Squared distance is :

$$d = 1 - 2r$$

where r is the Pearson Correlation defined above. (IOS., 2004)

6) Chebychev :

The Chebychev distance between two points is the maximum distance between the points $X = (x_1, x_2, \ldots, x_n)$ and

$Y = (y_1, y_2, \ldots, y_n)$ is computed using the formula :

$$\max_i |x_i - y_i|$$

where x_i and y_i are the values of the i^{th} variable at points X and Y, respectively.

The Chebychev distance may be appropriate if the difference between points is reflected more by differences in individual dimensions considered together.

Note that this distance measurement is very sensitive to outlaying measurements. (IOS., 2004)

7) Spearman Rank Correlation :

It measures the correlation between two sequences of values. The two sequences are ranked separately and the differences in rank are calculated at each position, i. the distance between sequences X = (x_1, x_2, …,x_n) and Y = (y_1, y_2 , …,y_n) is computed using the following formula :

$$1-\frac{6\sum_{i=1}^{n}(rank\,(x_i)-rank\,(y_i))^2}{n(n^2-1)}, \quad n \geq 2$$

Where x_i and y_i are the i th values of sequences x and y respectively.

The range of spearman Correlation is from -1 to 1 Spearman Correlation can detect certain linear and non-linear correlations.

However, Person Correlation may be more appropriate for finding

linear correlations. (IOS., 2004)

8) Minkowski Metric :

Another distance measure is the Minkowski metric :

$$d(x,y)=\left[\sum_{i=1}^{p}\left|x_i-y_i\right|^{m}\right]^{1/m}$$

For m = 1, d (x, y) measures the " city - block " distance between two points in p dimensions, (same as Manhatten). For m = 2, d (x, y) becomes the Euclidean distance. In general, varying m changes the weight given to large and smaller differences. (Saed , S., 2011).

Two additional popular measures of " distance " are given by the **Canberra Metric** and the **Czekanowski Coefficient**. Both of these measures are defined for nonnegative variables only.

We have;

9) Canberra Metric :

$$d(x,y)=\sum_{i=1}^{p}\frac{\left|x_i-y_i\right|}{(x_i+y_i)}$$

10) **Czekanowski Coefficient :**

$$d(x,y)=\frac{2\sum_{i=1}^{p}\min(x_i,y_i)}{\sum_{i=1}^{p}(x_i+y_i)}$$

(Richard A. J., Dean, W. W., 2007)

The Euclidean distance is the "ordinary " distance between two points the one would measure by using this formula as a distance, and the Euclidean norm or Euclidean length, or magnitude of a vector measures the length of the vector, ||.||. (Albert, C. J., Luo, 2010).

Since the norm of a vector gives a measure for the distance between an arbitrary vector and the zero vector, the distance between two vectors is defined as the norm of the difference of the vectors. (Burden, R. L., Faires, J. D., 2005, page, 421).

The common distance measure in the literatures is the Euclidean distance or squared Euclidean distance.

In our research we are going to use the Euclidean distance between two points.

An important step in most clustering is to select a distance measure, which will determine how the similarity of two elements are calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and farther away according to another.

For example, in a 2 – dimensional space, the distance between the point (x = 1, y = 0) and the origin (x = 0, y = 0) is always one according to usual norms, but the distance between the point (x= 1, y= 1) and the origin can be 2, $\sqrt{2}$, or 1, if you take respectively the 1 – norm, 2 – norm, or

infinity norm distance. (Velmrugan, T., and Santhanam ,T., 2011, page 19-30).

K - nearest neighbours regression is a simple algorithm that stores all available cases and predict the numerical target based on a similarity measures (e. g. distance functions). K – nearest neighbour has been used in statistical estimation and pattern recognition already in the beginning of 1970's as a non – parametric technique.

A simple implementation of k – NN regression is to calculate the average of the numerical target of the k– nearest neighbours.

Another approach uses an inverse distance weighted average of the k nearest neighbours is k – nearest neighbour regression which uses the same distance functions as k nearest neighbour classification.

(Saed , S., 2011).

There are distance measures which is only valid for continuous variables, like Euclidean distance, and Manhattan distance. So in the case of categorical variables one must use the Hamming distance, which is a measure of the number instance in which corresponding symbols are different in two strings of equal length. (Saed , S., 2011).

$$D_H = \sum_{i=1}^{k} \left| x_i - y_i \right|$$

$$X = Y \rightarrow D = 0$$

$$X \neq Y \rightarrow D = 1$$

Table (2. 4) : Hamming Distance

X	Y	Distance
Male	Male	0
Male	Female	1

2.7 : parametric and nonparametric tests

In statistical inferences, or hypothesis testing, the traditional tests are called parametric tests because they depend on the specification of a probability distribution, (such as the normal) except for a set of free parameters.

Parametric tests are said to depend on distribution assumptions. Nonparametric tests, on the other hand, do not require any strict distributional assumptions. Even if the data are distributed normally, and nonparametric methods are almost as powerful as parametric methods.

Many procedures perform nonparametric analysis. Some general references on nonparametric include Lehman (1975), Conover (1980), Hollander and Wolfe (1973), Hettmanspreger (1984), and Gibbons and Chakraborti (1992).

When some nonparametric tests for location and scale differences are performed, the data should consist of a random sample of observations from two different populations. The goal is either to compare the location parameters (medians) or the scale parameters of the two populations. For example, suppose your data consist of the number of days in the hospital for two groups of patients: those who received a standard surgical procedure, and those who received a new experimental surgical procedures. These patients are a random sample from the population of the patients who have received the two types of surgery. Your goal is to decide whether the median hospital stays differ for the two populations. (SAS/STAT 9.1 User's Guide, 2004)

When you test for independence, the question being answered is whether the two variables of interest are related in some way. For example, you might want to know if student scores on a standard test are related to whether students attended a public or private school. One way to think of this situation is to consider the data as a two – way table ; the hypothesis of interest is whether the rows and columns are independent. In the preceding example, the groups of students would form the two rows, and the scores would form the columns. The special case of two category response (Pass / Fail) leads to a 2×2 table ; the case of more than two categories for the response (A / B / C / D / F) leads to a 2×2 table, where C is the number of response categories. (SAS/STAT 9.1 User's Guide, 2004)

On goal in comparing k independent samples is to determine whether the location parameters (medians) of the populations are different. Another goal is to determine whether the scale parameters for the populations are different. For example, suppose new employees are randomly assigned to one of three training programs. At the end of the program, the employees receive a standard test that gives a rating score of there job ability. The goal of analysis is to compare the median scores for the three groups and decide whether the differences are real or due to chance alone. (SAS/STAT 9.1 User's Guide, 2004).

A central topic of modern statistics is statistical inference. Statistical inference is concerned with two types of problems: estimation of population parameters and tests of hypothesis.

Webster tells us that the verb " to infer " means " to derive as a consequence, conclusion, or probability ". When we see a woman who wears no ring on the third finger of her left hand, we may infer that she is unmarried. (Sidney Siegal, 1956,page 1)

In Statistical inference, we are concerned with how to draw a conclusion about a large number of events on the basis of observations of a portion of them. Statistics provides tools which formalize and standardize our procedures for drawing conclusions. For example, we might wish to determine which of three varieties of tomato sauce is most popular with Americans. Informally, we might gather information on this question by

stationing ourselves near to the tomato sauce counter at a grocery store and counting how many cans of each variety are purchased in the course of a day. Almost certainly the number of purchases of the three varieties will be unequal. But can we infer that the one most frequently chosen on that day in that store by that day is customers in really the most popular among Americans? whether we can make such an inference must depend on the margin of popularity held by the most frequently chosen brand, on the representativeness of the grocery store, and also on the representativeness of the group of purchases whom we observed. (Sidney Siegal, 1956,page 2)

A parametric statistical test is a test whose model specifies certain conditions about the parameters of the population from which the research sample was drawn. Since these conditions are not ordinarily tested, they are assumed to hold the meaningfulness of the results of a parametric test depends on the validity of these assumptions. Parametric tests require that scores under analysis result from measurement in the strength of at least an interval scale. (Sidney Siegal, 1956 page 30).

A nonparametric statistical test is a test whose model does not specify conditions about the parameters of the population from which the sample was drawn. Certain assumptions are associated with most nonparametric statistical tests, i.e., that the observations are independent and that the variables under study has underlying continuity, but these assumptions are fewer and much weaker than those associated with parametric tests. Moreover, nonparametric tests do not require measurements so strong as

the required for the parametric tests, most nonparametric tests apply to data in an ordinal scale, and some apply also to data in a nominal scale. Because the power of any nonparametric test may be increased by simply increasing the size of N, and because of behavioural scientists rarely achieve the sort of measurements which permits the meaningful use of parametric tests, nonparametric statistical tests deserve an increasingly prominent role in research in the behavioural sciences.

(Sidney Siegal, 1956,page 31).

2.8 : Advantages and Disadvantages of Nonparametric Statistical Tests

Advantages of Nonparametric Statistical Tests :

1) Probability statements obtained from most nonparametric statistical tests are exact probabilities (except in the case of large samples, where excellent approximations are available), regardless of the shape of the population distribution from which the random sample was drawn. The accuracy of the probability statement does not depend on the shape of the population, although some nonparametric tests may assume identity of shape of two or more population distributions, and some others assume symmetrical population distributions. In certain cases, the nonparametric tests do assume that the underlying distribution is continuous, an assumption which they share with parametric tests.

2) If the sample sizes as small as N = 6 are used, there is no alternative to using a nonparametric statistical tests unless the nature of the population distribution is known exactly.

3) There are suitable nonparametric statistical tests for treating samples made up of observations from several different populations. None of the parametric tests can handle such data without requiring us to make seemingly unrealistic assumptions.

