A NONPARAMETRIC MULTIVARIATE TEST FOR HOMOGENEITY BASED ON ALL NEAREST NEIGHBORS

by

Ali Said Barakat

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1866T

August 1989

A NONPARAMETRIC MULTIVARIATE TEST FOR HOMOGENEITY BASED ON ALL NEAREST NEIGHBORS

by

Ali Said Barakat

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics

Chapel Hill

1989

Approved by

Dana Quade
Advisor

A. T. Salam
Advisor

J. K. Sen
Reader

ABSTRACT

ALI SAID BARAKAT. A Nonparametric Multivariate Test For Homogeneity Based on All Nearest Neighbors. (Under the direction of DANA QUADE and IBRAHIM SALAMA.)

Schilling proposed a multivariate two-sample test using a fixed number of nearest neighbors, and we cannot tell how many nearest neighbors do we need to get the best results. In this research we propose a test related to that proposed by Schilling. Our test uses all nearest neighbors and it takes into account the position of each nearset neighbor.

The exact properties of the test are studied as well as the asymptotic ones. Also, as another proof, the nearest neighbors technique has been used to prove the normality of the Wilcoxon-Mann-Whitney statistic via the martingale central limit theorem.

Computer simulations are used to estimate the variance of the test, and Monte Carlo simulation is used to compute its power.

For detecting location shift differences between two populations, using Schilling's test, Hotelling's T^2 test, and our test, we have found that in many cases the proposed test compares favorably with Hotelling's T^2 test and in most cases it compares favorably with Schilling's test.

Finally, using a subset of the Fisher iris data, the three tests are used to test the hypothesis that the distribution of the two sepal measurements of the two species of iris are the same.

ACKNOWLEDGEMENT

It is a pleasure to express my sincere gratitude to my advisor Professor I. Salama, for suggesting the problem considered in this dissertation and for his confidence in me, his patience, and for many helpful discussions during its preperation. I gratefully acknowledge the guidance of my co-advisor Professor D. Quade during this research, I wish to express my deepest gratitude to him, for his patience and willingness to listen to me, and his suggestions and observations. It was a pleasure to work with him.

During my years in Chapel Hill, Professors Quade and Salama have been my teachers, coworkers, dissertation advisors, and my friends. I don't think I can ever repay the kindness that they and their families have shown me and my family throughout the years. I have appreciated and do appreciate it. I have made a great effort to choose the right words to express my feelings toward all of you, but I have not been completely succesful. My feelings are stronger than these words that I could find to express them.

I would also like to thank Professor P. K. Sen, for his helpful discussions and suggestions during this research. I would also like to thank Professors C. M. Suchindran and W. D. Kalsbeek, for their advice and encouragements during the course of this research. I would also like to thank Professor R. Bilsborrow for taking responsibilities of Professor Suchindran who is on leave.

My thanks also go to all of my friends, for their support and encouregements throughout my gradute studies in Chapel Hill.

Finally, I would like to thank my family and relatives, and especially my wife and my mother, for all the encouragement and support they have given me throughout the years.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

It is often necessary in data analysis to determine if observed measurements can be used as predictors for a binary property of the data. This involves comparing two multivariate samples to determine to what extent they are similar or different. For example, in medical diagnosis, the binary property of the data, the dependent variable, is the presence or absence of the disease. The two samples are compared using all clinical measurements performed on the patients. If the two samples are shown to be indistinguishable, then that set of measurements cannot be used to diagnose the disease. If they are different, then the set of measurements that are most likely to indicate the presence (or absence) of the disease can be identified. This information can be used to diagnose future patients.

1.2 Literature Review

The idea of making inferences about a new object from nearby objects appears to be a fundamental mechanism of human perception. Interest in statistical procedures based on " nearest neighbors " has grown and high-speed computers have made the application of nearest neighbor techniques practicable. Nearest neighbor procedures have been applied to many problems such as:

(i) nonparametric classification

(ii) density estimation

(iii) nonparametric regression

(iv) goodness-of-fit tests

(v) two-sample tests.

Definition:

let there be given a metric space upon which is defined a metric d and a random sample of observations $x_1, x_2, \ldots, x_n$ from a common population. The nearest neighbor to an arbitrary point x in the metric space 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 neighbor to x and it is called the k^{th} nearest neighbor to x if $d(x, x_i) > d(x, x_j)$ for exactly (k-1), $j \neq i$. The following is another way of defining k-nearest and k^{th} nearest neighbors:

Let $R_j(x_i)$ = rank of observation x_i with respect to distance from x_j, then x_i is the k^{th} nearest neighbor of x_j if $R_j(x_i) = k$ and a k-nearest neighbor if $R_j(x_i) \leq k$. Note that we assume that there are no ties, and therefore there will be only one k^{th} nearest neighbor.

(i) Nonparametric classification

In the classification problem, we wish to know to which class, or category, a new point x', with unobservable θ', belongs, given that the observations $(x_1, \theta_1), (x_2, \theta_2), \ldots, (x_n, \theta_n)$ are independently drawn from some unknown joint distribution $F(x, \theta)$, where the possible values of θ are finite and where θ_i is treated as the class, or category, to which x_i belongs. In this problem there are two extremes of knowledge which the statistician may posseses: complete statistical knowledge of the underlying joint distribution of the observation x and the true category θ, or no knowledge of the underlying distribution except what can be inferred from samples. In the first extreme, a standard Bayes analysis will yield an optimal decision procedure and the corresponding minimum (Bayes) probability of error of classification R^* — the minimum probability of error over all decision rules taking underlying probability structure into account. In the other extreme, a decision to classify x into category θ is allowed to depend only on a collection of n correctly classified observations (x_1, θ_1),

$(x_2, \theta_2), \ldots, (x_n, \theta_n)$. If it is assumed that the classified observations (x_i, θ_i) are independently and identically distributed according to the distribution of (x, θ), $F(x, \theta)$, certain heuristic arguments may be made about good decision procedures. For example, it is reasonable to assume that observations which are close together (in some appropriate metric) will have the same distribution, or at least will have almost the same posterior probability distributions on their respective classifications.

Thus to classify the unknown observation x, we may wish to weight the evidence of the nearby x_i's most heavily. Perhaps the simplest nonparametric decision procedure of this form is the nearest neighbor rule, which classifies x in the category of its nearest neighbor. Cover and Hart (1967) showed that the probability of error R of the nearest neighbor decision rule is at least as great as the Bayes probability of error R^*. They also showed that for any number of categories, the probability of error of the nearest neighbor rule is bounded above by twice the Bayes probability of error. So, it may be said that half the classification information in an infinite sample set is contained in the nearest neighbor. They also showed that

$$R_n = P[\text{misclassification}] \xrightarrow{n} R$$

where

$$R^* \leq R \leq 2R^*(1 - R^*),$$

and R^* is the Bayes probability of error. Rogers (1978) studied the rate of convergence of the k-nearest neighbor rule to its asymptote.

Wagner (1971) examined the random variable

$$L_n = P[\text{misclassification} \mid (x_1, \theta_1), (x_2, \theta_2), \ldots, (x_n, \theta_n)]$$

which is a function of the observed samples, and he showed that the nearest neighbor rule conditioned on the n known observations, L_n, so that $EL_n = R_n$, converges to R with probability 1. Fritz (1975) has also examined the behavior of the random variable L_n.

The k-nearest neighbor classifier can have substantial bias when there is a little class separation and the sample sizes are unequal. Goin (1984) examined this bias for the two-class situation and he presented formulas that allow selection of values of k that yield minimum bias.

(ii) Density estimation

Let $y_1, y_2, \ldots, y_n$ be independent observations on a d-dimensional random variable $Y = (Y_1, Y_2, \ldots, Y_d)$ with absolutely continuous distribution function F(y) where $y = (y_1, y_2, \ldots, y_d)$. Loftsgaarden and Quesenberry (1965) proposed the following multivariate estimate of f(y) at a point $x = (x_1, x_2, \ldots, x_d)$:

$$\hat{f}_n(x) = \frac{k-1}{n\, V^{(k)}(x)}$$

where $V^{(k)}(x)$ is the volume of the d-dimensional sphere whose radius extends from x to the k^{th} nearest neighbor to x. Also, they showed that

this density estimator is consistent when $k \longrightarrow \infty$ and $(k/n) \longrightarrow 0$ as $n \longrightarrow \infty$. Wagner (1973) gives consistency results for $\hat{f}_n(x)$ using similar conditions to those used by Loftsgaarden and Quesenberry. Fukunaga and Hostetler (1973) considered the problem of choosing the best k by developing a relation between the volume and the coverage of region and obtaining a functional form for the optimum k in term of the sample size, the dimensionality of the sample space, and the underlying probability distribution. Devroye and Wagner (1977) studied the consistency of $\hat{f}_n(x)$ using similar conditions to those used by Wagner (1973). Moore and Yackel (1977) studied cosistency properties of nearest neighbor density estimators. Mack and Rosenblatt (1979) described the asymptotic behavior of the bias and variance of k-nearest neighbor density estimates with weight functions. Fukunaga and Mantock (1984) proposed a nonparametric data reduction technique, based on the use of a criterion function and nearest neighbor density estimates, to select samples that are representative of the entire data set such that the nearest neighbor density estimates for the entire sample set and the selected sample set are close.

(iii) Nonparametric regression

In the regression problem, we wish to estimate the regression function $R(x) = E(Y|X=x)$ given a random sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ from an unknown joint distribution function, where (X, Y) is a pair of random variables such that X is in R^d and Y is in $R^{d'}$, let $\{W_i\}$ be a sequence of weight functions, $1 \leq i \leq n$, and $d(X, X_i)$ be a measure of distance on R^d. Then, the regression function

$E(Y|X=x)$ can be estimated by

$$\hat{E}_n(Y|X=x) = \sum_{i=1}^{n} W_i(x)\, Y_i$$

Stone (1977) has suggested the use of nearest neighbors for the problem of nonparametric regression. He gives sufficient conditions on these weights in order to make $\hat{E}_n(Y|X=x)$ consistent for $E(Y|X=x)$. Also, he showed that $\{W_i\}$ is consistent, using different forms of nearest neighbor weight-functions, if $k \longrightarrow \infty$ and $(k/n) \longrightarrow 0$ as $n \longrightarrow \infty$. He showed that if the weights are chosen to satisfy certain conditions then $\hat{E}_n(Y|X=x)$ is consistent in Bayes risk under squared error and weighted absolute error loss. Examples of k-nearest neighbor weights include

(a) Uniform weights

$$W_i = 1/k,\ i = 1, 2, \ldots, k \text{ and } W_i = 0 \text{ for } i > k$$

(b) Triangular weights

$$W_i = (k-i+1)/b_k,\ i = 1, 2, \ldots, k \text{ and } W_i = 0 \text{ for } i > k,$$

where $b_k = k(k+1)/2$.

(c) Quadratic weights

$$W_i = [k^2 - (i-1)^2]/b_k,\ i = 1, 2, \ldots, k \text{ and } W_i = 0 \text{ for } i > k,$$

where $b_k = k(k+1)(4k-1)/6$.

Devroye (1978, 1980) studied the asymptotic properties of the nearest neighbor regression function estimators. Cheng (1984) also studied the consistency of nearest neighbor regression function estimators under different conditions.

(iv) Goodness-of-fit test:

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 (1958) 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, $i = 1, 2, \ldots, n$, where $X_1, X_2, \ldots, X_n$ 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 $1/[ng(x)]$. The test compares the proportion of spheres containing exactly one point of the (n-1) points $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ to the proportion e^{-1} which is expected under the null hypothesis. Pyke (1965) studied the univariate goodness-of-fit problem using tests based on the spacing between consecutive order statistics of the sample. It is clear that spacings are a form of nearest neighbor. Rogers (1978) said that Friedman has proposed a goodness-of-fit test based on the volume of the k-nearest neighbor sphere $V_i^{(k)}$ centered at X_i,

$$V_i^{(k)} = \int_{S_i^{(k)}} dx = \int_{S_i^{(k)}} \frac{1}{f_0(x)} f_0(x)\, dx,$$

where $f_0(x)$ is the null distribution and $S_i^{(k)}$ is the sphere about X_i. Under the null hypothesis $V_i^{(k)}$ can be estimated by

$$\hat{V}_i^{(k)} = \frac{1}{k+1} \sum_{j=0}^{k} \frac{1}{f_0(x_i^{(j)})}$$

where $X_i^{(0)} = X_i$ and $X_i^{(j)}$ is the j^{th} nearest neighbor to X_i for $j = 1, 2, \ldots, k$. Then a comparison of $\hat{V}_i^{(k)}$ and $V_i^{(k)}$ could be made using standard univariate tests. Schilling (1979) considered a goodness-of-fit test based on nearest neighbors. Bickel and Breiman (1983) introduced a goodness-of-fit test based on the empirical distribution function of the variables

$$W_i = \exp\{ - n\, g(x_i)\, V(R_i)\}, \; i = 1, 2, \ldots, n,$$

where $X_1, X_2, \ldots, X_n$ is a random sample in R^d from unknown density, $g(x)$ is the hypothesized density, $V(R_i)$ represents the volume of the d-dimensional nearest neighbor sphere with radius R_i and center X_i, and $R_i = \min_{j \neq i} ||X_j - X_i||$ is the distance from X_i to its nearest neighbor. Schilling (1983a, b) introduced a weighted version of the test proposed by Bickel and Breiman and obtained the optimal weight

function. He also found an infinite-dimensional approximation to the asymptotic form of the weighted empirical distribution functions of the W_i.

(v) Two-sample test

Let $X_1, X_2, \ldots, X_n$ be a random sample in R^d from unknown distribution function $F(x)$ with a density $f(x)$ assumed to be continuous and $Y_1, Y_2, \ldots, Y_m$ be another random sample in R^d from unknown distribution function $G(x)$ with a continuous density $g(x)$. The two samples are assumed to be independent. The problem under consideration is to test if the two distributions are the same. The alternative is that the two distributions are different.

The two samples are combined into a single sample of size $N=n+m$, $Z_1, Z_2, \ldots, Z_N$, such that

$$Z_i = \begin{cases} X_i, & i = 1, 2, \ldots, n \\ Y_{i-n}, & i = n+1, n+2, \ldots, N \end{cases}$$

Note that in all chapters, the distance will be Euclidean and we assume that there will be no ties.

Weiss (1960) proposed k-variate two-sample tests based on the distances between observations from the same sample and observations from the different samples. He found the distance between each point

in the first sample X_i and its nearest neighbor from the same population, say $2R_i$. Then he counted the number of points, S_i, from the second sample, Y's, such that the distance between X_i and those Y's is less than R_i. His statistic is the proportion of X's with $S_i=0$. He showed that when $n \longrightarrow \infty$, $(n/m) \longrightarrow \alpha$, the probability under the null hypothesis that this count equals zero is equal to $2^d\alpha/(1+2^d\alpha)$ where d is the dimension of the space. He rejects H_0 if the sample proportion is too far above $2^d\alpha/(1+2^d\alpha)$. Note that for the univariate case S_i is equal to the number of Y's lying within a distance R_i on either side of X_i. Friedman and Steppel (1974) proposed a two-sample test based on the number of points C_i from, say, the first sample $X_1, X_2, \ldots, X_n$ which are among the k closest points to each point Z_i in the combined sample. Under the null hypothesis these counts are dependent hypergeometric variables and the frequency distributions have the same expectations. Let n(C) be the observed frequency distribution of C_i, i = 1, 2, . . ., N, and let $n_0(C)$ be the expected frequency under the null hypothesis. Asymptotically, the frequency distribution n(C) is compared to $n_0(C)$, the binomial distribution with parameters k and n/N, using any goodness-of-fit statistic, e.g., Pearson's statistic

$$\chi^2 = \sum \frac{[n(C) - n_0(C)]^2}{n_0(C)}$$

Or the comparison can be done as follows:

Let $n_1(C)$ be the frequency distribution of C_i, $i = 1, 2, \ldots, n$ and $n_2(C)$ be the frequency distribution of C_i, $i = n+1, n+2, \ldots, N$. Under the null hypothesis these two distributions are expected to be the same. Friedman and Steppel suggested comparing these frequency distributions by using a t-statistic of the form

$$t = \frac{\bar{n}_1(C) - \bar{n}_2(C)}{\sqrt{\frac{V_1}{n} + \frac{V_2}{m}}}$$

where

$$\bar{n}_1(C) = \frac{1}{n} \sum_{C=0}^{k} C\, n_1(C)$$

$$\bar{n}_2(C) = \frac{1}{m} \sum_{C=0}^{k} C\, n_2(C)$$

$$V_1 = \frac{1}{n} \sum_{C=0}^{k} [C - \bar{n}_1(C)]^2\, n_1(C)$$

and

$$V_2 = \frac{1}{m} \sum_{C=0}^{k} [C - \bar{n}_2(C)]^2\, n_2(C).$$

Rogers (1978) obtained further results under a different formulation. Let $C_i(X_i)$ be the count of class one neighbors among the k nearest neighbors to X_i. Let Y_i be the class of X_i. Any test statistic that can be written as

$$\sum_{i=1}^{n} h(Y_i, C_i)$$

can also be written as a linear function of the sums

$$S_{\alpha j} = \sum_{i=1}^{n} h_{\alpha j i}$$

where

$$h_{\alpha j i} = \begin{cases} 1 & \text{if exactly j out of k-nearest neighbors are from class } \alpha \text{ and } X_i \text{ is in class } \alpha \\ 0 & \text{otherwise} \end{cases}$$

$j = 1, 2, \ldots, k$ and $\alpha = 1, 2$.

Let S be the vector of $S_{\alpha j}$ values which are obtainable from the frequency distributions. Rogers showed that S is asymptotically normal. Also, he discussed tests based on linear combinations of the $S_{\alpha j}$'s.

Friedman and Rafsky (1979, 1983) were first to introduce procedures for the nonparametric two-sample problem based on the minimal spanning tree of the pooled sample points. These are multivariate generalizations of the Wald-Wolfowitz (1940) and Smirnov (1939) univariate two-sample tests. They derived null distribution results and estimated power performance. Also, they suggested tests for association based on the k-nearest neighbors graph and the k-minimal spanning tree. Schilling (1986b) proposed a multivariate two-sample test based on the proportion of all k nearest neighbor comparisons in which observations and their neighbors belong to the same sample. In order to test the hypothesis that $F(x) = G(x)$ against $F(x) \neq G(x)$, he proposed the following test statistic

$$T_{k,N} = \frac{1}{Nk} \sum_{i=1}^{N} \sum_{j=1}^{k} [1 - h(i, j)]$$

where

$$h(i, j) = \begin{cases} 1 & \text{if } Z_i \text{ and its } j^{th} \text{ nearest neighbor are from different samples} \\ 0 & \text{otherwise} \end{cases}$$

and the Z's and the k^{th} nearest neighbor are as before.

Schilling studied the unweighted and weighted versions of this statistic. He considered the ranks of nearest neighbors as weights and

concluded,theoretically, that taking them into account would not improve the test. The mean and variance of this statistic are

$$E(T_{k,N}) = \frac{n(n-1) + m(m-1)}{N(N-1)}$$

and

$$V(Nk\, T_{k,N}) = \left[\frac{knm}{n-1}\right]\left[1 - \frac{k}{n-1}\right]\left[\frac{(n-m)^2}{n-2} + 1\right] + \frac{nm}{n-3} \times$$

$$\left\{\frac{4(n-1)(m-1)}{n-2}\sum_{r=1}^{k}\sum_{s=1}^{k} P_1(r, s) + \left[\frac{(n-m)^2}{n-2} - 1\right]\sum_{r}^{k}\sum_{s}^{k} P_2(r, s)\right\}$$

where

$P_1(r, s) = P\{$ r^{th} nearest neighbor to Z_i is Z_j and the s^{th} nearest neighbor to Z_j is Z_i $\}$

$P_2(r, s) = P\{$ r^{th} nearest neighbor to Z_i = s^{th} nearest neighbor to $Z_j\}$ and $i \neq j = 1, 2, \ldots, N$.

This statistic can be written in terms of Rogers' statistic as

$$T_{k,N} = \frac{1}{Nk}\sum_{\alpha=1}^{2}\sum_{j=1}^{k} j\, S_{\alpha j}$$

where $S_{\alpha j}$ is the number of points Z_i for which exactly j out of k nearest neighbors are from the same sample as Z_i for $j = 1, 2, \ldots, k$ and $\alpha = 1, 2$. Hence $T_{k,N}$ has a limiting null distribution which is normal. Also, Henze (1987) proved that $T_{k,N}$ has a limiting null distribution which is normal without writing it in terms of Roger's statistic.

1.3 Summary of the Study

Schilling's test used a fixed number of nearest neighbors, k, and we cannot tell which value of k will give the best results. Moreover, his test does not take into account the ranks of the nearest neighbors with respect to nearness, given that they are among the first k, nor the positions of each one of these k nearest neighbors. In this research we propose a multivariate two-sample test related to that proposed by Schilling. Our test uses all values of k simultaneously. Moreover, it takes into account both the value of k and the position of the nearest neighbor. Our test has the following form:

$$T = \sum_{i=1}^{N} \sum_{k=1}^{N-1} T_{ik}$$

where

$$T_{ik} = \sum_{j=1}^{k} h(i, j)$$

and h(i, j) is as defined before, i=1, 2, . . ., N and k=1, 2, . . ., N-1.

Under the alternative hypothesis, we expect T to have smaller values because of a lack of complete mixing of the two samples but when the two parent distributions are identical we expect to have larger values. Hence smaller values of T are significant.

Some exact properties of our test are studied in Chapter 2, while the asymtotic distributions of T_i and T are in Chapter 3. In Chapter 4, Monte Carlo simulation study is done to compute the power of our test and compare it with Schilling's test (k = 1, . . ., 3); and then the power of both tests are compared to the theoretical power of Hotelling's T^2 test. In Chapter 5, a subset of Fisher's Iris data is used to apply the three tests. Finally, Chapter 6 contains a summary of the research and some topics for future research.

CHAPTER II

A MULTIVARIATE TEST FOR HOMOGENEITY BASED ON NEAREST NEIGHBORS

2.1 Introduction

In this Chapter we will define our test and the notations used in its calculation. Also, we will give a numerical example to illustrate how the test can be calculated. Then we will describe the test precisely by obtaining certain results concerning its exact distribution, and we will show how it is related to the well-known Wilcoxon-Mann-Whitney statistic.

2.2 Notation and test statistic

Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be independent random samples in R^d from unknown distributions $F(x)$ and $G(x)$, respectively, with corresponding continuous densities $f(x)$ and $g(x)$. The problem under consideration is to test the hypothesis H_0: $F(x) = G(x)$ against the completely general alternative H_a: $F(x) \neq G(x)$.

Construct the combined sample $Z_1, \ldots, Z_N$, where $N = n + m$, such that

$$Z_i = \begin{cases} X_i & i = 1, \ldots, n, \\ Y_{i-n} & i = n+1, \ldots, N. \end{cases}$$

Let $\| \cdot \|$ be the Euclidean norm, and define the k^{th} nearest neighbor to Z_i as that point Z_j satisfying $\| Z_{j'} - Z_i \| < \| Z_j - Z_i \|$ for exactly $(k-1)$ values of j' $(1 \leq j' \leq N,\ j' \neq i, j)$. Ties are neglected, since they occur with probability zero. Define also

$$h(i, k) = \begin{cases} 0 & \text{if } Z_i \text{ and its } k^{th} \text{ nearest neighbor, } Z_j, \text{ are from the same sample,} \\ 1 & \text{otherwise,} \end{cases}$$

and finally, for $i = 1, \ldots, N$ and $k = 1, \ldots, N-1$, define

$$T_{ik} = \sum_{j=1}^{k} h(i, j).$$

We will consider testing

$$H_0: F(x) = G(x) \quad \text{against} \quad H_a: F(x) \neq G(x)$$

by using the statistic

$$T = \sum_{i=1}^{N} \sum_{k=1}^{N-1} T_{ik}.$$

One would expect T to achieve a smaller value under H_a than under H_0 because of a lack of complete mixing of the two samples when the parent distributions are not identical; hence small values of T are significant.

Example:

The purpose of this example is to illustrate how the statistic T can be computed.

Let $X_1 = \begin{pmatrix} 3 \\ 1 \\ 9 \end{pmatrix}$, $X_2 = \begin{pmatrix} 2 \\ 5 \\ 8 \end{pmatrix}$, and $X_3 = \begin{pmatrix} 4 \\ 6 \\ 1 \end{pmatrix}$ be the first sample and

$Y_1 = \begin{pmatrix} 5 \\ 9 \\ 4 \end{pmatrix}$, $Y_2 = \begin{pmatrix} 1 \\ 10 \\ 6 \end{pmatrix}$, $Y_3 = \begin{pmatrix} 2 \\ 3 \\ 5 \end{pmatrix}$, and $Y_4 = \begin{pmatrix} 4 \\ 8 \\ 2 \end{pmatrix}$ be the second sample.

