K-SAMPLE ANALOGUES OF K-S STATISTICS AND GROUP TESTS

K-SAMPLE ANALOGUES OF

THE KOLMOGOROV-SMIRNOV STATISTICS AND BINOMIAL GROUP TESTS

By

LUCILLE LU KOW ZING, DIPL., M.A.

A thesis

Submitted to the School of Graduate Studies

in Partial Fulfilment of the Requirements

for the Degree

Doctor of Philosophy

McMaster University

May 1979

DOCTOR OF PHILOSOPHY (1979) McMASTER UNIVERSITY

(Mathematics) Hamilton, Ontario

TITLE: K-sample Analogues of the Kolmogorov-Smirnov Statistics and Binomial Group Tests

AUTHOR: Lucille Lu Kow Zing

Dipl. (Hong Kong Baptist College, 1971)

M.A. (University of New Brunswick, 1973)

SUPERVISOR: Dr. I. Z. Chorneyko

NUMBER OF PAGES: vi, 154.

ABSTRACT

The Kolmogorov-Smirnov tests of homogeneity or goodness-of-fit and the binomial group tests for eliminating defectives are of different nature. But they are both popular in applications. The former are widely used in nonparametric comparison, while the later are usually adopted in the group screening problems. When the experimenter has k populations, k-sample statistics should be considered for the testing of homogeneity or goodness-of-fit. On the other hand, when there are k experimenters available for performing group testing on a given population, a k-sample group testing procedure should be used.

In this thesis, the distribution functions of k-sample analogues of the Kolmogorov-Smirnov statistics have been found under certain conditions and a k-sample group testing procedure has been defined. This procedure has also been shown to be optimal in the sense that it requires a minimum expected number of k-sample group tests for finding a single defective from a binomial population.

Our methods are mainly combinatorial: matrix enumeration, tree counting and construction algorithms are explored as a foundation of our study.

ACKNOWLEDGEMENTS

The author acknowledges with gratitude the suggestion of the problem, the help and guidance given to her by her supervisor, Dr. I. Z. Chorneyko throughout the course of this work. Special thanks are due to Dr. S. G. Mohanty for his critical and helpful comments without which the enumeration results of Chapter I would not have been improved and to Dr. A. Rosa for serving on the supervisory committee.

Finally, the author is grateful for the scholarships and assistantships awarded to her by the National Research Council of Canada and McMaster University.

TABLE OF CONTENTS

		Page
ABSTRACT		iii
ACKNOWLEDGEMENTS		iv
CHAPTER I -	ON THE ENUMERATION OF MATRICES UNDER RESTRICTIONS AND SOME VARIATIONS	1
	1.1 Introduction	1
	1.2 Enumeration of the matrices	2
	1.3 Evaluation of two Multiple Integrals	21
CHAPTER II -	K-SAMPLE ANALOGUES OF THE KOLMOGOROV-SMIRNOV STATISTICS	28
	2.1 Introduction	28
	2.2 K-sample Rank Statistics of Kolmogorov-Smirnov Type	31
	2.3 K-sample Order Statistics of Kolmogorov-Smirnov Type	45
CHAPTER III -	A STUDY OF ROOTED PLANE TREES	53
	3.1 Introduction	53
	3.2 Definitions and Representation Theorems	54
	3.2.1 Pseudo-search Code Representation	56
	3.2.2 Matrix Representation	63
	3.3 Enumeration Methods	66
	3.3.1 Generating Function Techniques	67
	3.3.2 Matrix Enumeration Techniques	75
	3.4 Optimal Alphabetic q-ary Trees	77
	3.4.1 Basic Properties of the Weighted Trees	78
	3.4.2 Algorithms of Construction	83
	3.4.3 Entropy Bounds for the Costs of the Optimal q-ary Trees	92

CHAPTER IV - K-SAMPLE OPTIMAL NESTED BINOMIAL GROUP TESTING 97

4.1 Introduction 97

4.2 (K+1)-ary Tree Representation 102

4.3 On Defective Populations of Small Sizes 107

4.4 On Binomial Populations of Small Sizes 111

4.5 On Defective Populations with Units from a Unique Binomial Distribution 113

4.6 On Binomial Populations with Units from a Unique Binomial Distribution 124

4.7 Asymptotic Properties of the Cost Function 133

CHAPTER V - CONCLUSIONS AND SUGGESTIONS FOR FUTURE STUDY 146

5.1 Contributions of the Thesis 146

5.2 Suggestions for Future Study 148

REFERENCES 149

CHAPTER I

ON THE ENUMERATION OF MATRICES UNDER RESTRICTIONS AND SOME VARIATIONS

1.1 Introduction

In this chapter we shall use a method initiated by Narayana (1955) to enumerate matrices whose rows satisfy certain boundary conditions. A simple example of such a problem is to enumerate the number of vectors $(x_1, x_2, \ldots, x_m)$ that satisfy the conditions $0 \leq x_1 \leq x_2 \leq \ldots \leq x_m$ and $a_i \leq x_i \leq b_i$ for $i = 1, 2, \ldots, m$, where x's, a's and b's are non-negative integers such that $0 \leq a_1 \leq a_2 \leq \ldots \leq a_m$, $0 \leq b_1 \leq b_2 \leq \ldots \leq b_m$ and $a_i \leq b_i$, $i = 1, \ldots, m$. Among others who have solved these types of problems are Kreweras (1965), Steck (1969), and Mohanty (1971,1973).

Here we first generalize the results to enumerate distinct k×m matrices whose rows are vectors of non-negative integers satisfying certain more general boundary conditions than those stated above. Enumeration of these matrices involves the simplification of a k·m fold summation of the number 1 to a determinant of size (m+k)×(m+k). This enables us to determine the null distribution functions of a k-sample analogue of the two-sample Kolmogorov-Smirnov statistics under certain conditions in Chapter II and also enables us to enumerate certain classes of trees in Chapter III.

Secondly, following the suggestion of Mohanty (1971), we employ a similar method to evaluate multiple integrals of k·m continuous variables in the form of a k×m matrix with row vectors satisfying the same boundary conditions as mentioned in the discrete case. This enables us to find a (k+m)×(k+m) determinant as an expression of the joint distribution function of a k-sample analogue of the one-sample Kolmogorov-Smirnov statistics.

1.2 Enumeration of the Matrices

Let

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{im} \\ \vdots & & \vdots & & \vdots \\ x_{k1} & \cdots & x_{kj} & \cdots & x_{km} \end{bmatrix} \tag{1.2.1}$$

be a kxm matrix whose entries are non-negative integers. Let $A = (a_1, a_2, \ldots, a_m)$, $B = (b_1, b_2, \ldots, b_m)$, $C = (c_1, c_2, \ldots, c_k)$, $H = (h_1, h_2, \ldots, h_k)$ and $D = (d_1, \ldots, d_{k-1})$ be vectors of non-negative integers such that $0 \leq a_1 \leq a_2 \leq \ldots \leq a_m$, $0 \leq b_1 \leq b_2 \leq \ldots \leq b_m$, $0 \leq c_1 \leq c_2 \leq \ldots \leq c_k$, $0 \leq h_1 \leq h_2 \leq \ldots \leq h_k$ and $b_j \leq a_j$ for $j = 1, 2, \ldots, m$.

Let N(k, m, A, B, C, D) be the total number of distinct matrices of the form X satisfying the conditions:

$$\left.\begin{array}{ll} \text{(a)} & c_i \leq x_{i1} \leq x_{i2} \leq \ldots \leq x_{im}, \quad i = 1, 2, \ldots, k. \\ *\text{(b)} & x_{ij} \leq x_{i+1,j} + d_i, \quad i = 1, \ldots, k-1 \text{ and } j = 1, \ldots, m \\ \text{(c)} & b_j \leq x_{1j} \qquad\qquad j = 1, \ldots, m \\ \text{(d)} & x_{kj} \leq a_j \qquad\qquad j = 1, \ldots, m \end{array}\right\} \tag{1.2.2}$$

Similarly, for $b_1 \leq \sum_{t=1}^{k-1} d_t$, we let NB(k, m, A, B, H, D) be the total number of distinct matrices of the form X satisfying

$$\left.\begin{array}{l} \text{the conditions (b), (c), and (d) of (1.2.2) and} \\ \text{(a')} \ 0 \leq x_{i1} \leq x_{i2} \leq \ldots \leq x_{im} \leq h_i, \ i = 1, \ldots, k \end{array}\right\} \tag{1.2.3}$$

For the vectors A, B, C, H, and D considered here, we employ the convention that whenever all the components of the vectors are equal to a certain constant, we denote them by that constant. For example, if

$a_1 = a_2 = \ldots = a_m = a$ then we say that $A = a$. Also we use the conventions that $d_0 = d_k = 0$ and $\sum_{i=m}^{n} x_i = 0$ whenever $m > n$ for any x_i.

Note that when $C = D = 0$ and $h_i \geq a_m + \sum_{t=1}^{k-i} d_t$, $i = 1, \ldots, k$, we have $N(k, m, A, B, 0, 0) = NB(k, m, A, B, H, 0)$. The number $N(k, m, A, B, 0, 0)$ was originally found by Kreweras (1965). In particular, when $k = 1$, this is reduced to the problem stated in the introduction which was solved independently by Steck (1969) for finding the two-sample Kolmogorov-Smirnov statistics. Later, Mohanty (1971) provided a short proof for Steck's result. When $C = 0$ and $b \geq \sum_{t=1}^{k-1} d_t$, the number $N(k, m, A, B, 0, 0)$ was found by Mohanty (1973). Unfortunately, the conditions were wrongly stated in his original paper (see Mohanty (1977)).

Now the number $N(k, m, A, B, C, D)$ represents the class of $k \times m$ matrices X of non-negative integers, whose rows are bounded in the fashion of * and the i^{th} row is distributed within the region bounded above by the vector $(a_1^i, \ldots, a_j^i, \ldots, a_m^i)$ and bounded below by the vector $(u_1^i, \ldots, u_j^i, \ldots, u_m^i)$, where $a_j^i = a_j + \sum_{t=1}^{k-i} d_t$, $u_j^i = \max \{b_j - \sum_{t=1}^{i-1} d_t, c_i\}$ for $j = 1, \ldots, m$ and $i = 1, \ldots, k$.

On the other hand, the number $NB(k, m, A, B, H, 0)$ represents matrices X of non-negative integers whose rows are bounded in the fashion of * and the ith row is distributed within the region bounded above by the vector $(v_1^i, \ldots, v_j^i, \ldots, v_m^i)$ and bounded below by the vector $(b_1^i, \ldots, b_j^i, \ldots, b_m^i)$, where $v_j^i = \min \{a_j + \sum_{t=1}^{k} d_t, h_i\}$, $b_j^i = b_j - \sum_{t=1}^{i-1} d_t$ for $j = 1, \ldots, m$ and $i = 1, \ldots, k$. Notice that

the extra condition $b_1 \geq \sum_{t=1}^{k-1} d_t$ is necessary for the number NB(k, m, A, B, H, D) to be well defined because this guarantees that every entry of X is non-negative.