4) Nonparametric statistical test are available to treat data which are inherently in ranks as well as data whose seemingly numerical scores have the strength of ranks. That is the researcher may only be able to say of his subjects that one has more or less of the characteristic than another, without being able to say how much more or less. For example, in studying such a variable as anxiety, we may be able to state that subject A is more anxious than subject B without knowing at all exactly how much more anxious A is. If data are inherently in ranks, or even if they can only be categorized as plus or minus (more or less, better or worse), they can be treated by nonparametric methods, where as they cannot be treated by parametric methods unless precarious and perhaps an realistic assumptions are made a bout the underlying distributions. (Sidney Siegal, 1956, page 32).

5) Nonparametric methods are available to treat data which are simply classificatory. i.e., are measured in a nominal scale. No parametric technique applies to such data.

6) Nonparametric statistical tests are typically much easier to learn and to apply than are the parametric tests.

7) It will be remembered that if a nonparametric statistical test has power – efficiency of, say, 90 percent, this means that where all the conditions of the parametric test are satisfied the appropriate parametric test would be just as effective with a sample which is 10 percent smaller

than that used in the nonparametric analysis. (Sidney Siegal, 1956,

page 33).

Disadvantages of Nonparametric Statistical Tests :

If all the assumptions of the parametric statistical model are in fact met in the data, and if the measurement is of the required strength, then the nonparametric statistical tests are wasteful of data. The degree of wastefulness is expressed by the power – efficiency of the nonparametric test.

Chapter Three

A Multivariate Test for Two Sample Based on Weighted Nearest Neighbours

3. 1: Introduction

Scientific inquiry is an iterative learning process. Objectives pertaining to the explanation of a social or physical phenomenon must be specified and then tested by gathering and analyzing data. In turn, an analysis of the data gathered by experimentation or observation will usually suggest a modified explanation of the phenomenon. Through out this iterative learning process, variables are often added or deleted from the study. Thus, the complexities of most phenomena require an investigator to collect observations on many different variables. When the data include simultaneous measurements on many variables, this body of methodology is called multivariate analysis. (Richard A. J., Dean W. W., 2007).

Virtually in every area activity new procedures are invented and existing techniques revised. Advances occurs whenever a new technique proves to be better than the old one. So in real life most of times we need to compare between two samples.

A substantial number of nonparametric methods based on nearest neighbours have been developed in recent years of various multivariate situations. The popularity of these procedures has increased because of new theoretical developments, the expanding capabilities of modern high-speed computers and efficient algorithms for nearest neighbour calculations, which mitigate to a great extent of the computational obstacles involved. Classification, density, estimation and regression are areas that have

received particular attention; more recently, distribution –free tests for multivariate goodnees-of-fit based on nearest neighbours have been developed.

The need to understand the relationships between many variables makes multivariate analysis an inherently difficult subject. Often, thehuman mind is overwhelmed by the sheer bulk of the data. Additionally, more mathematics is required to derive multivariate statistical techniques for making inferences than in the univariate setting.(Richard A. J., Dean, W. W., 2007).

3. 2 : Test Statistic

Let us have two independent random samples in R^d, and suppose the first sample $X = (x_1, x_2, \ldots, x_{n1})$ from unknown distribution F(x), and the second sample $Y = (y_1, y_2, \ldots y_{n2})$ from unknown distribution G(x), with corresponding continuous densities f(x) and g(x), respectively. The two-sample problem treated here is to test the hypothesis

$$H_0: F(x) = G(x)$$

$$H_1: F(x) \neq G(x)$$

Label the pooled (combined) sample as $Z = (Z_1, Z_2, \ldots, Z_N)$ such that:

$$Z_i = \begin{cases} x_i, & i = 1,2,...,n_1 \\ y_{i-n_1}, & i = n_1+1, n_1+2,...,N \end{cases}$$

Such that, $N = n_1 + n_2$

Let the distance metric $\|.\|$ be the Euclidean norm between any two points, and let us define the k^{th} nearest neighbour to a point Z_i which will be Z_j satisfying $\| Z_{j'} - Z_i \| < \| Z_j - Z_i \|$ for exactly (k – 1) values of j' ($1 \leq j \leq N$, $j' \neq i$, j).

Since the ties occur with probability zero so we will ignore them.

Define the indicator function, h (m, k) =

$$\begin{cases} 0, & \text{if the } k^{th} \text{ nearest neighbour is from sample } Y. \\ 1, & \text{if the } k^{th} \text{ nearest neighbour is from sample } X. \end{cases}$$

Where m is the median of the combined sample.

We propose the following test statistic :

$$T_{mk} = \sum_{j=1}^{k} h(m, j)$$

We will use the statistic;

$$T_m = \sum_k T_{mk}$$

To show how the test statistic T_m can be computed see this example.

Example:

Let X_1 = (3, 1, 9), X_2 = (2, 5, 8), X_3 = (4, 6, 1) be the first sample, and let the second sample be Y_1 = (5, 9, 4), Y_2 =(1, 10, 6), Y_3 = (2, 3, 5), Y_4 = (4, 8, 2).

To solve this example we define the pooled (combined) sample which is Z_1, Z_2, …, Z_7, where,

$$Z_i = \begin{cases} X_i & i = 1,2,3. \\ Y_{i-3} & i = 4,5,6,7. \end{cases}$$

By taking the median as a starting point which is m = (3, 6, 5)

$$\| Z_m\text{-}Z_1 \| = \sqrt{(3-3)^2+(6-1)^2+(5-9)^2}$$

$$= \sqrt{41}$$

$$\| Z_m\text{-}Z_2 \| = \sqrt{(3-2)^2+(6-5)^2+(5-8)^2}$$

$$= \sqrt{11}$$

$$\| Z_m\text{-}Z_3 \| = \sqrt{(3-4)^2+(6-6)^2+(5-1)^2}$$

$$= \sqrt{17}$$

$$\| Z_m\text{-}Z_4 \| = \sqrt{(3-5)^2+(6-9)^2+(5-4)^2}$$

$$= \sqrt{14}$$

$$\| Z_m - Z_5 \| = \sqrt{(3-1)^2 + (6-10)^2 + (5-6)^2}$$

$$= \sqrt{21}$$

$$\| Z_m - Z_6 \| = \sqrt{(3-2)^2 + (6-3)^2 + (5-5)^2}$$

$$= \sqrt{10}$$

$$\| Zm - Z7 \| = \sqrt{(3-4)^2 + (6-8)^2 + (5-2)^2}$$

$$= \sqrt{14}$$

By this step we find the first nearest neighbour point to Z_m, but we need to find the nearest to the nearest so we do the Euclidean distance to find the closest point to the closest.

The combined ordered arrangement of $\| Z_j - Z_m \|$ from smallest to largest will give us the k^{th} nearest neighbour, Z_j, to Z_m, k =1,2,. . . , 7.

Table (3. 1): the k^{th} nearest neighbour to Z_m and the corresponding value of h (m, k).

K	**1**	**2**	**3**	**4**	**5**	**6**	**7**
M	$Z_6[0]$	$Z_2[1]$	$Z_1[1]$	$Z_3[1]$	$Z_7[0]$	$Z_4[0]$	$Z_5[0]$

Table (3. 2): the sample for each nearest neighbour and the responding T_{mk} value.

k	1	2	3	4	5	6	7
k^{th} NN to (m)	Z_6	Z_2	Z_1	Z_3	Z_7	Z_4	$\mathbf{Z_5}$
Sample	Y	X	X	X	Y	Y	**Y**
h(m, k)	0	1	1	1	0	0	0
T_{mk}	**0**	**1**	**2**	**3**	**3**	**3**	**3**

While,

$$T_{mk} = \sum_{j=1}^{k} h(m,j),\ k = 1,...,7$$

$$T_{m} = \sum_{k=1}^{7} T_{mk},\ k = 1,2,...,7.$$

$$= 0 + 1 + 2 + 3 + 3 + 3 + 3 = 15$$

Now we have under $H_{0,}$

Result 1:

i) $E[h(m,j)] = \frac{n_1}{N}$

since, $E(x) = \sum x\, p(x) = (1).\ \frac{\binom{n_1}{1}}{\binom{N}{1}} = \frac{n_1}{N}$

ii) $V[h(m,j)] = \frac{n_1 n_2}{(N)^2}$

since, $V = \sigma_x^2 = \sum x^2\, p(x) - \mu^2$

$$= (1)^2.\ \frac{n_1}{N} - \left(\frac{n_1}{N}\right)^2$$

$$= \frac{n_1}{N} - \frac{(n_1)^2}{N^2}$$

$$= \frac{Nn_1 - (n_1)^2}{N^2}$$

$$= \frac{n_1(N - n_1)}{N^2}$$

$$= \frac{n_1 n_2}{N^2} \quad , \quad \text{Since } N - n_1 = n_2.$$

iii) $Cov[h(m, j), h(m, j')] = \frac{-n_1 n_2}{(N)^2(N-1)}$

where j, j′ = 1, 2, . . . , N;

$j \neq 1 \neq j'$.