Then the combined sample is

$$Z_1, Z_2, \ldots, Z_7$$

where

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

Then for each i=1, . . ., 7 we calculate $||Z_j-Z_i||$, j=1, . . ., 7, $j \neq i$. The combined ordered arrangement of $||Z_j-Z_i||$ from smallest to largest will give us the k^{th} nearest neighbor, Z_j, to Z_i, k=1, . . ., 6. The k^{th} nearest neighbor to Z_i and the corresponding value of h(i,k) are given in Table 2.2.1. The values of $T_{ik} = \sum_{j=1}^{k} h(i,j)$, k=1,. . ., 6, $T_{i.}=\sum_{k=1}^{6} T_{ik}$, i=1, . . ., 7, $T_{.k}=\sum_{i=1}^{7} T_{ik}$, and the value of the test statistic T, are given in Table 2.2.2. Note that $T_{.k}$ is the number of instances in which a point and its k^{th} nearest neighbor are members of different samples.

Table 2.2.1

k^{th} nearest neighbor [h(i,k)]

i \ k	1	2	3	4	5	6
1	$Z_2[0]$	$Z_6[1]$	$Z_3[0]$	$Z_4[1]$	$Z_5[1]$	$Z_7[1]$
2	$Z_6[1]$	$Z_1[0]$	$Z_5[1]$	$Z_4[1]$	$Z_7[1]$	$Z_3[0]$
3	$Z_7[1]$	$Z_4[1]$	$Z_6[1]$	$Z_5[1]$	$Z_2[0]$	$Z_1[0]$
4	$Z_7[0]$	$Z_3[1]$	$Z_5[0]$	$Z_2[1]$	$Z_6[0]$	$Z_1[1]$
5	$Z_4[0]$	$Z_7[0]$	$Z_2[1]$	$Z_3[1]$	$Z_6[0]$	$Z_1[1]$
6	$Z_2[1]$	$Z_1[1]$	$Z_3[1]$	$Z_7[0]$	$Z_4[0]$	$Z_5[0]$
7	$Z_3[1]$	$Z_4[0]$	$Z_5[0]$	$Z_6[0]$	$Z_2[1]$	$Z_1[1]$

Table 2.2.2

T_{ik} values

i \ k	1	2	3	4	5	6	$T_{i.}$
1	0	1	1	2	3	4	11
2	1	1	2	3	4	4	15
3	1	2	3	4	4	4	18
4	0	1	1	2	2	3	9
5	0	0	1	2	2	3	8
6	1	2	3	3	3	3	15
7	1	1	1	1	2	3	9
$T_{.k}$	4	8	12	17	20	24	85

Result 1 Under H_0, we have

(i) $E[h(i,k)] = \frac{m}{N-1}$

(ii) $V[h(i,k)] = \frac{m(n-1)}{(N-1)^2}$

(iii) $Cov[h(i,k), h(i,k')] = \frac{-m(n-1)}{(N-1)^2(N-2)}$

where $i=1, \ldots, n$ and $k,k'=1, \ldots, N-1$; $k \neq i \neq k'$.

Proof:

Since

$$P[h(i,k)=s] = \begin{cases} \frac{m}{N-1} & \text{if } s=1 \\ \frac{n-1}{N-1} & \text{if } s=0 \\ 0 & \text{otherwise} \end{cases}$$

is the Bernoulli distribution, the mean and variance are

$$E[h(i,k)] = \frac{m}{N-1}$$

$$V[h(i,k)] = \frac{m}{(N-1)} \times \frac{(n-1)}{(N-1)}$$

For the joint moments, we have for $k \neq k'$

$$E[h(i,k),\ h(i,k')] = P[h(i,k) = 1 \cap h(i.k') = 1]$$

$$= \frac{\binom{m}{2}}{\binom{N-1}{2}}$$

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

So that

$$Cov[h(i,k),\ h(i,k')] = \frac{m(m-1)}{(N-1)(N-2)} - \frac{m}{N-1} \times \frac{m}{N-1}$$

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

2.3 The exact distribution of T_{ik}

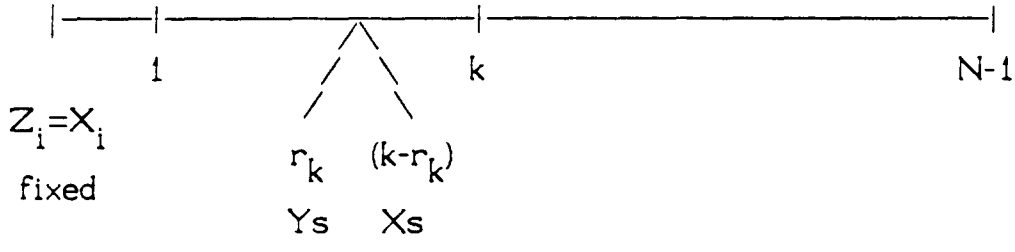
First, we note that for each i, i = 1, . . ., n, T_{ik} assumes nonnegative integer values only, and it is a monotonic nondecreasing function of k.

Result 2 Under H_0, we have

$$P(T_{ik} = r_k) = \frac{\binom{m}{r_k}\binom{n-1}{k-r_k}}{\binom{N-1}{k}},$$

where $\max(0, k-n+1) \le r_k \le \min(k, m)$; $i=1,\ldots,n$; $k=1,\ldots,N-1$.

Proof:



Consider the set of integers $\{ 1, \ldots, N \} - \{i\}$, $i = 1, \ldots, n$. For Z_i calculate $|| Z_i - Z_j ||$; then the result of calculations can be regarded as a permutation of this set of (N-1) integers, since Z_i is fixed. And, under H_0 the probability of having any such permutation is $1/(N-1)!$. Among all (N-1)! permutations, the number of permutations satisfying $\{ T_{ik} = r_k \}$ is the number of ways of choosing r_k Ys out of the m Ys and $(k-r_k)$ Xs out of the (n-1) Xs. This can be done in

$$\binom{m}{r_k}\binom{n-1}{k-r_k}$$

ways. But each way can be permuted k! (N-1-k)! ways in order to have r_k Ys and $(k-r_k)$ Xs in the first k values. Thus the number of ways to have $\{ T_{ik} = r_k \}$ is

$$\binom{m}{r_k}\binom{n-1}{k-r_k} k!\ (N-1-k)!.$$

So,

$$P(T_{ik} = r_k) = \frac{\binom{m}{r_k}\binom{n-1}{k-r_k} k!\ (N-1-k)!}{(N-1)!}$$

$$= \frac{\binom{m}{r_k}\binom{n-1}{k-r_k}}{\binom{N-1}{k}},$$

where

$$\max(0, k-n+1) \le r_k \le \min(k, m),\ i = 1, \ldots, n \text{ and}$$

$$k=1, \ldots, N-1.$$

Corollary Under H_0, we have

(i) $E(T_{ik}) = k \dfrac{m}{N-1}$

(ii) $E(T_{ik}^2) = k\,(k-1)\dfrac{m\,(m-1)}{(N-1)\,(N-2)} + k\dfrac{m}{N-1}$

(iii) $V(T_{ik}) = \dfrac{k\,m}{(N-1)^2\,(N-2)}\,(N-1-m)\,(N-1-k)$

(iv) $E[T_{ik}(T_{ik}-1)\ldots(T_{ik}-r+1)] = \dfrac{r!\begin{pmatrix} m \\ r \end{pmatrix}\begin{pmatrix} k \\ r \end{pmatrix}}{\begin{pmatrix} N-1 \\ r \end{pmatrix}}$

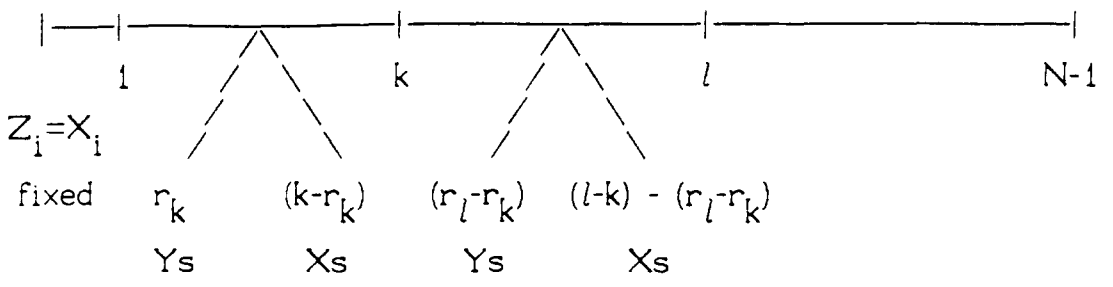
(v) $E[T_{ik}-E(T_{ik})]^3 = k\,\dfrac{m}{N-1}\,\dfrac{N-1-m}{N-1}\,\dfrac{N-1-2m}{N-1}\,\dfrac{N-1-k}{N-2}\,\dfrac{N-1-2k}{N-3}$

Proof: Immediate. (Feller (1966) and Mood, Graybill, and Boes (1974)).

Result 3 Under H_0, we have

$$P(T_{ik} = r_k\,,\ T_{il} = r_l) = \frac{\begin{pmatrix} m \\ r_k \end{pmatrix}\begin{pmatrix} n-1 \\ k-1-r_k \end{pmatrix}\begin{pmatrix} m-r_k \\ r_l-r_k \end{pmatrix}\begin{pmatrix} n-k+r_k \\ l-k-r_l+r_k \end{pmatrix}}{\begin{pmatrix} N-1 \\ k-1 \quad l-k \quad N-l \end{pmatrix}}, \quad l > k$$

Proof:



As in Result 2, fix Z_i and then calculate $||Z_i-Z_j||$, j=1, . . .,N. Among all (N-1)! permutations, the number of permutations satisfying $\{T_{ik}=r_k, T_{il}=r_l\}$ is the number of ways such that we have r_k Ys out of the m Ys and $(k-r_k)$ Xs out of the (n-1) Xs in the first k values providing that the first value is an X. This can be done in

$\binom{m}{r_k}\binom{n-1}{k-r_k}$ ways, and (r_l-r_k) Ys out of the remaining $(m-r_k)$ 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. This can be done in $\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}$ ways. Thus, the number of ways such that $\{T_{ik} = r_k, T_{il} = r_l\}$ is given by

$\binom{m}{r_k}\binom{n-1}{k-r_k}\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}$. But each way can be permuted k! (l-k)! (N-1-l)! times. Thus the number of ways to have $\{T_{ik} = r_k, T_{il} = r_l\}$ is

$$\binom{m}{r_k}\binom{n-1}{k-r_k}\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}\, k!(l-k)!(N-1-l)!$$

So,

$$P(T_{ik} = r_k,\ T_{il} = r_l) = \frac{\binom{m}{r_k}\binom{n-1}{k-r_k}\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}}{\binom{N-1}{k-1 \quad l-k \quad N-1-l}},$$

where

$\max(0,\ k-n+1) \le r_k \le r_l \le \min(l,\ m)$; $i = 1, \ldots, n$; and

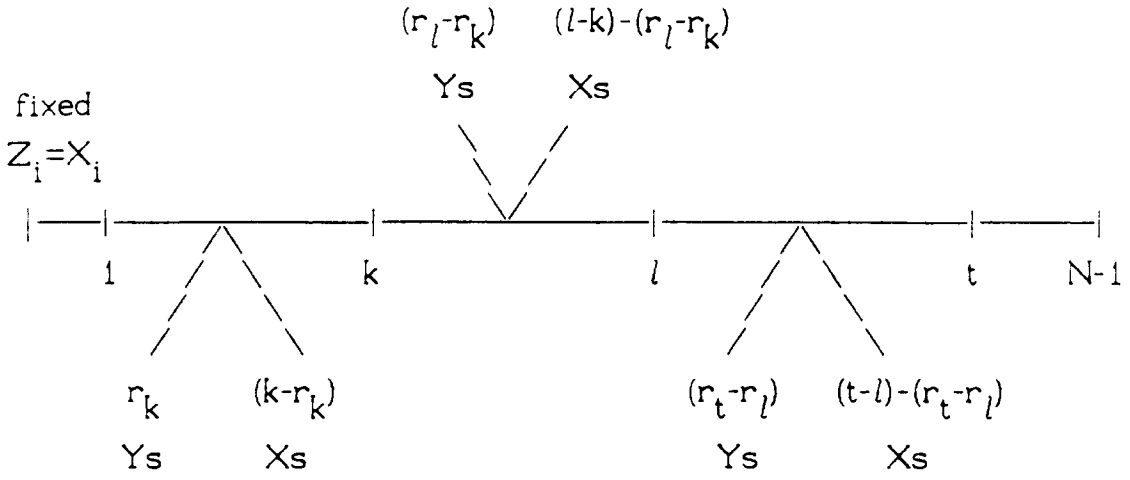
$1 \le k \le l \le N-1$.

Result 4 For $1 \le k < l < t \le N-1$, and under H_0, we have the following

$$P(T_{ik}=r_k,\ T_{il}=r_l,\ T_{it}=r_t) = \binom{m}{r_k}\binom{n-1}{k-r_k}\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}\binom{m-r_l}{r_l-r_k}$$

$$\times \binom{n-1-l+r_l}{t-l-r_t+r_l}\frac{k!(l-k)!(t-l)!(N-1-t)!}{(N-1)!}$$

Proof:



As before, among all (N-1)! permutations, the number of permutations satisfying $\{T_{ik}=r_k, T_{il}=r_l, T_{it}=r_t\}$ is the number of ways such that each way results in r_k Ys out of the m Ys and $(k-r_k)$ Xs out of the (n-1) Xs in the first k values providing that the first value, Z_i, is fixed. This can be done in $\binom{m}{r_k}\binom{n-1}{k-r_k}$ ways, (r_l-r_k) Ys out of the remaining $(m-r_k)$ 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. This can be done in $\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}$ ways, and (r_t-r_l) Ys out of the remaining $(m-r_l)$ Ys and $(t-l-r_t+r_l)$ Xs out of the remaining $(n-1-l+r_l)$ Xs in the following $(t-l)$ values. This can be done in $\binom{m-r_l}{r_t-r_l}\binom{n-1-l+r_l}{t-l-r_t+r_l}$ ways.

Thus the number of ways such that $\{T_{ik}=r_k,\ T_{il}=r_l,\ T_{it}=r_t\}$ is given by

$$\binom{m}{r_k}\binom{n-1}{k-r_k}\binom{m-r_k}{r_l-r_k}\binom{n-1-k+r_k}{l-k-r_l+r_k}\binom{m-r_l}{r_t-r_l}\binom{n-1-l+r_l}{t-l-r_t+r_l}.$$ Denote this by

$P(r_k, r_l, r_t)$. But each way can be permuted $k!(l-k)!(t-l)!(N-1-t)!$ times. Thus the number of ways to have $\{T_{ik}=r_k,\ T_{il}=r_l,\ T_{it}=r_t\}$ is

$$P(r_k, r_l, r_t)\ k!\ (l-k)!\ (t-l)!\ (N-1-t)!$$

Therefore, for $1 \le k < l < t \le N-1$, and

$$\max(0, k-n+1) \le r_k \le r_l \le r_t \le \min(t, m),$$

we have

$$P(T_{ik}=r_k, T_{il}=r_l, T_{it}=r_t) = \frac{P(r_k, r_l, r_t)}{\binom{N-1}{k-1 \quad l-k \quad t-l \quad N-1-t}}.$$

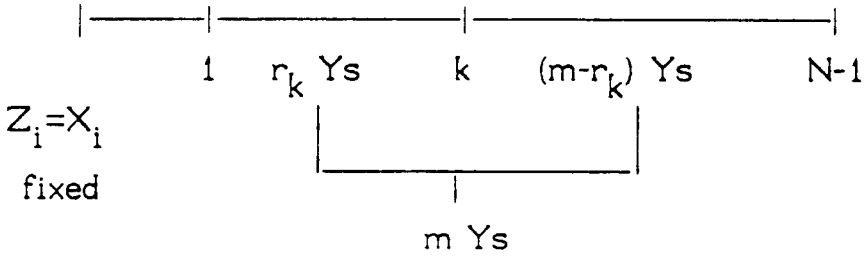
The following is another argument to prove the previous results. By some algebra it can be shown that each of the previous results, 2, 3, and 4, is equivalent to the corresponding one of the following results, 2′, 3′, and 4′. The proofs of these results are similar to that used by Salama and Quade (1981, 1982).

Result 2′ Under H_0, we have

$$P(T_{ik}=r_k) = \frac{\binom{k}{r_k}\binom{N-1-k}{m-r_k}}{\binom{N-1}{m}},$$

where $\max(0, k-n+1) \le r_k \le \min(k, m)$; $i = 1, \ldots, n$; and $k = 1, \ldots, N-1$.

Proof:



In order to have $\{T_{ik}=r_k\}$, the first k values must include r_k Ys and the remaining (N-1-k) values include the remaining $(m-r_k)$ Ys. But the first value, Z_i, must be an X, so to satisfy the condition $\{T_{ik}=r_k\}$ we need to select r_k Ys out of the first k values following Z_i and $(m-r_k)$ Ys out of the remaining (N-1-k) values. Thus the total number of ways to select the Ys such that $\{ T_{ik} = r_k \}$ is

$\binom{k}{r_k}\binom{N-1-k}{m-r_k}$ ways. But the m Ys can be selected from the combined

sample, (N-1) Xs and Ys, in $\binom{N-1}{m}$ ways, since the first X is fixed.

So,

$$P(T_{ik}=r_k) = \frac{\binom{k}{r_k}\binom{N-1-k}{m-r_k}}{\binom{N-1}{m}}.$$

Result 3′ Under H_0, we have

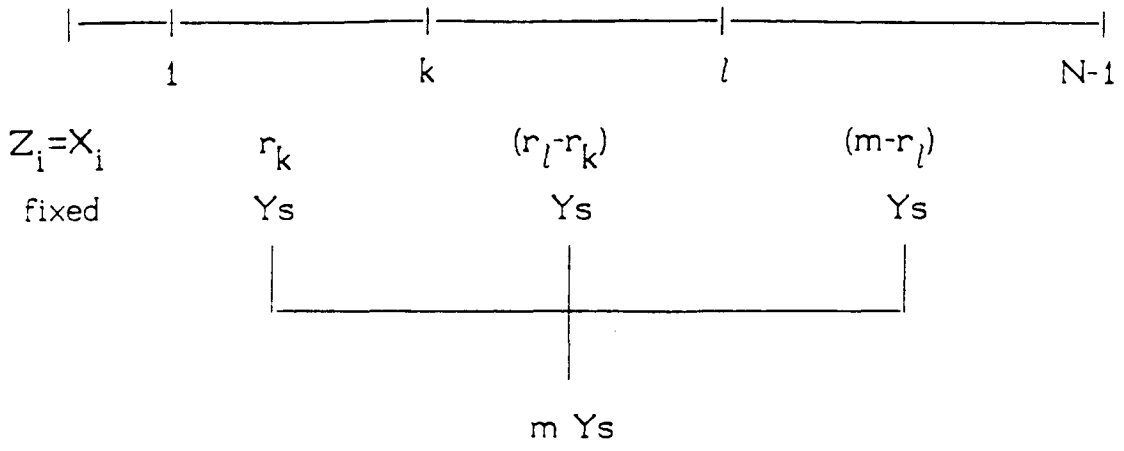
$$P(T_{ik}=r_k,\ T_{il}=r_l) = \frac{\binom{k}{r_k}\binom{l-k}{r_l-r_k}\binom{N-1-l}{m-r_l}}{\binom{N-1}{m}},$$

where $1 \le k \le l \le N-1$; $i = 1, \ldots, n$; and

$$\max(0,k-n+1) \le r_k \le r_l \le \min(l, m).$$

Proof:

As before, in order to have $\{T_{ik}=r_k,\ T_{il}=r_l\}$ the first (k-1) values must have r_k Ys, the next $(l-k)$ values must have (r_l-r_k) Ys, and the last $(N-l)$ will have the remaining $(m-r_l)$ Ys as illustrated in the following figure:



Thus the number of ways to select the Ys such that $\{T_{ik}=r_k,\ T_{il}=r_l\}$ is given by $\binom{k}{r_k}\binom{l-k}{r_l-r_k}\binom{N-1-l}{m-r_l}$. But the number of ways to choose the m Ys out of the combined sample providing that Z_i is fixed is $\binom{N-1}{m}$.

So,

$$P(T_{ik}=r_k,\ T_{il}=r_l) = \frac{\binom{k}{r_k}\binom{l-k}{r_l-r_k}\binom{N-1-l}{m-r_l}}{\binom{N-1}{m}},$$

where

$$1 \le k < l \le N-1, \text{ and}$$

$$\max(0,\ k-n+1) \le r_k \le r_l \le \min(l,\ m).$$

Result 4′ Under H_0, we have

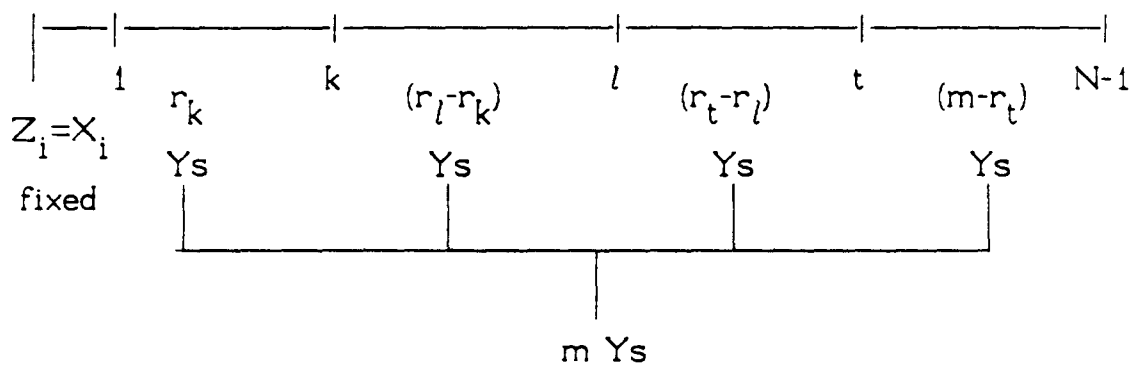
$$P(T_{ik}=r_k,\ T_{il}=r_l, T_{it}=r_t) = \frac{\binom{k}{r_k}\binom{l-k}{r_l-r_k}\binom{t-l}{r_t-r_l}\binom{N-1-t}{m-r_t}}{\binom{N-1}{m}},$$

where

$$\max(0,\ k-n+1) \le r_k \le r_l \le r_t \le \min(t,\ m);\ \text{and}$$

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

Proof:



As in results 2′ and 3′, in order to have $\{T_{ik}=r_k,\ T_{il}=r_l,\ T_{it}=r_t\}$ the first k values of the (N-1) values in the ordered combined sample, after we fix Z_i to be the first value, must have r_k Ys, the next $(l-k)$ values must have (r_l-r_k) Ys, the following $(t-l)$ values must have (r_t-r_l) Ys, and the remaining $(m-r_t)$ Ys must be in the last (N-1-t) values of the ordered arrangement of the combined two samples.

Thus the number of ways of choosing the m Ys such that

$\{T_{ik}=r_k,\ T_{il}=r_l,\ T_{it}=r_t\}$ is given by $\binom{k}{r_k}\binom{l-k}{r_l-r_k}\binom{t-l}{r_t-r_l}\binom{N-1-t}{m-r_t}$.

But the m Ys can be choosen out from the (N-1) Xs and Ys in

$\binom{N-1}{m}$ ways.

So,

$$P(t_{ik}=r_k,\ T_{il}=r_l,\ T_{it}=r_t) = \frac{\binom{k}{r_k}\binom{l-k}{r_l-r_k}\binom{t-l}{r_t-r_l}\binom{N-1-t}{m-r_t}}{\binom{N-1}{m}},$$

where

$\max(0,\ k-n+1) \le r_k \le r_l \le r_t \le \min(t,\ m)$, and

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

The following property of the multivariate hypergeometric distribution (Bishop, Fienberg, and Holland (1976)) will be used to derive the following results in this section.

If $X_1, X_2, \ldots, X_t$ have the multivariate hypergeometric distribution, i.e.,

$$P[X_1=x_1, X_2=x_2, \ldots, X_t=x_t] = \frac{\prod_{i=1}^{t} \binom{N_i}{x_i}}{\binom{N}{n}},$$

where

$0 \le x_i \le N_i$, $i=1, \ldots, t$, $\sum_i N_i = N$, and $\sum_i x_i = n$,

then the factorial moments can be computed as

$$E\left[\prod_{j=1}^{t} X_j^{(r_j)}\right] = \left[\frac{n^{(\sum r_j)}}{N^{(\sum r_j)}}\right] \prod_i N_i^{(r_i)},$$

where $X^{(r)} = X(X-1)(X-2) \ldots (X-r+1)$.

Now, by taking $N_1=k$, $N_2=l-k$, $N_3=t-l$, and $N_4=N-1-t$, $X_1=T_{ik}$, $X_2=T_{il}-T_{ik}$, $X_3=T_{it}-T_{il}$, and $X_4=T_{i,N-1}-T_{it}$, and $x_1=r_k$, $x_2=r_l-r_k$, $x_3=r_t-r_l$, and $x_4=m-r_t$, we have the following result:
(Note that $\sum N_i=N-1$ and $\sum x_i=m$).