Now we employ the technique presented by Narayana (1955) to express the numbers N(k, m, A, B, C, D) and NB(k, m, A, B, H, D) as km fold summations. Hence, we have

$$N(k, m, A, B, C, D) = \begin{array}{cccccc} \sum_{x_{k1}=u_k}^{a_1} & \sum_{x_{k2}=u_{k2}}^{a_2} & \cdots & \sum_{x_{kj}=u_{kj}}^{a_j} & \cdots & \sum_{x_{km}=u_{km}}^{a_m} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sum_{x_{i1}=u_i}^{a_{i1}} & \sum_{x_{i2}=u_{i2}}^{a_{i2}} & \cdots & \sum_{x_{ij}=u_{ij}}^{a_{ij}} & \cdots & \sum_{x_{im}=u_{im}}^{a_{im}} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sum_{x_{11}=u_1}^{a_{11}} & \sum_{x_{12}=u_{12}}^{a_{12}} & \cdots & \sum_{x_{1j}=u_{1j}}^{a_{1j}} & \cdots & \sum_{x_{1m}=u_{1m}}^{a_{1m}} 1 \end{array} \quad (1.2.4)$$

where $a_{ij} = x_{i+1,j} + d_i$ for $i = 1, \ldots, k-1$ and $j = 1, \ldots, m$;

$u_{ij} = \max \{b_j - \sum_{t=1}^{i-1} d_t, x_{i,j-1}\}$ for $i = 1, \ldots, k$ and $j = 2, \ldots, m$

and $u_i = \max \{b_1 - \sum_{t=1}^{i-1} d_t, c_i\}$ for $i = 1, \ldots, k$. Similarly,

$$NB(k, m, A, B, H, D) = \begin{array}{cccccc} \sum_{x_{1m}=b_m}^{v_1} & \sum_{x_{1m}=b_{m-1}}^{v_{1,m-1}} & \cdots & \sum_{x_{1j}=b_j}^{v_{ij}} & \cdots & \sum_{x_{11}=b_1}^{v_{11}} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sum_{x_{im}=b_{im}}^{v_1} & \sum_{x_{i,m-1}=b_{i,m-1}}^{v_{i,m-1}} & \cdots & \sum_{x_{ij}=b_{ij}}^{v_{ij}} & \cdots & \sum_{x_{il}=b_{il}}^{v_{il}} \\ \vdots & \vdots & & \vdots & & \vdots \\ \sum_{x_{km}=b_{km}}^{v_k} & \sum_{x_{k,m-1}=b_{k,m-1}}^{v_{k,m-1}} & \cdots & \sum_{x_{kj}=b_{kj}}^{v_{kj}} & \cdots & \sum_{x_{kl}=b_{kl}}^{v_{kl}} 1 \end{array} \quad (1.2.5)$$

where $b_{ij} = x_{i-1,j} - d_i$ for $i = 2, \ldots, k$ and $j = 1, \ldots, m$;

$$v_{ij} = \min \{a_j + \sum_{t=i}^{k-1} d_t, x_{i,j+1}\} \text{ for } i = 1, \ldots, k \text{ and } j = 1, \ldots, m-1;$$

$$v_i = \min \{a_m + \sum_{t=i}^{k-1} d_t, h_i\} \text{ for } i = 1, \ldots, k.$$

Our method of simplification is to replace the number 1 by a determinant of size (m+k)x(m+k) instead of a determinant of size m x m used by Kreweras (1965) and Mohanty (1973). By doing so, more generalized results are obtained in the two theorems of this chapter.

The following basic definitions and equalities will be used in the proofs of the theorems and are listed here for convenience.

For any two integers a and b, we write

$$\left\{ \begin{matrix} a \\ b \end{matrix} \right\}_+ = \begin{cases} \left(\begin{matrix} a \\ b \end{matrix} \right) & \text{if } a \geq b \geq 0 \\ 0 & \text{otherwise} \end{cases},$$

where $\left(\begin{matrix} a \\ b \end{matrix} \right)$ is a binomial coefficient. We call $\left\{ \begin{matrix} a \\ b \end{matrix} \right\}_+$ a positive binomial coefficient. The identity

$$\left\{ \begin{matrix} a \\ b \end{matrix} \right\}_+ = \left\{ \begin{matrix} a+1 \\ b+1 \end{matrix} \right\}_+ - \left\{ \begin{matrix} a \\ b+1 \end{matrix} \right\}_+ \qquad (1.2.6)$$

is well known for binomial coefficients. It is simple to check that it holds for positive binomial coefficients as well. Using (1.2.6), it is straightforward to verify that the formulation

$$\sum_{x=\max\{b_1,b_2\}}^{a} \left\{ \begin{matrix} x-b+t \\ r+t \end{matrix} \right\}_+ = \left\{ \begin{matrix} a-b+t+1 \\ r+t+1 \end{matrix} \right\}_+ - \left\{ \begin{matrix} b_2-b+t \\ r+t+1 \end{matrix} \right\}_+ \qquad (1.2.7)$$

holds for any non-negative integers a, b, b_1, b_2, t and r with the condition that $b_1 \leq b$.

Lemma 1.2.1 Let X be a matrix of size m x m, partitioned into the form $X = \begin{bmatrix} & \vdots & B \\ A & \vdots & \cdots \\ & \vdots & C \end{bmatrix}$, where B is a matrix of size r x s such that $r + s = m + 1$ and $1 \leq r, s \leq m$. Suppose every entry of the matrix B is zero, then det{X} equals to zero.

Proof: This can be proved by induction on the row number r of the matrix B. It is obviously true when r = 1 in which case we have s = m and the fact that all the entries in the first row of X are zeros implies that det{X} = 0. Suppose it is true for all $r \leq r_o$, where $1 \leq r_o < m$. We need to show that the Lemma is also true for $r = r_o + 1$. By expanding det{X} at its first row, we can obtain a sum of m determinants all of which equal to zero since the induction hypothesis is true when $r = r_o$. Therefore the lemma is also true for $r = r_o + 1$ and the proof is completed by induction.

Theorem 1.2.1 N(k, m, A, B, C, D) = det{E:F}, the determinant of an augmented matrix of an (m + k) x m matrix $E = \{e_{ij}\}$ and an (m + k) x k matrix $F = \{f_{ij}\}$ such that

$$e_{ij} = \left| a_{m-j+1} - b_{m-i+1} + \sum_{t=1}^{k-1} d_t + k \right|_{j-i+k} + \qquad \begin{matrix} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, m \end{matrix}$$

and

$$f_{ij} = \left| u_j - b_{m-i+1} + \sum_{t=1}^{j-1} d_t + j - 1 \right|_{m+j-i} + \qquad \begin{matrix} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, k, \end{matrix}$$

where we set $b_i = b_1$ for $i \leq 0$ and $u_i = \max\{b_1 - \sum_{t=1}^{i-1} d_t, c_i\}$ for $i = 1, \ldots, k$.

Proof: Let $\Delta^t = \det\{E^t : F\}$ be the determinant of an augmented matrix of an $(m+k) \times m$ matrix $E^t = \{e^t_{ij}\}$ and an $(m+k) \times k$ matrix $F = \{f_{ij}\}$ such that

$$e^t_{ij} = \left(x_{t,m-j+1} - b_{m-i+1} + \sum_{\ell=0}^{t-1} d_\ell + t \right)_{j-i+t}^{+} \qquad \begin{array}{l} i = 1, \ldots, m+k, \; j = 1, \ldots, m \\ \text{and } t = 0, 1, \ldots, k-1, \end{array}$$

where $d_0 = 0$, $\sum_{\ell=0}^{-1} d_\ell = 0$; and F is given in the statement of the theorem. Let

$$\sum\nolimits^i = \begin{cases} \sum_{x_{i1}=u_i}^{a_{i1}} \sum_{x_{i2}=u_{i2}}^{a_{i2}} \cdots \sum_{x_{ij}=u_{ij}}^{a_{ij}} \cdots \sum_{x_{im}=u_{im}}^{a_{im}} & \text{for } i = 1, \ldots, k-1 \\ \sum_{x_{k1}=u_k}^{a_1} \sum_{x_{k2}=u_{k2}}^{a_2} \cdots \sum_{x_{kj}=u_{kj}}^{a_j} \cdots \sum_{x_{km}=u_m}^{a_m} & \text{for } i = k \end{cases}$$

be used to simplify the expression (1.2.4). Furthermore, since

$$\Delta^o = 1,$$

we can replace the number 1 of expression (1.2.4) by the determinant Δ^o, so that

$$N(k, m, A, B, C, D) = \sum\nolimits^k \sum\nolimits^{k-1} \cdots \sum\nolimits^i \cdots \sum\nolimits^1 \Delta^o .$$

We claim that

$$\sum\nolimits^i \Delta^{i-1} = \begin{cases} \Delta^i & i = 1, \ldots, k-1 \\ \det\{E : F\} & i = k \end{cases} \tag{1.2.8}$$

Hence, the result of the theorem follows.

We illustrate the proof for the special case k = 3 and m = 3. The proof in general is similar but much more tedious to write. Now (1.2.4) becomes

$$N(k,m,A,B,C,D) = \sum_{x_{31}=u_3}^{a_1} \sum_{x_{32}=u_{32}}^{a_2} \sum_{x_{33}=u_{33}}^{a_3} \sum_{x_{21}=u_{21}}^{a_{21}} \sum_{x_{22}=u_{22}}^{a_{22}} \sum_{x_{23}=u_{23}}^{a_{23}} \sum_{x_{11}=u_1}^{a_{11}} \sum_{x_{12}=u_{12}}^{a_{12}} \sum_{x_{13}=u_{13}}^{a_{13}} 1$$

$$= \sum^3 \sum^2 \sum^1 \Delta^0$$

In this case, we have

$$\Delta^0 = \begin{vmatrix} \binom{x_{13}-b_3}{0}_+ & \binom{x_{12}-b_3}{1}_+ & \binom{x_{11}-b_3}{2}_+ & \binom{u_1-b_3}{3}_+ & \binom{u_2-b_3+d_1+1}{4}_+ & \binom{u_3-b_3+d_1+d_2+2}{5}_+ \\ 0 & \binom{x_{12}-b_2}{0}_+ & \binom{x_{11}-b_2}{1}_+ & \binom{u_1-b_2}{2}_+ & \binom{u_2-b_2+d_1+1}{3}_+ & \binom{u_3-b_2+d_1+d_2+2}{4}_+ \\ 0 & 0 & \binom{x_{11}-b_1}{0}_+ & \binom{u_1-b_1}{1}_+ & \binom{u_2-b_2+d_1+1}{2}_+ & \binom{u_3-b_1+d_1+d_2+2}{3}_+ \\ 0 & 0 & 0 & \binom{u_1-b_1}{0}_+ & \binom{u_2-b_2+d_1+1}{1}_+ & \binom{u_3-b_1+d_1+d_2+2}{2}_+ \\ 0 & 0 & 0 & 0 & \binom{u_2-b_2+d_1+1}{0}_+ & \binom{u_3-b_1+d_1+d_2+2}{1}_+ \\ 0 & 0 & 0 & 0 & 0 & \binom{u_3-b_1+d_1+d_2+2}{0}_+ \end{vmatrix} \qquad (1.2.9)$$