Since,

Cov (x, y) = E (x, y) – E (x). E (y) =

$$\frac{\binom{n_1}{2}}{\binom{N}{2}} - \left(\frac{n_1}{N} \cdot \frac{n_1}{N}\right)$$

=

$$\frac{n_1(n_1 - 1).1}{N(N-1)} - \frac{n_1^2}{N^2}$$

$$\frac{n_1^2 - n_1}{N(N-1)} - \frac{n_1^2}{N^2} = \frac{N(n_1^2 - n_1) - n_1^2(N-1)}{N(N-1)}$$

$$= \frac{n_1^2 - N n_1 + n_1}{N^2(N-1)} = \frac{n_1(n_1 - N + 1)}{N^2(N-1)}$$

$$= \frac{n_1(-n_2 + 1)}{N^2(N-1)} = \frac{-n_1(n_2 - 1)}{N^2(N-1)}$$

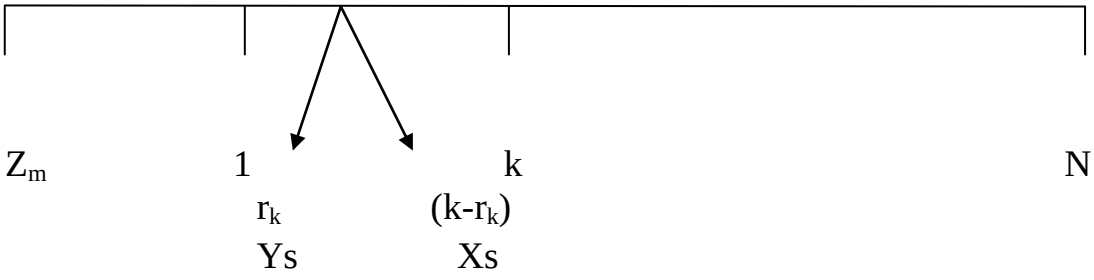
Result (2)

under H_0, we have

$$P\left(T_{mk} = r_k\right) = \frac{\binom{n_2}{r_k}\binom{n_1}{k-r_k}}{\binom{N}{k}}$$

Where $\max(0,\ k-n_1) \le r_k \le \min(k,\ n_2)$

Proof :



For the set of integers {1, 2, . . . , N} and for Z_m we calculate

$\| Z_m - Z_j \|$; then the result of the calculations is a permutation of this set of N integers. The probability of this permutation is (1 / N!).

The number of permutations satisfying $\{T_{mk} = r_k\}$ is the number of ways of choosing r_k Ys out of the n_2 Ys and $(k - r_k)$ Xs out of the n_1 Xs. which is done by

$$\binom{n_2}{r_k}\binom{n_1}{k-r_k}$$

ways. But each can be done by k!(N-k)! ways in order to have r_k Ys and ($k - r_k$) Xs in the first k values.

Therefore ،

the number of ways to have $\{T_{mk} = r_k\}$ is

$$\binom{n_2}{r_k}\binom{n_1}{k-r_k} k!(N-k)!$$

$$So,$$

$$P(T_{mk} = r_k) = \frac{\binom{n_2}{r_k}\binom{n_1}{k-r_k} k!(N-k)!}{N!}$$

$$= \frac{\binom{n_2}{r_k}\binom{n_1}{k-r_k}}{\binom{N}{k}}$$

Corollary

Under H_0 we have,

i) $E(T_{mk}) = k\frac{n_2}{N}$

ii) $E(T^2_{mk}) = k(k-1)\frac{n_2(n_2-1)}{N(N-1)} + k\frac{n_2}{N}$

iii) $V(T_{mk}) = \frac{kn_2}{N^2(N-1)}(N-n_2)(N-k)$

iv) $E[T_{mk}(T_{mk}-1)\ldots(T_{mk}-r+1)] = \frac{r!\binom{n_2}{r}\binom{k}{r}}{\binom{N}{r}}$

v) $E[T_{mk} - E(T_{mk})]^3 =$

$$k.\frac{n_2}{N}.\frac{N-n_2}{N}.\frac{N-2n_2}{N}.\frac{N-k}{N-1}.\frac{N-2k}{N-2}$$

Proof:

A theorem in hypergeometric distribution says:

If X is a hypergeometric distribution, then

$E[x] = n.\frac{K}{M}$ *and* $\text{var}[x] = n.\frac{K}{M}.\frac{M-K}{M}.\frac{M-n}{M-1}$

Where M is a positive integer, K is a nonnegative integer that is at most M. n is a positive integer that is at most M. X is a random variable which have a hypergeometric distribution, x = 1, 2, . . . , n.
(MOOD, GRAYBILL, and BOES, 1974,page 91)

the proof is immediate.

$$E[x] = \sum_{x=0}^{n} x\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}} = n.\frac{K}{M}\sum_{x=1}^{n}\frac{\binom{K-1}{x-1}\binom{M-K}{n-x}}{\binom{M-1}{n-1}}$$

$$= n.\frac{K}{M}\sum_{y=0}^{n-1}\frac{\binom{K-1}{y}\binom{M-1-K+1}{n-1-y}}{\binom{M-1}{n-1}}$$

$$= n.\frac{K}{M} \quad which \quad is \quad in \quad our \quad case \quad k.\frac{n_2}{N}$$

Using,

$$\sum_{i=0}^{m}\binom{a}{i}\binom{b}{m-i}=\binom{a+b}{m}$$

$$\text{E[X (X-1)]} = \sum_{x=0}^{n} x\,(x-1)\frac{\binom{K}{x}\binom{M-K}{n-x}}{\binom{M}{n}}$$

$$= \text{n (n - 1)}\ \frac{K\,(K-1)}{M\,(M-1)}\sum_{x=2}^{n}\frac{\binom{K-2}{x-2}\binom{M-K}{n-x}}{\binom{M-2}{n-2}}$$

$$= n(n-1)\frac{K(K-1)}{M(M-1)}\sum_{y=0}^{n-2}\frac{\binom{K-2}{y}\binom{M-2-K+2}{n-2-y}}{\binom{M-2}{n-2}}$$

$$= n(n-1)\frac{K(K-1)}{M(M-1)}$$

Hence,

Var [X] = E [X^2]-(E [X])^2 = E [X (X-1)] + E [X] - (E [X])^2

$$= n(n-1)\frac{K(K-1)}{M(M-1)} + n\frac{K}{M} - n^2\frac{K^2}{M^2}$$

$$= = n\frac{K}{M}\left[(n-1)\frac{K(K-1)}{M(M-1)} + 1 - \frac{nK}{M}\right]$$

$$= \frac{nK}{M}\left[\frac{(M-K)(M-n)}{M(M-1)}\right]$$

(MOOD, GRAYBILL, BOES, 1974)

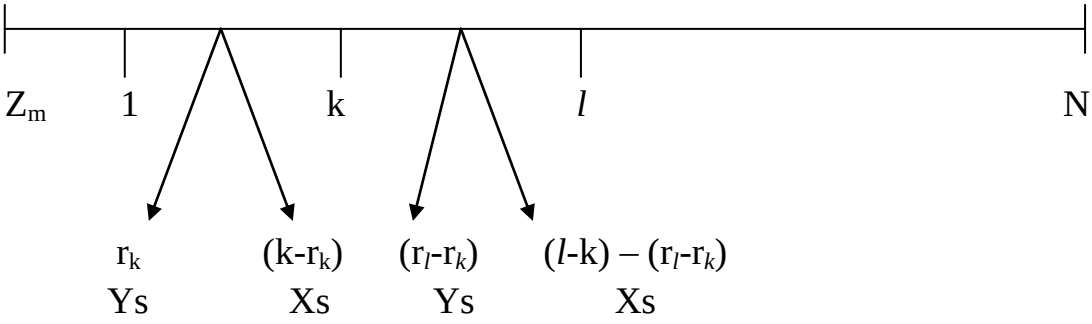
Similarly, (iv) and (v) can be proved easily.

Result 3:

We have under H_0,

$$P(T_{mk}=r_k,\ T_{ml}=r_l)=\frac{\binom{n_2}{r_k}\binom{n_1}{k-r_k}\binom{n_2-r_k}{r_l-r_k}\binom{n_1-k+r_k}{l-k-r_l+r_k}}{\binom{N}{k\quad l-k\quad N-l}},l>k$$

Proof:



As in result 2, we have Z_m fixed, For the set of integers{1, 2,. . . , N} and for Z_m we calculate $\| Z_m - Z_j \|$; then the result of the calculations is a permutation of this set of N integers. The probability of this permutation is (1 / N!).

The number of permutations satisfying {$T_{mk} = r_k$, $T_{ml} = r_l$ } is the number of ways of choosing r_k Ys out of the n_2 Ys and ($k - r_k$) Xs out of the n_1 Xs. in the first k values which can be done by

$$\binom{n_2}{r_k}\binom{n_1}{k-r_k}$$

ways. And ($r_l - r_k$) Ys out of the remaining ($n_2 - r_k$) Ys and (l-k) - ($r_l - r_k$) Xs out of the remaining ($n_1 - k + r_k$) Xs in the next ($l - k$) values which can be done by

$$\binom{n_2 - r_k}{r_l - r_k}\binom{n_1 - k + r_k}{l - k - r_l + r_k}$$

ways.