Result 5 Under H_0, we have

$$E[T_{ik}(T_{ik}-1)\ldots(T_{ik}-r_1+1)(T_{il}-T_{ik})(T_{il}-T_{ik}-1)\ldots(T_{il}-T_{ik}-r_2+1)$$

$$\times (T_{it}-T_{il})(T_{it}-T_{il}-1)\ldots(T_{it}-T_{il}-r_3+1)]$$

$$= \frac{m(m-1)\ldots[m-(r_1+r_2+r_3)+1]}{(N-1)(N-2)\ldots[N-1-(r_1+r_2+r_3)+1]}\,[k(k-1)\ldots(k-r_1+1)(l-k)$$

$$\times (l-k-1)\ldots(l-k-r_2+1)(t-l)(t-l-1)\ldots(t-l-r_3+1)]$$

where i=1, . . ., n.

Result 6 Under H_0, we have

(i) $$E[T_{ik}(T_{il}-T_{ik})] = \frac{m(m-1)}{(N-1)(N-2)}\,k(l-k)$$

(ii) $$E[T_{ik}(T_{il}-T_{ik})(T_{it}-T_{il})] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)}\,k(l-k)(t-l)$$

(iii) $$E[T_{ik}^2(T_{it}-T_{il})] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)}\,k(k-1)(t-l)$$

(iv) $$E[T_{ik}(T_{il}-T_{ik})^2] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)}\,k(l-k)(l-k-1)$$

(v) $$E[T_{ik}^{2}(T_{il}-T_{ik})] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(k-1)(l-k)$$

(vi) $$E[T_{ik}^{3}] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(k-1)(k-2)$$

Proof: Immediate.

Result 7 Under H_0, we have

(i) $$E[T_{ik}T_{il}] = \frac{m(m-1)}{(N-1)(N-2)} k(l-1) + \frac{m}{(N-1)} k$$

(ii) $$E[T_{ik}^{2}T_{il}] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(k-1)(l-2)$$

(iii) $$E[T_{ik}T_{il}^{2}] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(l-1)(l-2)$$

(iv) $$E[T_{ik}T_{il}T_{it}] = \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(l-1)(t-2)$$

where $1 \le k \le l \le t \le (N-1)$.

Proof:

(i) $E[T_{ik}T_{il}] = E[T_{ik}(T_{il}-T_{ik})]+E[T_{ik}^2]$

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

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

(ii) $E[T_{ik}^2T_{il}] = E[T_{ik}^2(T_{il}-T_{ik})] + E[T_{ik}^3]$

$$= \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} [k(k-1)(l-k) + k(k-1)(k-2)]$$

$$= \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(k-1)(l-2)$$

(iii) $E[T_{ik}^2T_{il}^2] = E[T_{ik}(T_{il}-T_{ik})^2] - E[T_{ik}^3] + 2E[T_{ik}^2T_{il}]$

$$= \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} [k(l-k)(l-k-1) - k(k-1)(k-2) + 2k(k-1)$$

$$\times(l-2)]$$

$$= \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)} k(l-1)(l-2)$$

(iv) $E[T_{ik}T_{il}T_{it}] = E[T_{ik}(T_{il}-T_{ik})(T_{it}-T_{il})]+E[T_{ik}T_{il}^2]+E[T_{ik}^2(T_{it}-T_{il})]$

$$= \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)}[k(l-k)(t-l)+k(l-1)(l-2)+k(k-1)(t-l)]$$

$$= \frac{m(m-1)(m-2)}{(N-1)(N-2)(N-3)}k(l-1)(t-2)$$

Result 8 Under H_0, we have

$$Cov(T_{ik},T_{il}) = \frac{mk(n-1)(N-l-1)}{(N-1)^2(N-2)}, \qquad k < l.$$

Proof:

$$Cov(T_{ik}, T_{il}) = Cov[T_{ik}, (T_{il}-T_{ik})+T_{ik}]$$

$$= Cov[T_{ik}, (T_{il}-T_{ik})] + V(T_{ik}).$$

From the properties of the multivariate hypergeometric distribution

$$Cov[T_{ik}, (T_{il}-T_{ik})] = -\frac{mk(l-k)(N-1-m)}{(N-1)^2(N-2)}.$$

Therefore,

$$\mathrm{Cov}(T_{ik}, T_{il}) = -\frac{mk(l-k)(N-1-m)}{(N-1)^2(N-2)} + \frac{mk(N-1-k)(N-1-m)}{(N-1)^2(N-2)}$$

$$= \frac{mk(n-1)(N-1-l)}{(N-1)^2(N-2)} .$$

Note: The previous results were derived for i=1, . . ., n. The same proof can be used to derive the results for i=n+1, . . ., N. This can be done by interchanging m and n.

2.4 The exact distribution of T_i

The statistic $T_i = \sum_{k=1}^{N-1} T_{ik}$ has a minimum value of $\frac{m(m+1)}{2}$ when we have the following combined ordered arrangement of the two samples with $Z_i=X_i$ fixed as the first variable, i.e., i=1, . . ., n.

$$\underbrace{X\,X\,\ldots\,X}_{(n-1)\text{ values}}\;\underbrace{Y\,Y\,\ldots\,Y}_{m\text{ values}}$$

where X in the k^{th} position means that the k^{th} nearest neighbor to Z_i is an X and Y means that it is a Y. But T_i has a maximum value of $\frac{m(m+2n-1)}{2}$ when we have the following combined ordered arrangement of the two samples with Z_i as the fixed variable

$$\underbrace{Y\,Y\,\ldots\,Y}_{m\text{ values}}\;\underbrace{X\,X\,\ldots\,X}_{(n-1)\text{ values}}$$

Similarly, the test statistic T_i has a minimum value of $\frac{n(n+1)}{2}$ when we have the following ordered arrangement of the two samples with $Z_i=Y_i$ fixed as the first variable, i.e., i=n+1, . . ., N.

$$\underbrace{Y\,Y\ldots Y}_{(m-1)\text{ values}}\;\underbrace{X\,X\ldots X}_{n\text{ values}}$$

But the maximum value of T_i, which is equal to $\dfrac{n(n+2m-1)}{2}$, occurs when we have the following ordered arrangement

$$\underbrace{X\,X\ldots X}_{n\text{ values}}\;\underbrace{Y\,Y\ldots Y}_{(m-1)\text{ values}}$$

So, for i=1, . . ., N, we have

$$\min\left[\frac{m(m+1)}{2}, \frac{n(n+1)}{2}\right] \le T_i \le \max\left[\frac{m(m+2n-1)}{2}, \frac{n(n+2m-1)}{2}\right].$$

Result 9 Under H_0, we have

(i) $E(T_i) = \dfrac{mN}{2}$

(ii) $V(T_i) = \dfrac{mN(n-1)}{12}$

where i = 1, . . ., n

Proof:

(i) $$E(T_i) = E\left[\sum_{k=1}^{N-1} T_{ik}\right]$$

$$= \sum_{k=1}^{N-1} E[T_{ik}]$$

$$= \frac{m}{N-1} \sum_{k=1}^{N-1} k$$

$$= \frac{mN}{2}$$

(ii) $$V(T_i) = V\left[\sum_{k=1}^{N-1} T_{ik}\right]$$

$$= V\left[\sum_{k=1}^{N-1} \sum_{j=1}^{k} h(i,j)\right]$$

$$= V\left[\sum_{j=1}^{N-1} (N-j)h(i,j)\right]$$

$$V(T_i) = V\left[\sum_{j=1}^{N-1} a_j h(i,j)\right], \qquad \text{where } a_j = N-j$$

$$= \sum_{j=1}^{N-1} a_j^2 V[h(i,j)] + \sum\sum_{j \neq j'} a_j a_{j'} Cov[h(i,j), h(i,j')]$$

$$= \frac{m(n-1)}{(N-1)^2} \sum_j a_j^2 - \frac{m(n-1)}{(N-1)^2(N-2)} \sum\sum_{j \neq j'} a_j a_{j'}$$

$$= \frac{m(n-1)}{(N-1)^2(N-2)}\left[(N-2)\sum_j a_j^2 - \sum\sum_{j \neq j'} a_j a_{j'}\right]$$

$$= \frac{m(n-1)}{(N-1)^2(N-2)}\left[(N-1)\sum_j a_j^2 - (\sum_j a_j)^2\right]$$

$$= \frac{m(n-1)}{(N-1)^2(N-2)}\left[\frac{(N-1)^2 N(2N-1)}{6} - \frac{N^2(N-1)^2}{4}\right]$$

$$V(T_i) = \frac{m(n-1)N}{12(N-2)}[2(2N-1) - 3N]$$

$$= \frac{m(n-1)N}{12}$$

Similarly, for i=n+1, . . ., N, we have

(i) $E(T_i) = \frac{nN}{2}$, and

(ii) $V(T_i) = \frac{n(m-1)N}{12}$.

Now, we rewrite the statistic T_i in another form in order to find its exact distribution. Choose the first variable to be Z_i (fixed), i=1, . . ., N, then calculate $||Z_j - Z_i||$, j=1, . . ., N, j≠i. Then the combined ordered arrangement of the two samples can be denoted by a vector of indicator random variables Z_{ik}, where $Z_{ik}=1$ if the point Z_i and its k^{th} nearest neighbor, Z_j, belong to different samples and $Z_{ik}=0$ if both points belong to the same sample, for k = 1, . . ., N-1, with N=n+m. The rank of the observation for which Z_{ik} is an indicator is k, and therefore the vector $\mathbf{Z_i}$ indicates the rank-order statistic of the combined ordered arrangement of the two samples and in addition identifies the sample to which each observation belongs.

T_i can be easily expressed in terms of this notation. This kind of statistic is called a linear rank statistic, (Gibbons (1985)) which is defined as

$$T_{N-1}(\mathbf{Z_i}) = \sum_{k=1}^{N-1} a_k Z_{ik}$$

where the a_k are given numbers. For example, a_k can be taken as a weight or score. Note that $T_{N-1}(\mathbf{Z_i})$ is linear in the indicator variables but no similar restriction on the constants is implied.

Result 10 Under H_0, we have

T_i is symmetric about its mean $\mu = mN/2$, i=1, . . ., n.

Proof:

One of the properties of the linear rank statistic

$$T_{N-1}(\mathbf{Z_i}) = \sum_{j=1}^{N-1} a_j Z_{ij}$$

is that it is symmetric about its mean whenever

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

But T_i can be written as

$$T_i = \sum_{k=1}^{N-1} \sum_{j=1}^{k} h(i,j)$$

$$T_i = \sum_{k=1}^{N-1} \sum_{j=1}^{k} Z_{ij}$$

$$= \sum_{j=1}^{N-1} (N-j)\, Z_{ij}$$

Take $a_j = N-j$, then $a_{N-j} = j$.

Therefore,

$$a_j + a_{N-j} = N = \text{constat.}$$

Hence,

T_i is symmetric about its mean.

Note that, for $i=n+1, \ldots, N$, T_i is symmetric about its mean $nN/2$.

The exact null probability of T_i, $i=1, \ldots, N$, can be obtained systematically by enumeration using the previous results. For example, suppose $n=3$ and $m=4$, then, for $i=1, 2, 3$, there are $\binom{6}{4}=15$ possible distinguishable configurations of 1s and 0s in the vector $\mathbf{Z_i}$, but these need not be enumerated individually. T_i ranges between 10 and 18, and is symmetric about 14. But for $i=4, \ldots, 7$, there are $\binom{6}{3}=20$ possible distinguishable configurations of 1s and 0s in the

vector $\mathbf{Z_i}$. T_i ranges between 6 and 15, and is symmetric about 10.5.

So, for i=1, . . ., 7, the total possible distinguishable configurations of 1s and 0s in the vector $\mathbf{Z_i}$ are 35. T_i ranges between 6 and 18. The values occurring in conjunction with the combined ordered arrangement of the two samples are in Table 2.4.1, from which the complete probability distribution is easily found.

Since each possible combined ordered arrangement of the two samples occurs with probability $\dfrac{1}{\binom{N-1}{m} + \binom{N-1}{n}}$, the exact null probability distribution of any linear statistic, T_i(say), can always be found by direct enumeration. The values of T_i are calculated for each possible combined ordered arrangement of the two samples, and the probability of a particular value r is the number of combined ordered arrangements of the two samples which lead to that number, divided by $\left[\binom{N-1}{m} + \binom{N-1}{n} \right]$ (Randles and Wolfe (1979)).

Table 2.4.1

Probability distribution of T_i

n=3 and m=4

Value of T_i	Combined ordered arrangement the two samples	Frequency f_i	Probability p_i
6	000111	1	1/35
7	001011	1	1/35
8	001101; 010011	2	2/35
9	100011; 001110; 010101	3	3/35
10	001111; 100101; 011001; 010110	4	4/35
11	010111; 100110; 101001; 011010	4	4/35
12	011011; 100111; 110001; 101010; 011100	5	5/35
13	011101; 101011; 110010; 101100	4	4/35
14	011110; 110011; 101101; 110100	4	4/35
15	110101; 101110; 111000	3	3/35
16	111001; 110110	2	2/35
17	111010	1	1/35
18	111100	1	1/35

For larger samples generation of the exact probability distribution is rather irksome. However, for $n \rightarrow \infty$ and $m \rightarrow \infty$ in such a way that m/n remains constant, an approximation exists which is applicable to the distribution of almost all linear rank statistics. Since T_i is a linear combination of the Z_{ik}, which are identically distributed (though dependent) random variables, a generalization of the central limit theorem allows us to conclude that the probability distribution of the standarized linear rank statistic

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

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

The following illustrates the relation between T_i and the sum of ranks of the nearest neighbors to Z_i in the combined ordered arrangement of the two samples when arranged from largest to smallest in magnitude of distance from Z_i.

Suppose we have the following combined ordered arrangement of the two samples with Z_i as the fixed variable.

k :	1	2	3	4	5	6
k^{th} nearest to Z_i :	Y	X	Y	Y	Y	X
$Z_{ik}=h(i,k)$:	1	0	1	1	1	0
T_{ik} :	1	1	2	3	4	4
R_{ik} :	6	5	4	3	2	1

where R_{ik} is the rank of the k^{th} nearest neighbor to Z_i when the combined arrangement of the two samples is ordered from largest to smallest. Then

$$T_{i.} = \sum_{k=1}^{6} T_{ik} = 15 = R_{i.} = \sum_{k=1}^{6} R_{ik}$$

Result 11 Under H_0, we have

Under H_0, we have for each i $(1 \le i \le N)$, $T_i = \sum_{k=1}^{N-1} T_{ik}$ has the Wilcoxon-Mann-Whitney distribution.

Proof:

Since $Z_{ik} = h(i, k)$, where h(i,k) is defined in Section 2.2 and Z_{ik} is defined at the beginning of this section,

$$T_{ik} = \sum_{j=1}^{k} h(i, j) = \sum_{j=1}^{k} Z_{ij}$$

and

$$T_i = \sum_{k=1}^{N-1} T_{ik} = \sum_{k=1}^{N-1} \sum_{j=1}^{k} Z_{ij}$$

If we take $a_j = (N-j)$ then the linear rank statistic $T_{N-1}(\mathbf{Z_i})$ can be written in terms of T_{ik} as follows:

$$T_{N-1}(\mathbf{Z_i}) = \sum_{k=1}^{N-1} (N-k)Z_{ik} = (N-1)Z_{i1} + (N-2)Z_{i2} + \ldots + 2Z_{iN-2} + Z_{iN-1}$$

$$= Z_{i1} + (Z_{i1}+Z_{i2}) + (Z_{i1}+Z_{i2}+Z_{i3}) + \ldots$$

$$+ (Z_{i1}+Z_{i2}+Z_{i3}+\ldots+Z_{iN-2})$$

$$+ (Z_{i1}+Z_{i2}+Z_{i3}+\ldots+Z_{iN-1})$$

$$= T_{i1} + T_{i2} + T_{i3} + \ldots + T_{iN-1}$$

$$= \sum_{k=1}^{N-1} T_{ik}$$

$$= T_i$$

Therefore,

$$T_i = \sum_{j=1}^{N-1} (N-j)\, Z_{ij}$$

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

$$= \begin{cases} mN - W_i, & \text{if } i = 1, \ldots, n, \\ \\ nN - W_i, & \text{if } i = n+1, \ldots, N \end{cases}$$

where, $W_i = \sum_{j=1}^{N-1} jZ_{ij}$ is the Wilcoxon rank sum statistic. So,

T_i has a Wilcoxon-Mann-Whitney distribution.

Thus T_i is actually the same as the Wilcoxon-Mann-Whitney rank sum test, since a linear relationship exists between the two test ststistics. Therefore, all the properties of the tests are the same, including consistency and the minimum ARE of 0.864 (Gibbons (1985)) relative to the t test.

Note: To rewrite the previous results for $i = 1, \ldots, N$, the indicator function $I(.)$ will be used. Let $\Omega_1 = \{ 1, 2, \ldots, n \}$, $\Omega_2 = \{ n+1, n+2, \ldots, N \}$, $n_1 = n$, $n_2 = m$, and $N = n + m$. For example, under H_0, the first two moments of T_{ik} and T_i can be written as:

$$\text{(i)} \quad E(T_{ik}) = \sum_{\alpha=1}^{2} \frac{k n_{3-\alpha}}{N-1} I\{ i \in \Omega_\alpha \}$$

$$\text{(ii)} \quad V(T_{ik}) = \sum_{\alpha=1}^{2} \frac{k n_{3-\alpha}}{(N-1)^2 (N-2)} (n_\alpha - 1)(N-1-k) I\{ i \in \Omega_\alpha \}$$

$$\text{(iii)} \quad Cov(T_{ik}, T_{il}) = \sum_{\alpha=1}^{2} \frac{k n_{3-\alpha} (n_\alpha - 1)(N-1-l)}{(N-1)^2 (N-2)} I\{ i \in \Omega_\alpha \}, \qquad k < l$$

$$\text{(iv)} \quad E(T_i) = \sum_{\alpha=1}^{2} \frac{n_{3-\alpha} N}{2} I\{ i \in \Omega_\alpha \}$$

$$\text{(v)} \quad V(T_i) = \sum_{\alpha=1}^{2} \frac{n_{3-\alpha} (n_\alpha - 1) N}{12} I\{ i \in \Omega_\alpha \}$$

where $k = 1, \ldots, N-1$.

2.5 The mean and variance of T

Result 13 Under H_0, we have

(i) $E(T) = mnN$

(ii) $$V(T) = \frac{mnN}{12(N-2)}[(3N-1)(4mn-N^2) + 4(mn+1-N)]$$

$$+ \frac{4mn(m-1)(n-1)}{(N-2)(N-3)} \sum_{k=1}^{N-1} \sum_{l=1}^{N-1} (N-k)(N-l)P_1(k,l)$$

$$- \frac{mn}{(N-2)(N-3)}[(N-2)(N-3) - 4(m-1)(n-1)] \sum_{k=1}^{N-1} \sum_{l=1}^{N-1} (N-k)$$

$$\times (N-l)\, P_2(k,\, l)$$

where $P_1(k,l) = P(k^{th}$ nearest neighbor to $Z_i = Z_j$ and l^{th} nearest neighbor to $Z_j = Z_i)$,

$P_2(k,l) = P(k^{th}$ nearest neighbor to $Z_i = l^{th}$ nearest neighbor to $Z_j)$,

and $i \neq j = 1, 2, \ldots, N$.

Proof:

(i) $$E(T) = E\left[\sum_{i=1}^{N} \sum_{k=1}^{N-1} T_{ik}\right]$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{N-1} E(T_{ik})$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{N-1} \sum_{\alpha=1}^{2} \frac{k n_{3-\alpha}}{N-1} I\{ i \epsilon \Omega_{\alpha} \}$$

$$= \sum_{i=1}^{N} \sum_{k=1}^{N-1} \frac{k}{N-1} [m I\{i \epsilon \Omega_1\} + n I\{i \epsilon \Omega_2\}]$$

$$= \frac{2mn}{N-1} \sum_{k=1}^{N-1} k$$

$$= mnN.$$

(ii) $$V(T) = V\left(\sum_{i=1}^{N} T_i \right)$$

$$= \sum_{i=1}^{N} V(T_i) + \sum_{i \neq j}^{N} \sum^{N} Cov(T_i, T_j)$$

$$= \underbrace{\sum_{i}^{N} V(T_i)}_{(a)} - \underbrace{\sum_{i \neq j}^{N} \sum^{N} E(T_i) E(T_j)}_{(b)} + \underbrace{\sum_{i \neq j}^{N} \sum^{N} E(T_i T_j)}_{(c)}$$

But

$$\text{(a)}\quad \sum_{i=1}^{N} V(T_i) = \sum_{i=1}^{N} \sum_{\alpha=1}^{2} \frac{Nn_{3-\alpha}(n_\alpha - 1)}{12} I\{i \epsilon \Omega_\alpha\}$$

$$= \frac{N}{12} \sum_{i} [m(n-1) I\{i \epsilon \Omega_1\} + n(m-1) I\{i \epsilon \Omega_2\}]$$

$$= \frac{N}{12} [m(n-1)n + n(m-1)m]$$

$$= \frac{mnN(N-2)}{12} .$$

$$\text{(b)}\quad \sum_{i \neq j}^{N} \sum^{N} E(T_i)E(T_j) = \sum_{i \neq j} \sum \left[\sum_{\alpha=1}^{2} \frac{Nn_{3-\alpha}}{2} I\{i \epsilon \Omega_\alpha\} \right] \left[\sum_{\alpha=1}^{2} \frac{Nn_{3-\alpha}}{2} \times I\{j \epsilon \Omega_\alpha\} \right]$$

$$= \frac{N^2}{4} \sum_{i \neq j} \sum \sum_{\alpha}^{2} [n_{3-\alpha} I\{i, j \epsilon \Omega_\alpha\} + n_\alpha n_{3-\alpha} I\{i \epsilon \Omega_\alpha, j \epsilon \Omega_{3-\alpha}\}]$$

$$= \frac{N^2}{4} \sum_{i \neq j} \sum [m^2 I\{i, j \epsilon \Omega_1\} + n^2 I\{i, j \epsilon \Omega_2\}$$

$$+ mnI\{i \epsilon \Omega_1, j \epsilon \Omega_2\} + nmI\{i \epsilon \Omega_2, j \epsilon \Omega_1\}]$$

$$= \frac{N^2}{4} [m^2 n(n-1) + n^2 m(m-1) + m^2 n^2 + n^2 m^2]$$

$$= \frac{mnN^2}{4} (4mn - N).$$

and

$$\text{(c)} \quad \sum_{i \neq j}^{N} \sum^{N} E(T_i T_j) = \sum_{i \neq j}^{N} \sum^{N} \sum_{t=1}^{N-1} \sum_{t'=1}^{N-1} \sum_{k=1}^{t} \sum_{l=1}^{t'} E[h(i, k)h(j, l)]$$

Thus, to find the variance of T, it is necessary to compute $E[h(i, k)h(j, l)]$. When $i \neq j$ various nearest neighbor geometries come into play. There are five mutually exclusive and exhaustive cases:

(1) k^{th} nearest neighbor to $Z_i = Z_j$, l^{th} nearest neighbor to $Z_j = Z_i$,

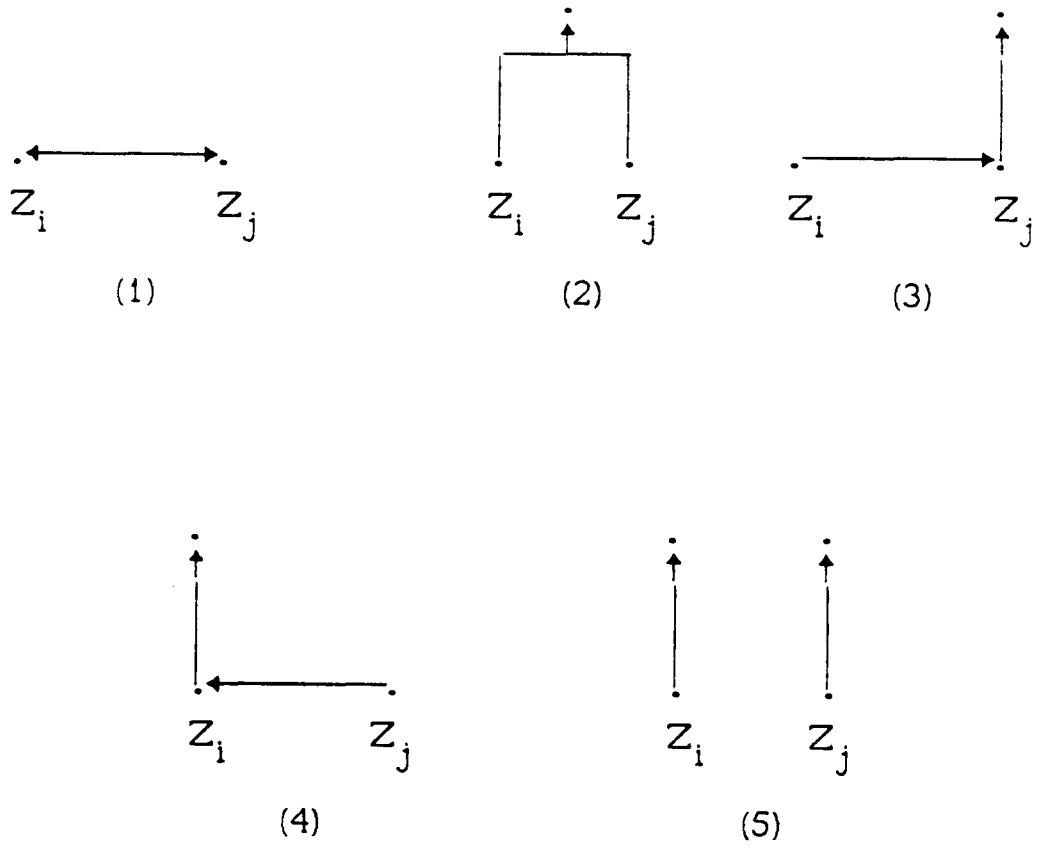
(2) k^{th} nearest neighbor to $Z_i = l^{th}$ nearest neighbor to Z_j,

(3) k^{th} nearest neighbor to $Z_i = Z_j$, l^{th} nearest neighbor to $Z_j \neq Z_i$,

(4) k^{th} nearest neighbor to $Z_i \neq Z_j$, l^{th} nearest neighbor to $Z_j = Z_i$, and

(5) k^{th} nearest neighbor to $Z_i \neq Z_j$, l^{th} nearest neighbor to $Z_j \neq Z_i$, and k^{th} nearest neighbor to $Z_i \neq l^{th}$ nearest neighbor to Z_j.