which is obviously equal to 1. Now

$$\sum_{x_{13}=u_{13}}^{x_{23}+d_1} \Delta^{\circ}$$

$$= \begin{vmatrix}
\binom{x_{23}+d_1-b_3+1}{1}_+ - \binom{x_{12}-b_3}{1}_+ & \binom{x_{12}-b_3}{1}_+ & \binom{x_{11}-b_3}{2}_+ & \binom{u_1-b_3}{3}_+ & \cdots & \binom{u_3-b_3+d_1+d_2+2}{5}_+ \\
\binom{x_{23}+d_1-b_2+1}{0}_+ - \binom{x_{12}-b_2}{0}_+ & \binom{x_{12}-b_2}{0}_+ & \binom{x_{11}-b_2}{1}_+ & \binom{u_1-b_2}{2}_+ & \cdots & \binom{u_3-b_2+d_1+d_2+2}{4}_+ \\
0 & 0 & \binom{x_{11}-b_1}{0}_+ & \binom{u_1-b_1}{1}_+ & \cdots & \binom{u_3-b_1+d_1+d_2+2}{3}_+ \\
0 & 0 & 0 & \binom{u_1-b_1}{0}_+ & \cdots & \binom{u_3-b_1+d_1+d_2+2}{2}_+ \\
0 & 0 & 0 & 0 & \cdots & \binom{u_3-b_1+d_1+d_2+2}{1}_+ \\
0 & 0 & 0 & 0 & \cdots & \binom{u_3-b_1+d_1+d_2+2}{0}_+
\end{vmatrix}$$

$$= \begin{vmatrix}
\binom{x_{23}+d_1-b_3+1}{1}_+ & \binom{x_{12}-b_3}{1}_+ & \binom{x_{11}-b_3}{2}_+ & \binom{u_1-b_3}{3}_+ & \binom{u_2-b_3+d_1+1}{4}_+ & \binom{u_3-b_3+d_1+d_2+2}{5}_+ \\
\binom{x_{23}+d_1-b_2+1}{0}_+ & \binom{x_{12}-b_2}{0}_+ & \binom{x_{11}-b_2}{1}_+ & \binom{u_1-b_2}{2}_+ & \binom{u_2-b_2+d_1+1}{3}_+ & \binom{u_3-b_2+d_1+d_2+2}{4}_+ \\
0 & 0 & \binom{x_{11}-b_1}{0}_+ & \binom{u_1-b_1}{1}_+ & \binom{u_2-b_1+d_1+1}{2}_+ & \binom{u_3-b_1+d_1+d_1+2}{3}_+ \\
0 & 0 & 0 & \binom{u_1-b_1}{0}_+ & \binom{u_2-b_1+d_1+1}{1}_+ & \binom{u_3-b_1+d_1+d_2+2}{2}_+ \\
0 & 0 & 0 & 0 & \binom{u_2-b_1+d_1+1}{0}_+ & \binom{u_3-b_1+d_1+d_2+2}{1}_+ \\
0 & 0 & 0 & 0 & 0 & \binom{u_3-b_1+d_1+d_2+2}{0}_+
\end{vmatrix}$$

(1.2.10)

where we have added over the first column using (1.2.7) and simplified the determinant using the fact that $\begin{pmatrix} x_{23}+d_1-b_2+1 \\ 0 \end{pmatrix}_+ - \begin{pmatrix} x_{12}-b_2 \\ 0 \end{pmatrix}_+ = 1 - 1 = 0.$ Note that the last three columns of all the determinants involved in this proof remain the same as the corresponding columns of the determinant Δ^o. From now on, every entry of the fourth and fifth columns of the determinants will be denoted by "..." for the purpose of saving space.

If we add over the second column, we obtain a similar result in the second column and so forth, so that the end result is:

$$\sum\nolimits^1 \Delta^o$$

$$= \begin{vmatrix}
\begin{pmatrix} x_{23}+d_1-b_3+1 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} x_{22}+d_1-b_3+1 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} x_{21}+d_1-b_3+1 \\ 3 \end{pmatrix}_+ & -\begin{pmatrix} u_1-b_3 \\ 3 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_3 \\ 3 \end{pmatrix}_+ & \cdots & \cdots \\
\begin{pmatrix} x_{23}+d_1-b_2+1 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} x_{22}+d_1-b_2+1 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} x_{21}+d_1-b_2+1 \\ 2 \end{pmatrix}_+ & -\begin{pmatrix} u_1-b_2 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_2 \\ 2 \end{pmatrix}_+ & \cdots & \cdots \\
0 & \begin{pmatrix} x_{22}+d_1-b_1+1 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} x_{21}+d_1-b_1+1 \\ 1 \end{pmatrix}_+ & -\begin{pmatrix} u_1-b_1 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_1 \\ 1 \end{pmatrix}_+ & \cdots & \cdots \\
0 & 0 & \begin{pmatrix} x_{21}+d_1-b_1+1 \\ 0 \end{pmatrix}_+ & -\begin{pmatrix} u_1-b_1 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_1 \\ 0 \end{pmatrix}_+ & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & \cdots & \cdots \\
0 & 0 & 0 & 0 & 0 & \cdots & \cdots
\end{vmatrix}$$

$$= \begin{vmatrix} \binom{x_{23}+d_1-b_3+1}{1}_+ & \binom{x_{22}+d_1-b_3+1}{2}_+ & \binom{x_{21}+d_1-b_3+1}{3}_+ & \binom{u_1-b_3}{3}_+ & \cdots & \cdots \\ \binom{x_{23}+d_1-b_2+1}{0}_+ & \binom{x_{22}+d_1-b_2+1}{1}_+ & \binom{x_{21}+d_1-b_2+1}{2}_+ & \binom{u_1-b_2}{2}_+ & \cdots & \cdots \\ 0 & \binom{x_{22}+d_1-b_1+1}{0}_+ & \binom{x_{21}+d_1-b_1+1}{1}_+ & \binom{u_1-b_1}{1}_+ & \cdots & \cdots \\ 0 & 0 & \binom{x_{21}+d_1-b_1+1}{0}_+ & \binom{u_1-b_1}{0}_+ & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \end{vmatrix}$$

$$= \Delta^1.$$

Thus we have shown that $\sum^1 \Delta^o = \Delta^1$. Note that the fourth column of the determinant Δ^o plays the role of simplifying the resultant determinant after the first three summations $\sum^1$ so that the next three summations $\sum^2$ can be carried out smoothly in a similar pattern. However, notice that Lemma 1.2.1 is essential in simplifying the expression $\sum^2 \Delta^1$.

Since

$$\sum_{x_{23}=u_{23}}^{x_{33}+d_2} \Delta^1$$

$$= \begin{vmatrix} \binom{x_{33}+d_2+d_1-b_3+2}{2}_+ & \binom{x_{22}+d_1-b_3+1}{2}_+ & \binom{x_{21}+d_1-b_3+1}{3}_+ & \binom{u_1-b_3}{3}_+ & \cdots & \cdots \\ \binom{x_{33}+d_2+d_1-b_2+2}{1}_+ & \binom{x_{22}+d_1-b_2+1}{1}_+ & \binom{x_{21}+d_1-b_2+1}{2}_+ & \binom{u_1-b_2}{2}_+ & \cdots & \cdots \\ \binom{x_{33}+d_2+d_1-b_1+2}{0}_+ & \binom{x_{22}+d_1-b_1+1}{0}_+ & \binom{x_{21}+d_1-b_1+1}{1}_+ & \binom{u_1-b_1}{1}_+ & \cdots & \cdots \\ 0 & 0 & \binom{x_{21}+d_1-b_1+1}{0}_+ & \binom{u_1-b_1}{0}_+ & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \end{vmatrix}$$

$$- \begin{vmatrix} \binom{u_{23}+d_1-b_3+1}{2}_+ & \binom{x_{22}+d_1-b_3+1}{2}_+ & \binom{x_{21}+d_1-b_3+}{3}_+ & \binom{u_1-b_3}{3}_+ & \cdots & \cdots \\ \binom{u_{23}+d_1-b_2+1}{1}_+ & \binom{x_{22}+d_1-b_2+1}{1}_+ & \binom{x_{21}+d_1-b_2+1}{2}_+ & \binom{u_1-b_2}{2}_+ & \cdots & \cdots \\ \binom{u_{23}+d_1-b_1+1}{0}_+ & \binom{x_{22}+d_1-b_1+1}{0}_+ & \binom{x_{21}+d_1-b_1+1}{1}_+ & \binom{u_1-b_1}{1}_+ & \cdots & \cdots \\ 0 & 0 & \binom{x_{21}+d_1-b_1+1}{0}_+ & \binom{u_1-b_1}{0}_+ & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \end{vmatrix}$$

where we have added over the first column of Δ^1 using equality (1.2.6) and decomposed the resultant determinant at the first column so that the above two determinants are obtained. Now we claim that the second determinant in the above expression equals zero. But, equality (1.2.7) is inadequate to enable us to replace every $u_{23}=\max\{b_3-d_1, x_{22}\}$ involved in the first column of the second determinant by x_{22} as we have done in the summations of $\sum^1$. For those entries below the main diagonal, the conditions that r being non-negative in equality (1.2.7) is violated. Now we observe that when $b_3-d_1 > x_{22}$, in the second determinant, all the entries with row number less than or equal to "one" (in this case) and column number greater than or equal to "one" vanishes. Thus, Lemma 1.2.1 implies that the determinant equals zero. Hence, we conclude that the second determinant of the above expression vanishes no matter whether u_{23} equals to b_3-d_1 or x_{22}.

If we add over the second column, we obtain a similar result in the second column and so forth, so that the end result is:

$\sum^2 \Delta^1$

$$= \begin{vmatrix} \begin{pmatrix} x_{32}+d_2+d_1-b_3+2 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} x_{32}+d_2+d_1-b_3+2 \\ 3 \end{pmatrix}_+ & \begin{pmatrix} x_{31}+d_2+d_1-b_3+2 \\ 4 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_3 \\ 3 \end{pmatrix}_+ & \cdots & \cdots \\ \begin{pmatrix} x_{32}+d_2+d_1-b_2+2 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} x_{32}+d_2+d_1-b_2+2 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} x_{31}+d_2+d_1-b_2+2 \\ 3 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_2 \\ 2 \end{pmatrix}_+ & \cdots & \cdots \\ \begin{pmatrix} x_{32}+d_2+d_1-b_1+2 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} x_{32}+d_2+d_1-b_1+2 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} x_{31}+d_2+d_1-b_1+2 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_1 \\ 1 \end{pmatrix}_+ & \cdots & \cdots \\ 0 & \begin{pmatrix} x_{32}+d_2+d_1-b_1+2 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} x_{31}+d_2+d_1-b_1+2 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_1 \\ 0 \end{pmatrix}_+ & \cdots & \cdots \\ 0 & 0 & \begin{pmatrix} x_{31}+d_2+d_1-b_1+2 \\ 0 \end{pmatrix}_+ & 0 & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \end{vmatrix}$$

$$- \begin{vmatrix} \begin{pmatrix} x_{32}+d_2+d_1-b_3+2 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} x_{32}+d_2+d_1-b_3+2 \\ 3 \end{pmatrix}_+ & \begin{pmatrix} u_2-b_3+d_1+1 \\ 4 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_3 \\ 3 \end{pmatrix}_+ & \cdots & \cdots \\ \begin{pmatrix} x_{32}+d_2+d_1-b_2+2 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} x_{32}+d_2+d_1-b_2+2 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} u_2-b_2+d_1+1 \\ 3 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_2 \\ 2 \end{pmatrix}_+ & \cdots & \cdots \\ \begin{pmatrix} x_{32}+d_2+d_1-b_1+2 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} x_{32}+d_2+d_1-b_1+2 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} u_2-b_1+d_1+1 \\ 2 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_1 \\ 1 \end{pmatrix}_+ & \cdots & \cdots \\ 0 & \begin{pmatrix} x_{32}+d_3+d_1-b_1+2 \\ 0 \end{pmatrix}_+ & \begin{pmatrix} u_2-b_1+d_1+1 \\ 1 \end{pmatrix}_+ & \begin{pmatrix} u_1-b_1 \\ 0 \end{pmatrix}_+ & \cdots & \cdots \\ 0 & 0 & \begin{pmatrix} u_2-b_1+d_1+1 \\ 0 \end{pmatrix} & 0 & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & \cdots \end{vmatrix}$$

Note that the second determinant in the above expression vanishes since the third column is identical with the fifth column. Hence, we have proved that $\sum^2 \Delta^1 = \Delta^2$.