Therefore,

the number of ways to have $\{T_{mk} = r_k,\ T_{ml} = r_l\}$ is given by

$$\binom{n_2}{r_k}\binom{n_1}{k - r_k}\binom{n_2 - r_k}{r_l - r_k}\binom{n_1 - k + r_k}{l - k - r_l + r_k}.$$

But each can be permuted k!(l-k)!(N-l)! times. Thus the number of ways to have $\{ T_{mk} = r_k,\ T_{ml} = r_l \}$ is

$$\binom{n_2}{r_k}\binom{n_1}{k - r_k}\binom{n_2 - r_k}{r_l - r_k}\binom{n_1 - k + r_k}{l - k - r_l + r_k} k!(l - k)!(N - l)!$$

So,

$$P(T_{mk} = r_k,\ T_{ml} = r_l) = \frac{\binom{n_2}{r_k}\binom{n_1}{k - r_k}\binom{n_2 - r_k}{r_l - r_k}\binom{n_1 - k + r_k}{l - k - r_l + r_k}}{\binom{N}{k \quad l - k \quad N - l}}, l > k$$

Where

$\max(0,\ k-n_1) \le r_k \le r_l \le \min(l,\ n_2)$, and

$1 \le k < l \le N$

Result (4)

We have under H_0 another result for $1 \le k < l < t \le N$

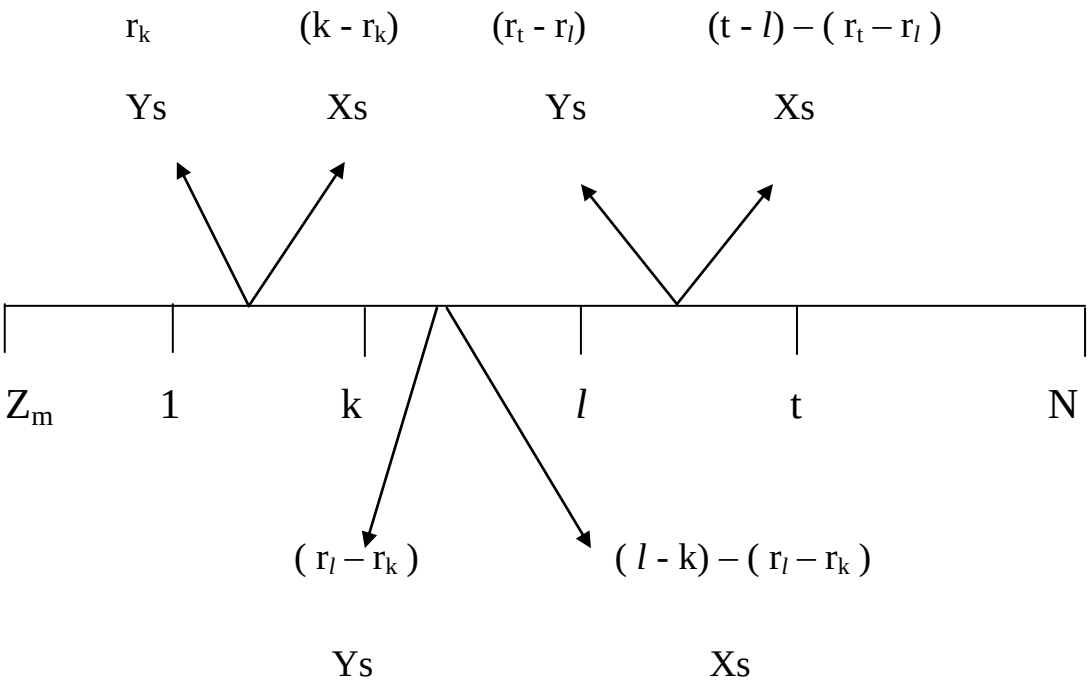
which is

$$P(T_{mk} = r_k,\ T_{ml} = r_l,\ T_{mt} = r_t) =$$

$$\binom{n_2}{r_k}\binom{n_1}{k-r_k}\binom{n_2-r_k}{r_l-r_k}\binom{n_1-k+r_k}{l-k-r_l+r_k}\binom{n_2-r_l}{r_l-r_k}\binom{n_1-l+r_l}{t-l-r_t+r_l}$$

$$\times \frac{k!(l-k)!(t-l)!(N-t)!}{N!}$$

Proof :



As in results (2) and (3), we have (N!) permutation.

we have Z_m fixed, For the set of integers {1, 2, . . . , N} and for Z_m we calculate $\| Z_m - Z_j \|$; then the result of the calculations is a permutation of this set of N integers. The probability of this permutation is (1 / N!).

The number of permutations satisfying $\{T_{mk} = r_k,\ T_{ml} = r_l,\ T_{mt} = r_t\}$ is the number of ways of choosing r_k Ys out of the n_2 Ys and $(k - r_k)$ Xs out of the n_1 Xs. in the first k values which can be done by

$$\binom{n_2}{r_k}\binom{n_1}{k - r_k}$$

ways. And ($r_l - r_k$) Ys out of the remaining ($n_2 - r_k$) Ys and

(l-k) - (r_l - r_k) Xs out of the remaining ($n_1 - k + r_k$) Xs in the next

($l - k$) values which can be done by

$$\binom{n_2 - r_k}{r_l - r_k}\binom{n_1 - k + r_k}{l - k - r_l + r_k}$$

ways, and (r_t - r_l) Ys out of the remaining ($n_2 - r_l$) Ys and

($t - l - r_t + r_l$) Xs out of the remaining ($n_1 - l + r_l$) Xs in the remaining

($t - l$) values, which can be done in

$$\binom{n_2 - r_l}{r_t - r_l}\binom{n_1 - l + r_l}{t - l - r_t + r_l} \text{ ways.}$$

Therefore,

the number of ways to have {$T_{mk} = r_k$, $T_{ml} = r_l$, $T_{mt} = r_t$ } is given by

$$\binom{n_2}{r_k}\binom{n_1}{k - r_k}\binom{n_2 - r_k}{r_l - r_k}\binom{n_1 - k + r_k}{l - k - r_l + r_k}\binom{n_2 - r_l}{r_t - r_l}\binom{n_1 - l + r_l}{t - l - r_t + r_l}.$$

now lets denote this by $\wp$ (r_k, r_l, r_t).

But each can be permuted k!(l-k)!(t-l)!(N-t)! times. Thus the number of ways to have { $T_{mk} = r_k$, $T_{ml} = r_l$, $T_{mt} = r_t$} is

$\wp$ (r_k, r_l, r_t) k! (l-k)! (t-l)!(N-t)!

Then,

for $1 \le k < l < t \le N$, and $\max(0, k-n_1) \le r_k \le r_l \le r_t \le \min(t, n_2)$

So,

$$P(T_{mk} = r_k,\ T_{ml} = r_l,\ T_{mt} = r_t) = \frac{\wp(r_k, r_l, r_t)}{\binom{N}{K\ \ l-k\ \ t-l\ \ N-t}}$$

Result (5) :

Under H_0 we have :

I) $$E[T_{mk}(T_{ml} - T_{mk})] = \frac{n_2(n_2-1)}{N(N-1)}k(l-k)$$

II) $$E[T_{mk}(T_{ml} - T_{mk})(T_{mt} - T_{ml})] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(t-l)$$

III) $$E[T_{mk}^2(T_{mt} - T_{ml})] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(t-l)$$

IV) $$E[T_{mk}(T_{ml} - T_{mk})^2] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(l-k-1)$$

V) $$E[T_{mk}^2(T_{ml} - T_{mk}) = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-k)$$

VI) $$E[T_{mk}^3] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(k-2)$$

Proof:

I) $E[T_{mk}(T_{ml}-T_{mk})]=\frac{n_2(n_2-1)}{N(N-1)}k(l-k)$

$$E=\sum x.p(x)$$

$$E[T_{mk}(T_{ml}-T_{mk})]=\frac{\binom{n_2}{2}\binom{n_1}{0}}{\binom{N}{2}}.\binom{k}{1}\binom{l-k}{1}$$

$$=\frac{n_2(n_2-1)}{N(N-1)}.k(l-k)$$

II) $E[T_{mk}(T_{ml}-T_{mk})(T_{mt}-T_{ml})]=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(t-l)$

$$\frac{\binom{n_2}{3}}{\binom{N}{3}}.\binom{k}{1}\binom{l-k}{1}\binom{t-l}{1}=$$

$$\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(t-l)$$

III) $E[T_{mk}^{2}(T_{mt}-T_{ml})]=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(t-l)$

$$\frac{\binom{n_2}{3}}{\binom{N}{3}}\binom{k}{2}\binom{t-l}{1}=$$

$$\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(t-l)$$

IV) $E[T_{mk}(T_{ml}-T_{mk})^2]=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(l-k-1)$

$$\frac{\binom{n_2}{3}}{\binom{N}{3}}\cdot\binom{k}{1}\cdot\binom{l-k}{2}=$$

$$\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(l-k-1)$$

V) $E[T_{mk}^{2}(T_{ml}-T_{mk})=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-k)$

$$\frac{\binom{n_2}{3}}{\binom{N}{3}}\cdot\binom{k}{2}\cdot\binom{l-k}{1}=$$

$$\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-k)$$

VI) $E[T_{mk}^{3}] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} k(k-1)(k-2)$

$$\frac{\binom{n_2}{3}}{\binom{N}{3}} \cdot \binom{k}{3} =$$

$$\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} . k(k-1)(k-2)$$

Result (6)

Again under H0 we have,

i) $E[T_{mk}T_{ml}] = \frac{n_2(n_2-1)}{N(N-1)} k(l-1) + \frac{n_2}{N} k$

ii) $E[T_{mk}^{2}T_{ml}] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} k(k-1)(l-2)$

iii) $E[T_{mk}T_{ml}^{2}] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} k(l-1)(l-2)$

iv) $E[T_{mk}T_{ml}T_{mt}] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)} k(l-1)(t-2)$

where $1 \le k < l < t \le N$.