These five cases can be illustrated by the following figure:



The arrows from Z_i and Z_j point to the k^{th} nearest neighbor to Z_i and the l^{th} nearest neighbor to Z_j respectively.

Let these five events and their respective probabilities be denoted by E_a and $P_a(k, l)$, a=1, 2, . . ., 5. These probabilities are independent of i and j, since the Zs are interchangeable. So, for $i \neq j$, it is required to compute

$$\sum_{a=1}^{5} \{ E[h(i, k)h(j, l)] | E_a \} P_a(k, l).$$

Using the interchangeability of the Z_is the number of distinct conditional expectations is fairly small and these values will be feasible to compute. Thus the determination of the P_as is of paramount importance. Note however that

$$P_1(k, l) = P(k^{th} \text{ nearest neighbor to } Z_1=Z_2, \ l^{th} \text{ nearest neighbor to } Z_2=Z_1)$$

$$= P(l^{th} \text{ nearest neighbor to } Z_2=Z_1 \,|\, k^{th} \text{ nearest neighbor to } Z_1=Z_2) P(k^{th} \text{ nearest neighbor to } Z_1=Z_2)$$

$$= P(l^{th} \text{ nearest neighbor to } Z_2=Z_1 \,|\, k^{th} \text{ nearest neighbor to } Z_1=Z_2) \times \frac{1}{N-1}$$

from which it follows that

$$P_3(k, l) = P(k^{th} \text{ nearest neighbor to } Z_1=Z_2, \ l^{th} \text{ nearest neighbor to } Z_2 \neq Z_1)$$

$$= P(l^{th} \text{ nearest neighbor to } Z_2 \neq Z_1 \mid k^{th} \text{ nearest neighbor to } Z_1 = Z_2) P(k^{th} \text{ nearest neighbor to } Z_1 = Z_2)$$

$$= [1 - (N-1)P_1(k, l)] \times \frac{1}{N-1}$$

$$= \frac{1}{N-1} - P_1(k, l)$$

$$P_4(k, l) = P(k^{th} \text{ nearest neighbor to } Z_1 \neq Z_2, \; l^{th} \text{ nearest neighbor to } Z_2 = Z_1)$$

$$= [1 - (N-1)P_1(k, l)] \times \frac{1}{N-1}$$

$$= \frac{1}{N-1} - P_1(k, l)$$

and

$$P_5(k, l) = P(k^{th} \text{ nearest neighbor to } Z_1 \neq Z_2, \; l^{th} \text{ nearest neighbor to } Z_2 \neq Z_1, \text{ and } k^{th} \text{ nearest neighbor} \neq l^{th} \text{ nearest neighbor})$$

$$= 1 - \sum_{a=1}^{4} P_a(k, l)$$

$$= 1 - P_1(k, l) - P_2(k, l) - \frac{1}{N-1} + P_1(k, l) - \frac{1}{N-1} + P_1(k, l)$$

$$= \frac{N-3}{N-1} + P_1(k, l) - P_2(k, l).$$

Thus, it is only necessary to calculate $P_1(k, l)$ and $P_2(k, l)$ for each k and l.

Therefore,

$$E[h(i, k)h(j, l)] = P[h(i, k)=1 \cap h(j, l)=1]$$

$$= \sum_{a=1}^{5} C_a P_a(k, l)$$

where

$$C_1 = \sum_{\alpha=1}^{2} I\{i\epsilon\Omega_\alpha, j\epsilon\Omega_{3-\alpha}\}$$

$$C_2 = \sum_{\alpha=1}^{2} I\{i, j\epsilon\Omega_\alpha\} \times \frac{n_{3-\alpha}}{N-2}$$

$$C_3 = \sum_{\alpha=1}^{2} I\{i\epsilon\Omega_\alpha, j\epsilon\Omega_{3-\alpha}\} \times \frac{n_\alpha - 1}{N-2}$$

$$C_4 = \sum_{\alpha=1}^{2} I\{i\epsilon\Omega_\alpha, j\epsilon\Omega_{3-\alpha}\} \times \frac{n_{3-\alpha} - 1}{N-2}$$

and

$$C_5 = \sum_{\alpha=1}^{2} \left[I\{i, j\epsilon\Omega_\alpha\} \times \frac{n_{3-\alpha}(n_{3-\alpha} - 1)}{(N-2)(N-3)} + I\{i\epsilon\Omega_\alpha, j\epsilon\Omega_{3-\alpha}\} \times \frac{(n_\alpha - 1)(n_{3-\alpha} - 1)}{(N-2)(N-3)} \right]$$

Therefore,

$$\sum_{i \neq j} \sum E(T_i T_j) = \sum_{i \neq j} \sum_{k} \sum_{l} (N-k)(N-l) P[h(i, k)=1=h(j, l)]$$

$$= \sum_{k} \sum_{l} (N-k)(N-l) \underbrace{\sum_{i \neq j} \sum P[h(i, k)=1=h(j, l)]}_{(d)}$$

$$(d) \quad \sum_{i \neq j} \sum P[h(i, k)=1=h(j, l) = \sum_{i \neq j} \sum_{a} \sum C_a P_a(k, l)$$

$$(d.1) \quad \sum_{i \neq j} \sum C_1 P_1(k, l) = \sum_{i \neq j} \sum_{\alpha} \sum I\{i \epsilon \Omega_\alpha, j \epsilon \Omega_{3-\alpha}\} P_1(k, l)$$

$$= \sum_{i \neq j} \sum [I\{i \epsilon \Omega_1, j \epsilon \Omega_2\} + I\{i \epsilon \Omega_2, j \epsilon \Omega_1\}] P_1(k, l)$$

$$= (nm + mn) P_1(k, l)$$

$$= 2mn P_1(k, l)$$

$$(d.2) \quad \sum_{i \neq j} \sum C_2 P_2(k, l) = \sum_{i \neq j} \sum_{\alpha} \sum I\{i, j \epsilon \Omega_\alpha\} \frac{n_{3-\alpha}}{N-2} P_2(k, l)$$

$$= \left(\frac{n(n-1)m}{N-2} + \frac{m(m-1)n}{N-2} \right) P_2(k, l)$$

$$= mnP_2(k, l)$$

(d.3) $$\sum_{i \neq j} \sum C_3 P_3(k, l) = \sum_{i \neq j} \sum_{\alpha} \sum I\{i \epsilon \Omega_\alpha, j \epsilon \Omega_{3-\alpha}\} \frac{n_\alpha - 1}{N-2} P_3(k, l)$$

$$= \left(\frac{nm(n-1)}{N-2} + \frac{mn(m-1)}{N-2} \right) \left(\frac{1}{N-1} - P_1(k, l) \right)$$

$$= \frac{mn}{N-1} - mnP_1(k, l)$$

(d.4) $$\sum_{i \neq j} \sum C_4 P_4(k, l) = \sum_{i \neq j} \sum_{\alpha} \sum I\{i \epsilon \Omega_\alpha, j \epsilon \Omega_{3-\alpha}\} \frac{n_{3-\alpha} - 1}{N-2} P_4(k, l)$$

$$= \left(\frac{nm(m-1)}{N-2} + \frac{mn(n-1)}{N-2} \right) \left(\frac{1}{N-1} - P_1(k, l) \right)$$

$$= \frac{mn}{N-1} - mnP_1(k, l)$$

$$(d.5)\quad \sum_{i\neq j}\sum C_5 P_5(k,\ l) = \underbrace{\sum_{i\neq j}\sum_{\alpha} I\{1,\ j\epsilon\Omega_\alpha\}\frac{n_{3-\alpha}(n_{3-\alpha}-1)}{(N-2)(N-3)}P_5(k,\ l)}_{A1}$$

$$+\underbrace{\sum_{i\neq j}\sum_{\alpha} I\{i\epsilon\Omega_\alpha,\ j\epsilon\Omega_{3-\alpha}\}\frac{(n_\alpha-1)(n_{3-\alpha}-1)}{(N-2)(N-3)}P_5(k,\ l)}_{A2}$$

$$A1 = \frac{1}{(N-2)(N-3)}\left[n(n-1)m(m-1) + m(m-1)n(n-1)\right]\left[\frac{N-3}{N-1} + P_1(k,\ l) - P_2(k,\ l)\right]$$

$$=\frac{2mn(m-1)(n-1)}{(N-1)(N-2)}+\frac{2mn(m-1)(n-1)}{(N-2)(N-3)}P_1(k,\ l)-\frac{2mn(m-1)(n-1)}{(N-2)(N-3)}P_2(k,\ l)$$

$$A2 = \frac{1}{(N-2)(N-3)}\left[nm(n-1)(m-1) + mn(m-1)(n-1)\right]\left[\frac{N-3}{N-1} + P_1(k,\ l) - P_2(k,\ l)\right]$$

$$=\frac{2mn(m-1)(n-1)}{(N-1)(N-2)}+\frac{2mn(m-1)(n-1)}{(N-2)(N-3)}P_1(k,\ l)-\frac{2mn(m-1)(n-1)}{(N-2)(N-3)}P_2(k,\ l)$$

Therefore,

$$\sum_{i \neq j} \sum C_5 P_5(k, l) = \frac{4mn(m-1)(n-1)}{(N-1)(N-2)} + \frac{4mn(m-1)(n-1)}{(N-2)(N-3)} P_1(k, l)$$

$$- \frac{4mn(m-1)(n-1)}{(N-2)(N-3)} P_2(k, l).$$

Hence, to get the final result of (d), the results of (d.1) - (d.5) will be added together. So,

$$\sum_{i \neq j} \sum P[h(i, k)=1=h(j, l)] = 2mnP_1(k, l) + mnP_2(k, l) + \frac{2mn}{N-1}$$

$$- 2mnP_1(k, l) + \frac{4mn(m-1)(n-1)}{(N-1)(N-2)}$$

$$+ \frac{4mn(m-1)(n-1)}{(N-2)(N-3)} [P_1(k, l) - P_2(k, l)]$$

$$= \frac{2mn}{(N-1)(N-2)} [(N-2) + 2(m-1)(n-1)]$$

$$+ \frac{4mn(m-1)(n-1)}{(N-2)(N-3)} P_1(k, l) - \frac{mn}{(N-2)(N-3)}$$

$$\times [(N-2)(N-3) - 4(m-1)(n-1)] P_2(k, l)$$

Thus,

$$\sum_{i \neq j}^{N}\sum^{N} E(T_i T_j) = \sum_{i \neq j}^{N}\sum^{N} \sum_{t}^{N-1} \sum_{t'}^{N-1} \sum_{k}^{t} \sum_{l}^{t'} E[h(i, k)h(j, l)]$$

$$= \frac{2mn}{(N-1)(N-2)} [(N-2) + 2(m-1)(n-1)] \sum_{k=1}^{N-1} \sum_{l=1}^{N-1} (N-k)(N-l)$$

$$+ \frac{4mn(m-1)(n-1)}{(N-2)(N-3)} \sum_{k=1}^{N-1} \sum_{l=1}^{N-1} (N-k)(N-l)\, P_1(k, l)$$

$$- \frac{mn}{(N-2)(N-3)} [(N-2)(N-3) - 4(m-1)(n-1)] \sum_{k=1}^{N-1} \sum_{l=1}^{N-1} (N-k)(N-l)$$

$$\times P_2(k, l)$$

But

$$\sum_{k=1}^{N-1} \sum_{l=1}^{N-1} (N-k)(N-l) = \left[\sum_{k=1}^{N-1} (N-k) \right]^2$$

$$= \left[\frac{N(N-1)}{2} \right]^2$$

$$= \frac{N^2(N-1)^2}{4}$$

Therefore,

$$V(T) = \frac{mnN(N-2)}{12} - \frac{mnN^2}{4}(4mn - N) + \frac{2mn}{(N-1)(N-2)}[(N-2) + 2(m-1)$$

$$\times (n-1)]\frac{N^2(N-1)^2}{4} + \frac{4mn(m-1)(n-1)}{(N-2)(N-3)}\sum_k \sum_l (N-k)(N-l)\hat{P}_1(k, l)$$

$$- \frac{mn}{(N-2)(N-3)}[(N-2)(N-3) - 4(m-1)(n-1)]\sum_k \sum_l (N-k)(N-l)P_2(k, l)$$

Thus,

$$V(T) = \frac{mnN}{12(N-2)}[(3N-1)(4mn-N^2) + 4(mn + 1 - N)]$$

$$+ \frac{4mn(m-1)(n-1)}{(N-2)(N-3)}\sum_{k=1}^{N-1}\sum_{l=1}^{N-1}(N-k)(N-l)\,P_1(k, l)$$

$$- \frac{mn}{(N-2)(N-3)}[(N-2)(N-3)-4(m-1)(n-1)]\sum_{k=1}^{N-1}\sum_{l=1}^{N-1}(N-k)(N-l)\,P_2(k, l)$$

where $P_1(k, l) = P(k^{th}$ nearest neighbor to $Z_i = Z_j$, l^{th} nearest neighbor to $Z_j = Z_i)$,

$P_2(k, l) = P(k^{th}$ nearest neighbor to $Z_i = l^{th}$ nearest neighbor to $Z_j)$,

and $i \neq j = 1, 2, \ldots, N$.

CHAPTER III

A SYMPTOTIC NORMALITY OF T_i AND T

3.1 Introduction

In this Chapter we will prove the asymptotic normality of T_i. In order to prove that we will show that T_{ik} is a Markov Chain. Then this property as well as the martingale central limit theorem will be used to prove the asymptotic normality of T_i (Sen and Salama (1983) and Puri and Sen (1971)). This proof can be taken as another method to prove the asymptotic normality of the Wilcoxon-Mann-Whitney statistic. Finally, we will rewrite T as a U-statistic in order to investigate its asymptotic normality using Lehmann's theorem.

3.2 A symptotic normality of T_i

First we will prove the Markovian property of T_{ik}.

Result 3.2.1 A Markovian property of T_{ik}

For every i ($1 \le i \le n$), the sequence $\{T_{ik}; k = 1, \ldots, N-1\}$ is a Markov chain, i.e., for every k ($\le N-1$) and $\max(0, k-n+1) \le r_1 \le \ldots \le r_k \le r_{k+1} \le \min(k+1, m)$

$$P(T_{i,k+1} = r_{k+1} \mid T_{ij} = r_j;\ j \le k) = P(T_{i,k+1} = r_{k+1} \mid T_{ik} = r_k).$$

Proof:

Let P be the set of all permutations of $(\{1, \ldots, N\} - \{i\})$ satisfying the condition $\{ T_1 = r_1, \ldots, T_{ik} = r_{ik} \}$. It is clear that for any $p \epsilon P$, $T_{i,k+1}$ can assume only the values r_k and r_{k+1}. If $\{ \alpha_1, \ldots, \alpha_N \} \epsilon P$, then in the set $\{ \alpha_1, \ldots, \alpha_k \}$, we have r_k elements of the set $\{ n+1, \ldots, N \}$ and $(k - r_k)$ elements of the set $\{ 1, \ldots, n \}$. Then we may have either of the following:

(i) $k+1 \epsilon \{ n+1, \ldots, N \}$. This happens with (conditional) probability

$$\frac{m-r_k}{N-1-k}$$

(ii) $k+1 \notin \{ n+1, \ldots, N \}$. This happens with (conditional) probability

$$1 - \frac{m-r_k}{N-1-k} = \frac{n-1-k+r_k}{N-1-k}$$

In case (i), $T_{i,k+1}$ can assume only the value r_{k+1} with probability $\frac{m-r_k}{N-1-k}$, while in case (ii), $T_{i,k+1}$ can assume only the value r_k with probability $\frac{n-1-k+r_k}{N-1-k}$.

Thus, the assumable values of $T_{i,k+1}$ (viz. r_k, r_{k+1}) and their respective (conditional) probabilities (given the T_{ij}, $j \leq k$) depend only on the value r_k assumed by T_{ik}.

We may note that, by the previous result,

$$P(T_{i,k+1} = s | T_{ik} = r) = \frac{P(T_{i,k+1} = s,\ T_{ik} = r)}{P(T_{ik} = r)}$$

$$= \frac{\binom{k}{r}\binom{k+1-k}{s-r}\binom{N-k-2}{m-s}\binom{N-1}{m}}{\binom{N-1}{m}\binom{k}{r}\binom{N-1-k}{m-r}}$$

$$= \frac{\binom{N-2-k}{m-s}}{\binom{N-1-k}{m-r}}$$

Therefore,

$$P(T_{i,k+1} = s \mid T_{ik} = r) = \begin{cases} \dfrac{m-r}{N-1-k} & \text{if } s=r+1, \\ \dfrac{n-1-k+r}{N-1-k} & \text{if } s=r, \\ 0 & \text{if } s \geq r+2 \text{ or } s < r. \end{cases}$$

Hence, from the distribution of $\{ T_{i,k+1} \mid T_{ik} \}$ we have

$$E(T_{i,k+1} \mid T_{ik}=r) = (r+1) \frac{m-r}{N-1-k} + (r) \frac{n-1-k+r}{N-1-k}.$$

So,

$$E(T_{i,k+1} \mid T_{ik}) = \left[\frac{N-k-2}{N-1-k} \right] T_{ik} + \frac{m}{N-1-k},$$

where,

$$k = 1, \ldots, N-2.$$

Note: For $i = n+1, \ldots, N$, we have the same result with m and n interchanged.

Also,

$$E(T^2_{i,k+1} \mid T_{ik} = r) = (r+1)^2 \frac{m-r}{N-1-k} + (r^2) \frac{n-1-k+r}{N-1-k}$$

$$= \frac{1}{N-1-k} [\, r^2m + 2rm + m - r^3 - 2r^2 - r + r^2n - r^2 - r^2k + r^3]$$

$$= \frac{1}{N-1-k} [\, r^2N - 3r^2 - r^2k + 2rm - r + m \,]$$

$$= \frac{1}{N-1-k} [\, r^2(N-3-k) + r(2m-1) + m]$$

So,

$$E(T^2_{i,k+1} \mid T_{ik}) = \frac{N-3-k}{N-1-k} T^2_{ik} + \frac{2m-1}{N-1-k} T_{ik} + \frac{m}{N-1-k}$$

Note: As before for $i = n+1, \ldots, N$, we have the same result with n and m interchanged.

Result 3.2.2 Asymptotic permutational normality of T_i

For every $i = 1, \ldots, n$, as $N \longrightarrow \infty$

$$T_i^{\star} = \frac{T_i - ET_i}{[V(T_i)]^{1/2}} \longrightarrow_D N(0, 1).$$

Proof:

For $k = 1, \ldots, N-1$, let

$$d_{ik} = \frac{m}{(N-1-k)(N-1)}$$

$$d_{ik}^{\star} = \sum_{i=1}^{k} d_{ij}$$

$$Y_{ik} = \frac{1}{N-1-k} T_{ik} - d_{ik}^{\star}, \qquad Y_{i0} = 0$$

and let

$B_{ik} = B(T_{ij};\ j \leq k)$ be a σ-algebra.

Then

$$E(Y_{ik}|B_{i,k-1}) = \frac{1}{N-1-k} E(T_{ik}|B_{i,k-1}) - d^{\star}_{ik}$$

$$= \frac{1}{N-1-k} E(T_{ik}|T_{i,k-1}) - d^{\star}_{ik}$$

$$= \frac{1}{N-1-k} \left[\frac{N-1-k}{N-k} T_{i,k-1} + \frac{m}{N-1} \right] - d^{\star}_{ik}$$

$$= \frac{1}{N-k} T_{i,k-1} + \frac{m}{(N-1-k)(N-1)} - d^{\star}_{ik}$$

$$= \frac{1}{N-k} T_{i,k-1} - d^{\star}_{i,k-1}$$

$$= Y_{i,k-1}$$

By letting

$$Z_{ik} = Y_{ik} - Y_{i,k-1}$$

the Z_{ik} are martingale differences.