The procedure of showing that $\sum^3 \Delta^2 = \det\{E:F\}$ is entirely similar. This gives the result of the theorem for k = 3 and m = 3. A similar but more lengthy argument yields the theorem.

The number NB(k, m, A, B, H, D) can also be computed by using the above theorem according to the following lemma.

Lemma 1.2.2 If $b_1 \geq \sum_{t=1}^{k=1} d_t$, then the expression (1.2.5) can be transformed so that NB(k, m, A, B, H, D) = N(k, m, A',B', C', D'), where $A' = (a'_1, \ldots, a'_m)$, $B' = (b'_1, \ldots, b'_m)$, $C' = (c'_1, \ldots, c'_k)$ and $D' = (d'_1, \ldots, d'_{k-1})$ are vectors such that $a'_i = h_k - b_{m-i+1}$, $b'_i = h_k - a_{m-i+1}$ for i = 1, ... , m, $c'_i = h_k - h_{k-i+1}$, for i = 1, ... , k and $d'_i = d_{k-i}$ for i = 1, ... , k-1.

Proof: In determining the number NB(k, m, A, B, H, D), we find that the overall upper bound for every entry of the matrix X satisfying the condition of (1.2.3) is h_k. Thus, if we transform the origin (0, 0) to the point (m, h_k) and rotate 180° clockwise with the new origin as a center, then the matrix $X = \{x_{ij}\}$ is transformed to the matrix $X' = \{x'_{ij}\}$ where $x'_{ij} = h_k - x_{k-i+1,m-j+1}$ for i = 1, ... , k and j = 1, ..., m. It can easily be checked that the matrix X' satisfies all the conditions of (1.2.2). This completes the proof of the lemma.

Accordingly, we have

$$u'_i = \max \{b'_1 - \sum_{t=1}^{i-1} d'_t, c'_i\}$$

$$= \max \{h_k - a_{m-1+1} - \sum_{t=1}^{i-1} d_{k-t}, h_k - h_{k-i+1}\}$$

$$= \max \{h_k - a_m - \sum_{j=k-i+1}^{k-1} d_j,\ h_k - h_{k-i+1}\}$$

$$= h_k + \max \{-a_m - \sum_{j=k-i+1}^{k-1} d_j,\ -h_{k-i+1}\}$$

$$= h_k - \min \{a_m + \sum_{j=k-i+1}^{k-1} d_j,\ h_{k-i+1}\}$$

$$= h_k - v_{k-i+1}$$

for $i = 1, \ldots, k$. Then the following theorem can be obtained as a result of Theorem 1.2.1 and Lemma 1.2.2.

Theorem 1.2.2 $NB(k, m, A, B, H, D) = \det\{E' : F'\}$, the determinant of an augmented matrix of a $(m+k) \times m$ matrix $E' = \{e'_{ij}\}$ and a $(m+k) \times k$ matrix $F' = \{f'_{ij}\}$ such that

$$e'_{ij} = \left| a_i - b_j + \sum_{t=1}^{k-1} d_t + k \right|_{j-i+k}^{+} \qquad \begin{matrix} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, m \end{matrix}$$

and

$$f'_{ij} = \left| a_i - v_{k-i+1} + \sum_{t=1}^{k-1} d_t + j - 1 \right|_{m+j-i}^{+} \qquad \begin{matrix} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, k \end{matrix}$$

where we set $a_i = a_m$ for $i \geq m$ and $v_i = \min \{a_m + \sum_{t=i}^{k-1} d_t,\ h_i\}$ for $i = 1, \ldots, k$.

We have already mentioned that in some special cases, the numbers $N(k, m, A, B, C, D)$ and $NB(k, m, A, B, H, D)$ have been determined by various authors in terms of relatively simple expressions. Thus, our determinant can be reduced to a rather simple form provided that additional

conditions are satisfied.

Corollary 1.2.1 The following equalities hold for determinants with non-negative integral entries satisfying the specified conditions.

(i) When $C = 0$ and $b_1 \geq \sum_{t=1}^{k-1} d_t$, the determinant $\det\{E \vdots F\}$ of Theorem 1.2.1 equals to

$$\begin{vmatrix} \begin{pmatrix} a_m - b_m + \sum_{t=1}^{k-1} d_t + k \\ k \end{pmatrix}_+ & \cdots & \begin{pmatrix} a_{m-j+1} - b_m + \sum_{t=1}^{k-1} d_t + k \\ k+j-1 \end{pmatrix}_+ & \cdots & \begin{pmatrix} a_1 - b_m + \sum_{t=1}^{k-1} d_t + k \\ k+m-1 \end{pmatrix}_+ \\ \cdots & & \cdots & & \cdots \\ \begin{pmatrix} a_m - b_{m-i+1} + \sum_{t=1}^{k-1} d_t + k \\ k+1-i \end{pmatrix}_+ & \cdots & \begin{pmatrix} a_{m-j+1} - b_{m-i+1} + \sum_{t=1}^{k-1} d_t + k \\ k+j-i \end{pmatrix}_+ & \cdots & \begin{pmatrix} a_1 - b_{m-i+1} + \sum_{t=1}^{k-1} d_t + k \\ k+m-i \end{pmatrix}_+ \\ \cdots & & \cdots & & \cdots \\ \begin{pmatrix} a_m - b_1 + \sum_{t=1}^{k-1} d_t + k \\ k+1-m \end{pmatrix}_+ & \cdots & \begin{pmatrix} a_{m-j+1} - b_{m-i+1} + \sum_{t=1}^{k-1} d_t + k \\ k+j-i \end{pmatrix}_+ & \cdots & \begin{pmatrix} a_1 - b_1 + \sum_{t=1}^{k-1} d_t + k \\ k \end{pmatrix}_+ \end{vmatrix}$$

(ii) When $A = a$, $B = C = D = 0$, $k = 1$, the determinant in (i) becomes

$$\begin{vmatrix} \binom{a+1}{1}_+ & \binom{a+1}{2}_+ & \cdots & \binom{a+1}{j}_+ & \cdots & \binom{a+1}{m}_+ \\ \binom{a+1}{0}_+ & \binom{a+1}{1}_+ & \cdots & \binom{a+1}{j-1}_+ & \cdots & \binom{a+1}{m+1}_+ \\ 0 & \binom{a+1}{0}_+ & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \binom{a+1}{j-i+1}_+ & \cdots & \binom{a+1}{m-i+1}_+ \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \binom{a+1}{1}_+ \end{vmatrix} = \binom{a+m}{m}_+$$

Proof: (i) The determinant given here equals the number $N(k, m, A, B, C, D)$ when $C = 0$ and $b_1 \geq \sum_{t=1}^{k-1} d_t$ according to Mohanty (1973). Thus, the equality holds.

This can also be shown directly by expanding the determinant $\det\{E:F\}$ of Theorem 1.2.1 at its $(m+1)^{th}$ column and simplifying it using Lemma 1.2.1.

(ii) This is a result of Pólya (1947) (cf. Kaucký (1975)). it also follows readily from the enumeration argument, see Steck (1969), Mohanty (1977), and Narayana (to appear).

Remark 1.2.1 A meaningful refinement of our results in this section can be considered as follows.

Let $N(k,m,A,B,C,D,H)$ be the total number of distinct matrices of the form X given by (1.2.1) which satisfy

$$\left.\begin{array}{l} \text{all the conditions of (1.2.2) and} \\ (\bar{a})\quad c_i \leq x_{i1} \leq x_{i2} \leq \ldots \leq x_{im} \leq h_i,\ i = 1, \ldots, k \\ (\bar{b})\quad b_j \leq x_{ij} \leq a_j \quad i = 1, \ldots, k \text{ and } j = 1, \ldots, m \end{array}\right\} \qquad (1.2.11)$$

Similarly, if $b_1 \geq \sum_{t=1}^{k-1} d_t$, we let $NB(k,m,A,B,H,D,C)$ be the total number of distinct matrices of the form X given by (1.2.1) which satisfy

$$\left.\begin{array}{l} \text{all the conditions of (1.2.3) and} \\ (\bar{a}) \text{ and } (\bar{b}) \text{ of (1.2.11)} \end{array}\right\} \qquad (1.2.12)$$

It follows from the definition that $N(k,m,A,B,C,D,H)$ can be expressed as a km fold summation of the number 1 which is of the same form as the right hand side of (1.2.4) except that here we have $a_{ij}=\min\{x_{i+1,j}+d_i, h_i, a_j\}$,

$u_{ij} = \max\{b_1, x_{i,j-1}, c_i\}$ and $u_i = \max\{b_1, c_i\}$.

Similarly, NB(k,m,A,B,H,D,C) can be expressed as a km fold summation of the number 1 which is of the same form as the right hand side of (1.2.5) except that here we have $b_{ij} = \max\{x_{i-1,j}-d_i\ c_i, b_j\}$, $v_{ij}=\min\{a_m, x_{i,j+1}, h_i\}$ and $v_i = \min\{a_m, h_i\}$.

Unhappily, we are unable to simplify the above expressions by using the method employed in the proof of Theorem 1.2.1 and Theorem 1.2.2. The basic reason is that when the value of a_{ij} is changed from $x_{i+1,j} + d_i$ to $\min\{x_{i+1,j} + d_i, h_i, a_j\}$,etc.,the relation $\sum^1 \Delta^o = \Delta^1$ of (1.2.8) no longer holds. To illustrate this, let us consider again the special case that k = 3 and m = 3 which was chosen to illustrate the proof of Theorem 1.2.1. Here Δ^o is the determinant given by expression (1.2.9). But

$$\sum_{x_{13}=u_{13}}^{x_{21}+d_1} \Delta^o$$

$$= \begin{vmatrix} \binom{x_{23}+d_1-b_3+1}{1}_+ - \binom{u_{13}-b_3}{1}_+ & \binom{x_{12}-b_3}{1}_+ & \binom{x_{11}-b_3}{2}_+ & \binom{u_1-b_3}{3}_+ & \cdots & \binom{u_3-b_3+d_1+d_2+2}{5}_+ \\ \binom{x_{23}+d_1-b_2+1}{0}_+ - \binom{u_{13}-b_2}{0}_+ & \binom{x_{12}-b_2}{0}_+ & \binom{x_{11}-b_2}{1}_+ & \binom{u_1-b_2}{2}_+ & \cdots & \binom{u_3-b_2+d_1+d_2+2}{4}_+ \\ 0 & 0 & \binom{x_{11}-b_1}{0}_+ & \binom{u_1-b_1}{1}_+ & \cdots & \binom{u_3-b_1+d_1+d_2+2}{3}_+ \\ 0 & 0 & 0 & \binom{u_1-b_1}{0}_+ & \cdots & \binom{u_3-b_1+d_1+d_2+2}{2}_+ \\ 0 & 0 & 0 & 0 & \cdots & \binom{u_3-b_1+d_1+d_2+2}{1}_+ \\ 0 & 0 & 0 & 0 & \cdots & \binom{u_3-b_1+d_1+d_2+2}{0}_+ \end{vmatrix}.$$

the above determinant cannot be reduced to a form like the one given by expression (1.2.10), because neither equation (1.2.7) nor Lemma 1.2.1 can be applied in this case.