Proof:

I) $$E[T_{mk}T_{ml}]=\frac{n_2(n_2-1)}{N(N-1)}k(l-1)+\frac{n_2}{N}k$$

$$E[T_{mk}-T_{ml}]=E[T_{mk}(T_{ml}-T_{mk})]+E[T_{mk}^2]$$

$$=\frac{\binom{n_2}{2}}{\binom{N}{2}}\cdot\binom{k}{1}\binom{l-k}{1}+\frac{\binom{n_2}{2}}{\binom{N}{2}}\cdot\binom{k}{2}+\frac{\binom{n_2}{1}}{\binom{N}{1}}\cdot\binom{k}{1}$$

$$=\frac{n_2(n_2-1)}{N(N-1)}k(l-1)+\frac{n_2(n_2-1)}{N(N-1)}k(k-1)+\frac{n_2}{N}k$$

$$=\frac{n_2(n_2-1)}{N(N-1)}\left[k(l-k)+k(k-1)\right]+\frac{n_2}{N}k$$

$$=\frac{n_2(n_2-1)}{N(N-1)}\left[k(l-k+k-1)\right]+\frac{n_2}{N}k$$

$$=\frac{n_2(n_2-1)}{N(N-1)}k(l-1)+\frac{n_2}{N}k$$

II) $E[T_{mk}^{2}T_{ml}]= \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-2)$

$E[T_{mk}^{2}T_{ml}] = E[T_{mk}^{2}(T_{ml}-T_{mk})] + E[T_{mk}^{3}]$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-k)+\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(k-2)$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}\left[k(k-1)(l-k)+k(k-1)(k-2)\right]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)\left[l-k+k-2\right]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-2)$$

III) $E[T_{mk}T_{ml}^{2}]= \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-1)(l-2)$

$E[T_{mk}(T_{ml}-T_{mk})^2]-E[T_{mk}^{3}]+2E[T_{mk}^{2}T_{ml}]$

$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(l-k-1) \; -\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(k-2)$

$+ \; 2\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(l-2)$

$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}\left[k(l-k)(l-k-1)-k(k-1)(k-2)+2k(k-1)(l-2)\right]$

$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k\left[l^2-lk-l-lk+k^2+k-k^2+\right.$

$\left.k+2k-2+2lk-2l-4k+4\right]$

$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k\left[l^2-3l+2\right]$

$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-1)(l-2)$

IV) $E[T_{mk}T_{ml}T_{mt}] = \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-1)(t-2)$

$$E\left[T_{mk}(T_{ml}-T_{mk})(T_{mt}-T_{ml})\right]+E\left[T_{mk}T_{ml}^{2}\right]+E\left[T_{mk}^{2}(T_{mt}-T_{ml})\right]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-k)(t-l) \; + \; \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-1)(l-2)$$

$$+ \; \frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(k-1)(t-l)$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k[(l-k)(t-l)+(l-1)(l-2)+(k-1)(t-l)]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k[(t-l)((l-k)+(k-1))+(l-1)(l-2)]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k[(t-l)(l-1)+(l-1)(l-2)]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-1)[(t-l)(l-2)]$$

$$=\frac{n_2(n_2-1)(n_2-2)}{N(N-1)(N-2)}k(l-1)(t-2)$$

Result (7)

Under H_0, we have

$$Cov(T_{mk}, T_{ml}) = \frac{n_1 n_2 k (N - l)}{N^2 (N - 1)}, \qquad k < l.$$

Proof:

$$Cov(T_{mk}, T_{ml}) = Cov[T_{mk}, (T_{ml} - T_{mk}) + T_{mk}]$$

$$= Cov[T_{mk}, (T_{ml} - T_{mk})] + V(T_{mk})$$

From Multivariate Hypergeometric distribution properties we have;

$$Cov[T_{mk}, (T_{ml} - T_{mk})] = \frac{-n_2 k (l - k)(N - n_2)}{N^2 (N - 1)}$$

Then we have,

$$Cov[T_{mk},(T_{ml}-T_{mk})]+V(T_{mk}) = \frac{-n_2k(l-k)(N-n_2)}{N^2(N-1)}+V(T_{mk})$$

$$= \frac{-n_2k(l-k)(N-n_2)}{N^2(N-1)} + \frac{n_2k(N-k)(N-n_2)}{N^2(N-1)}$$

$$= \frac{n_2k(N-n_2)[-(l-k)+(N-k)]}{N^2(N-1)}$$

$$= \frac{n_2k(N-n_2)[-l+k+N-k]}{N^2(N-1)}$$

$$= \frac{n_2k(N-n_2)(-l+N)}{N^2(N-1)}$$

$$= \frac{n_2kn_1(N-l)}{N^2(N-1)} = \frac{n_1n_2k(N-l)}{N^2(N-1)}$$

(3.3)The exact distribution of T_m:

When the test statistic $T_m = \sum_{k=1}^{N} T_{mk}$ have the following combined arrangement of the two samples X and Y, with a fixed one equals to Z_m

$$Z_m \underbrace{X \quad X \quad \ldots \quad X}_{n_1 \text{ values}} \underbrace{Y \quad Y \quad \ldots \quad Y}_{n_2 \text{ values}}$$

which Y means that the nearest neighbour to Z_m is from sample Y and X means from sample X, then the test statistic Tm have a maximum value which is

$$\frac{n_1(n_1+2n_2+1)}{2}$$

and when the test have the order of

$$Z_m \underbrace{Y \quad Y \quad \ldots Y}_{\substack{n_2 \\ values}} \underbrace{X \quad X \quad \ldots X}_{\substack{n_1 \\ values}}$$

then the test will have the minimum value of T_m which is

$$\frac{n_1(n_1+1)}{2}$$

So the test statistic of T_m will have a value which is

$$\min\left(\frac{n_1(n_1+1)}{2}\right) \leq T_m \leq \max\left(\frac{n_1(n_1+2n_2+1)}{2}\right)$$

Result (8)

Under H_0, we have the following;

i) $E(T_m) = \frac{n_2(N+1)}{2}$

ii) $V(T_m) = \frac{n_2 n_1(N+1)}{12}$

Proof:

I) $E(T_m) = \frac{n_2(N+1)}{2}$

$$E(T_m) = E\left[\sum_{k=1}^{N} T_{mk}\right]$$

$$= E\left[\sum_{k=1}^{N}\sum_{j=1}^{k} h(m,j)\right]$$

$$= \sum_{k=1}^{N} k\,\frac{n_2}{N} = \frac{n_2}{N}\sum_{k=1}^{N} k$$

$$= \frac{n_2(N+1)}{2}$$

II) $V(T_m) = \frac{n_2 n_1(N+1)}{12}$

$$V(T_m) = V\left[\sum_{k=1}^{N} T_{mk}\right] = V\left[\sum_{k=1}^{N}\sum_{j=1}^{k} h(m,j)\right]$$

$$= V\left[\sum_{j=1}^{N}(N-j+1)^2 V\left[h(m,j)\right]\right] + \sum_{j\neq j'}\sum (N-j+1)(N-j+1)$$

$$\times\, Cov\left[h(m,j), h(m,j')\right]$$

$$= \frac{n_1 n_2}{N^2}\sum (N-j+1)^2 - \frac{n_1 n_2}{N^2(N-1)}\sum_{j\neq}\sum_{j'}(N-j+1)(N-j+1)$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[N\sum_j (N-j+1)^2 - (\sum_j (N-j+1))^2 \right]$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[\frac{N^2(N+1)(2N+1)}{6} - \frac{N^2(N+1)^2}{4} \right]$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[N^2(N+1)\left[\frac{2N+1}{6} - \frac{N+1}{4} \right] \right]$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[N^2(N+1)\left[\frac{4N+2-3N-3}{12} \right] \right]$$

$$= \frac{n_1 n_2}{N^2(N-1)}\left[N^2(N+1).\frac{(N-1)}{12} \right]$$

$$= \frac{n_1 n_2 (N+1)}{12}$$

(Barakat, A. S., 2003)

In our example (see page 50) ,

We have the following;

$T_m = 15$

$E(T_m) = \underline{3(8)} = 12$
2

$V(T_m) = \underline{(3)(4)(8)} = 8$
12

$Z_c = \underline{15 - 12} = 1.06$
$\sqrt{8}$

p-value = 2 . p ($Z \geq |Z_c|$) = 2(1 – 0.8554) = 0.2892

we do not reject H_0, and they are homogenous .

To find the exact distribution for T_m we can write the test statistic in another way. Again Z_m is the fixed variable which we said before is the median of the combined sample, then we calculate $\| Z_m - Z_j \|$,

j = 1, 2, …, N. Then the Z sample (pooled sample) which have an arrangement of the two samples can be denoted by a vector which is 0 if the nearest neighbour to Z_m is from sample Y, and equals 1 if the nearest neighbour to Z_m is from sample X. Now the rank of the observation for which Z_{mk} is an indicator is k, therefore Z_m indicates the rank-order of the arrangement of the combined sample and identifies the to which sample each belongs.

This is called a linear rank statistic which is defined as

$$T_N(Z_m) = \sum_{k=1}^{N} a_k Z_{mk}$$

Where the a_k are given numbers, which can be score or weight, and $T_N(Z_m)$ is linear in the indicator variables.

(Gibbons, 1985)

Result (9)

Under H_0, we have

T_m is symmetric about its mean $\mu = \frac{n_2 N}{2}$

Proof:

Linear rank statistic,

$$T_N(Z_m) = \sum_{j=1}^{N} a_j Z_{mj}$$

have a property which is that its symmetric about its mean whenever

$a_j + a_{N-j} = c$ $\quad$ c = constant, for $j = 1,2,...,N$

But we can write T_m as the following

$$T_m = \sum_{k=1}^{N}\sum_{j=1}^{k} h(m,j)$$

$$T_m = \sum_{k=1}^{N}\sum_{j=1}^{k} Z_{mj}$$

$$= \sum_{j=1}^{N} (N - j + 1) Z_{mj}$$

Take $a_j = N - j + 1$ $\quad$ *then* a_{N-j+1}

Therefore,

$$a_j + a_{N-j+1} = N + 1$$

But N+1 = constant, so

T_m is symmetric about its mean.

Using the previous results we can obtain the exact null probability of T_m systematically by enumeration. For example, suppose we have two samples X with $n_1 = 3$, and sample Y with $n_2 = 4$, so there are

$\binom{N}{n_1} = \binom{N}{n_2} = \binom{7}{3} = \binom{7}{4} = 35$ possible distinguishable configurations

of 1s and 0s in the combined variable Z_m. (Randles, and Wolfe, 1979).