Now,

$$T_i - ET_i = \sum_{k=1}^{N-1} T_{ik} - E\left[\sum_{k=1}^{N-1} T_{ik} \right]$$

$$= \sum_{k=1}^{N-2} [(N-1-k)Y_{ik} + (N-1-k)\, d_{ik}] -$$

$$\sum_{k=1}^{N-2} [(N-1-k)EY_{ik} + (N-1-k)d_{ik}]$$

$$= \sum_{k=1}^{N-2} (N-1-k)Y_{ik} - \sum_{k=1}^{N-2} (N-1-k)EY_{ik}$$

Note that

$$EY_{i1} = E[E(Y_{i1}|B_{i,0})] = E[Y_{i0}] = 0,$$

$$EY_{i2} = E[E(Y_{i2}|B_{i1})] = E[Y_{i1}) = 0,$$

and so on. Hence $EY_{ik} = 0$ for $k = 1,\ldots, N-2$,

and

$$T_i - ET_i = \sum_{k=1}^{N-2} (N-1-k)Y_{ik}.$$

But,

$$Y_{ik} = Z_{i1} + Z_{i2} + \ldots + Z_{ik}$$

$$= \sum_{j=1}^{k} Z_{ij}$$

So,

$$T_i - ET_i = \sum_{k=1}^{N-2} (N-1-k)\left(\sum_{j=1}^{k} Z_{ij}\right)$$

$$= \sum_{k=1}^{N-2} Z_{ik} \left(\sum_{j=k}^{N-2} (N-1-j)\right.$$

$$= \sum_{k=1}^{N-2} \frac{(N-k-1)(N-k)}{2} Z_{ik}$$

Let

$$C_{ij} = \frac{(N-1-j)(N-j)}{2[m(n-1)N/12]^{1/2}}, \qquad j = 1, \ldots, N-2$$

and

$$U_{ij} = C_{ij} Z_{ij}, \qquad j = 1, \ldots, N-2$$

then

$$T_i^{\star} = \frac{T_i - ET_i}{[V(T_i)]^{1/2}}$$

$$= \sum_{k=1}^{N-2} \frac{(N-1-k)(N-k)}{2[m(n-1)N/12]^{1/2}}$$

$$= \sum_{k=1}^{N-2} U_{ik}$$

But,

$$E(U_{ik} | B_{i,k-1}) = C_{ik}\, E(Z_{ik} | B_{i,k-1})$$

$$= C_{ik}.0$$

$$= 0$$

and

$$V(T_i^{\star}) = 1$$

$$= \mathrm{Var}\left(\sum_{k=1}^{N-1} U_{ik}\right)$$

$$= \sum_{k=1}^{N-1} \mathrm{Var}\,(U_{ik})$$

$$= \sum_{k=1}^{N-1} EU_{ik}^{2}$$

Thus, $T_i^{\star}$ relates to a martingale array, normalized by $\Sigma\ EU_{ik}^2 = 1$, and to show that $T_i^{\star}$ has an N(0,1) distribution as $N \longrightarrow \infty$, the following conditions (Shiryayev (1984)) must be satisfied:

$$\text{(i)}\quad \tilde{U}_i = \sum \tilde{U}_{ik} = \sum E(U_{ik}^2 | B_{i,k-1}) \xrightarrow{p} 1$$

and for every $\epsilon > 0$,

$$\text{(ii)}\quad \sum_{k=1} E[U_{ik}^2\, I(|U_{ik}| > \epsilon)\, |B_{i,k-1}] \xrightarrow{p} 0.$$

(i) By virtue of $\sum_k EU_{ik}^2 = 1$ for all values of k, to prove (i) it suffices to show that

$$E(\tilde{U}_i - 1)^2 \longrightarrow 0 \text{ as } N \longrightarrow \infty.$$

Towards this, for every k, $1 \le k \le N-1$, we have

$$\tilde{U}_{ik} = E(U_{ik}^2 | B_{i,k-1})$$

$$= E(C_{ik}^2 Z_{ik}^2 | B_{i,k-1})$$

$$= C_{ik}^2 [E(Y_{ik}^2 | B_{i,k-1}) - Y_{i,k-1}^2]$$

$$\tilde{U}_{ik} = C_{ik}^2\{E[(\frac{1}{N-1-k} T_{ik} - d_{ik}^{\star})^2 | B_{i,k-1}] - [\frac{1}{N-k} T_{i,k-1} - d_{i,k-1}]^2\}$$

$$= C_{ik}^2\{\frac{1}{(N-1-k)^2} E(T_{ik}^2 | B_{i,k-1}) - \frac{2d_{ik}^{\star}}{N-1-k} E(T_{ik} | B_{i,k-1}) + d_{ik}^{\star 2} -$$

$$\frac{1}{(N-k)^2} T_{i,k-1}^2 + \frac{2d_{i,k-1}^{\star}}{N-k} T_{i,k-1} - d_{i,k-1}^{\star 2}\} .$$

But,

$$E(T_{ik}^2 | B_{i,k-1}) = \frac{N-2-k}{N-k} T_{i,k-1}^2 + \frac{2m-1}{N-k} T_{i,k-1} + \frac{m}{N-k}$$

$$E(T_{ik} | B_{i,k-1}) = \frac{N-1-k}{N-k} T_{i,k-1} + \frac{m}{N-k}$$

and

$$d_{ik}^{\star 2} - d_{i,k-1}^{\star 2} = (d_{ik}^{\star} + d_{i,k-1}^{\star})(d_{ik}^{\star} - d_{i,k-1}^{\star})$$

$$= 2\, d_{1,k-1} d_{ik} + d_{ik}^2$$

$$= \frac{2m}{(N-1-k)(N-1)} d_{i,k-1}^{\star} + [\frac{m}{(N-1-k)(N-1)}]^2 .$$

So,

$$\tilde{U}_{ik} = C_{ik}^2 \{ \frac{1}{(N-1-k)^2} [\frac{N-2-k}{N-k} T_{i,k-1}^2 + \frac{2m-1}{N-k} T_{i,k-1} + \frac{m}{N-k}]$$

$$- \frac{2d_{ik}^{\star}}{N-1-k} [\frac{N-1-k}{N-k} T_{i,k-1} + \frac{m}{N-k}] - \frac{1}{(N-k)^2} T_{i,k-1}^2 +$$

$$\frac{2d_{i,k-1}^{\star}}{N-k} T_{i,k-1} + \frac{2md_{i,k-1}^{\star}}{(N-1-k)(N-1)} + [\frac{m}{(N-1-k)(N-1)}]^2 \}$$

$$= C_{ik}^2 \{ [\frac{N-2-k}{(N-1-k)^2(N-k)} - \frac{1}{(N-k)^2}] T_{i,k-1}^2 +$$

$$[\frac{2m-1}{(N-1-k)^2(N-k)} - \frac{2d_{ik}^{\star}}{N-k} + \frac{2d_{i,k-1}^{\star}}{N-k}] T_{i,k-1} + [\frac{m}{(N-1-k)(N-k)}$$

$$- \frac{2md_{ik}^{\star}}{(N-1-k)(N-k)} + \frac{2md_{i,k-1}^{\star}}{(N-1-k)(N-1)} + (\frac{m}{(N-1-k)(N-1)})^2] \}$$

$$= a_{ik} T_{i,k-1}^2 + b_{ik} T_{i,k-1}^2 + g_{ik}$$

where

$$a_{ik} = \frac{3(N-1-k)^2(N-k)^2}{m(n-1)N}\left[\frac{(N-2-k)(N-k)-(N-1-k)^2}{(N-1-k)^2(N-k)^2}\right]$$

$$= \frac{-3}{m(n-1)N}$$

$$= \frac{-3}{(m/N)((n-1)/N)N^3}$$

$$= O(N^{-3}), \qquad m/N,\ n/N \longrightarrow \lambda_1, \lambda_2 \text{(constants) as } m,\ n \longrightarrow \infty$$

$$b_{ik} = \frac{3(N-1-k)^2(N-k)^2}{m(n-1)N}\left[\frac{2m-1}{(N-1-k)^2(N-k)} - \frac{2d^{\star}_{ik}}{N-k} + \frac{2d^{\star}_{i,k-1}}{N-k}\right]$$

$$= \frac{3(N-1-k)^2(N-k)^2}{m(n-1)N}\left[\frac{2m-1}{(N-1-k)^2(N-k)} - \frac{2}{N-k}\,\frac{m}{(N-1-k)(N-1)}\right]$$

$$= \frac{3(N-k)}{m(n-1)N}\left[\frac{(2m-1)(N-1) - 2m(N-1-k)}{(N-1)}\right]$$

$$= \frac{3(N-k)}{m(n-1)N}\left[\frac{2mk}{N-1} - 1\right]$$

$$b_{ik} = \frac{6(N-k)k}{(n-1)N(N-1)} - \frac{3(N-k)}{m(n-1)N}$$

$$= O(N^{-1}) + O(N^{-2})$$

$$= O(N^{-1}),$$

$$g_{ik} = \frac{3(N-1-k)^2(N-k)^2}{m(n-1)N} \left[\frac{m}{(N-1-k)^2(N-k)} - \frac{2md^{\star}_{ik}}{(N-1-k)(N-k)} + \frac{2md^{\star}_{i,k-1}}{(N-1-k)(N-1)} \right.$$

$$\left. + \frac{m^2}{(N-1-k)^2(N-1)^2} \right]$$

$$= \frac{3(N-k)}{(n-1)N(N-1)^2} \left[(N-1)^2 - 2d^{\star}_{ik}(N-1-k)(N-1)^2 + 2d^{\star}_{i,k-1}(N-1)(N-k) \times \right.$$

$$\left. (N-1-k) + m(N-k) \right]$$

$$= \frac{3(N-k)}{(n-1)N(N-1)^2} \left[(N-1)^2 - 2m(N-1)(N-1-k) \sum_{j=1}^{k} \frac{1}{(N-1-j)} + \right.$$

$$\left. 2m(N-k)(N-1-k) \sum_{j=1}^{k-1} \frac{1}{(N-1-j)} + m(N-k) \right]$$

$$g_{ik} = \frac{3(N-k)}{(n-1)N(N-1)^2}\left[(N-1)^2 - 2m(N-1-k)(k-1)\sum_{j=1}^{k-1}\frac{1}{(N-1-j)} - \right.$$

$$\left. 2m(N-1)(N-1-k)\frac{1}{N-1-k} + m(N-k)\right]$$

$$= \frac{3(N-k)}{(n-1)N(N-1)^2}\left[(N-1)^2 - 2m(N-1-k)(k-1)\sum_{j=1}^{k-1}\frac{1}{(N-1-j)} - m(N-2-k)\right]$$

$$= O(N^{-1}),$$

where,

$$E(\tilde{U}_i) = E[\sum E(U_{ik}^2 | B_{i,k-1})]$$

$$= \sum E[E(U_{ik}^2 | B_{i,k-1})$$

$$= \sum EU_{ik}^2$$

$$= 1.$$

So,

$$\tilde{U}_i - 1 = \tilde{U}_i - E\tilde{U}_i$$

and,

$$E(\tilde{U}_i - 1)^2 = E(\tilde{U}_i - E\tilde{U}_i)^2$$

$$= E[\sum \tilde{U}_{ik} - E(\sum \tilde{U}_{ik})]^2$$

$$= E[\sum (\tilde{U}_{ik} - E\tilde{U}_{ik})]^2$$

$$= E\{\sum [\, a_{ik}(T^2_{i,k-1} - ET^2_{i,k-1}) + b_{ik}(T_{i,k-1} - ET_{i,k-1}) + (g_{ik} - Eg_{ik})]\}^2$$

$$= \sum_k a^2_{ik}\, E(T^2_{i,k-1} - ET^2_{i,k-1})^2$$

$$+ \sum_k b_{ik}\, E(T_{i,k-1} - ET_{i,k-1})^2$$

$$+ 2 \sum_k a_{ik} b_{ik}\, E(T^2_{i,k-1} - ET^2_{i,k-1})(T_{i,k-1} - ET_{i,k-1})$$

$$+ \sum_{k \neq k'} \sum b_{ik} b_{ik'}\, E(T_{i,k-1} - ET_{i,k-1})(T_{i,k'-1} - ET_{i,k'-1})$$

$$+ 2 \sum_{k<k'} \sum a_{ik} b_{ik'} E(T^2_{i,k-1} - ET^2_{i,k-1})(T_{i,k'-1} - ET_{i,k'-1})$$

$$+ 2 \sum_{k<k'} \sum a_{ik} a_{ik'} E(T^2_{i,k-1} - ET^2_{i,k-1})(T^2_{i,k'-1} - ET^2_{i,k'-1})$$

where,

$$E(T^2_{i,k-1} - ET^2_{i,k-1})^2 = ET^4_{i,k-1} - (ET^2_{i,k-1})^2.$$

But

$$ET_{i,k-1} = \frac{(k-1)m}{N-1}$$

$$ET^2_{i,k-1} = \frac{(k-1)(k-2)m(m-1)}{(N-1)(N-2)} + \frac{(k-1)m}{N-1}$$

$$ET^3_{i,k-1} = \frac{(k-1)(k-2)(k-3)m(m-1)(m-2)}{(N-1)(N-2)(N-3)}$$

and

$$ET^4_{i,k-1} = ET_{i,k-1}(T_{i,k-1} -1)(T_{i,k-1} - 2)(T_{i,k-1} -3) + 6ET^3_{i,k-1}$$
$$- 11ET^2_{i,k-1} + 6ET_{i,k-1}$$

Hence,

$$E(T^2_{i,k-1} - ET^2_{i,k-1})^2 = \frac{4! \binom{m}{4}\binom{k-1}{4}}{\binom{N-1}{4}} + \frac{6(k-1)(k-2)(k-3)m(m-1)(m-2)}{(N-1)(N-2)(N-3)}$$

$$- \frac{11(k-1)(k-2)m(m-1)}{(N-1)(N-2)} - \frac{11(k-1)m}{N-1} + \frac{6(k-1)m}{N-1}$$

$$- \left[\frac{(k-1)(k-2)m(m-1)}{(N-1)(N-2)} + \frac{(k-1)m}{N-1}\right]^2$$

$$= \frac{(k-1)(k-2)(k-3)(k-4)m(m-1)(m-2)(m-3)}{(N-1)(N-2)(N-3)(N-4)}$$

$$- \frac{6(k-1)(k-2)(k-3)m(m-1)(m-2)}{(N-1)(N-2)(N-3)} - \frac{11(k-1)(k-2)m(m-1)}{(N-1)(N-2)}$$

$$- \frac{11(k-1)m}{N-1} + \frac{6(k-1)m}{N-1} - \frac{(k-1)^2(k-2)^2m^2(m-1)^2}{(N-1)^2(N-2)^2}$$

$$- \frac{(k-1)^2m^2}{(N-1)^2} - \frac{2(k-1)^2(k-2)m^2(m-1)}{(N-1)^2(N-2)}$$

$$= \frac{(k-1)(k-2)(k-3)m(m-1)(m-2)}{(N-1)(N-2)(N-3)(N-4)} [(k-4)(m-3) + 6(N-4)]$$

$$- \frac{(k-1)m}{(N-1)(N-2)} [11(k-2)(m-1) + 5(N-2)]$$

$$- \frac{(k-1)^2 m^2}{(N-1)^2(N-2)^2} [(k-2)^2(m-1)^2 + (N-2)^2 + 2(k-2)(m-1)(N-2)]$$

$$= O(N^4).$$

$$E(T_{i,k-1} - ET_{i,k-1})^2 = V(T_{i,k-1})$$

$$= \frac{(k-1)m(n-1)(N-k)}{(N-1)^2(N-2)}$$

$$= O(N).$$

$$E(T^2_{i,k-1} - ET^2_{i,k-1})(T_{i,k-1} - ET_{i,k-1})$$

$$= E[T^3_{i,k-1} - T^2_{i,k-1}ET_{i,k-1} - T_{i,k-1}ET^2_{i,k-1} + ET^2_{i,k-1}\, ET_{i,k-1}]$$

$$= \frac{(k-1)(k-2)(k-3)m(m-1)(m-2)}{(N-1)(N-2)(N-3)} - \left[\frac{(k-1)(k-2)m(m-1)}{(N-1)(N-2)} + \frac{(k-1)m}{N-1} \right] \times$$

$$\frac{(k-1)m}{N-1}$$

$$= \frac{(k-1)(k-2)m(m-1)}{(N-1)^2(N-2)(N-3)} [(k-3)(m-2)(N-1) - (N-3)] + \frac{(k-1)^2 m^2}{(N-1)^2}$$

$$= O(N^3).$$

$$E(T_{i,k-1} - ET_{i,k-1})(T_{i,k'-1} - ET_{i,k'-1})$$

$$= Cov(T_{i,k-1}, T_{i,k'-1})$$

$$= \frac{m(k-1)(n-1)(N-k')}{(N-1)^2(N-2)}$$

$$= O(N).$$

$$E(T^2_{i,k-1} - ET^2_{i,k-1})(T_{i,k'-1} - ET_{i,k'-1})$$

$$= ET^2_{i,k-1} T_{i,k'-1} - ET^2_{i,k-1} ET_{i,k'-1}$$

$$= ET_{i,k-1}(T_{i,k-1} - 1)(T_{i,k'-1} - T_{i,k-1}) + ET^3_{i,k-1} - ET^2_{i,k-1}$$

$$+ E(T_{i,k-1}T_{i,k'-1}) - ET^2_{i,k-1}ET_{i,k'-1}$$

$$= \frac{m(m-1)(m-2)(k-1)(k-2)(k'-k)}{(N-1)(N-2)(N-3)} + \frac{(k-1)(k-2)(k-3)m(m-1)(m-2)}{(N-1)(N-2)(N-3)}$$

$$- \frac{(k-1)(k-2)m(m-1)}{(N-1)(N-2)} - \frac{(k-1)m}{N-1} + Cov(T_{i,k-1}, T_{i,k'-1})$$

$$+ ET_{i,k-1}ET_{i,k'-1} - \left[\frac{(k-1)(k-2)m(m-1)}{(N-1)(N-2)} + \frac{(k-1)m}{N-1}\right]\frac{(k'-1)m}{N-1}$$

$$= \frac{m(m-1)(m-2)(k-1)(k-2)}{(N-1)(N-2)(N-3)}[(k'-k) + (k-3)] + Cov(T_{i,k-1}, T_{i,k'-1})$$

$$- \frac{(k-1)(k-2)m(m-1)}{(N-1)^2(N-2)}[(N-1) + m(k'-1)]$$

$$- \frac{(k-1)m}{(N-1)^2}[(k'-1)m + (N-1) - (k'-1)m]$$

$$= O(N^3).$$

$$E(T^2_{i,k-1} - ET^2_{i,k-1})(T^2_{i,k'-1} - ET^2_{i,k'-1})$$

$$= ET^2_{i,k-1}T^2_{i,k'-1} - ET^2_{i,k-1}ET^2_{i,k'-1}$$

$$= ET_{i,k-1}(T_{i,k-1}-1)(T_{i,k'-1}-T_{i,k-1})(T_{i,k'-1}-T_{i,k-1}-1)$$

$$- ET^4_{i,k-1} + 2ET^3_{i,k-1}T_{i,k'-1} - ET^2_{i,k-1}T_{i,k'-1} + ET_{i,k-1}T^2_{i,k'-1}$$

$$- ET_{i,k-1}T_{i,k'-1} + ET^2_{i,k-1} - ET^2_{i,k-1}ET^2_{i,k'-1}$$

$$= O(N^4).$$

Therefore, for every $k = 1, \ldots, N-1$, we have

$$E(\tilde{U}_i - 1)^2 = O(N^{-1}),$$

and hence, the first condition holds.

(ii) $\sum_k E\{U^2_{ik}\, I(|U_{ik}| > \epsilon)\,|B_{i,k-1}\} \xrightarrow{p} 0$

To verify this condition, note that T_{ik} are nonnegative, $T_{ik} \leq k$ for every $k < N$.

For every $k = 1, \ldots, N-1$

$$|U_{ik}| = |C_{ik}\, Z_{ik}|$$

$$= |C_{ik}|\ |Y_{ik} - Y_{i,k-1}|$$

$$= \left| \frac{(N-1-k)(N-k)}{\{m(n-1)N/3\}^{1/2}} \right| \left| \frac{1}{N-1-k} T_{ik} - d^{\star}_{ik} - \frac{1}{N-k} T_{i,k-1} + d^{\star}_{i,k-1} \right|$$

$$= \left| \frac{(N-1-k)(N-k)}{\{m(n-1)N/3\}^{1/2}} \right| \left| \frac{1}{N-1-k} T_{ik} - \frac{1}{N-k} T_{i,k-1} - \frac{m}{(N-1-k)(N-1)} \right|$$

$$\leq \left| \frac{(N-1-k)(N-k)}{\{m(n-1)N/3\}^{1/2}} \right| \left| \frac{1}{N-1-k} T_{i,k-1} - \frac{1}{N-k} T_{i,k-1} - \frac{m}{(N-1-k)(N-1)} \right|$$

$$= \left| \frac{(N-1-k)(N-k)}{\{m(n-1)N/3\}^{1/2}} \right| \left| \frac{1}{(N-1-k)(N-k)} T_{i,k-1} - \frac{m}{(N-1-k)(N-1)} \right|$$

$$= \left| \frac{1}{\{m(n-1)N/3\}^{1/2}} T_{i,k-1} - \frac{(N-k)m}{\{m(n-1)N/3\}^{1/2}(N-1)} \right|$$

$$\leq \left| \frac{k-1}{\{m(n-1)N/3\}^{1/2}} - \frac{(N-k)m}{\{m(n-1)N/3\}^{1/2}(N-1)} \right|$$

$$\leq \left| \frac{k-1}{\{m(n-1)N/3\}^{1/2}} \right| + \left| \frac{(N-k)m}{\{m(n-1)N/3\}^{1/2}(N-1)} \right|$$

$$\leq \left| \frac{N}{N^{3/2}\{\frac{m}{N}\frac{n-1}{N}\frac{1}{3}\}^{1/2}} \right| + \left| \frac{(N-1)N(m/N)}{N^{3/2}(N-1)\{\frac{m}{N}\frac{n-1}{N}\frac{1}{3}\}^{1/2}} \right|$$

$$= C_1 N^{-1/2} + C_2 N^{-1/2}$$

$$= CN^{-1/2}$$

if (m/N) and (n/N) tend to λ_1 and λ_2 (constants) as $n, m \longrightarrow \infty$.

So,

$$|U_{ik}| \leq CN^{-1/2} \text{ with probability 1,}$$

for every k: $1 \leq k \leq N-1$, when C does not depend on N $(0 < C < \infty)$.

$$|U_{ik}| \longrightarrow 0 \text{ as } N \longrightarrow \infty$$

$$I(|U_{ik}| > \epsilon) \longrightarrow 0 \text{ as } N \longrightarrow \infty$$

$\Longrightarrow$

$$E\{U_{ik}^2 \, I(|U_{ik}| > \epsilon) \, | B_{i,k-1}\} \longrightarrow 0 \text{ as } N \longrightarrow \infty \text{ for every k}$$

$$\Longrightarrow$$

$$\sum_{k} E\{U_{ik}^2 \, I(|U_{ik}| > \epsilon) \, | B_{i,k-1}\} \xrightarrow{p} 0$$

Hence, the second condition holds for N sufficiently large.

This completes the proof of the theorem. i.e.

$$T_i^{\star} = \frac{T_i - ET_i}{[V(T_i)]^{1/2}} \longrightarrow_D N(0, 1).$$

3.3 Asymptotic Distribution of T

In order to investigate the asymptotic distribution of T, first we will express $T_x = \sum_{i=1}^{n} T_i$, $T_y = \sum_{i=n+1}^{N} T_i$, and T as U-statistics, then if the variance of the U-statistic is finite as n and m $\longrightarrow \infty$, we can apply Lehmann's theorem (Randles and Wolf (1979)) to prove that T_x, T_y, or T is asymptotically normal.

Result 3.3.1 Expressing T_x as a U-statistic

Let $\phi(x_1, x_2, y_1) = I\{D(x_1, x_2) < D(x_1, y_1)\} + I\{D(x_1, x_2) < D(x_2, y_1)\}$ where D(u, v) = distance function, D(u, v) = $||u - v||$, and

$\Phi_x = \Phi(Z_i, Z_j, Z_k) = \frac{2}{n(n-1)m} \sum_{i<j}^{n} \sum_{k=n+1}^{N} \phi(Z_i, Z_j, Z_k)$ be the

generalized U-statistic, i, j=1, . . ., n and $i \neq j$, then

$$T_x = \frac{nm}{2}[(N+n-1) - (n-1)\Phi_x]$$

i.e., T_x is equivalent to a U-statistic.

Proof:

Let R_{ij} = rank of Z_j on nearness to Z_i, for j=1, . . ., n+m and $i \neq j$. $1 \leq R_{ij} \leq N-1$.

Note that

$$I\{D(X_i, X_j) < D(X_i, Y_k)\} = I\{R_{ij} < R_{ik}\}$$

and

$$I\{D(X_i, X_j) < D(X_j, Y_k)\} = I\{R_{ji} < R_{jk}\}$$

Let $\Phi_x^* = \sum_{i<j} \sum_{k=n+1} \phi(Z_i, Z_j, Z_k)$, then

$$\Phi_x^* = \sum_{i<j} \sum_{k=n+1} [I\{R_{ij} < R_{ik}\} + I\{R_{ji} < R_{jk}\}]$$

$$= \sum_{i \neq j} \sum_{k=n+1} I\{R_{ij} < R_{ik}\}$$

Note that, for fixed Z_i and Z_j

$\sum_{k=n+1} I\{R_{ij} < R_{ik}\}$ = number of Ys (Z_k's) which are larger than Z_j with respect to the ranks relative to Z_i

= m - number of Ys smaller than Z_j with respect to the ranks relative to Z_i

Then

$$\sum_{k=n+1} I\{ R_{ij} < R_{ik} \} = m - T_{iR_{ij}}$$

Then

$$\Phi_x^{\star} = \sum_{i \neq j}^{n} (m - T_{iR_{ij}})$$

$$= n(n-1)m - \sum_{i \neq j} T_{iR_{ij}}$$

$$= nm(n-1) - \sum_{k=1}^{N-1} \sum_{i=1}^{n} T_{ik}(1 - h(i, k))$$

$$= nm(n-1) - \sum_{i=1}^{n} \sum_{k=1}^{N-1} T_{ik} + \sum_{i=1}^{n} \sum_{k=1}^{N-1} h(i, k) T_{ik}$$

$$= nm(n-1) - T_x + \sum_{i=1}^{n} \frac{m(m+1)}{2}$$

$$= \frac{nm}{2} (N+n-1) - T_x$$

Therefore,

$$\Phi_x = \frac{2}{n(n-1)m} \Phi_x^{\star}$$

$$= \frac{2}{n(n-1)m} \left[\frac{nm}{2} (N+n-1) - T_x \right]$$

Hence,

$$T_x = \frac{nm}{2}(N+n-1) - \frac{n(n-1)m}{2}\Phi_x$$

$$= \frac{nm}{2}[(N+n-1) - (n-1)\Phi_x]$$

i.e., T_x is equivalent to a U-statistic.

Result 3.3.2 Expressing T_y as a U-statistic

By interchanging n and m and x and y in Result 3.3.1, we have

$$\Phi_y = \frac{2}{m(m-1)n}\sum_{k<l}^{N}\sum_{i=1}^{n}\phi(Z_k, Z_l, Z_i), \quad k, l=n+1, \ldots, N \text{ and } k \neq l$$

and

$$T_y = \frac{nm}{2}[(N+m-1) - (m-1)\Phi_y]$$

i.e., T_y is equivalent to a U-statistic.

Proof: As in Result 3.3.1.

Result 3.3.3 Expressing T as a U-statistic

Let $\phi'(x_1, x_2, y_1, y_2) = \phi(x_1, x_2, y_1) + \phi(y_1, y_2, x_1) + \phi(x_1, x_2, y_2) + \phi(y_1, y_2, x_2)$

where ϕ is defined in Result 3.3.1, and let

$$\Phi = \Phi(Z_i, Z_j, Z_k, Z_l) = \frac{4}{n(n-1)m(m-1)}\sum_{i<j}\sum_{k<l}\phi'(Z_i, Z_j, Z_k, Z_l)$$

where i, j=1, . . ., n; $i \neq j$ and k, l=n+1, . . ., N; $k \neq l$.

Then, if m=n, we have

$$T = \frac{n^2}{4}[4(3n-1) - (n-1)\Phi]$$

i.e., T is equivalent to a U-statistic.