Hence we conclude that further investigation is required in order to find simple expressions for the numbers N(k, m, A, B, C, D, H) and NB(k, m, A, B, H, D, C) defined in this remark, except for the special cases that

$$N(k, m, A, B, C, D, h_k) = N(k, m, A, B, C, D)$$

amd

$$NB(k, m, A, B, H, D, 0) = NB(k, m, A, B, H, D),$$

the simple expressions in terms of $(m+k)\times(m+k)$ determinants are given by Theorem 1.2.1 and Theorem 1.2.2, respectively.

1.3 Evaluation of Two Multiple Integrals

The enumeration method developed in the previous section can readily be extended to the continuous case. Here we would like to determine a measure of km continuous variables in the form of a k x m matrix X given by (1.2.1).

Let I(k, m, A, B, C, D) be a measure represented by a multiple integral of the number 1 over the region of X specified by the conditions of (1.2.2). Therefore,

$$I(k,m,A,B,C,D) = \int_{x_{k1}=u_k}^{a_1} \int_{x_{k2}=u_{k2}}^{a_2} \cdots \int_{x_{kj}=u_{kj}}^{a_j} \cdots \int_{x_{km}=u_{km}}^{a_m}$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$\int_{x_{i1}=u_i}^{a_{i1}} \int_{x_{i2}=u_{i2}}^{a_{i2}} \cdots \int_{x_{ij}=u_{ij}}^{a_{ij}} \cdots \int_{x_{im}=u_{im}}^{a_{im}}$$
$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$
$$\int_{x_{11}=u_{11}}^{a_{11}} \int_{x_{12}=u_{12}}^{a_{12}} \cdots \int_{x_{1j}=u_{1j}}^{a_{1j}} \cdots \int_{x_{1m}=u_{1m}}^{a_{1m}} 1$$
$$dx_{1m}..dx_{1j}..dx_{11}....dx_{im}..dx_{ij}..dx_{i1}....dx_{km}..dx_{kj}..dx_{k1} \qquad (1.3.1)$$

where $a_{ij} = x_{i+1,j} + d_i$ for i = 1, ..., k-1 and j = 1, ..., m,

$u_{ij} = \max\{b_j - \sum_{t=1}^{i-1} d_t, x_{i,j-1}\}$, for i = 1, ..., k and j = 1, ..., m,

$u_i = \max\{b_1 - \sum_{t=1}^{i-1} d_t, c_i\}$ for i = 1, ..., k.

Similarly, if $b_1 \geq \sum_{t=1}^{k-1} d_t$, we let IB(k, m, A, B, H, D) be a measure represented by a multiple integral of the number 1 over the region of X specified by the conditions of (1.2.3). Therefore,

$$IB(k,m,A,B,H,D) = \int_{x_{1m}=b_m}^{v_m} \int_{x_{1,m-1}=b_{m-1}}^{v_{1,m-1}} \cdots \int_{x_{1j}=b_j}^{v_{1j}} \cdots \int_{x_{11}=b_1}^{v_{11}}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\int_{x_{im}=b_{im}}^{v_i} \int_{x_{i,m-1}}^{v_{i,m-1}} \cdots \int_{x_{ij}=b_{ij}}^{v_{ij}} \cdots \int_{x_{il}=b_{il}}^{v_{il}}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$\int_{x_{km}=b_{km}}^{v_1} \int_{x_{k,m-1}=b_{k,m-1}}^{v_{k,m-1}} \cdots \int_{x_{kj}=b_{kj}}^{v_{kj}} \cdots \int_{x_{kl}=b_{kl}}^{v_{kl}} 1$$

$$dx_{kl}..dx_{kj}..dx_{km}....dx_{il}..dx_{il}..dx_{im}....dx_{ll}..dx_{lj}..dx_{lm} \qquad (1.3.2)$$

where $b_{ij} = x_{i-1,j} - d_i$ for i = 2, ..., k and j = 1, ..., m,

$v_{ij} = \min\{a_j, x_{i,j+1}\}$ for i = 1, ..., k and j = 1, ..., m-1,

$v_i = \min\{a_m + \sum_{t=1}^{k-1} d_t, h_i\}$ for i = 1, ..., m.

It is clear from the definitions that when C = D = 0 and $k_i \geq a_m + \sum_{t=1}^{k-1} d_t$, i = 1, ..., k, we have

$$I(k, m, A, B, 0, 0) = IB(k, m, A, B, H, 0).$$

Evaluation of the integrals I(k, m, A, B, C, D) and IB(k, m, A, B, H, D) is analogous to the summation problem of determining the numbers N(k, m, A, B, C, D) and NB(k, m, A, B, H, D). For if we replace the symbol $\int$ by the symbol $\sum$ in (1.3.1) and (1.3.2), then they become the same as (1.2.4) and (1.2.5) respectively.

The following definition and equalities have been suggested by Mohanty (1971) as analogues of (1.2.6) and (1.2.7).

Let x and a be any two real numbers, we write

$$(x)_+ = \max (0, x)$$

Note that for real numbers a and b such that $a \geq b \geq 0$, we have

$$\int_b^a (x)_+^r \, dx = \frac{(a)_+^{r+1}}{r+1} - \frac{(b)_+^{r+1}}{r+1} \tag{1.3.3}$$

It is also simple to check that

$$\int_{\max\{b_1,b_2\}}^a (x - b)_+^r \, dx = \frac{(a - b)_+^{r+1}}{r+1} - \frac{(b_2 - b)_+^{r+1}}{r+1} \tag{1.3.4}$$

for non-negative real numbers a, b, b_1 and b_2, such that $a \geq b \geq 0$ and $b_1 \leq b$.

Now if we view the function $\frac{(x - b)_+^{(r)}}{(r)!}$ as $\binom{x - b}{r}_+$ etc. and replace $\int$ by $\sum$, then equality (1.3.4) becomes the same as equality (1.2.7) except that the term a - b on the right hand side of the expression in equality (1.3.4) corresponds to the term a - b + 1 of equality (1.2.7).

Therefore, if we adjust properly the symbols used in Section 1.2 and take into account the difference mentioned above, we can obtain the corresponding results in the continuous case also. Since the arguments of proof are similar, we only state the results below without repeating the same pattern of proof again.

Theorem 1.3.1 $I(k, m, A, B, C, D) = \det\{\bar{E}:\bar{F}\}$, the determinant of an augmented matrix of an $(m + k)$ x m matrix $\bar{E} = \{\bar{e}_{ij}\}$ and an $(m + k)$ x k matrix $\bar{F} = \{\bar{f}_{ij}\}$ such that

$$\bar{e}_{ij} = \frac{\left(a_{m-j+1} - b_{m-i+1} + \sum_{t=1}^{k-1} d_t\right)_+^{(j-i+k)}}{(j - i + k)\ !} \qquad \begin{matrix} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, m \end{matrix}$$

and

$$\bar{f}_{ij} = \frac{\left(u_j - b_{m-i+1} + \sum_{t=1}^{k-1} d_t\right)_+^{(m+j-i)}}{(m + j - i)\ !} \qquad \begin{matrix} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, k \end{matrix}$$

where we define $b_i = b_1$ for $i \leq 0$ and

$$u_i = \max\{b_1 - \sum_{t=1}^{i-1} d_t,\ c_i\} \text{ for } i = 1, \ldots, k.$$

Lemma 1.3.1 If $b_1 \geq \sum_{t=1}^{k-1} d_t$, then the expression (1.3.2) can be transformed so that IB(k, m, A, B, H, D) = I(k, m, A, B, C, D) where $A' = (a'_1, \ldots, a'_m)$, $B' = (b'_1, \ldots, b'_m)$ and $C' = (c'_1, \ldots, c'_m)$ are vectors such that $a'_i = h_k - b_{m-i+1}$, $b'_i = h_k - a_{m-i+1}$ for $i = 1, \ldots, m$ and $c'_i = h_k - c_{k-i+1}$ for $i = 1, \ldots, k$.

Theorem 1.3.2 IB(k, m, A, B, H, D) = $\det\{\bar{E}' : \bar{F}'\}$, the determinant of an augmented matrix of an $(m + k) \times m$ matrix $\bar{E}' = \{\bar{e}'_{ij}\}$ and an $(m + k) \times k$ matrix $\bar{F}' = \{\bar{f}'_{ij}\}$ such that

$$\bar{e}'_{ij} = \frac{\left(a_i - b_j + \sum_{t=1}^{k-1} d_t\right)_+^{(j-i+k)}}{(j - i + k)!} \qquad \begin{array}{l} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, m \end{array}$$

and

$$\bar{f}'_{ij} = \frac{\left(a_i - v_{k-j+1} + \sum_{t=1}^{k-1} d_t\right)_+^{(m+j-i)}}{(m + j - i)!} \qquad \begin{array}{l} i = 1, \ldots, m+k \text{ and} \\ j = 1, \ldots, k. \end{array}$$

where we define $a_i = a_m$ for $i \geq m$ and $v_i = \min\{a_m + \sum_{t=1}^{k-1} d_t, h_i\}$ for $i = 1, \ldots, k$.

The equalities in the following corollary are obtained from the boundary cases of the theorems.

Corollary 1.3.1 The following equalities hold for determinants of non-negative integral entries satisfying the specified conditions.

(i) When $C = 0$ and $b_1 \geq \sum_{t=1}^{k-1} d_t$, the determinant $\det\{E:F\}$ of Theorem 1.3.1 equals the determinant $\det\{E'\}$ where $E' = \{e'_{ij}\}$ is a $m \times m$ matrix, such that

$$e'_{ij} = \frac{\left(a_{m-j+1} - b_{m-i+1} + \sum_{t=1}^{k-1} d_t\right)_+^{(j-i+k)}}{(j-i+k)!} \qquad i,j = 1, \ldots, m$$

(ii) When $A = a$, $B = C = D = 0$ and $k = 1$, the determinant in (i) becomes

$$\begin{vmatrix} a^1 & \frac{a^2}{2!} & \cdots & \frac{a^j}{j!} & \cdots & \frac{a^{m-1}}{(m-1)!} & \frac{a^m}{m!} \\ 1 & \frac{a^1}{1!} & \cdots & \frac{a^{j-1}}{(j-1)!} & \cdots & \frac{a^{m-2}}{(m-2)!} & \frac{a^{m-1}}{(m-1)!} \\ 0 & 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \frac{a^{(j-i+1)}}{(j-i+1)!} & \cdots & \frac{a^{(m-i)}}{(m-i)!} & \frac{a^{(m-i+1)}}{(m-i+1)!} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & 1 & a \end{vmatrix} = \frac{a^m}{m!}$$

Remark 1.3.1 If we define $\overline{I}(k,m,A,B,C,D,H)$ (resp. $\overline{IB}(k,m,A,B,H,D,C)$) to be a measure represented by a multiple integral of the number 1 over the region of X specified by the conditions of (1.2.11) (resp. (1.2.12)). Hence $\overline{I}(k,m,A,B,C,D,H)$ (resp. $\overline{IB}(k,m,A,B,H,D,C)$) equals to a $k\cdot m$ fold integral of the number 1 which is of the same form as the right hand side of (1.3.1) (resp. (1.3.2)), except that here we have $a_{ij} = \min\{x_{i+1,j} + d_i, h_i, a_j\}$, $u_{ij} = \max\{b_j, x_{i,j-1}, c_i\}$ and $u_i = \max\{b_1, c_i\}$ (resp. $b_{ij}=\max\{x_{i-1,j}-d_i, c_i, b_j\}$, $v_{ij} = \min\{a_m, x_{i,j+1}, h_i\}$ and $v_i = \min\{a_m, h_i\}$).