But for the larger samples generation of the exact probability distribution is harder, so for $n_1 \to \infty$ and for $n_2 \to \infty$ which n_2/n_1 remains constant, an approximation exists which is applicable to the distribution of almost all linear rank statistics.

Since T_m is a linear combination of the Z_{mk}, which are identically distributed random variables, a generalization of the central limit theorem allows us to conclude that the probability distribution of the linear rank statistic

$$\frac{T_m - E(T_m)}{\sigma(T_m)}$$

approaches the standard normal probability distribution subject to certain regularity conditions (Gibbons, 1985)

we want to show the relationship between T_m and the sum of ranks of the nearest neighbours to Z_m in the pooled ordered arrangement of the two samples when they are arranged from the largest to the smallest that's to give the nearest neighbour to Z_m the larger rank and so on.

Suppose we have the following arrangement;

k	:	1	2	3	4	5	6	7
k^{th} nearest neighbour to Z_m	:	Y	X	X	X	Y	Y	Y
Z_{mk} = h (m, k)	:	0	1	1	1	0	0	0
T_{mk}	:	0	1	2	3	3	3	3
R_{mk}	:	7	6	5	4	3	2	1

Where R_{mk} is the rank of the k^{th} nearest neighbour to Z_m when we have the arrangement of the pooled sample ordered from the largest to the smallest. Therefore we have,

$$T_m = \sum_{k=1}^{7} T_{mk} = 15 = R_m = \sum_{k=1}^{7} R_{mk}$$

Result (10)

Under H_0, for

$$T_m = \sum_{k=1}^{N} T_{mk}$$

It has the Wilcoxon – Mann – Whitney distribution.

(Ali S. barakat, 2003)

proof:

Since $Z_{mk} = h(m, k)$ so,

$$T_{mk} = \sum_{j=1}^{k} h(m, j) = \sum_{j=1}^{k} Z_{mj}$$

and

$$T_m = \sum_{k=1}^{N} T_{mk} = \sum_{k=1}^{N} \sum_{j=1}^{k} Z_{mj}$$

If we take $a_j = (N - j + 1)$ then the linear rank statistic $T_N(Z_m)$ can be written in terms of T_{mk} as follows:

$$T_N(Z_m) = \sum_{k=1}^{N} (N - k + 1) Z_{mk} =$$

$$NZ_{m1} + (N-1)Z_{m2} + \ldots + 2Z_{mN-1} + Z_{mN}$$

$$= Z_{m1} + (Z_{m1} + Z_{m2}) + (Z_{m1} + Z_{m2} + Z_{m3}) + \ldots$$

$$+ (Z_{m1} + Z_{m2} + Z_{m3} + \ldots + Z_{mN-1})$$

$$+ (Z_{m1} + Z_{m2} + Z_{m3} + \ldots + Z_{mN})$$

$$= \quad T_{m1} + T_{m2} + T_{m3} + \ldots + T_{mN}$$

$$= \sum_{k=1}^{N} T_{mk}$$

$$= \mathrm{T_m}$$

Therefore,

$$T_m = \sum_{j=1}^{N} (N - j + 1) Z_{mj}$$

$$= N \sum_{j=1}^{N} Z_{mj} - \sum_{j=1}^{N} j Z_{mj}$$

$$= n_1 (N+1) - W_m$$

While,

$$W_m = \sum_{j=1}^{N} j Z_{mj}$$

is the Wilcoxon rank sum statistic. So,

T_m has a Wilcoxon – Mann – Whitney distribution.

Thus T_m is actually the same as the Wilcoxon – Mann – Whitney rank sum test, since a linear relationship exists between the two test statistics, and because of that all the properties of the two test are the same, including consistency and the minimum asymptotic relative efficiency (ARE) of 0.864 relative to the t test.

The most important property we need is that the Wilcoxon – Mann – Whitney test is normally distributed, and the main result of this work is the following theorem.

Theorem:

Since our test has the same properties of the Wilcoxon – Mann – Whitney test so T_m is normally distributed.

Chapter Four

Conclusions and Suggestions for Future Work

4.1: Conclusion

For our test statistic which is a mulultivariate two – sample test based on weighted nearest neighbours taking the nearest to the nearest technique , we found its distribution and we proved that it is normally distributed .

This test is important for our real life , a lot of multivariate samples needs to be compared to know which way is better or if they have the same result

Our test is to provide researchers with an easy computer program to apply the test of homogeneity sample , which may be used in many applications such as medical diagnoses , so this test statistic can be used to diagnose future patients .

4.2: Suggestions For The Future Work

The work of this thesis can be extended in the following areas :

1) Comparing our test with Barakat and Schilling tests to find which is stronger .

2) Generalizing the nearest neighbour tests to more than two populations.

3) Using another distance metric than the Euclidean distance and doing our test or Barakat test.

References

1. Albert, C. J., Luo, (2010), " ***Dynamical systems*** ", Springer Science and Business Media LLC.USA.p402 pp464.

2. Anjanita, D., (2010), " ***A useful classification technique - k-nearest neighbours*** ", analyticbridge, http://www.analyticbridge.com/profiles/blogs/a-useful-classification

3. Barakat, A. S., (2003), " ***Two-sample Multivariate Test of Homogeniety*** ", An-Najah University, J. Research, Department of Statistics, Vol. 17(1)., Palestine, pp 26-31.

4. Barakat, A. S., Quade, D., and Salama, I. A., (1996), " ***Multivariate homogeneity testing using an extended concept of nearest neighbors*** ", Biometrical Journal, 38, pp 605-612.

5. Bremner, D., Demaine, E., Erickson, J., Iacono , J., Langerman, S., Morin, P., Toussaint, G., (2005). " ***Output-sensitive algorithms for computing nearest-neighbor decision boundaries*** ". Discrete and Computational Geometry Vol. 33 (4): pp, 593–604.

6. Bruce, E. H., (2009), "***Lecture Notes on nonparametrics*** ", University of Wisconsin, USA, from the world wide web, http://www.ssc.wisc.edu/~bhansen/718/NonParametrics10.pdf

7. Burden, R. L., Faires, J. D., (2005), " ***Numerical analysis*** " 8th Edition, International Student Edition, USA, pp, 421.

8. Christian, A., Eirik, F., Per, L., (2009), "***Nearest Neighbor Classifiers*** ", TNM033Data Mining Technique, Linköping University.
9. Classle, (05/23/2009), " ***covariance*** ", from the world wide web, http://www.classle.net/book/covariance
10. Conover, W. J., (1980), " ***Practical Nonparametric Statistics*** ", Second Edition, New York: John Wiley & Sons, Inc
11. Coomans, D., Massart,D.L. (1982). " ***Alternative k-nearest neighbour rules in supervised pattern recognition : Part 1. k-Nearest neighbour classification by using alternative voting rules***". Analytica Chimica Acta Vol. 136: pp,15–27.
12. Douglas, D., Jeffrey, C., (2010)," ***Business Statistics*** ", 5th edition, Barron educational series Inc., USA, pp 2,545.
13. Ga'bor, J. S., Maria, L. R., and Nail, K. B., (2007), " ***Mesuring and Testing Dependence by Correlation of Distances*** ", The Annals of Statistics, Institute of Mathematical Statistics, Vol. 35, No. 6, 2769–2794.
14. Gibbons, J. D., (1985), " ***Nonparametric Statistical Inference*** ", Marcel Dekker, New York.
15. Gibbons, J. D., and Chakraborti, S., (1992), " ***Nonparametric Statistical Inference*** ", Third Edition, New York: Marcel Dekker, Inc.
16. Hamid, P., Hosein, A., and Behrouz, M. B., (2008), " ***MKNN:***

Modified K-Nearest Neighbor ", WCECS, San Francisco, USA.

17. Hettmansperger, T. P., (1984), " ***Statistical Inference Based on Ranks*** ", New York: JohnWiley & Sons, Inc

18. Hill, T., Lewicki, P., (2006), " ***Statistics: Methods and applications, A comprehensive reference for science, industry, and data mining*** ", 1st Edition, Stat Soft, Inc, USA, pp 324.

19. Hollander, M., and Wolfe, D. A., (1973), ***" Nonparametric Statistical Methods*** ", New York: John Wiley & Sons, Inc.

20. Ian, F. B., (1979), " ***An Introduction To Applied Probability*** ", John Wiley and sons, USA, pp 1, 27-128.

21. IOS, Improved Outcomes Software Inc, (2004), from the world wide web,

http://www.improvedoutcomes.com/docs/ WebSiteDocs/Clustering Clustering_Parameters/Distance_Metrics_Overview.htm.

22. Jean-Marie, D., Abdeljelil, F., (2001), "***Exact Nonparametric Two-Sample Homogeneity Tests for Possibly Discrete Distributions*** ", De Montreal University, CRDE and CIRANO, Canada.

23. Jerome, H. F., (2003), "***On Multivariate Goodness-Of-Fit and Two-Sample*** ", Department of Statistics and SLAC, Stanford, California.

24. Jin, Z., (2006), " ***Powerful Two-Sample Tests Based on the Likelihood Ratio*** " Department of Statistics, University of

Manitoba, Winnipeg, Manitoba, R3T 2N2, Canada, 48(1): 95-103.