Proof:

$$\Phi^* = \sum_{i<j} \sum_{k<l} \phi'(Z_i, Z_j, Z_k, Z_l)$$

$$= (m-1) \sum_{i<j} \sum_{k=n+1}^{N} \phi(Z_i, Z_j, Z_k) + (n-1) \sum_{k<l} \sum_{i=1}^{n} \phi(Z_k, Z_l, Z_i)$$

$$= (m-1) \sum_{i \neq j} \sum_{k=n+1} I\{ R_{ij} < R_{ik} \} + (n-1) \sum_{k \neq l} \sum_{i=1} I\{ R_{kl} < R_{ki} \}$$

$$= (m-1) \frac{n(n-1)m}{2} \Phi_x + (n-1) \frac{m(m-1)n}{2} \Phi_y$$

$$= (m-1)[\frac{nm}{2}(N+n-1) - T_x] + (n-1)[\frac{nm}{2}(N+m-1) - T_y]$$

$$= \frac{nm}{2}[(m-1)(N+n-1) + (n-1)(N+m-1)] - [(m-1)T_x + (n-1)T_y]$$

Therefore,

$$\Phi = \frac{4}{n(n-1)m(m-1)} \Phi^{\star}$$

$$= \frac{2(N+n-1)}{(n-1)} + \frac{2(N+m-1)}{(m-1)} - \frac{4}{nm(n-1)} T_x - \frac{4}{nm(m-1)} T_y$$

So, if m=n, we have

$$\Phi = \frac{4(3n-1)}{(n-1)} - \frac{4}{n^2(n-1)} T$$

Therefore,

$$T = \frac{n^2}{4} [4(3n-1) - (n-1)\Phi]$$

i.e., T is equivalent to a U-statistic.

Asymptotic normality of T when expressed as a U-statistic

To show that T is asymptotically normal we have to prove that $0 < V(\Phi) < \infty$ as $n, m \longrightarrow \infty$. Then we can apply Lehmann's theorem (Randles and Wolfe (1979)). Since T is a linear combination of two U-statistics, then T is asymptotically normal if each of the U-statistics is asymptotically normal. To prove that, we need to show that $0 < V(\Phi_x), V(\Phi_y) < \infty$ as $n, m \longrightarrow \infty$. So, we need to derive $V(T_x)$ and $V(T_y)$.

Result 3.3.4 Under the null hypothesis, we have

$$V(T_x) = \frac{nm(n-1)N(1-3mN)}{12} + \frac{nm(n-1)(m-1)}{(N-2)(N-3)} \frac{N^2(N-1)^2}{4} +$$

$$\frac{nm(n-1)(n-2)}{(N-2)(N-3)} \sum_{k}^{N-1} \sum_{l}^{N-1} (N-k)(N-l)P_2(k, l)$$

where $P_2(k, l)$ is defined in Result 13, Chapter 2.

Proof:

The same as in Result 13, Chapter 2 with $C_1 = C_2 = C_3 = 0$ and $i, j=1, \ldots, n$.

$$V(T_x) = V(\sum_{i=1}^{n} T_i)$$

$$= \sum_{i} V(T_i) - \sum_{i \neq j} E(T_i)E(T_j) + \sum_{i \neq j} E(T_iT_j)$$

$$= \frac{nm(n-1)N}{12} - n(n-1) \frac{m^2N^2}{4} + \sum_{i \neq j} E(T_iT_j)$$

As in Result 13, Chapter 2

$$\sum_{i \neq j} E(T_iT_j) = \sum_{i \neq j} \sum_{k}^{N-1} \sum_{l}^{N-1} (N-k)(N-l)P[h(i, k)=1=h(j, l)]$$

$$\sum_{i\neq j} E(T_i T_j) = \sum_{k} \sum_{l} (N-k)(N-l) \underbrace{\sum_{i\neq j} P[h(i,k)=1=h(j,l)]}_{A}$$

$$A = \sum_{i\neq j} [C_2 P_2(k,l) + C_5 P_5(k,l)]$$

$$= n(n-1)\left[\frac{m}{N-2} P_2(k,l) + \frac{m(m-1)}{(N-2)(N-3)}(1 - P_2(k,l))\right]$$

$$= \frac{n(n-1)m(m-1)}{(N-2)(N-3)} + \frac{nm(n-1)(n-2)}{(N-2)(N-3)} P_2(k,l)$$

Therefore,

$$V(T_x) = \frac{nm(n-1)N(1-3mN)}{12} + \frac{nm(n-1)(m-1)}{(N-2)(N-3)} \sum_{k}^{N-1} \sum_{l}^{N-1} (N-k)(N-l) +$$

$$\frac{nm(n-1)(n-2)}{(N-2)(N-3)} \sum_{k} \sum_{l} (N-k)(N-l) P_2(k,l)$$

$$= \frac{nmN(n-1)(1-3mN)}{12} + \frac{nm(n-1)(m-1)}{(N-2)(N-3)} \frac{N^2(N-1)^2}{4} +$$

$$\frac{nm(n-1)(n-2)}{(N-2)(N-3)} \sum_{k} \sum_{l} (N-k)(N-l) P_2(k,l).$$

Note: $V(T_y)$ is the same as $V(T_x)$ with n and m interchanged.

Result 3.3.5 Under the null hypothesis, we have

$$V(\Phi_x),\ V(\Phi_y) \longrightarrow 0 \text{ as } n=m \longrightarrow \infty.$$

Proof:

From Result 3.3.1 we know that

$$\Phi_x = \frac{2}{nm(n-1)}\left[(nm/2)(N+n-1) - T_x\right]$$

So,

$$V(\Phi_x) = \frac{4}{n^2m^2(n-1)^2}\,V(T_x)$$

$$= \frac{N(1-3mN)}{3nm(n-1)} + \frac{(m-1)N^2(N-1)^2}{nm(n-1)(N-2)(N-3)} +$$

$$\frac{4(n-2)}{nm(n-1)(N-2)(N-3)} \sum_{k}^{N-1} \sum_{l}^{N-1} (N-k)(N-l)P_2(k,\ l)$$

If n=m, we have

$$V(\Phi_x) = \frac{2(1-6n^2)}{3n(n-1)} + \frac{2(2n-1)^2}{(n-1)(2n-3)} + \frac{2(n-2)}{n^2(n-1)^2(2n-3)} \sum_{k} \sum_{l} (N-k) \times$$

$$(N-l)P_2(k,\ l)$$

Note: It is clear that $P_2(k,\ l)$ is $O(N^{-1})$.

$$V(\Phi_x) = \frac{2}{3n(n-1)} - \frac{4n}{n-1} + \frac{8n^2}{(n-1)(2n-3)} - \frac{8n}{(n-1)(2n-3)} + \frac{2}{(n-2)(n-3)} +$$

$$= \frac{2(n-2)}{n^2(n-1)^2(2n-3)} \sum_{k}^{N-1} \sum_{l}^{N-1} (N-k)(N-l)P_2(k, l)$$

Therefore,

As $n \longrightarrow \infty$, $V(\Phi_x) \longrightarrow 0-4+4-0+0+0 = 0$.

Similarly, $V(\Phi_y) \longrightarrow 0$ as $n \longrightarrow \infty$.

Result 3.3.6 Under the null hypothesis, we have

$V(\Phi) \longrightarrow 0$ as $n=m \longrightarrow \infty$.

Proof:

From Result 3.3.3, we have

$$\Phi = \frac{4(3n-1)}{(n-1)} - \frac{4}{n^2(n-1)} T$$

So,

$$V(\Phi) = \frac{16}{n^4(n-1)^2} V(T)$$

Using Result 13, Chapter 2 and the same algebra as in the proof of Result 3.3.5, we can show that

$V(\Phi) \longrightarrow 0$ as $n \longrightarrow \infty$.

From the results in this section, it is clear that we cannot use Lehmann's theorem to prove that T is asymptotically normal under the null hypothesis.

3.4 Conclusions

Since $V(T) \longrightarrow 0$ as $n=m \longrightarrow \infty$, using Lehmann's theorem for generalized U-statistic, we cannot say that T has an asymptotic normal distribution. But for a fixed i, $i = 1, \ldots, N$, T_i is proved to be asymptotically normal. So, as $N (=n+m) \longrightarrow \infty$ we expect the dependency between the T_i's to be small then we can say that $T = \sum T_i$, which is the sum of N independent normal variates, has an asymptotic normal distribution. Moreover T can be expressed as a sum of Schilling tests where Schilling's test is asymptotically normal. Hence, we still conjecture that the distribution of T is asymptotically normal. Therefore, it is still reasonable to use the normal approximation to calculate the empirical power of T. Also, computer simulation will be used to estimate the 5^{th} percentile of T then this value will be used as the critical value of the test T, i.e., the null hypothesis will be rejected if the value of T is less than this critical value. The empirical power of the test T will be calculated again based on the critical value. From the simulation study in Chapter 4, in most cases the relative difference between the empirical powers when the normal distribution and the critical value were used is small, which supports the idea of using the normal approximation.

CHAPTER IV

MONTE CARLO ESTIMATION OF THE POWER OF T

4.1 Introduction

In order to use the normal distribution approximation of the test statistic T, we need to compute V(T), which has the two parameters $P_1(r, s)$ and $P_2(r, s)$. But the values of these parameters in finite samples depend on the underlying density and are extremely difficult to compute.

In this Chapter, we use simulation to estimate the means and variances of $T_{k,N}$ and T and to estimate the 5^{th} percentile, critical value, of the test T. The estimated values are used in Monte Carlo simulation to estimate the powers of $T_{k,N}$ and T, which are then compared with the theoretical power of Hotelling's T^2 test.

4.2 Monte Carlo Simulation Parameters

The method of estimating the powers of $T_{k,N}$ and T using computer simulation is similar to the method used by Whaley and Quade (1985) for estimating the power of the multidimensional runs test.

Suppose we have samples of sizes n and m from two d-dimensional populations with densities f(x) and g(x), where $f(x) = g(x+\Delta)$, so there

is a location shift of Δ between the two populations. A test of homogeneity of these two populations can be performed by calculating T, using the nearest neighbors, rejecting the null hypothesis that the two populations are homogeneous if T is too small. Since the two populations are not homogeneous, the power to detect this shift is the probability of rejecting the null hypothesis. This probability depends on the following factors:

1) The type I error. We set $\alpha = .05$.

2) The distribution of each population. We used multivariate normal distributions, so we can compare the power of the nearest neighbor tests with that of Hotelling's T^2 test.

3) The sample sizes. We let n = m = 5, 10, 25, and 50.

4) The number of dimensions. We let d = 2, 5, 10, and 20.

5) The magnitude and direction of the shift Δ.

4.3 The Procedure Used to Compute Power

4.3.1 Mean and variance computation

Since it is extremely difficult to compute the variances of $T_{k,N}$ and T, we estimated the means and variances of both tests using the following simulation procedure: For each size-dimension-correlation we generated 1000 sets of N(=n+m) d-dimensional uniform random numbers between zero and one using the IMSL program GGUBT. Then we converted the numbers to standard normal deviates. For the case

when $\rho \neq 0$ we generated an extra deviate Q for each observation, and added γQ to each component, and divided the sum by $(1 + \gamma^2)^{1/2}$ in order to have N(0, V) where

$$V = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ \rho & \rho & \cdots & 1 \end{bmatrix} \quad \text{and } \rho = \frac{\gamma^2}{1 + \gamma^2}.$$

We let $\gamma = .75$ so that $\rho = .36$. Then $T_{k,N}$ (k = 1, 2, 3) and T were calculated for each set and their means and standard deviations were computed. Moreover, the 5^{th} percentile of T was calculated. The results are in Appendix I.

4.3.2 Power computation

To compute the power of each test, for each size-dimension-correlation-shift we generated 200 sets of N d-dimensional uniform random numbers between zero and one using the IMSL program GGUBT. Then we converted the numbers to standard normal deviates. Then we added the appropriate shift value to each dimension of the same set of m observations in each set of n+m. When $\rho = 0$, the power depends only on the amount of shift. But when $\rho = .36$, it depends on both the direction and the amount of shift. For $\rho = 0$, we let $\Delta = (\Delta, \Delta, \ldots, \Delta)$ while for $\rho = .36$, first we will use the same

direction shift, SDS, Δ as when $\rho = .00$ and then we will use the opposite direction shift, ODS, by letting $\Delta = (\Delta, \ldots, \Delta, -\Delta, \ldots, -\Delta)$ if d even and $\Delta = (\Delta, \ldots, \Delta, -\Delta, \ldots, -\Delta, 0)$ if d is odd, so we use the same amount of shift in every dimension. Moreover, we choose Δ such that the power using Hotelling's T^2 test is .70 or .90. The procedure for finding the amount of shift and the value of Δ for each size-dimension-correlation-shift combination that we investigated are presented in Appendices II and III.

For any particular size-dimension-correlation-shift combination, we determined whether to accept or reject the null hypothesis by using first the normal distribution approximation and then the critical value. We then repeated this for 199 more trials, and estimated the power for a given size-dimension-correlation-shift combination by the proportion of trials for which the null hypothesis was rejected.

4.4 Power Results

4.4.1 Samples of size 5

The estimated power of $T_{k,N}$ (k = 1, 2, 3), T using the normal distribution approximation, T_z, and T using the critical value, T_c, when the theoretical power of Hotelling's T^2 is .70 or .90 are presented in Table 4.4.1. As we expected, the power of the proposed test T is always greater than or equal that of Schilling's test $T_{k,N}$ (k = 1, 2, 3). In general, we can say that $T_{k,N}$ and T have greater power than that of T^2 when d = 5 while $T_{k,N}$ has poorer power but T still has greater power than both $T_{k,N}$ and T^2 when d = 2 except for the case

when using the critical value, T_c, and ODS. Also, for the case $\rho = .36$, the power using the same direction shift is greater than that when using the opposite direction shift. Moreover, the powers corresponding to the case $\rho = .36$ and ODS are poorer than those when $\rho = .00$. Finally, increasing the dimension increases the power.

4.4.2 Samples of size 10

The estimated powers of $T_{k,N}$ (k = 1, 2, 3), T using the normal distribution approximation, T_z, and T using the critical value, T_c, when the theoretical power of Hotelling's T^2 is .70 or .90 are presented in Table 4.4.2. For the case $\rho = .36$ the powers of $T_{k,N}$ (k = 1, 2, 3) and T are greater when using SDS than when using ODS. The powers of $T_{k,N}$ (k = 1, 2, 3) and T are poorer than the theoretical power of Hotelling's T^2 when d = 2 except the case with $\rho = .36$ and SDS. The worst power of T, compared to the theoretical power of Hotelling's T^2, .70, is for the case d = 2 and ODS when the critical value is used..

4.4.3 Samples of size 25

The estimated powers of $T_{k,N}$ (k = 1, 2, 3), T using the normal distribution approximation, T_z, and T using the critical value, T_c, when the theoretical power of Hotelling's T^2 is .70 or .90 are presented in Table 4.4.3. Comparing with the theoretical power of Hotelling's T^2, .70, the worst powers of T are when $\rho = .36$ and ODS while the best powers are when $\rho = .36$ and SDS. The minimum power of T, comparing with .70, the theoretical power of Hotelling's T^2, is for the

case d = 5 , ρ =.36, and ODS while the maximum power is for the case d = 20, ρ = .36, and SDS. Also, the minimum power of T, comparing with .90, the theoretical power of Hotelling's T^2, is for the case d = 5, ρ = .36, and ODS while the maximum power is for the case d = 10, 20, ρ = .36, and SDS. Moreover, for the case ρ = .36 and ODS the powers of both $T_{k,N}$ (k = 1, 2, 3) and T are poorer than the theoretical power of Hotelling's T^2. Also, when the components are uncorrelated or correlated (SDS) the powers of $T_{k,N}$ (k = 1, 2, 3) and T are poorer than the theoretical power of Hotelling's T^2 when d = 2 but they are better when d = 20.

4.4.4 Samples of size 50

The estimated powers of $T_{k,N}$ (k = 1, 2, 3), T using the normal distribution approximation, T_z, and T using the critical value, T_c, when the theoretical power of Hotelling's T^2 is .70 or .90 are presented in Table 4.4.4. The powers of $T_{k,N}$ (k = 1, 2, 3) are very poor comparing with both the theoretical power of Hotelling's T^2 and the empirical power of T. But the powers of T are poorer than the theoretical power of Hotelling's T^2 when ρ = .36 and ODS and better when ρ = .36, SDS, and larger dimensions. For the case ρ = .00, T is better for larger dimensions and poorer for smaller ones when compared to the theoretical power of Hotelling's T^2 test.

Table 4.4.1

Estimated power of $T_{k,N}$ and T when the theoretical power of Hotelling's T^2 is .70 or .90

Samples of size 5

ρ	d	Power of T^2 = .70					Power of T^2 = .90				
		$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c	$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c
.00	2	.475	.610	.660	.840	.755	.745	.845	.900	.955	.925
	5	.915	.965	.985	.995	.995	.985	.995	1.00	1.00	1.00
.36 SDS	2	.555	.660	.725	.885	.835	.770	.895	.920	.975	.905
	5	.960	1.00	1.00	1.00	1.00	.995	1.00	1.00	1.00	1.00
.36 ODS	2	.455	.580	.600	.720	.675	.770	.895	.920	.975	.955
	5	.835	.945	.920	.960	.945	.995	1.00	1.00	1.00	1.00

Table 4.4.2

Estimated power of $T_{k,N}$ and T when the theoretical power of Hotelling's T^2 is .70 or .90

Samples of size 10

ρ	d	Power of T^2 = .70					Power of T^2 = .90				
		$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c	$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c
	2	.270	.415	.475	.685	.655	.415	.645	.705	.905	.875
.00	5	.295	.550	.610	.820	.790	.515	.785	.840	.960	.935
	10	.690	.830	.870	.960	.940	.915	.920	.985	1.00	1.00
.36	2	.375	.535	.550	.755	.700	.595	.750	.790	.940	.920
SDS	5	.410	.710	.855	.965	.950	.690	.895	.970	1.00	.995
	10	.790	.965	.975	1.00	1.00	.950	.995	1.00	1.00	1.00
.36	2	.370	.460	.465	.690	.615	.585	.680	.725	.895	.855
ODS	5	.325	.520	.595	.740	.660	.525	.745	.835	.930	.915
	10	.550	.690	.735	.890	.865	.795	.930	.940	.990	.990

Table 4.4.3

Estimated power of $T_{k,N}$ and T when the theoretical power of Hotelling's T^2 is .70 or .90

Samples of size 25

ρ	d	Power of T^2 = .70					Power of T^2 = .90				
		$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c	$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c
.00	2	.210	.240	.290	.700	.650	.375	.475	.555	.880	.865
	5	.205	.360	.415	.775	.760	.390	.570	.640	.940	.915
	10	.260	.385	.425	.850	.805	.415	.595	.685	.975	.965
	20	.350	.480	.550	.910	.900	.560	.730	.795	.990	.990
.36 SDS	2	.165	.225	.280	.755	.690	.305	.415	.505	.925	.915
	5	.305	.445	.485	.930	.900	.475	.680	.775	.980	.975
	10	.435	.575	.690	.985	.980	.650	.790	.875	1.00	1.00
	20	.700	.875	.930	1.00	1.00	.900	.965	.985	1.00	1.00
.36 ODS	2	.180	.240	.285	.610	.545	.280	.390	.455	.850	.795
	5	.185	.310	.290	.555	.515	.340	.485	.505	.790	.745
	10	.230	.265	.330	.570	.525	.365	.455	.490	.840	.815
	20	.300	.395	.465	.690	.655	.400	.610	.695	.915	.910

Table 4.4.4

Estimated power of $T_{k,N}$ and T when the theoretical power of Hotelling's T^2 is .70 or .90

Samples of size 50

ρ	d	Power of T^2 = .70					Power of T^2 = .90				
		$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c	$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c
.00	2	.160	.215	.290	.655	.620	.250	.385	.420	.830	.805
	5	.135	.185	.210	.660	.615	.220	.300	.400	.880	.840
	10	.155	.260	.280	.755	.715	.210	.380	.430	.930	.925
	20	.195	.300	.415	.795	.780	.270	.460	.615	.960	.955
.36 SDS	2	.155	.165	.195	.665	.600	.265	.315	.380	.890	.860
	5	.160	.255	.305	.835	.810	.295	.435	.520	.955	.895
	10	.255	.390	.455	.930	.915	.395	.540	.670	.995	.985
	20	.560	.690	.765	1.00	1.00	.665	.800	.830	1.00	1.00
.36 ODS	2	.175	.225	.235	.595	.505	.245	.315	.395	.825	.765
	5	.130	.195	.210	.550	.460	.215	.340	.390	.780	.705
	10	.200	.265	.270	.510	.465	.285	.350	.425	.730	.660
	20	.175	.215	.250	.395	.350	.260	.385	.415	.650	.600

4.5 Conclusions

From the results in Section 4.4, we noticed that the power of the proposed test T is always better than that proposed by Schilling $T_{k,N}$ (k=1, 2, 3) except in a few cases the powers of $T_{3,N}$ and T are the same. For small sample sizes the powers of T are better than the theoretical power of T^2 in most cases while $T_{k,N}$ is better in some cases especially when d is larger. However, the theoretical power of T^2 in most cases is better than the empirical powers of both $T_{k,N}$ and T when the sample sizes are large especially for the case $\rho = .36$ and ODS. Moreover, when the theoretical power of T^2 and d were fixed the powers of $T_{k,N}$ and T decreased when the sample size increased. So, based on the empirical power calculations it is better to use T when the sample size is small while it is better to use Hotelling's T^2 when the sample size is large.

The same as Whaley (1983), we designed our simulation so that we could use Hotelling's T^2 as a benchmark under the mistaken impression that, since Hotelling's T^2 is a uniformly most powerful test when the data are normal, the power of our test, T, could not exceed its power. So, the fact that our estimated power sometimes exceeded the Hotelling's T^2 power was surprising. However, Hotelling's T^2 is optimal among tests which are invariant with respect to nonsingular linear transformations (Anderson (1958), Theorem 5.5.2, pp 116) or which have power depending only on the noncentrality parameter (Anderson (1958), Theorem 5.5.4, pp 118).

Our test depends on the ranking of distances between points. But

linear transformations do not necessarily preserve that.

For example, in two-dimensional space, let there be three points

$X = (0, 1)$ $Y = (1, 0)$ and $Z = (2, 0)$

Distance $XY = \sqrt{2}$ $XZ = \sqrt{5}$

Apply linear transformation $\begin{pmatrix} 1 & 4 \\ 4 & 7 \end{pmatrix}$

(The matrix doesn't have to be positive definite -- just nonsingular)

Then $X \longrightarrow (4, 7)$ $Y \longrightarrow (1, 4)$ and $Z \longrightarrow (2, 8)$

Now, distance $XY = \sqrt{18}$ and distance $XZ = \sqrt{5}$ still. The ordering or ranking of the distances has reversed. So, our test is not within the first class for which Hotelling's T^2 test is best, the invariant class. Moreover, we have strong evidence that our test does not depend on the noncentrality parameter only, because we have deliberately chosen our location shifts between populations for a given sample size and dimension such that $(nm/N)\Delta' V^{-1}\Delta$ is constant in order to fix the power of Hotelling's T^2. (See Appendices II and III). Yet our power varies, sometimes enormously, within a particular size-dimension combination. So, our test is also not within the second class for which Hotelling's T^2 test is best, the noncentrality parameter class.

CHAPTER V

AN APPLICATION OF THE NEAREST NEIGHBORS TEST

5.1 A description of the data

Fisher's iris data , containing the sepal and petal widths and lengths of three species of iris, have been used by many authors to illustrate various statistical procedures. In this Chapter we will compare Iris virginica to Iris versicolor with respect to the two sepal measurements. A listing of these data is in Table 5.1.1. The mean, standard deviation, median, minimum, and maximum sepal width and length for each of the two species are presented in Table 5.1.2.

Whaley (1983) used the Shapiro-Wilk (1965) statistic to test the normality of the sepal width and sepal length in each species, and found p-values ranging from .35 to .45. This indicates that the marginal distributions of sepal width and sepal length are not far from normal, and we assume that the joint distributions are bivariate normal.

The Pearson correlation between sepal width and sepal length for the Iris virginica data is .46, while it is .53 for the Iris versicolor data. Under the null hypothesis that the distributions of sepal measurements of Iris virginica and Iris versicolor are the same, the

Pearson correlation for the two data combined is .55. Thus, there is a strong positive correlation between the two measurements.