With a similar reason as the one explained in Remark 1.2.1, we are unable to simplify the above expressions any further.

CHAPTER II

K-SAMPLE ANALOGUES OF THE KOLMOGOROV-SMIRNOV STATISTICS

2.1 Introduction

Let $X^i_1 \le X^i_2 \le \ldots \le X^i_{n_i}$ be the order statistics from a sample of n_i independent identically distributed (i.i.d.) random variables with a continuous cumulative distribution function (c.d.f.) F^i and a sample empirical distribution $F^i_{n_i}$, that is,

$$F^i_{n_i}(z) = \begin{cases} 0 & z < X^i_1 \\ j/n_i & X^i_j \le z < X^i_{j+1} \\ 1 & X^i_{n_i} \le z \end{cases} \qquad i = 1, \ldots, k.$$

For $k = 1$, Kolmogorov (1933, 1941) proposed the one-sample statistic

$$\sup_z \left| F^1_{n_1}(z) - G(z) \right|$$

for testing the goodness-of-fit hypothesis H_1: $F^1 = G$, where G is some specified c.d.f.

For $k = 2$, Smirnov (1939, 1948) proposed the two-sample statistic

$$\sup_z \left| F^1_{n_1}(z) - F^2_{n_2}(z) \right|$$

for testing the homogeneity hypothesis H_2: $F^1 = F^2$.

The asymptotic null distributions of these statistics were also found by the above authors. Among others, the exact null distributions were investigated by van der Waerden (1971) and Epanechnikov (1968) in the one sample case and by Gnedenko and Korolyuk (1951), Drion (1952) and Massey (1951) in the two sample case. However, simple and closed forms of both the one-sample and two-sample statistics in general setting were not known until Steck (1969, 1971). Recently, Govindarajulu, Alter and Gragg (1975) have used the generating function technique to obtain a closed form expression of the exact distribution of the one-sample Kolmogorov statistic. Interestingly, these results turn out to be special cases of the formulas developed in Chapter I.

K-sample analogues of the Kolmogorov-Smirnov statistics have been established by various authors. For example, David (1958) derived the null distribution of a one-sided three-sample statistic of the form

$$\max \{\sup_z(F_n^2(z) - F_n^1(z)),\ \sup_z(F_n^3(z) - F_n^2(z)),\ \sup_z(F_n^1(z) - F_n^3(z))\}$$

where $n_1 = n_2 = n_3 = n$. Kiefer (1955, 1959) considered statistics for testing the homogeneity hypothesis H_2: $F^1 = F^2 = \ldots = F^k$ or the goodness-of-fit hypothesis H_1: $F^1 = F^2 = \ldots = F^k = G$, where G is some specified c.d.f.. For testing H_1, the statistic is

$$U_2 = \sup_{\substack{z,i,j \\ i \neq j}} c_{ij} \left| F^i_{n_i}(z) - F^j_{n_j}(z) \right|, \qquad (2.1.1)$$

where c_{ij} is some fixed constant, for i,j = 1, ..., k and $i \neq j$. The statistic U_1 for testing H_1 may be obtained by writing G for $F^j_{n_j}$ and c_i for c_{ij}, j = 1, ..., k. Kiefer also showed that these statistics are

consistent against all alternatives and have good power properties. Dwass (1960) studied statistics of the same nature. However, the null distributions of the above statistics remained unknown. Conover (1967) found the exact distribution of the statistic

$$\sup_{z,i<k} (F_n^i(z) - F_n^{i+l}(z)),$$

where $n_1 = n_2 = \ldots = n_k = n$. In his paper, Conover stated that such a testing statistic would be useful in situations where the experimenter has k populations, $k > 2$ and may legitimately assume, from biological or other non-mathematical considerations, that $F^1(x) \geq F^2(x) \geq \ldots \geq F^k(x)$ for all x. Wolf and Naus (1973) provided tables of critical values based on Conover's result and showed that for certain alternatives, the test has reasonable power relative to parametric and other distribution-free competitors.

In this chapter, we obtain the null distribution functions of the statistics U_1 and U_2 for the special case of certain fixed constants c_{ij} in the definition. Furthermore, subject to some ordering conditions (see Sections 2.2 and 2.3) the conditional null distributions of the same statistics are also obtained. These conditional statistics are expected to be useful in situations like those specified by Conover(1967) that we have mentioned in the previous paragraph.

All the random variables considered in this chapter are univariate. Extension to the multivariate case is possible according to Bickel (1969) and Ahmad (1977).

The research work achieved under this subject is enormous (see Hájeck (1967)), our introduction only includes those references that are closely related to the topics discussed later.

2.2 K-sample Rank Statistics of Kolmogorov-Smirnov Type

In this section, we consider the statistic U_2 (cf. (2.1.1)) defined by Kiefer (1959) for testing the homogeneity hypothesis H_2: $F^1 = F^2 = \ldots = F^k$. It has been suggested by Anderson (1962) that the two-sample statistics of Kolmogorov-Smirnov type can be expressed in terms of statistically equivalent blocks, namely the ranks. Therefore, using the formulas of Section 1.2, we are able to determine the null distributions of the statistic for certain fixed constants c_{ij} in the definition. Furthermore, we also obtain the conditional null distributions of the same statistic subject to the restrictions on the ordering of the ranks. When $k = 2$, the conditions become degenerated and the result is due to Steck (1969).

Now we define the following statistics which constitute the U_1 statistic:

$$D_+^i(n_k, n_i) = \sup_z (F_{n_k}^k(z) - F_{n_i}^i(z))$$

$$D_-^i(n_k, n_i) = \sup_z (F_{n_i}^i(z) - F_{n_k}^k(z))$$

and

$$D^i(n_k, n_i) = \max \{D_+^i(n_k, n_i), D_-^i(n_k, n_i)\}$$

are the two-sample statistics for comparing the i^{th} sample with the k^{th} sample, $i = 1, \ldots, k-1$.

$$D_+(n_k; n_1, \ldots, n_{k-1}) = \sup_{1 \le i \le k-1} \{n_k n_i D_+^i(n_k, n_i)\}$$

$$D_-(n_k; n_1, \ldots, n_{k-1}) = \sup_{1 \le i \le k=1} \{n_k n_i D_-^i(n_k, n_i)\}$$

and

$$D(n_k;\ n_1,\ \ldots,\ n_{k-1}) = \sup_{1\le i\le k-1} \{n_k n_i D^i(n_k,\ n_i)\}$$

are the k-sample statistics measuring the supremum of the weighted maximum distance between the i^{th} sample and the k^{th} sample, for $1 \le i \le k-1$.

Note that those statistics with a subscript '+' or '-' are one-sided while those without a subscript are two-sided. In addition to the above, we write

$$D^{ij}(n_i,\ n_j) = \sup_z \left(F^i_{n_i}(z) - F^j_{n_j}(z)\right)$$

for every i, j = 1, ..., k-1 and $i \neq j$, and

$$D'(n_1,\ \ldots,\ n_{k-1}) = \sup_{\substack{1\le i,j\le k-1 \\ i\neq j}} \left\{\frac{n_i n_j n_k}{n_i+n_j} D^{ij}(n_i,\ n_j)\right\}$$

Then, the U_2 statistic can be expressed as

$$U_2(n_k;\ n_1,\ \ldots,\ n_{k-1}) = \max\{D(n_k,\ n_1,\ \ldots,\ n_{k-1}),\ D'(n_1,\ \ldots,\ n_{k-1})\}$$

for the special case that

$$c_{ij} = \begin{cases} n_i n_k & i = 1,\ \ldots,\ k-1,\ j = k \\ n_j n_k & j = 1,\ \ldots,\ k-1,\ i = k \\ \dfrac{n_i n_j n_k}{n_i+n_j} & i,j = 1,\ \ldots,\ k-1 \text{ and } i \neq j \end{cases}$$

Definition 2.2.1 Let i be an integer such that $1 \le i \le k-1$. Let $Z^i_1 \le Z^i_2 \le \ldots \le Z^i_{n_i+n_k}$ be the combined and ordered sample of $X^i_1, \ldots, X^i_{n_i}$ and $X^k_1, \ldots, X^k_{n_k}$. The rank of X^k_j in that sample, denoted by R^i_j, is the total number of Z^i_t, $1 \le t \le n_i+n_k$, which is less than or equal to X^k_j, for every $j = 1, \ldots, n_k$.

It is clear from the definition that

$$0 \leq R_1^i - 1 \leq R_2^i - 2 \leq \ldots \leq R_j^i - j \leq \ldots \leq R_{n_k}^i - n_k \leq n_i \qquad (2.2.1)$$

for every i = 1, ..., k-1.

The following is a result used by Steck (1969) who quoted the statement from a paper by Maag and Stephens (1968). However, the basic idea can be found in Anderson's (1962) book. For the sake of completeness, we provide the proof here.

Lemma 2.2.1 $$D_+^i(n_k, n_i) = \sup_{1 \leq j \leq n_k} \left(\frac{j}{n_k} - \frac{R_j^i - j}{n_i} \right),$$

for every i = 1, ..., k-1.

Proof: Without loss of generality, we may assume that i = 1. For any fixed j, $1 \leq j \leq n_k$, we have

$$F_{n_k}^k(x_j) - F_{n_1}^1(x_j) = \frac{j}{n_k} - \frac{R_j^1 - j}{n_1}.$$

Now

$$D_+^1(n_k, n_1) = \sup (F_{n_k}^k(z) - F_{n_1}^1(z))$$

$$= \sup_{z \in \{x_1^k, x_2^k, \ldots, x_{n_k}^k\}} (F_{n_k}^k(z) - F_{n_1}^1(z))$$

$$= \sup_{1 \leq j \leq n_k} \left(\frac{j}{n_k} - \frac{R_j^1 - j}{n_1} \right).$$

Thus, the lemma is proved.

By ordering the random variables from the largest to the smallest, that is $\tilde{X}^{i}_{1} \geq \tilde{X}^{i}_{2} \geq \ldots \geq \tilde{X}^{i}_{n_i}$, $i = 1, \ldots, k$, Steck (1969) obtained the following lemma as an analogue of Lemma 2.2.1.

Lemma 2.2.2 $$D^{i}_{-}(n_k, n_i) = \sup_{1 \leq j \leq n_k} \left(\frac{R^{i}_{j} - j}{n_i} - \frac{j-1}{n_k} \right),$$

for every $i = 1, \ldots, k-1$.