25. Lehmann, E. L., (1975), " ***Nonparametrics: Statistical Methods Based on Ranks*** ", San Francisco: Holden-Day
26. Maria, L. R., (2002), "***A Test of Homogeneity for Two Multivariate Populations***", Ohio University, Department of Mathematics, Athens.
27. **MATLAB**
28. Mills, Peter., (2011) " ***Efficient statistical classification of satellite measurements*** ". International Journal of Remote Sensing (in press).
29. Mood, A. M., Graybill, F. A., and Boes, D. C., (1974), " ***Introduction to the Theory of Statistics*** ", 3rd Edition, McGraw-Hill,Inc., USA, pp 91-92, 520-521.
30. Pesarin, F., and Luigi, S., (2006), " ***Permutation Tests for Univariate and Multivariate Ordered Categorical Data***", Austrian Journal of Statistics, University of Padova, Italy, Vol (35) Number (2&3), pp, 315–324.
31. Philip, M. D., (2001), " ***Nearest Neighbor Methods*** ", Department of Statistics, Iowa State University. pp 1.
32. Randles, R. H. and Walfe, D. A. (1979), ***" Introduction to the the Theory of Nonparametric Statistics "***, Wiley, New York.
33. Richard A. J., Dean, W. W., (2007), " ***Applied Multivariate Statistical Analysis*** ", 6th Edition, pp 3, 673-674.

34. Saed , S., (2011), " ***An Introduction to Data Mining*** ", from the world wide web, http://chemeng.utoronto.ca/~datamining/dmc/k_nearest_neighbors

35. Sargur N. Srihari, (Fall,2009), " ***Nearest Neighbor Classification***", SUNY Buffalo school, CSE 555, pp.25, from the world wide web, http://www.cedar.buffalo.edu/~srihari/CSE555

36. SAS , SAS/STAT 9.1 User's Guide, (2004) , " ***Nonparametric Tests*** ", Chapter 12, SAS Institute Inc., USA, pp(193,1319).

37. Schilling, M. F., (1986), " ***Multivariate two-sample tests based on nearest neighbours*** ", Journal of the American Statistical Association, 81, pp 799-809.

38. Sidney, S., (1956), " ***Nonparametric Statistics for the behavioral science*** ", international student Edition, McGraw-Hill Kogakusha, LTD., Japan, pp 1-2,30-33.

39. Smirnov, N. V., (1939), " ***On the estimation of the discrepancy between empirical curves of distribution for two independent samples*** ", Bull. Math. University, Moscow, pp 2, 3-16.

40. Surbhi, C., Xiabing, X., Nancy, A., (2008), "***Nearest Neighbor Search Methods*** ", Technical Report, Parasol Lab., Department of Computer Science, Texas A & M University.

41. Taylor, Charles (1997). "Classification and kernel density

estimation". Vistas in Astronomy, Vol. 41 (3): pp,441–417.

42. Terrell, D. G., Scott, D. W., (1992). " ***Variable kernel density estimation*** ", Annals of Statistics Vol. 20 (3): pp,1236–1265.

43. Velmrugan, T., and Santhanam ,T., (15, May, 2011)," ***A comparative Analysis between K- Medoid And Fuzzy C-Means Clustering Algorithms For statistically distributed Data points*** ", Journal Of Theoretical And Applied Information technology, Vol (27),No(1).pp19-30.

44. Weiss, L., (1958), "***A test of fit for multivariate distributions***.", Annals of Mathematical Statistics, 29, pp 595-599.

45. **Wikipedia**, The Free Encyclopidia. http://en.wikipedia.org/wiki/k-nearest-neighbor-algorithm

Appendices

Appendix (I)

Real life application of the nearest neighbour

Fisher Iris Data ,which contains the Sepal and the Petal widths and lengths of three three species of iris ,(virginica, setosa, versicolor) . have been studied by so many scientists to illustrate various statistical procedures, in this section we will compare Iris verginica and Iris versicolor with respect to the two sepal measurements . and we list these data in this table.

Fisher's Iris Data

Sepal Width	Sepal Length	Species
3.2	7.0	Versicolor
3.2	6.4	Versicolor
3.1	6.9	Versicolor
2.3	5.5	Versicolor
2.8	6.5	Versicolor
2.8	5.7	Versicolor
3.3	6.3	Versicolor
2.4	4.9	Versicolor
2.9	6.6	Versicolor
2.7	5.2	Versicolor
2.0	5.0	Versicolor
3.0	5.9	Versicolor
2.2	6.0	Versicolor
2.9	6.1	Versicolor
2.9	5.6	Versicolor
3.1	6.7	Versicolor
3.0	5.6	Versicolor
2.7	5.8	Versicolor
2.2	6.2	Versicolor
2.5	5.6	Versicolor
3.2	5.9	Versicolor

2.8	6.1	Versicolor
2.5	6.3	Versicolor
2.8	6.1	Versicolor
2.9	6.4	Versicolor
3.0	6.6	Versicolor
2.8	6.8	Versicolor
3.0	6.7	Versicolor
2.9	6.0	Versicolor
2.6	5.7	Versicolor
2.4	5.5	Versicolor
2.4	5.5	Versicolor
2.7	5.8	Versicolor
2.7	6.0	Versicolor
3.0	5.4	Versicolor
3.4	6.0	Versicolor
3.1	6.7	Versicolor
2.3	6.3	Versicolor
3.0	5.6	Versicolor
2.5	5.5	Versicolor
2.6	5.5	Versicolor
3.0	6.1	Versicolor
2.6	5.8	Versicolor
2.3	5.0	Versicolor
2.7	5.6	Versicolor
3.0	5.7	Versicolor
2.9	5.7	Versicolor
2.9	6.2	Versicolor
2.5	5.1	Versicolor
2.8	5.7	Versicolor
3.3	6.3	Virginica
2.7	5.8	Virginica
3.0	7.1	Virginica
2.9	6.3	Virginica
3.0	6.5	Virginica
3.0	7.6	Virginica
2.5	4.9	Virginica
2.9	7.3	Virginica
2.5	6.7	Virginica
3.6	7.2	Virginica
3.2	6.5	Virginica

2.7	6.4	Virginica
3.0	6.8	Virginica
2.5	5.7	Virginica
2.8	5.8	Virginica
3.2	6.4	Virginica
3.0	6.5	Virginica
3.8	7.7	Virginica
2.6	7.7	Virginica
2.2	6.0	Virginica
3.2	6.9	Virginica
2.8	5.6	Virginica
2.8	7.7	Virginica
2.7	6.3	Virginica
3.3	6.7	Virginica
3.2	7.2	Virginica
2.8	6.2	Virginica
3.0	6.1	Virginica
2.8	6.4	Virginica
3.0	7.2	Virginica
2.8	7.4	Virginica
3.8	7.9	Virginica
2.8	6.4	Virginica
2.8	6.3	Virginica
2.6	6.1	Virginica
3.0	7.7	Virginica
3.4	6.3	Virginica
3.1	6.4	Virginica
3.0	6.0	Virginica
3.1	6.9	Virginica
3.1	6.7	Virginica
3.1	6.9	Virginica
2.7	5.8	Virginica
3.2	6.8	Virginica
3.3	6.7	Virginica
3.0	6.7	Virginica
2.5	6.3	Virginica
3.0	6.5	Virginica
3.4	6.2	Virginica
3.0	5.9	Virginica

Suppose Iris virginica the sample X ($n_1 = 50$) , and Iris versicolor the sample Y ($n_2 = 50$).
We have under H_0,
$E(T_m) = n_1(N+1)/2$

$V(T_m) = n_1.n_2.(N+1)/12$

$$Z_c = \frac{T_m - E(T_m)}{\sqrt{V(T_m)}}$$

In Fisher Iris Data example , we have $T_m = 2110$,

$E(T_m) = \frac{(50)(101)}{2} = 2525$

$V(T_m) = \frac{(50)(50)(101)}{12} = 21041.7$

$Z_c = \frac{2110 - 2525}{145} = -2.86$

P value = $2.\ p(Z \geq |Z_c|) = 2.(1 - .9979) = .0042 < .05$

So we reject H_0, and they are not homogeneous.

Appendix (II)

Computer Program

```
clc
clear
rty=1;
x =[3.3 6.3; 2.7 5.8; 3.0 7.1; 2.9 6.3; 3.0 6.5; 3.0 7.6; 2.5 4.9; 2.9 7.3;
 2.5 6.7; 3.6 7.2; 3.2 6.5; 2.7 6.4; 3.0 6.8; 2.5 5.7; 2.6 6.1; 2.8 5.8;
3.2 6.4; 3.0 6.5; 3.8 7.7; 2.6 7.7; 2.2 6.0; 3.2 6.9; 2.8 5.6; 2.8 7.7; 2.7 6.3;
3.3 6.7; 3.2 7.2; 2.8 6.2; 3.0 6.1; 2.8 6.4; 3.0 7.2; 2.8 7.4; 3.8 7.9; 2.8 6.4;
2.8 6.3; 3.0 7.7; 3.4 6.3; 3.1 6.4; 3.0 6.0; 3.1 6.9; 3.1 6.7; 3.1 6.9; 2.7 5.8;
3.2 6.8; 3.3 6.7; 3.0 6.7; 2.5 6.3; 3.0 6.5; 3.4 6.2; 3.0 5.9]; %data of sample
one
y =[3.2 7.0; 3.2 6.4; 3.1 6.9; 2.3 5.5; 2.8 6.5; 2.8 5.7; 3.3 6.3; 2.4 4.9; 2.9
6.6; 2.7 5.2; 2.0 5.0; 3.0 5.9; 2.2 6.0; 2.9 6.1; 3.0 5.4; 2.9 5.6; 3.1 6.7; 3.0
5.6; 2.7 5.8; 2.2 6.2; 2.5 5.6; 3.2 5.9; 2.8 6.1; 2.5 6.3; 2.8 6.1; 2.9 6.4; 3.0
6.6; 2.8 6.8; 3.0 6.7; 2.9 6.0; 2.6 5.7; 2.4 5.5; 2.4 5.5; 2.7 5.8; 2.7 6.0; 3.4
6.0; 3.1 6.7; 2.3 6.3; 3.0 5.6; 2.5 5.5; 2.6 5.5; 3.0 6.1; 2.6 5.8; 2.3 5.0; 2.7
5.6; 3.0 5.7; 2.9 5.7; 2.9 6.2; 2.5 5.1; 2.8 5.7]; %data of sample two
t12=input('Enter number shift data for x = ');
t13=input('Enter number shift data for y = ');
x=t12+x;
y=t13+y;
%'nh=Input the number sample for one group'
%'gh=SIGMA'
[nh gh]=size(x)
[nh2 gh2]=size(y)
nh3=nh+nh2;
if gh==gh2
q=[x;y];
'***************  sample of x *********************************'
x
'***************  sample of y  *********************************'
y
'***************  sample x and y *********************************'

sample=q
'***************  median for all sample **********************'
med=median(q)
```

```
mn=med;
for i=1:nh3
   for j=1:nh3
      dis2=0;
      for t=1:gh
         dis1=(mn(1,t)-q(j,t))^2;
         dis2=dis2+dis1;
      end
      dis(j,1)=sqrt(dis2);
   end
   dis; % Show distance
   [yt e]=min(dis);
   mn(1,:)=q(e,:);
   q(e,:)=inf; % Show data
   tr(i,1)=e;
   if e<=nh
      tr(i,2)=1;
   else
      tr(i,2)=0;
   end
   if i>1
      tr(i,3)=tr(i-1,3)+tr(i,2);
   elseif i==1
      tr(i,3)=tr(i,2);
   end
end
tr; % Show
sample;
g=sum(tr);
g(:,3);
T(rty)=g(:,3);
'**************** Tm-test statistc*****************************'
Tm
'**************** End ***********************************'
else
   'the size of sample is not same'
end
```