Table 5.1.1

A Subset of the Fisher Iris Data to be Used in Subsequent Analysis

Iris virginica (n = 50)		Iris versicolor (m = 50)	
Sepal Width (cm)	Sepal Length (cm)	Sepal Width (cm)	Sepal Length (cm)
3.3	6.3	3.2	7.0
2.7	5.8	3.2	6.4
3.0	7.1	3.1	6.9
2.9	6.3	2.3	5.5
3.0	6.5	2.8	6.5
3.0	7.6	2.8	5.7
2.5	4.9	3.3	6.3
2.9	7.3	2.4	4.9
2.5	6.7	2.9	6.6
3.6	7.2	2.7	5.2
3.2	6.5	2.0	5.0
2.7	6.4	3.0	5.9
3.0	6.8	2.2	6.0
2.5	5.7	2.9	6.1
2.6	6.1	3.0	5.4
2.8	5.8	2.9	5.6
3.2	6.4	3.1	6.7
3.0	6.5	3.0	5.6
3.8	7.7	2.7	5.8
2.6	7.7	2.2	6.2
2.2	6.0	2.5	5.6

Table 5.1.1 (continued)

Iris virginica		Iris versicolor	
Sepal Width	Sepal Length	Sepal Width	Sepal Length
3.2	6.9	3.2	5.9
2.8	5.6	2.8	6.1
2.8	7.7	2.5	6.3
2.7	6.3	2.8	6.1
3.3	6.7	2.9	6.4
3.2	7.2	3.0	6.6
2.8	6.2	2.8	6.8
3.0	6.1	3.0	6.7
2.8	6.4	2.9	6.0
3.0	7.2	2.6	5.7
2.8	7.4	2.4	5.5
3.8	7.9	2.4	5.5
2.8	6.4	2.7	5.8
2.8	6.3	2.7	6.0
3.0	7.7	3.4	6.0
3.4	6.3	3.1	6.7
3.1	6.4	2.3	6.3
3.0	6.0	3.0	5.6
3.1	6.9	2.5	5.5
3.1	6.7	2.6	5.5
3.1	6.9	3.0	6.1
2.7	5.8	2.6	5.8
3.2	6.8	2.3	5.0
3.3	6.7	2.7	5.6
3.0	6.7	3.0	5.7
2.5	6.3	2.9	5.7
3.0	6.5	2.9	6.2
3.4	6.2	2.5	5.1
3.0	5.9	2.8	5.7

Table 5.1.2

Descriptive Statistics - Fisher Iris Data

(n = m = 50)

		Iris virginica	Iris versicolor
Sepal Width (cm)	Mean	2.97	2.77
	St. dev.	.32	.31
	Median	3.00	2.80
	Minimum	2.20	2.00
	Maximum	3.80	3.40
Sepal Length (cm)	Mean	6.59	5.94
	St. dev.	.64	.52
	Median	6.5	5.9
	Minimum	4.90	4.90
	Maximum	7.90	7.00

5.2 Analysis

The hypothesis to be tested is that the distributions of the sepal measurements of Iris virginica and Iris versicolor are the same. Since we assume that the data are bivariate normal, we can use Hotelling's T^2 test to test this hypothesis.

$$T^2 = \frac{nm}{n+m} (\bar{x}_1 - \bar{x}_2)' S^{-1} (\bar{x}_1 - \bar{x}_2)$$

where

n = number of Iris virginica observations in the sample,

m = number of Iris versicolor observations in the sample,

$\bar{x}_1$ = vector of the mean sepal width and length in the Iris virginica sample,

$\bar{x}_2$ = vector of the mean sepal width and length in the Iris versicolor sample,

S = pooled estimate of the common covariance matrix

and

$$\frac{n+m-d-1}{(n+m-2)\,d} T^2 \sim F_{d,n+m-d-1}$$

where

d = number of measurements (dimensions).

In our case, $n = m = 50$, $d = 2$, $\bar{x}_1 = (2.974, 6.588)'$, $\bar{x}_2 = (2.77, 5.936)'$, and

$$S = \begin{bmatrix} .101 & .089 \\ .089 & .335 \end{bmatrix}$$

So, we have

$$T^2 = 25(.204, .652) \begin{vmatrix} 12.925 & -3.448 \\ -3.448 & 3.901 \end{vmatrix} \begin{vmatrix} .204 \\ .652 \end{vmatrix}$$

$$= 31.98,$$

and

$$\frac{97}{196} T^2 = 15.827 \simeq F_{.999999,2,97}.$$

Thus, the p-value is .000001 and we reject the hypothesis that the two species have the same distribution of sepal measurements.

Now, to test the hypothesis using the nearest neighbors tests, we have to estimate the means and variances of $T_{k,N}(k = 1, 2, 3)$ and T. Under the null hypothesis that the two species have the same distribution of sepal measurements, Pearson correlation between sepal width and length for the combined data is .55, from which $\gamma = 1.11$. As in Chapter IV, we generated 1000 sets of 100 two-dimensional uniform random numbers between zero and one using the IMSL Fortran

program GGUBT. Then we converted the numbers to standard normal deviates. Moreover, we generated an extra deviate Q for each observation added γQ to each measurement, and divided the sum by $(1 + \gamma^2)^{1/2}$ in order to have N(0, V) where

$$V = \begin{bmatrix} 1 & .55 \\ .55 & 1 \end{bmatrix}$$

Then, the means and standard deviations of $T_k = NkT_{k,N}$ (k = 1, 2, 3) and T are computed. These values are presented in Table 5.2.1

Table 5.2.1

Estimated mean and standard deviation of T_k, and T

Two-dimensional samples of size 50

	T_1	T_2	T_3	T
Mean	49.80	99.52	149.05	249953.00
St. dev.	6.41	8.96	11.00	1813.91

The p-value when using $T_{1,N}$, $T_{3,N}$, and T is $< .000001$ and it is .00080 when using $T_{2,N}$. Therefore, using the nearest neighbors tests, the hypothesis that the two species have the same distribution of sepal measurements is rejected. Also, using the critical value, T_c, we

have $T = 174075$ and $T_c = 246484$. So, using the critical value, T_c, we have $T < T_c$ therefore we reject the null hypothesis.

Now, recall from Chapter 4 that we have estimated the power of the nearest neighbors tests for the case when the sample sizes are 50 and the number of dimensions is 2, which corresponds to the number of each species and the number of measurements in our analysis. So, it is of interest to look at the results of our simulations noting that we had $\rho = .36$, which is not too far from the Pearson correlation between sepal width and sepal length, .55. The estimated power of $T_{k,N}$ ($k = 1, 2, 3$) and T when the theoretical power of Hotelling's T^2 is .70 or .90, $d = 2$, $\rho = .36$, and $n = m = 50$ are presented in Table 5.2.2.

Because sepal width and length are positively correlated for both species, since the mean difference between Iris virginica and Iris versicolor is positive for both measurements, we have a same direction shift. Alternatively, if the difference had been positive for one measurement and negative for the other, we would had an opposite direction shift.

Table 5.2.2

Estimated power of $T_{k,N}$ and T when the theoretical power of Hotelling's T^2 is .70 or .90

Two-dimensional samples of size 50

ρ	Power of T^2 = .70					Power of T^2 = .90				
	$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c	$T_{1,N}$	$T_{2,N}$	$T_{3,N}$	T_z	T_c
.36 SDS	.155	.165	.195	.665	.600	.265	.315	.380	.890	.860
.36 ODS	.175	.225	.235	.595	.505	.245	.315	.395	.825	.765

CHAPTER VI

SUMMARY AND SUGGESTIONS FOR FUTURE RESEARCH

6.1 Summary

Schilling (1986) proposed a test of homogeneity of two populations based on the first k-nearest neighbors. His test does not take into account the ranks of the k nearest neighbors with respect to nearness, given that they are among the first k. Furthermore, we do not know which value of k gives the best results. So, we defined our test based on all nearest neighbors and not only the first k neighbors. Moreover, our test takes into account the ranks. We studied both exact and asymptotic properties of the test. Also, we used the nearest neighbors technique to prove the normality of WilcoxonMann-Whitney statistic via the martingale central limit theorem.

Because of the difficulty in computing the variance of the test, we used computer simulations to estimate the mean and variance of the proposed test. Monte Carlo simulation was used to compute the power of the test. Then, the computed power of our test was compared with that of Schilling's test, which was also computed using Monte Carlo simulation. Then, the power of both tests were compared with the theoretical power of Hotelling's T^2 test.

For detecting location shift differences between two populations using Schilling's test for k = 1, 2, 3, and Hotelling's T^2 test, we found that in many cases our test compares quite favorably with Hotelling's T^2 test and in most cases it compares favorably with Schilling's test.

Finally, using a subset of the Fisher iris data, the three tests were used to test the hypothesis that the distribution of the two sepal measurements of the two species of iris are the same.

6.2 Suggestion for future research

There are many directions that future research in this area can take. Some of these directions are:

1. Studying in detail the exact computation of the variance of the proposed test, i.e., without using computer simulations.

2. Using computer simulations to find the value of k which gives the best results when using Schilling's test.

3. Investigating alternative methods to prove the normality of the proposed test.

4. Performing computer simulations to find the smallest sample size required such that the distribution of the test which is proposed by Schilling is asymptotically normal.

5. Performing Monte Carlo simulations using other sample sizes, numbers of dimensions, variance matrices, distributions, etc.

6. Studying in detail the power and other asymptotic properties of the tests proposed by Schilling and us theoretically.

7. Investigating alternative methods, other than the power, to compare the tests to each other, such as expected significance level (Dempster and Schatzoff (1965) and Silva and Quade (1980)).

8. Introducing weights in the proposed test.

9. Constructing confidence regions for the nearest neighbor statistic. How practical is it to construct confidence regions, and if it is practical, how useful are they?

10. Generalizing the nearest neighbor tests to more than two populations.

APPENDICES

Appendix I: Empirical means and standard deviations, the exact means of Schilling's test and our test, and the empirical 5^{th} percentile of our test T.

Appendix II: Finding the location shift differences between two multivariate normal populations that yields a specified power using Hotelling's T^2 statistic.

Appendix III: The location shift between two multivariate normal populations such that the power to detect such a difference using Hotelling's T^2 is .70 or .90.

Appendix IV: Computer programs.

Appendix I

The empirical means and standard deviations and the exact means of $T_k = NkT_{k,N}$ (k = 1, 2, 3) and T for different size-dimension-correlation are presented in the Tables I.1 - I.4 while the empirical 5^{th} percentile, the critical value, of T is presented in Table I.5.

Table I.1

Empirical Mean(standard deviation) and

exact mean of T_k and T

Samples of size 5

ρ	d	T_1	T_2	T_3	T
.00	2	4.40 (1.97)	8.86 (2.63)	13.32 (2.88)	249.84 (15.05)
	5	4.42 (1.86)	8.85 (2.55)	13.26 (2.71)	250.23 (14.80)
.36	2	4.49 (2.03)	8.94 (2.73)	13.37 (3.03)	249.85 (15.38)
	5	4.37 (1.91)	8.78 (2.53)	13.30 (2.89)	249.76 (15.31)
Exact mean		4.44	8.89	13.33	250

Table I.2

Empirical Mean(standard deviation) and

exact mean of T_k and T

Samples of size 10

ρ	d	T_1	T_2	T_3	T
.00	2	9.50 (2.90)	19.00 (3.89)	28.45 (4.63)	1999.49 (65.41)
	5	9.55 (2.75)	19.07 (3.76)	28.52 (4.70)	1997.43 (66.25)
	10	9.57 (2.65)	19.11 (3.54)	28.51 (4.21)	1998.13 (61.80)
.36	2	9.30 (2.82)	18.72 (3.81)	28.03 (4.43)	2003.26 (58.57)
	5	9.55 (2.77)	19.04 (3.80)	28.58 (4.49)	1998.02 (66.26)
	10	9.57 (2.73)	19.04 (3.88)	28.51 (4.60)	1999.26 (63.31)
Exact mean		9.47	18.95	28.42	2000

Table I.3

Empirical Mean(standard deviation) and

exact mean of T_k and T

Samples of size 25

ρ	d	T_1	T_2	T_3	T
.00	2	24.54 (4.45)	48.96 (6.47)	73.43 (7.67)	31248.40 (405.70)
	5	24.43 (4.19)	48.96 (5.93)	73.40 (7.35)	31261.80 (407.90)
	10	24.51 (4.14)	49.01 (5.95)	73.65 (7.15)	31255.60 (399.25)
	20	24.68 (4.21)	49.12 (5.95)	73.55 (7.22)	31266.00 (397.07)
.36	2	24.64 (4.44)	49.20 (6.44)	73.62 (7.73)	31247.20 (406.87)
	5	24.27 (4.33)	48.77 (6.07)	73.23 (7.40)	31252.10 (418.31)
	10	24.62 (4.20)	49.33 (5.97)	73.90 (7.36)	31250.70 (420.02)
	20	24.53 (4.05)	49.07 (5.72)	73.59 (6.96)	31259.70 (404.63)
Exact mean		24.49	48.98	73.47	31250

Table I.4

Empirical Mean(standard deviation) and

exact mean of T_k and T

Samples of size 50

ρ	d	T_1	T_2	T_3	T
.00	2	49.49 (6.17)	98.85 (9.06)	148.34 (11.19)	249908.00 (1715.06)
	5	49.84 (6.18)	99.68 (8.83)	149.24 (11.01)	249882.00 (1696.40)
	10	49.32 (5.89)	98.82 (8.30)	148.07 (10.37)	249974.00 (1607.57)
	20	49.63 (5.98)	99.13 (8.28)	148.55 (9.96)	249974.00 (1613.37)
.36	2	49.57 (6.27)	99.18 (9.29)	148.83 (11.29)	249956.00 (1777.23)
	5	49.64 (5.98)	99.38 (8.73)	148.86 (10.48)	249977.00 (1693.29)
	10	49.49 (6.01)	99.12 (8.69)	148.44 (10.43)	249942.00 (1752.88)
	20	49.27 (5.52)	98.70 (8.21)	148.34 (10.22)	249932.00 (1794.44)
Exact mean		49.49	98.99	148.48	250000

Table I.5

Empirical 5th percentile, critical value, of T

d	n=m	T_c	
		ρ = .00	ρ = .36
2	5	221	221
	10	1870	1880
	25	30427	30442
	50	246700	246169
5	5	224	221
	10	1868	1878
	25	30512.5	30493
	50	246759	246715
10	10	1884	1881
	25	30531	30494.5
	50	246874	246633
20	25	30568	30570.5
	50	247140	246590

Appendix II

Finding The Location Shift Difference Between Two Multivariate Normal Populations That Yields A Specified Power Using Hotelling's T^2 Statistic

Whaley (1983) found the location shift difference that yields a specified power as follows:

If observations from the first population are $N(\mu_1, V)$ and observations from the second population are $N(\mu_2, V)$, the usual statistic that tests whether μ_1 and μ_2 are equal is Hotelling's T^2 which has the distribution

$$\frac{n+m-d-1}{(n+m-2)d} T^2 \sim F_{d,n+m-d-1,\delta} \quad ,$$

where $\delta = \frac{nm}{N} (\mu_1-\mu_2)' V^{-1} (\mu_1-\mu_2)$ is the noncentrality parameter, and under H_0, $\delta = 0$. Since the power is defined as $P(\text{Reject } H_0 | H_0 \text{False})$, the power at α level of significance is $P(X > F_{1-\alpha,d,N-d-1,0})$ where $X \sim F_{d,N-d-1,\delta}$. So, we can set δ to achieve the power that we desire. So, we fix δ, and since n and m are known, we can solve for μ_1 and μ_2 for a given V.

If we let $\Delta = \mu_1-\mu_2 = (\Delta, \ldots, \Delta)'$, then $\delta = \frac{nm}{N} \Delta^2 \text{sum}(V^{-1})$ where $\text{sum}(V^{-1})$ is the sum of the elements of V^{-1}. In our case $V = I$,

so $V^{-1} = I$ and $\text{sum}(V^{-1}) = d$. As a result $\Delta = \{N\delta/(nmd)\}^{1/2}$.

Now suppose $V = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$. Then

$$V^{-1} = \frac{1}{1-\rho}\{ I - \frac{1}{1+(d-1)\rho} \mathbf{1}'\mathbf{1}\} \text{ and}$$

$$\text{sum}(V^{-1}) = \frac{d}{1-\rho}\{1 - \frac{\rho}{1+(d-1)\rho}\} + \frac{d(d-1)}{1-\rho}\{- \frac{\rho}{1+(d-1)\rho}\} = \frac{d}{1+(d-1)\rho}.$$

So, in this case $\Delta = [N\delta\{1+(d-1)\rho\}/(nmd)]^{1/2}$.

If we let $\Delta_1=\Delta_2= \cdots =\Delta_{d/2}=\Delta$ and $\Delta_{(d/2)+1}= \cdots =\Delta_d=-\Delta$, where d is even, then

$$\Delta'V^{-1}\Delta = \Delta^2[\frac{d}{1-\rho}\{ I - \frac{\rho}{1+(d-1)\rho}\} + \frac{d(d-2)}{2(1-\rho)}\{ - \frac{\rho}{1+(d-1)\rho}\}$$

$$- \frac{d^2}{2(1-\rho)}\{\frac{-\rho}{1+(d-1)\rho}\}] = \frac{d\Delta^2}{1-\rho}.$$

Then, $\Delta = \{N\delta(1-\rho)(nmd)\}^{1/2}$

Finally, if $\Delta_1=\Delta_2= \cdots =\Delta_{(d-1)/2}=\Delta$, $\Delta_{(d+1)/2}= \cdots =\Delta_{d-1}=-\Delta$, and $\Delta_d = 0$, where d is odd, then

$$\Delta' V^{-1} \Delta = \Delta^2 [\frac{d-1}{1-\rho} \{1 - \frac{\rho}{1+(d-1)\rho}\} + \frac{(d-1)(d-3)}{2(1-\rho)} \{- \frac{\rho}{1+(d-1)\rho}\}$$

$$- \frac{(d-1)^2}{2(1-\rho)} \{ - \frac{\rho}{1+(d-1)\rho}\}] = \frac{(d-1)\Delta^2}{1-\rho} .$$

Then $\Delta = [N\delta(1-\rho)/\{nm(d-1)\}]^{1/2}$. We will use the results from these last three cases in Appendix III.

Appendix III

The location shift between two multivariate normal populations such that the power to detect such a difference using Hotelling's T^2 test is .70 or .90

In Appendix II we have

$$\frac{n+m-d-1}{(n+m-2)d} T^2 \sim F_{d,n+m-d-1,\delta}$$

where $\delta = \frac{nm}{N} \Delta' V^{-1} \Delta$ is the noncentrality parameter. The power at $\alpha = .05$ is $P(X > F_{.95,N-d-1,0})$ where $X \sim F_{d,N-d-1,\delta}$. Furthermore, the relationship between Δ and the values of n, m, ρ, d, and δ is as follows:

Table III.1

ρ	d	Δ
.00	2,5,10,20	$\{N\delta/(nmd)\}^{1/2}$
.36 (SDS)	2,5,10,20	$[N\delta\{1+(d-1)\rho\}/(nmd)]^{1/2}$
.36 (ODS)	2,10,20	$\{N\delta(1-\rho)/(nmd)\}^{1/2}$
.36 (ODS)	5	$[N\delta(1-\rho)/\{nm(d-1)\}]^{1/2}$

The following table lists the values of $F_{.95,d,N-d-1}$, δ, and Δ needed to achieve a power of .70 and .90 for each size-dimension-correlation combination used in the simulation.

Table III.2

n=m	d	ρ	$F_{.95,d,N-d-1}$	$\delta_{.70}$	$\delta_{.90}$	$\Delta_{.70}$	$\Delta_{.90}$
5	2	.00	4.74	12.10	20.26	1.56	2.01
		.36 SDS				1.81	2.35
		.36 ODS				1.24	1.61
	5	.00	6.26	35.79	61.30	1.69	2.21
		.36 SDS				2.64	3.46
		.36 ODS				1.51	1.98
10	2	.00	3.59	9.22	15.20	.96	1.23
		.36 SDS				1.12	1.44
		.36 ODS				.77	.99
	5	.00	2.96	15.23	24.24	.78	.98
		.36 SDS				1.22	1.54
		.36 ODS				.70	.88

Table III.2 (continued)

n=m	d	ρ	$F_{.95,d,N-d-1}$	$\delta_{.70}$	$\delta_{.90}$	$\Delta_{.70}$	$\Delta_{.90}$
10	10	.00	3.14	30.14	47.80	.78	.98
		.36 SDS				1.60	2.01
		.36 ODS				.62	.78
25	2	.00	3.20	8.21	13.50	.57	.73
		.36 SDS				.67	.86
		.36 ODS				.46	.59
	5	.00	2.43	11.83	18.66	.44	.55
		.36 SDS				.68	.85
		.36 ODS				.39	.49
	10	.00	2.08	16.68	25.63	.37	.45
		.36 SDS				.75	.93
		.36 ODS				.29	.36
	20	.00	1.94	27.20	41.01	.33	.41
		.36 SDS				.92	1.13
		.36 ODS				.26	.32

Table III.2(continued)

n=m	d	ρ	$F_{.95,d,N-d-1}$	$\delta_{.70}$	$\delta_{.90}$	$\Delta_{.70}$	$\Delta_{.90}$
50	2	.00	3.09	7.94	13.05	.40	.51
		.36 SDS				.46	.60
		.36 ODS				.32	.41
	5	.00	2.31	11.08	17.46	.30	.37
		.36 SDS				.47	.58
		.36 ODS				.27	.33
	10	.00	1.94	14.81	22.70	.24	.30
		.36 SDS				.50	.62
		.36 ODS				.19	.24
	20	.00	1.70	20.99	31.44	.20	.25
		.36 SDS				.57	.70
		.36 SDS				.16	.20

APPENDIX IV

COMPUTER PROGRAMS

PROGRAM IV.1

```
//DISS001 JOB UNC.B.S625C,BARAKAT,PRTY=2
/*XEQ UNC
//*PW=XXXXXX
// EXEC SAS
//SYSIN DD *

      ************************************************************
      *                                                          *
      * THIS IS A SAS PROGRAM USED TO COMPUTE THE LOCATION       *
      * SHIFT DIFFERENCE BETWEEN TWO MULTIVARIATE NORMAL         *
      * POPULATIONS THAT YIELDS A SPECIFIED POWER, .70 OR        *
      * .90, USING HOTELLING'S T**2 STATISTIC.                   *
      *                                                          *
      ************************************************************;

   *** P1 = ALPHA
   *** NC = NONCENTRALITY PARAMETER
   *** N1=N2 IS THE NUMBER OF OBSERVATIONS FROM EACH POPULATION
   *** D = THE NUMBER OF DIMENSIONS
   *** DF1, DF2 ARE THE DEGREES OF FREEDOM
   *** NCP70 = NONCENTRALITY PARAMETER THAT YIELDS A POWER OF .70
   *** SH570 = LOCATION SHIFT THAT YIELDS A POWER OF .70 WHEN THE
   ***          CORRELATION BETWEEN DIMENSIONS IS ZERO
   *** SDS570 = LOCATION SHIFT THAT YIELDS A POWER OF .70 WHEN THE
   ***           CORRELATION IS .36 (SAME DIRECTION SHIFT)
   *** ODS570 = LOCATION SHIFT THAT YIELDS A POWER OF .70 WHEN THE
   ***           CORRELATION IS .36 (OPPOITE DIRECTION SHIFT) ***;

DATA ONE;
 INPUT P1 NC;
 CARDS;
 .05 0
 ;
DATA TWO;
 SET ONE;
 DO J=2,5,10,20;
 DO K=5,10,25,50;
 D=J;
 N1=K;
 N2=K;
 N=N1+N2;
```

```
DF1=D;
DF2=N-D-1;
P=1-P1;
F05=FINV(P,DF1,DF2,NC);
NCP70=FNONCT(F05,DF1,DF2,.30);
NCP90=FNONCT(F05,DF1,DF2,.10);

SH570=(N*NCP70/(N1*N2*D))**.5;
SH590=(N*NCP90/(N1*N2*D))**.5;

SDS570=(N*NCP70*(1+(D-1)*.36)/(N1*N2*D))**.5;
SDS590=(N*NCP90*(1+(D-1)*.36)/(N1*N2*D))**.5;

IF D=2 OR D=10 OR D=20 THEN DO;
ODS570=(N*NCP70*(1-.36)/(N1*N2*D))**.5;
ODS590=(N*NCP90*(1-.36)/(N1*N2*D))**.5;
OUTPUT;
END;

ELSE IF D=5 THEN DO;
ODS570=(N*NCP70*(1-.36)/(N1*N2*(D-1)))**.5;
ODS590=(N*NCP90*(1-.36)/(N1*N2*(D-1)))**.5;
OUTPUT;
END;

END;
END;
PROC PRINT;
VAR D N1 N2 F05 NCP70 SH570 SDS570 ODS570
                NCP90 SH590 SDS590 ODS590;
//
```