Proof: For any fixed integer i, such that $1 \leq i \leq k-1$, we write

$$D^{i}_{-}(n_k, n_i) = \sup_{z} \left(\tilde{F}^{k}_{n_k}(z) - \tilde{F}^{i}_{n_i}(z) \right)$$

where we let $F^{t}_{n_i}(z) = 1 - F^{t}_{n_i}(z)$, $X^{t}_{j} = X^{t}_{n_t - j + 1}$ for $j = 1, \ldots, n_t$ and $\tilde{Z}^{t}_{j} = Z^{t}_{n_t + n_k - j + 1}$ for $j = 1, \ldots, n_t + n_k$, so that $\tilde{X}^{t}_{1} \geq \tilde{X}^{t}_{2} \geq \ldots \geq \tilde{X}^{t}_{n_t}$ and $\tilde{Z}^{t}_{1} \geq \tilde{Z}^{t}_{2} \geq \ldots \geq \tilde{Z}^{t}_{n_t + n_k}$ for every $t = 1, \ldots, k$. Also let $\tilde{R}^{i}_{j}$ be the total number of $\tilde{Z}^{i}_{s}$, $1 \leq s \leq n_i + n_k$ such that $\tilde{Z}^{i}_{s} \geq X^{k}_{j}$ for $j = 1, \ldots, n_k$. Thus, $\tilde{R}^{i}_{j} = n_i + n_k - R^{i}_{n_k - j + 1} + 1$. By the definition,

$$\tilde{F}^{t}_{n_t}(z) = \begin{cases} 0 & \tilde{X}^{t}_{1} \leq z \\ r/n_t & \tilde{X}^{t}_{r+1} \leq z < \tilde{X}^{t}_{r} \\ 1 & z < \tilde{X}^{t}_{n_t} \end{cases}$$

where $r = n_j - s$, $t = 1, \ldots, k$. Thus by a similar argument as the proof of Lemma 2.2.1, we have

$$D_{-}^{i}(n_k, n_i) = \sup_{1 \le \ell \le n_k} \left(\frac{\ell}{n_k} - \frac{R_{\ell}^{i} - \ell}{n_i} \right)$$

Writing in terms of the original ranks

$$D_{-}^{i}(n_k, n_i) = \sup_{1 \le j \le n_k} \left(\frac{n_k - j + 1}{n_k} - \frac{(n_k + n_i - R_{n_k - j + 1}^{i}) - (n_k - j + 1)}{n_i} \right)$$

where $j = n_k - \ell + 1$. This can be reduced to the result of the lemma.

Let $\lfloor x \rfloor$ denote the largest integer less than or equal to x and let $\langle x \rangle = -\lfloor -x \rfloor$ denote the smallest integer greater than or equal to x. Using Lemma 2.2.1 and Lemma 2.2.2 and the fact that the ranks defined in Definition 2.2.1 are integers satisfying the inequalities (2.2.1), we can easily prove the following theorem which has been used by Steck (1969).

Theorem 2.2.1 The following hold for every integer i, $1 \le i \le k-1$.

(1) $\text{Prob}\{n_k n_i D_{+}^{i}(n_k, n_i) < r\}$

$$= \text{Prob}\left\{ \max\left\{0, \left\lfloor j\frac{n_i}{n_k} - \frac{r}{n_k} + 1 \right\rfloor\right\} \le R_j^i - j \le n_i,\ j = 1, \ldots, n_k \right\}$$

(2) $\text{Prob}\{ n_k n_i D_{+}^{i}(n_k, n_i)\} \le r$

$$= \text{Prob}\left\{ \max\left\{0, \left\langle j\frac{n_i}{n_k} - \frac{r}{n_k} \right\rangle\right\} \le R_j^i - j \le n_i,\ j = 1, \ldots, n_k \right\}$$

(3) $\text{Prob}\{ n_k n_i D_{-}^{i}(n_k, n_i) < r \}$

$$= \text{Prob}\left\{ 0 \le R_j^i - j \le \min\left\{ n_i, \left\langle \frac{r}{n_k} + j\frac{n_i}{n_k} - \frac{n_i}{n_k} - 1 \right\rangle \right\},\ j=1,\ldots,n_k \right\}$$

(4) $\text{Prob}\{n_k n_i D^i(n_k, n_i) \leq r\}$

$$= \text{Prob}\left\{0 \leq R_j^i - j \leq \min\left\{n_i, \left\lfloor \frac{r}{n_k} + j\frac{n_i}{n_k} - \frac{n_i}{n_k} \right\rfloor\right\} \quad j = 1, \ldots, n_k\right\}$$

(5) Suppose $n_i > 1$ and $r > \max\left\{n_k, \sup_{1 \leq i \leq k-1}\left(\frac{n_i + n_k}{2}\right)\right\}$, then

$\text{Prob}\{n_k n_i D^i(n_k, n_i) < r\}$

$$= \text{Prob}\left\{\max\left\{0, \left\langle j\frac{n_i}{n_k} - \frac{r}{n_k} + 1\right\rangle\right\} \leq R_j^i - j \leq \min\left\{n_i, \left\lfloor \frac{r}{n_k} + j\frac{n_i}{n_k} - \frac{n_i}{n_k} - 1 \right\rfloor\right\} \right\}$$
$$j = 1, \ldots, n_k$$

(6) Suppose $n_i \leq 1$ and $r \leq \max\left\{n_k, \sup_{1 \leq i \leq k-1}\left(\frac{n_i + n_k}{2}\right)\right\}$, then

$\text{Prob}\{n_k n_i D^i(n_k, n_i) \leq r\}$

$$= \text{Prob}\left\{\max\left\{0, \left\langle j\frac{n_i}{n_k} - \frac{r}{n_k}\right\rangle\right\} \leq R_j^i - j \leq \min\left\{n_i, \left\lfloor \frac{r}{n_k} + j\frac{n_i}{n_k} - \frac{n_i}{n_k} \right\rfloor\right\} \right\}$$
$$j = 1, \ldots, n_k$$

Remark 2.2.1 Notice that the extra conditions imposed on n_j and r that appears in (5) and (6) of Theorem 2.2.1 are to guarantee that the given inequalities hold without contradiction.

Given k populations, an experiment $\mathcal{E}$ for testing the null hypothesis $H_2 : F^1 = F^2 = \ldots = F^k$ may be performed in the following steps:

(1) a sample of size n_i, namely, $X_1^i, X_2^i, \ldots, X_{n_i}^i$ such that $X_1^i \leq X_2^i \leq \ldots \leq X_{n_i}^i$ is drawn from the i^{th} population, $i = 1, \ldots, k$.

(2) The sample $X_1^i, X_2^i, \ldots, X_{n_i}^i$ is combined with the sample $X_1^k, X_2^k, \ldots, X_{n_k}^k$ so that the ranks $R_1^i, R_2^i, \ldots, R_{n_k}^i$ are computed, for $i = 1, \ldots, k-1$.

(3) Compute the statistics defined at the beginning of this section by using Lemma 2.2.1 and Lemma 2.2.2.

It is clear that every outcome of the experiment $\mathcal{E}$ can be written in the form of a $(k-1) \times n_k$ matrix

$$R = \begin{bmatrix} R_1^1 & \ldots & R_j^1 & \ldots & R_{n_k}^1 \\ \vdots & & \vdots & & \vdots \\ R_1^i & \ldots & R_j^i & \ldots & R_{n_k}^i \\ \vdots & & \vdots & & \vdots \\ R_1^{k-1} & \ldots & R_j^{k-1} & \ldots & R_{n_k}^{k-1} \end{bmatrix} \qquad (2.2.2)$$

which is called a rank matrix. Therefore, the sample space of the experiment $\mathcal{E}$, that is, the collection of every possible outcome of $\mathcal{E}$, is the set of all rank matrices of the form R.

Suppose we consider a modified experiment of the experiment $\mathcal{E}$, namely, the experiment $\mathcal{E}'$, described as the follows. For every $i = 1, 2, \ldots, k-3$, after the ranks $R_1^i, R_2^i, \ldots, R_{n_k}^i$ has been determined, the sample $X_1^k, X_2^k, \ldots, X_{n_k}^k$

is replaced in the k^{th} population and a new sample of the same size is chosen from the k^{th} population in order to determine $R_1^{i+1}, R_2^{i+1}, \ldots, R_3^{i+1}$. Then it is obvious that the row vectors $R^i = (R_1^i, R_2^i, \ldots, R_{n_k}^i)$ and $R^j = (R_1^j, R_2^j, \ldots, R_{n_k}^j)$ of the rank matrix R are independent of each other whenever $i \neq j$ and thus

$$\text{Prob}\{R^i = U, R^j = V\} = \text{Prob}\{R^i = U\} \times \text{Prob}\{R^j = V\}$$

for any vectors $U = (U_1, U_2, \ldots, U_{n_k})$ and $V = (V_1, V_2, \ldots, V_{n_k})$ consisting of integers satisfying $0 \leq U_1 \leq U_2 \leq \ldots \leq U_{n_k} \leq n_i$ and $0 \leq V_1 \leq V_2 \leq \ldots \leq V_{n_k} \leq n_j$.

Thus each row vector of the rank matrix R can be regarded as an outcome of a two-sample experiment and the experiment $\mathcal{E}$ consists of k-1 independent two-sample experiments.

Since the statistic $D^i(n_k, n_i)$ is only a function of the ranks $(R_1^i, R_2^i, \ldots, R_{n_k}^i)$, $i = 1, \ldots, k-1$, we conclude that

$$(1)\quad \text{Prob}\{D(n_k; n_1, \ldots, n_{k-1}) < r\} = \prod_{i=1}^{k-1} \text{Prob}\{n_k n_i D^i(n_k, n_i) < r\}$$

and

$$(2)\quad \text{Prob}\{D(n_k; n_1, \ldots, n_{k-1}) \leq r\} = \prod_{i=1}^{k-1} \text{Prob}\{n_k n_i D^i(n_k, n_i) \leq r\}$$

Recall that N(k, m, A, B, C, D) and NB(k, m, A, B, C, H, D) are functions whose values have been determined in Section 1.2 for given positive integers k and m and vectors A, B, C, D and H satisfying the conditions specified by (1.2.2) and (1.2.3) respectively.

It is not hard to find the null distribution of the statistic $D(n_k; n_1, \ldots, n_{k-1})$ obtained from the experiment $\mathcal{E}'$ since the null distributions of the two-sample statistics $D^i(n_k,n_i)$, $i = 1,2,\ldots,k-1$ are known due to Steck (1979). In terms of our function $N(1,n_k,A,B,C,D)$, Steck's result can be formulated as the following:

(1) $\text{Prob}\{n_k n_i D^i(n_k, n_i) < r\} = N(1,n_k,E^i,F^i,0,0) \,/\, N(1,n_k,n_i,0,0,0)$

where $E^i = (e^i_1, \ldots, e^i_{n_k})$, $F^i = (f^i_1, \ldots, f^i_{n_k})$ are vectors such that

$$e^i_j = \min\left\{n_i, \left\langle \frac{r}{n_k} - j\frac{n_i}{n_k} - \frac{n_i}{n_k} - 1 \right\rangle\right\} \text{ and } f^i_j = \max\left\{0, \left\lfloor j\frac{n_i}{n_k} - \frac{r}{n_k} + 1 \right\rfloor\right\} \text{ for}$$

$j = 1, \ldots, n_k$ and $i = 1, \ldots, k-1$.

(2) $\text{Prob}\{n_k n_i D^i(n_k, n_i) \leq r\} = N(1,n_k,E'^i,F'^i,0,0) \,/ N(1,n_k,n_i,0,0,0)$

where $E'^i = (e'^i_1, \ldots, e'^i_{n_k})$ and $F' = (f'^i_1, \ldots, f'^i_{n_k})$ are vectors such that

$$e'^i_j = \min\left\{n_i, \left\lfloor \frac{r}{n_k} + j\frac{n_i}{n_k} - \frac{n_i}{n_k} \right\rfloor\right\} \text{ and } f'^i_j = \max\left\{0, \left\langle j\frac{n_i}{n_k} - \frac{r}{n_k} \right\rangle\right\} \text{ for}$$

$j = 1, \ldots, n_k$ and $i = 1, \ldots, k-1$.