Enter number shift data for x = 0
Enter number shift data for y = 1.5

ans =

********** sample of x **********

x =

3.3000	6.3000
2.7000	5.8000
3.0000	7.1000
2.9000	6.3000
3.0000	6.5000
3.0000	7.6000
2.5000	4.9000
2.9000	7.3000
2.5000	6.7000
3.6000	7.2000
3.2000	6.5000
2.7000	6.4000
3.0000	6.8000
2.5000	5.7000
2.6000	6.1000
2.8000	5.8000
3.2000	6.4000
3.0000	6.5000
3.8000	7.7000
2.6000	7.7000
2.2000	6.0000
3.2000	6.9000
2.8000	5.6000
2.8000	7.7000
2.7000	6.3000
3.3000	6.7000
3.2000	7.2000
2.8000	6.2000
3.0000	6.1000
2.8000	6.4000
3.0000	7.2000
2.8000	7.4000

```
    3.8000    7.9000
    2.8000    6.4000
    2.8000    6.3000
    3.0000    7.7000
    3.4000    6.3000
    3.1000    6.4000
    3.0000    6.0000
    3.1000    6.9000
    3.1000    6.7000
    3.1000    6.9000
    2.7000    5.8000
    3.2000    6.8000
    3.3000    6.7000
    3.0000    6.7000
    2.5000    6.3000
    3.0000    6.5000
    3.4000    6.2000
    3.0000    5.9000

ans =

**********  sample of y  **********
y =

    4.7000    8.5000
    4.7000    7.9000
    4.6000    8.4000
    3.8000    7.0000
    4.3000    8.0000
    4.3000    7.2000
    4.8000    7.8000
    3.9000    6.4000
    4.4000    8.1000
    4.2000    6.7000
    3.5000    6.5000
    4.5000    7.4000
    3.7000    7.5000
    4.4000    7.6000
    4.5000    6.9000
    4.4000    7.1000
    4.6000    8.2000
```

```
    4.5000    7.1000
    4.2000    7.3000
    3.7000    7.7000
    4.0000    7.1000
    4.7000    7.4000
    4.3000    7.6000
    4.0000    7.8000
    4.3000    7.6000
    4.4000    7.9000
    4.5000    8.1000
    4.3000    8.3000
    4.5000    8.2000
    4.4000    7.5000
    4.1000    7.2000
    3.9000    7.0000
    3.9000    7.0000
    4.2000    7.3000
    4.2000    7.5000
    4.9000    7.5000
    4.6000    8.2000
    3.8000    7.8000
    4.5000    7.1000
    4.0000    7.0000
    4.1000    7.0000
    4.5000    7.6000
    4.1000    7.3000
    3.8000    6.5000
    4.2000    7.1000
    4.5000    7.2000
    4.4000    7.2000
    4.4000    7.7000
    4.0000    6.6000
    4.3000    7.2000

ans =

**********  sample x and y**********
sample =

    3.3000    6.3000
    2.7000    5.8000
```

3.0000	7.1000
2.9000	6.3000
3.0000	6.5000
3.0000	7.6000
2.5000	4.9000
2.9000	7.3000
2.5000	6.7000
3.6000	7.2000
3.2000	6.5000
2.7000	6.4000
3.0000	6.8000
2.5000	5.7000
2.6000	6.1000
2.8000	5.8000
3.2000	6.4000
3.0000	6.5000
3.8000	7.7000
2.6000	7.7000
2.2000	6.0000
3.2000	6.9000
2.8000	5.6000
2.8000	7.7000
2.7000	6.3000
3.3000	6.7000
3.2000	7.2000
2.8000	6.2000
3.0000	6.1000
2.8000	6.4000
3.0000	7.2000
2.8000	7.4000
3.8000	7.9000
2.8000	6.4000
2.8000	6.3000
3.0000	7.7000
3.4000	6.3000
3.1000	6.4000
3.0000	6.0000
3.1000	6.9000
3.1000	6.7000
3.1000	6.9000
2.7000	5.8000

3.2000	6.8000
3.3000	6.7000
3.0000	6.7000
2.5000	6.3000
3.0000	6.5000
3.4000	6.2000
3.0000	5.9000
4.7000	8.5000
4.7000	7.9000
4.6000	8.4000
3.8000	7.0000
4.3000	8.0000
4.3000	7.2000
4.8000	7.8000
3.9000	6.4000
4.4000	8.1000
4.2000	6.7000
3.5000	6.5000
4.5000	7.4000
3.7000	7.5000
4.4000	7.6000
4.5000	6.9000
4.4000	7.1000
4.6000	8.2000
4.5000	7.1000
4.2000	7.3000
3.7000	7.7000
4.0000	7.1000
4.7000	7.4000
4.3000	7.6000
4.0000	7.8000
4.3000	7.6000
4.4000	7.9000
4.5000	8.1000
4.3000	8.3000
4.5000	8.2000
4.4000	7.5000
4.1000	7.2000
3.9000	7.0000
3.9000	7.0000
4.2000	7.3000

```
    4.2000    7.5000
    4.9000    7.5000
    4.6000    8.2000
    3.8000    7.8000
    4.5000    7.1000
    4.0000    7.0000
    4.1000    7.0000
    4.5000    7.6000
    4.1000    7.3000
    3.8000    6.5000
    4.2000    7.1000
    4.5000    7.2000
    4.4000    7.2000
    4.4000    7.7000
    4.0000    6.6000
    4.3000    7.2000

ans =
**********  median for all sample**********
med =

    3.7000    7.1000
ans =
**********  T_m-test statistic**********
T_m =
        2110
**********  End  **********
```

جامعة النجاح الوطنية

كلية الدراسات العليا

اختبار متعدد الأبعاد للتجانس بين عينتين باستخدام الجوار المرجح الأقرب

إعداد

أريج علي سعيد بركات

إشراف

د. محمد نجيب أسعد

قدمت هذه الأطروحة استكمالاً لمتطلبات درجة الماجستير في الرياضيات المحوسبة بكلية الدراسات العليا في جامعة النجاح الوطنية في نابلس ، فلسطين.

2012

اختبار متعدد الأبعاد للتجانس بين عينتين باستخدام الجوار المرجح الأقرب

إعداد

أريج علي سعيد بركات

إشراف

د. محمد نجيب أسعد

الملخص

قدمت هذه الرسالة اختباراً جديداً متعدد الأبعاد للتجانس بين عينتين بحيث تم اعتماد الأوزان وإعطاء كل نقطة وزنا مختلفاً حسب بعد هذه النقطة عن نقطة البداية (ا لوسيط) بحيث أخذت النقطة الأقرب لنقطة البداية وزناً أكثر من النقطة الأبعد وهكذا حسب بعد كل نقطة عن الوسيط لهذه النقاط والتي تم اتخاذها نقطة بداية في هذا الاختبار ، ثم تم ترتيب هذه النقاط حسب تكنيك الأقرب للأقرب ، حيث تم حساب النقطة الأقرب لنقطة البداية ثم النقطة الأقرب لتلك النقطة وهكذا.

لقد أثبتت الرسالة أن توزيع هذا الاختبار

$$T_m = \sum_k T_{mk}$$

حيث أن

$$T_{mk} = \sum_{j=1}^{k} h(m, j)$$

له توزيع طبيعي علماً بأن هذا الاختبار امتداداً لاختبار Schilling للتجانس الذي استخدم k = 1 , 2 , or 3 وليس كل النقاط ، كما انه امتداداً لاختبار Barakat الذي أخذ جميع النقاط وبُعدها عن نقطة البداية (الوسيط) ولكن لم يتخذ تكنيك الأقرب للأقرب .

كما أن بركات وقويد وسلامة قدموا اختباراً آخر آخذين بعين الاعتبار المسافة لجميع النقاط وبعدها عن نقطة البداية ولكن توزيع هذا الاختبار لم يُعرف بعد.

كما قدمت هذه الرسالة برنامجاً في لغة MATLAB استُخدم في حل الاختبار، وتم تطبيق البرنامج على مثال من واقع الحياة .