PROGRAM IV.2

```
//DISS002 JOB UNC.B.S625C,BARAKAT,PRTY=1,TIME=15
/*XEQ UNC
//*PW=XXXXXX
// EXEC FTVCLG,REGION=3500K,IMSL=DBLE
//C.SYSIN DD *
C
C     ********************************************************
C     *                                                      *
C     * THIS IS A SIMULATION PROGRAM USED TO COMPUTE THE     *
C     * POWER OF OUR TEST T AND SCHILLING'S TEST T1, T2,     *
C     * AND T3 CORRESPONDING TO K=1,2,3, WHEN THE THEOR-     *
C     * ITICAL POWER OF HOTELLING'S T**2 IS .90.  IT IS      *
C     * A FORTRAN PROGRAM WHICH USES THE IMSL SUBROUTINE     *
C     * LIBRARY.                                             *
C     *                                                      *
C     ********************************************************
C
      DIMENSION DRN(100,20,100),DRNS(100,100),TT05(100),
     1DRNN(100,100),H(100,100),TIK(100,100),TI(100),
     2DIFF(100),CS(100),CSN(100),R(100),RN(100,20),
     3T(100),TS1(100),TS2(100),TS3(100),T1(100),T2(100),T3(100)
      REAL*8 ESEED,DSEED,ZVAL,PVAL,ZVAL1,PVAL1,ZVAL2,PVAL2,
     1ZVAL3,PVAL3,IPW05,I1PW05,I2PW05,I3PW05,JPW05
C
      ESEED=123495678D0
      DSEED=123456789D0
C
C     IF THE CORRELATION BETWEEN DIMENSIONS WITHIN AN
C     OBSERVATION IS ZERO, SET GAMMA=0.
C
      GAMMA=.75
C
C     ID IS THE NUMBER OF DIMENSIONS IN THE DATA.
C     N1=N1 IS THE NUMBER OF OBSERVATIONS FROM EACH
C     POPULATION.  IS (OR XIS) IS THE NUMBER OF SAMPLES
C     GENERATED.
C
      N1=25
      N2=25
      XN1=25.
      XN2=25.
      XN=XN1+XN2
      N=N1+N2
      IS=100
      M=N1+1
      ID=5
C
C
C
```

```
C     SETTING THE POWER TO ZERO . WRITING THE ESTIMATED VALUES
C     OF THE MEAN AND STANDARD DEVIATION OF EACH TEST, AND
C     WRITING THE ESTIMATED CRITICAL VALUE OF OUR TEST.
C
      TC05=30493.
      JPW05=0.0
      ET=31252.1
      SDT=418.308
      IPW05=0.0
      ET1=24.266
      SDT1=4.33221
      I1PW05=0.0
      ET2=48.765
      SDT2=6.06882
      I2PW05=0.0
      ET3=73.227
      SDT3=7.39569
      I3PW05=0.0
C
      DO 101 KR=1,IS
C
C     DEFINING THE SHIFT DIFFERENCE BETWEEN POPULATIONS.
C
      DO 211 IZ=1,N1
      DIFF(IZ)=0.
211   CONTINUE
      DO 212 IZ=M,N
      DIFF(IZ)=.853
C
C     IF WE HAVE OPPOSITE DIRECTION SHIFT(ODS) THEN
C     DIFF(IZ)=.48870
C
212   CONTINUE
C
C     A STANDARD NORMAL RANDOM NUMBER IS GENERATED FOR
C     EACH OBSERVATION IN EACH TRIAL.  THIS NUMBER IS USED
C     TO CREATE STANDARD NORMAL RANDOM NUMBERS THAT ARE
C     CORRELATED BETWEEN DIMENSIONS WITHIN AN OBSERVATION.
C     THE INTER-DIMENSION CORRELATION IS
C     GAMMA**2/(1+GAMMA**2)=.36.
C
      CALL GGUBT(ESEED,N,CS)
      DO 143 ICS=1,N
      CALL MDNRIS(CS(ICS),CSN(ICS),IER)
143   CONTINUE
      DO 42 JR=1,20
      IF (JR .GT. ID) GO TO 42
      CALL GGUBT(DSEED,N,R)
      DO 43 IR=1,N
      CALL MDNRIS(R(IR),RX,IER)
C
```

```
C      THE SHIFT BETWEEN POPULATIONS IS POSITIVE IN ALL
C      DIMENSIONS (SAME DIRECTION SHIFT).
C
       RN(IR,JR)=((RX + GAMMA*CSN(IR))/SQRT(1. + GAMMA**2)) +
      1DIFF(IR)
C
C       ******************************************************
C       IF WE HAVE OPPOSITE DIRECTION SHIFT(ODS) THEN THE
C       PREVIOUS TWO LINES WILL BE REPLACED BY THE FOLLOWING
C       LINES:
C          IF (JR.GT.2) GO TO 243
C          RN(IR,JR)=((RX+GAMMA*CSN(IR))/SQRT(1.+GAMMA**2))+
C         1DIFF(IR)
C          GO TO 43
C       243 IF (JR.EQ.5) GO TO 244
C          RN(IR,JR)=((RX+GAMMA*CSN(IR))/SQRT(1.+GAMMA**2))-
C         1DIFF(IR)
C          GO TO 43
C       244 RN(IR,JR)=((RX+GAMMA*CSN(IR))/SQRT(1.+GAMMA**2))
C       ******************************************************
C
43     CONTINUE
42     CONTINUE
C
C      COMPUTING THE TEST STATISTICS T, TS1, TS2, AND TS3.
C
       DO 10 I=1,N
       DO 20 K=1,N
       DRNS(I,K)=0.0
       H(I,K)=0.0
       DO 30 J=1,ID
       DRN(I,J,K)=(RN(I,J) - RN(K,J))**2
       DRNS(I,K)=DRNS(I,K) + DRN(I,J,K)
30     CONTINUE
       DRNN(I,K)=SQRT(DRNS(I,K))
20     CONTINUE
10     CONTINUE
C
C
       DO 40 I=1,N1
       DO 50 J=M,N
       IT=0
       DO 60 L=1,N
       IF (DRNN(I,L) .GT. DRNN(I,J)) GO TO 60
       IT=IT+1
60     CONTINUE
       H(I,IT)=1.0
50     CONTINUE
40     CONTINUE
C
```

```
C
      DO 41 I=M,N
      DO 51 J=1,N1
      IT=0
      DO 61 L=1,N
      IF(DRNN(I,L) .GT. DRNN(I,J)) GO TO 61
      IT=IT+1
61    CONTINUE
      H(I,IT)=1.0
51    CONTINUE
41    CONTINUE
C
C
      TIK(1,1)=0.0
      T(KR)=0.0
      TS1(KR)=0.0
      TS2(KR)=0.0
      TS3(KR)=0.0
C
C
      DO 70 I=1,N
      TI(I)=0.0
      DO 80 K=2,N
      TIK(I,K)=TIK(I,K-1) + H(I,K)
      TI(I)=TI(I) + TIK(I,K)
80    CONTINUE
      T(KR)=T(KR) + TI(I)
      TS1(KR)=TS1(KR)+TIK(I,2)
      TS2(KR)=TS2(KR)+TIK(I,3)
      TS3(KR)=TS3(KR)+TIK(I,4)
70    CONTINUE
      TT05(KR)=T(KR)-TC05
      T1(KR)=XN-TS1(KR)
      T2(KR)=2.*XN-TS2(KR)
      T3(KR)=3.*XN-TS3(KR)
C
C     DETERMINING WHETHER OR NOT TO REJECT THE (FALSE)
C     NULL HYPOTHESIS THAT THE TWO POPULATIONS ARE THE
C     SAME AND COMPUTING THE ESTIMATED POWER OF THE TESTS
C     FOR ALPHA=.05.
C
      IF (TT05(KR).LT..0) JPW05=JPW05+1.0
      ZVAL=(T(KR)-ET)/SDT
      IF (ZVAL.LT.-1.645) IPW05=IPW05+1.0
      ZVAL1=(T1(KR)-ET1)/SDT1
      IF (ZVAL1.GT.1.645) I1PW05=I1PW05+1.0
      ZVAL2=(T2(KR)-ET2)/SDT2
      IF (ZVAL2.GT.1.645) I2PW05=I2PW05+1.0
      ZVAL3=(T3(KR)-ET3)/SDT3
      IF (ZVAL3.GT.1.645) I3PW05=I3PW05+1.0
C
```

```
C
101    CONTINUE
       WRITE(6,100) JPW05,IPW05,I1PW05,I2PW05,I3PW05
100    FORMAT(5(F12.6,4X))
       STOP
       END
//
```

PROGRAM IV.3

```
//DISS003 JOB UNC.B.S625C,BARAKAT,PRTY=2
/*XEQ UNC
//*PW=XXXXXX
// EXEC FTVCLG,IMSL=DBLE
//C.SYSIN DD *
C
C     ********************************************************
C     *                                                      *
C     * THIS FORTRAN PROGRAM COMPUTES THE P-VALUE FOR        *
C     * TESTING THE NULL HYPOTHESIS THAT IRIS VIRGINICA      *
C     * AND IRIS VERSICOLOR ARE THE SAME USING OUR TEST,     *
C     * T, AND SCHILLING'S TEST (K=1,2,3), T1, T2, AND       *
C     * T3.                                                  *
C     *                                                      *
C     ********************************************************
C
      DIMENSION RN(100,2),DRN(100,2,100),DRNS(100,100),
     1DRNN(100,100),H(100,100),TIK(100,100),TI(100)
      REAL*8 ZVAL,PVAL,ZVAL1,PVAL1,ZVAL2,PVAL2,ZVAL3,
     1PVAL3,T05,T,T1,T2,T3,TS1,TS2,TS3
C
C     ID IS THE NUMBER OF DIMENSIONS IN THE DATA.
C     N1=N2 IS THE NUMBER OF OBSERVATIONS FROM EACH
C     POPULATION.
C
      N1=50
      N2=50
      XN1=50.
      XN2=50.
      XN=XN1+XN2
      N=N1+N2
      M=N1+1
      ID=2
C
C     THE ESTIMATED MEAN AND STANDARD DEVIATION OF OUR
C     TEST AND SCHILLING'S TEST (K=1,2,3) AND THE CRITICAL
C     VALUE OF OUR TEST T.
C
      TC05=246484.
      ET=249953.
      SDT=1813.91
      ET1=49.804
      SDT1=6.41098
      ET2=99.524
      SDT2=8.96436
      ET3=149.048
      SDT3=11.0026
C
C
```

```
C
C     READING IN THE FISHER IRIS DATA.
C
      DO 101 I=1,100
      READ(5,200) RN(I,1),RN(I,2)
200   FORMAT(3X,F3.1,1X,F3.1)
101   CONTINUE
C
C     COMPUTING THE TEST STATISTICS T, TS1, TS2, TS3.
C
      DO 10 I=1,N
      DO 20 K=1,N
      DRNS(I,K)=0.0
      H(I,K)=0.0
      DO 30 J=1,ID
      DRN(I,J,K)=(RN(I,J) - RN(K,J))**2
      DRNS(I,K)=DRNS(I,K) + DRN(I,J,K)
30    CONTINUE
      DRNN(I,K)=SQRT(DRNS(I,K))
20    CONTINUE
10    CONTINUE
C
C
      DO 40 I=1,N1
      DO 50 J=M,N
      IT=0
      DO 60 L=1,N
      IF (DRNN(I,L) .GT. DRNN(I,J)) GO TO 60
      IT=IT+1
60    CONTINUE
      H(I,IT)=1.0
50    CONTINUE
40    CONTINUE
C
C
      DO 41 I=M,N
      DO 51 J=1,N1
      IT=0
      DO 61 L=1,N
      IF(DRNN(I,L) .GT. DRNN(I,J)) GO TO 61
      IT=IT+1
61    CONTINUE
      H(I,IT)=1.0
51    CONTINUE
41    CONTINUE
C
      TIK(1,1)=0.0
      T=0.0
      TS1=0.0
      TS2=0.0
      TS3=0.0
```

```
C
      DO 70 I=1,N
      TI(I)=0.0
      DO 80 K=2,N
      TIK(I,K)=TIK(I,K-1) + H(I,K)
      TI(I)=TI(I) + TIK(I,K)
80    CONTINUE
      T=T + TI(I)
      TS1=TS1+TIK(I,2)
      TS2=TS2+TIK(I,3)
      TS3=TS3+TIK(I,4)
70    CONTINUE
      TO5=T-TCO5
      T1=XN-TS1
      T2=2.*XN-TS2
      T3=3.*XN-TS3
C
C     CALCULATING THE Z-VALUE AND THE CORRESPONDING
C     P-VALUE FOR TESTING THE NULL HYPOTHESIS USING
C     THE NEAREST NEIGHBORS TESTS.
C
      ZVAL=(T-ET)/SDT
      CALL MDNORD(ZVAL,PVAL)
      ZVAL1=(T1-ET1)/SDT1
      CALL MDNORD(ZVAL1,PVAL1)
      PVAL1=1.0-PVAL1
      ZVAL2=(T2-ET2)/SDT2
      CALL MDNORD(ZVAL2,PVAL2)
      PVAL2=1.0-PVAL2
      ZVAL3=(T3-ET3)/SDT3
      CALL MDNORD(ZVAL3,PVAL3)
      PVAL3=1.0-PVAL3
C
      WRITE(6,100) T,T1,T2,T3
      WRITE(6,100) ZVAL,ZVAL1,ZVAL2,ZVAL3
      WRITE(6,100) PVAL,PVAL1,PVAL2,PVAL3
100   FORMAT(4(F20.6,4X))
      WRITE(6,*) TO5
      STOP
      END
//G.FT05F001 DD *
   3.3 6.3
   2.7 5.8
(LIST OF DATA )
//
```

PROGRAM IV.4

```
//DISS004 JOB UNC.B.S625C,BARAKAT,PRTY=2
/*XEQ UNC
//*PW=XXXXXX
// EXEC FTVCLG,IMSL=DBLE
//C.SYSIN DD *
C
C     ************************************************
C     *                                              *
C     * THIS FORTRAN PROGRAM COMPUTES THE P-VALUE FOR *
C     * TESTING THE NULL HYPOTHESIS THAT THE TWO POP- *
C     * ULATIONS (IRIS VIRGINICA AND IRIS VERSICOLOR) *
C     * ARE THE SAME USING HOTELLING'S T**2 TEST.     *
C     *                                              *
C     ************************************************
C
      DIMENSION RN(100,2)
      REAL*8 SUM1(2),SUM2(2),MEAN1(2),MEAN2(2),MEANDF(2),
     1A1(2,2),A2(2,2),S(2,2),SINV(2,2),WKAREA(10),H(2),HT,HT2,FVAL,
     2PVALN,PVAL
C
C     ID (OR XID) ARE THE NUMBER OF DIMENSIONS IN THE DATA.
C     N1=N2 IS THE NUMBER OF OBSERVATIONS FROM EACH
C     POPULATION.
C
      N=100
      N1=50
      N2=50
      XN=100.
      XN1=50.
      XN2=50.
      M=N1+1
      ID=2
      IID=N-ID-1
      XID=2.0
C
C     READING IN THE FISHER IRIS DATA.  THE FIRST 50
C     OBSERVATIONS ARE IRIS VERSICOLOR, THE SECOND 50 ARE
C     IRIS VIRGINICA.  EACH OBSERVATION HAS 4 VARIABLES -
C     SEPAL LENGTH, SEPAL WIDTH, PETAL LENGTH, AND PETAL
C     WIDTH.
C     ONLY THE SEPAL VARIABLES ARE USED.
C
      DO 101 I=1,100
      READ(5,100) RN(I,1),RN(I,2)
100   FORMAT(3X,F3.1,1X,F3.1)
101   CONTINUE
C
```

```
C     CALCULATING THE MEAN OF EACH VARIATE FOR EACH
C     SPECIES AND THE DIFFERENCE BETWEEN THE MEANS OF EACH
C     VARIATE, MEANDF(JD).  NOTE THAT THIS IS CALCULATED
C     BY SUBTRACTING IRIS VIRGINICA FROM IRIS VERSICOLOR.
C
      DO 51 JD=1,ID
      SUM1(JD)=0.
      SUM2(JD)=0.
      DO 61 I=1,N
      IF (I.GT.N1) GO TO 62
      SUM1(JD)=SUM1(JD)+RN(I,JD)
      GO TO 61
62    SUM2(JD)=SUM2(JD)+RN(I,JD)
61    CONTINUE
      MEAN1(JD)=SUM1(JD)/XN1
      MEAN2(JD)=SUM2(JD)/XN2
      MEANDF(JD)=MEAN1(JD)-MEAN2(JD)
      WRITE(6,200) JD,MEAN1(JD),MEAN2(JD),MEANDF(JD)
200   FORMAT((1X,I2,3(F12.6,3X))//)
51    CONTINUE
C
C     CALCULATING THE MATRIX OF SQUARES AND CROSS-PRODUCTS
C     FROM THE TWO SPECIES AND THEN THE POOLED COVARIANCE
C     MATRIX, S.
C
      DO 71 J2=1,ID
      DO 81 J1=1,J2
      A1(J1,J2)=0.
      A2(J1,J2)=0.
      DO 91 I=1,N
      IF (I.GT.N1) GO TO 92
      A1(J1,J2)=A1(J1,J2)+RN(I,J1)*RN(I,J2)-MEAN1(J1)*
     1MEAN1(J2)
      GO TO 91
92    A2(J1,J2)=A2(J1,J2)+RN(I,J1)*RN(I,J2)-MEAN2(J1)*
     1MEAN2(J2)
91    CONTINUE
      S(J1,J2)=(A1(J1,J2)+A2(J1,J2))/(XN-2.)
      S(J2,J1)=S(J1,J2)
      WRITE(6,300) J1,J2,A1(J1,J2),A2(J1,J2),S(J1,J2)
300   FORMAT((1X,I2,2X,I2,10X,3(F12.6,2X))//)
81    CONTINUE
71    CONTINUE
C
C     COMPUTING THE INVERSE OF THE POOLED COVARIANCE
C     MATRIX.
C
      IA=ID
      IDGT=0
      CALL LINV2F(S,ID,IA,SINV,IDGT,WKAREA,IER)
```

```
      WRITE(6,400) SINV(1,1),SINV(1,2),SINV(2,1),SINV(2,2)
400   FORMAT((10X,2(F12.6,3X))///)
C
C     CALCULATING THE HOTELLING'S T**2 STATISTIC AND
C     THE F-VALUE THEN THE P-VALUE FOR TESTING THE NULL
C     HYPOTHESIS.
C
      DO 11 J=1,ID
      H(J)=0.0
      DO 21 I=1,ID
      H(J)=H(J)+MEANDF(I)*SINV(I,J)
21    CONTINUE
11    CONTINUE
      HT=0.0
      DO 111 K=1,ID
      HT=HT+H(K)*MEANDF(K)
111   CONTINUE
      HT2=HT*XN1*XN2/XN
      FVAL=((XN-XID-1.0)*HT2)/((XN-2)*XID)
      WRITE(6,500) HT2,FVAL
500   FORMAT(10X,2(F12.6,3X)//)
      CALL MDFD(FVAL,ID,IID,PVALN,IER)
      PVAL=1.0 - PVALN
      WRITE(6,500) FVAL,PVAL
      STOP
      END
//G.FT05F001 DD *
   3.3 6.3
   2.7 5.8
( LIST OF DATA )
//
```

References

Anderson, T. W. (1958) , *Introduction to Multivariate Statistical Analysis.* Wiley, New York.

Bickel, P. J. and Breiman, L. (1983) , "Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test." *Annals of Probability*, 11, 185-214.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1976) "*Discrete Multivariate Analysis: Theory and Practice.*" MIT Press, Massachusetts.

Cheng, P. E. (1984) "Strong consistency of nearest neighbor regression function estimators." *Journal of Multivariate Analysis*, 15, 63-72.

Cover, T. M. and Hart, P. E. (1967) "Nearest neighbor pattern classification." *IEEE Transactions on Information Theory*, IT-13, 21-27.

Dempster, A. P. and Schatzoff, M. (1965) "Expected significance level as a sensitivity index for test statistics." *Journal of the American Statistical Association*, 60, 420-436.

Devroye, L. P. (1978) "The uniform convergence of nearest regression function estimators and their application in optimization." *IEEE Transactions on Information Theory*, IT-24, 142-151.

Devroye, L. P. (1980) "Consistency of a recursive nearest neighbor regression function estimate." *Journal of Multivariate Analysis*, 10, 539-550.

——————— and Wagner,T. J. (1977) "The strong uniform consistency of nearest neighbor density estimates." *Annals of Statistics*, 5, 536-540.

Feller, W. (1966) *An Introduction to Probability Theory and its Applications.* Vol. II. Wiley, New Tork.

Friedman, J. H. and Rafsky, L. (1979) "Multivariate generalization of the Wald-Wolfowitz and Smirnov two-sample tests." *Annals of Statistics*, 7, 697-717.

——————————— (1983) "Graph-theoretic measures of multivariate association and prediction." *Annals of Statistics*, 11, 377-391.

Friedman, J. H. and Steppel, S. (1974) "A nonparametric procedure for comparing multivariate point sets." Unpublished manuscript.

Fritz, J. (1975) "Distribution-free exponential error bound for nearest neighbor pattern classification." *IEEE Transactions on Information Theory*, IT-21, 552-557.

Fukunaga, K. and Hostetler, L. D. (1973) "Optimization of k-nearest neighbor density estimates." *IEEE Transactions on Information Theory*, IT-19, 320-326.

————— and Mantock, J. M. (1984) "Nonparametric data reduction." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6, 115-118.

Gibbons, J. D. (1985) *Nonparametric Statistical Inference.* Marcel Dekker, New York.

Goin, J. E. (1984) "Classification bias of the k-nearest neighbor algorithm." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6,379-381.

Henze, N. (1988) "A multivariate two-sample test based on the number of nearest neighbor type coincidences." *Annals of Statistics*, 16(2), 772-783.

Loftsgaarden, D. O. and Quesenberry, C. P. (1965) "A nonparametric estimate of a multivariate density function." *Annals of Mathematical Statistics*, 36, 1049-1051.

Mack, Y. P. and Rosenblatt, M. (1979) "Multivariate k-nearest neighbor density estimates." *Journal of Multivarite Analysis*, 9, 1-15.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, New York.

Moore, D. S. and Yackel, J. W. (1977) "Consistency properties of nearest neighbor density function estimators." *Annals of Statistics*, 5, 143-154.

Puri, M. L. and Sen, P. K. (1971) *Nonparametric Methods in Multivariate Analysis.* Wiley, New York.

Pyke, R. (1965) "Spacings." *Journal of the Royal Statistical Society*, Series B, 27, 395-436.

Randles, R. H. and Wolfe, D. A. (1979) *Introduction to the Theory of Nonparametric Statistics.* Wiley, New York.

Rogers, W. H. (1978) "Some convergence properties of k-nearest neighbor estimates." Ph.D dissertation. Dept of Statistics, Stanford University.

Salama, I. A. and Quade, D. (1981) "Using weighted ranking to test against ordered alternatives in complete blocks." *Communications in Statistics*, Theory and Methods, A10(4), 385-399.

——————————— (1982) "A nonparametric comparison of two multiple regressions by means of a weighted measure of correlation." *Communications in Statistics*, Theory and Methods, 11(11), 1185-1195.

Schilling, M. F. (1979) "Testing for goodness of fit based on nearest neighbors." Ph.D dissertation. Statistics, University of California, Berkeley.

——————————— (1983a) "Goodness of fit testing based on the weighted empirical distribution of certain nearest neighbor statistics." *Annals of Statistics*, 11, 1-12.

——————————— (1983b) "An infinite-dimensional approximation for nearest neighbor goodness of fit tests." *Annals of Statistics*, 11, 13-24.

——————————— (1986) "Mutual and shared neighbor probabilities: Finite and infinite dimensional results." *Advances in Applied Probability*, 18, 388-405.

——————————— (1986) "Multivariate two-sample tests based on nearest neighbors." *Journal of the American Statistical Association*, 81, 799-806.

Sen, P. K. and Salama, I. A. (1983) "The Spearman footrule and a Markov chain property." *Statistics and Probability Letters*, 1, 285-289.

Shapiro, S, S. and Wilk, M. B. (1965) "An analysis of variance test for normality (complete samples)." *Biometrika*, 52, 591-611.

Silva, C. and Quade, D. (1980) "Evaluation of weighted ranking using expected significance level." *Communications in Statistics*, Theory and Methods, A9(10), 1087-1096.

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

Shiryayev, A. N. (1984) *Probability*. Springer-Verlag, New York.

Stone, C. J. (1977) "Consistent nonparametric regression." *Annals of Statistics*, 5, 595-645.

Wagner, T. J. (1971) "Convergence of the nearest neighbor rule." *IEEE Transactions on Information Theory*, IT-17, 566-571.

———— (1973) "Strong consistency of a nonparametric estimate of a density function." *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3, 289-290.

Wald, A. and Wolfowitz,, J. (1940) "On a test whether two samples are from the same population." *Annals of Mathematical Statistics*, 11, 147-162.

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

———— (1960) "Two-sample tests for multivariate distributions." *Annals of Mathematical Statistics*, 31, 159-164.

Whaley, F. S. (1983) "Some properties of the two-sample multidimensional runs statistic." Ph.D dissertation. Dept of Biostatistics, University of North Carolina at Chapel Hill.

———— and Quade, D. (1985) "Optimizing the power of the two-sample multidimensional runs statistic: Guidelines based on computer simulation." *Communications in Statistics*, Simulation and Computation, 14, 1-11.

———————————— (1987) "Optimizing the Wald-Wolfowitz runs statistic using a linkage tolerance: Guidelines based on computer simulation." *Communications in Statistics*, Theory and Methods, 16(7), 2125-2138.