However, for the statistics obtained from the output of the experiment $\mathcal{E}$, we are only able to find several conditional distribution functions in terms of the functions $N(k-1,n_k,A,B,C,D)$ or $NB(k-1,n_k,A,B,H,D)$.

Theorem 2.2.2 Under the conditions that $k \geq 2$, $R^i_j \leq R^{i+1}_j + d_i$ for $j = 1, \ldots, n_k$ and $i = 1, \ldots, k-1$, where d_i's are non-negative integers such that $0 \leq d_i \leq n_i - n_{i+1}$, $i = 1, \ldots, k-2$, the following are true when the null hypothesis H_2 holds.

(1) Prob $\{n_k n_1 D_+^1(n_k, n_1) < r_1, n_k n_{k-1} D_-^{k-1}(n_k, n_{k-1}) < r_2\}$

$= N(k-1, n_k, A, B, 0, D)/N(k-1, n_k, n_{k-1}, 0, 0, D)$,

where $A = (a_1, \ldots, a_{n_k})$, $B = (b_1, \ldots, b_{n_k})$, $D = (d_1, \ldots, d_{k-2})$,

$$a_j = \min\left\{ n_{k-1}, \left\langle \frac{r_2}{n_k} - j\frac{n_{k-1}}{n_k} - \frac{n_{k-1}}{n_k} - 1 \right\rangle \right\} \text{ and } b_j = \max\left\{ 0, \left\lfloor j\frac{n_1}{n_k} - \frac{r_1}{n_k} + 1 \right\rfloor \right\}$$

for $j = 1, \ldots, n_k$.

(2) Prob $\{n_k n_1 D_+^1(n_k, n_1) \leq r_1, n_k n_{k-1} D_-^{k-1}(n_k, n_{k-1}) \leq r_2\}$

$= N(k-1, n_k, A', B', 0, D)/N(k-1, n_k, n_{k-1}, 0, 0, D)$,

where $A' = (a_1', \ldots, a_{n_k}')$, $B' = (b_1', \ldots, b_{n_k}')$ and

$$a_j' = \min\left\{ n_{k-1}, \left\lfloor \frac{r_2}{n_k} + j\frac{n_{k-1}}{n_k} - \frac{n_{k-1}}{n_k} \right\rfloor \right\}, \quad b_j = \max\left\{ 0, \left\langle j\frac{n_1}{n_k} - \frac{r_1}{n_k} \right\rangle \right\}, \text{ for}$$

$j=1, \ldots, n_k$.

Proof: Under the null hypothesis H_2, each distinct rank matrix of the form R (cf. (2.2.2)) happens equally likely as an outcome of the experiment $\mathcal{E}$. Since each row vector of R satisfies the inequalities (2.2.1) by definition, we know that the total number of distinct matrices of the form R which satisfies the conditions stated in the theorem is equal to the number $N(k-1, n_k, n_{k-1}, 0, 0, D)$ and the total number of distinct rank matrices which also satisfy the restrictions $n_k n_1 D_+^1(n_k, n_1) < r_1$, and $n_k n_{k-1} D_-^{k-1}(n_k, n_{k-1}) < r_2$ is equal to the number $N(k-1, n_k, A, B, 0, D)$. Thus the result of the first part of the theorem follows. The second part can also be proved analogously.

Remark 2.2.2 For k = 2, the conditions of Theorem 2.2.2 degenerate and the two-sample result by Steck (1969) follows.

Similar arguments can be used to prove the following theorem.

Theorem 2.2.3 Under the conditions that $k \geq 2$, $n_1 \leq n_2 \leq \ldots \leq n_k$, and that $R^i_j \leq R^{i+1}_j + d_i$, $j = 1, \ldots, n_k$ and $i = 1, \ldots, k-2$, where d_i's are non-negative integers such that $\sum_{i=1}^{k-2} d_i \leq \max\left\{0, \left\lfloor \frac{n_1}{n_k} - \frac{r}{n_k} + 1 \right\rfloor\right\}$, the following is true when the null hypothesis H_2 holds.

(1) $\text{Prob}\{n_k n_1 D^1_+(n_k,n_1) < r_1,\ n_k n_{k-1} D^{k-1}_-(n_k,n_{k-1}) < r_2\}$

$= NB(k-1,n_k,A,B,H,D)/NB(k-1,n_k,n_{k-1},0,H,D)$,

where A, B are the same as those defined in Theorem 2.2.2, $D = (d_1,\ldots,d_{k-2})$ and $H = (n_1, \ldots, n_{k-1})$.

In addition, if the inequality $\sum_{i-1}^{k-2} d_i \leq \max\left\{0, \frac{n_1}{n_k} - \frac{r}{n_k}\right\}$ is also satisfied, then the following is true when the null hypothesis H_2 holds.

(2) $\text{Prob}\{n_k n_1 D^1_+(n_k,n_1) \leq r_1,\ n_k n_{k-1} D^{k-1}_-(n_k,n_{k-1}) \leq r_2\}$

$= NB(k-1,n_k,A',B',H,D)/NB(k-1,n_k,n_{k-1},0,H,D)$,

where A', B' are the same as those defined in Theorem 2.2.2.

Corollary 2.2.1 Under the conditions that $k \geq 2$, $R^i_j \leq R^{i+1}_j$, $j=1,\ldots,n_k$ and $i = 1, \ldots, k-1$ and that $n_1 = \ldots = n_{k-1} = n$, then the following is true when the null hypothesis H_2 holds.

(1) Prob $\{D_+(n_k; n_1, \ldots, n_{k-1}) < r_1, D_-(n_k; n_1, \ldots, n_{k-1}) < r_2\}$

$= N(k-1,n_k,A,B,0,0)/N(k-1,n_k,n_{k-1},0,0,0)$.

(2) Prob $\{D_+(n_k; n_1,\ldots, n_{k-1}) \leq r_1, D_-(n_k; n_1, \ldots, n_{k-1}) \leq r_2\}$

$= N(k-1,n_k,A',B',0,0)/N(k-1,n_k,n_{k-1},0,0,0)$.

Proof: It suffices to show that in this case,

Prob $\{D_+(n_k; n_1, \ldots n_{k-1}) < r_1, D_-(n_k; n_1, \ldots, n_{k-1}) \leq r_2\}$

$=$ Prob $\{n_k n_1 D_+^1(n_k, n_1) < r_1, n_k n_{k-1} D_-^{k-1}(n_k, n_{k-1}) < r_2\}$

and the above is also true when "<" is replaced everywhere by "$\leq$". Then the corollary follows from either Theorem 2.2.2 or Theorem 2.2.3 since $N(k-1,n_k,A,B,0,0) = NB(k-1,n_k,A,B,H,0)$ under the given conditions.

Our assertion is true because $R_j^i \leq R_j^{i+1}$, $j = 1, \ldots, n_k$, $i=1,\ldots,k-1$ and Lemma 2.2.1 implies that

$$n_k n_i D_+^i(n_k,n_i) \leq n_k n_1 D_+^1(n_k,n_1) \qquad i = 1, \ldots, k-1$$

and

$$n_k n_i D_-^i(n_k,n_i) \leq n_k n_{k-1} D_-^{k-1}(n_k,n_{k-1}) \qquad i = 1, \ldots, k-1$$

when $n_1 = \ldots = n_k = n$. Therefore, in this case, the conditions

$$D_+(n_k; n_1, \ldots, n_{k-1}) < r_1 \text{ and } D_-(n_k; n_1, \ldots, n_{k-1}) < r_2$$

are equivalent to the conditions

$$D_+^1(n_k, n_1) < r_1 \text{ and } D_-^{k-1}(n_k, n_{k-1}) < r_2$$

when $n_1 = \ldots = n_k = n$. Thus the result of the corollary follows.

Under the null hypothesis H_2, we have already found the null distribution and the conditional null distribution of the statistic $D(n_k;n_1,\ldots,n_{k-1})$. In the next theorem we show that the distribution of the statistic $U_2(n_k,n_1,\ldots,n_{k-1})$ is the same as the distribution of the statistic $D(n_k;n_1,\ldots,n_{k-1})$.

Theorem 2.2.4 **The cumulative probability** distribution function of the statistic $U_2(n_k;n_1,\ldots,n_{k-1})$ is given by

(1) $\text{Prob}\{U_2(n_k;n_1,\ldots,n_{k-1}) < r\} = \text{Prob}\{D(n_k;n_1,\ldots,n_{k-1}) < r\}$

(2) $\text{Prob}\{U_2(n_k;n_1,\ldots,n_{k-1}) \leq r\} = \text{Prob}\{D(n_k;n_1,\ldots,n_{k-1}) \leq r\}$

Proof: Since

$$\sup_z \left|(F^i_{n_i}(z) - F^j_{n_j}(z))\right| \leq \sup_z \left|F^k_{n_k}(z) - F^j_{n_j}(z)\right| + \sup_z \left|F^k_{n_k}(z) - F^i_{n_i}(z)\right|,$$

writing in terms of the statistics, we have

$$D^{ij}(n_i,n_j) \leq D^j(n_k,n_j) + D^i(n_k,n_i)$$

$$\frac{n_k n_i n_j}{n_i+n_j} D^{ij}(n_i,n_j) \leq \frac{n_i}{n_i+n_j} n_k n_j D^j(n_k,n_j) + \frac{n_j}{n_i+n_j} n_k n_i D^i(n_k,n_i)$$

for every $i,j = 1, \ldots, k-1$ and $i \neq j$. Taking sup on both sides of the above inequality, we find that the right hand side is bounded by the statistic $D(n_k; n_1,\ldots, n_{k-1})$, while the left hand side becomes $D'(n_1,\ldots,n_{k-1})$. Therefore, $D(n_k;n_1,\ldots,n_{k-1}) < r$ implies that $D'(n_1,\ldots,n_{k-1}) < r$. Thus,

$$\text{Prob}\{D(n_k;n_1,\ldots,n_{k-1}) < r\} = \text{Prob}\{D(n_k;n_1,\ldots,n_{k-1}) < r,\ D'(n_1,\ldots,n_{k-1}) < r\}$$
$$= \text{Prob}\{U_2(n_k;n_1,\ldots,n_{k-1}) < r\}.$$

This proves the first part of the theorem. The second part follows analogously.

Remark 2.2.3 The conditions imposed on the rank matrices when we find the null districtions in Theorem 2.2.2 and Theorem 2.2.3 arise due to the restrictions given on the matrices enumerated by using Theorem 1.2.1 and Theorem 1.2.2. Therefore, if we can improve the results of Chapter I by finding simple expressions for enumerating the classes of matrices satisfying the conditions (1.2.11) or (1.2.12) mentioned in Remark 1.2.1, the corresponding results on the conditional null distributions obtained in this chapter can also be improved.

2.3 K-sample Order Statistics of Kolmogorov-Smirnov Type

In this section, we consider statistics of the form U_1 defined by Kiefer (1959) (cf. Section 2.1) for testing the goodness-of-fit hypothesis H_1: $F^1 = F^2 = \ldots = F^k = G$. When k = 1, van der Waerden (1971) wrote the statistic as a function of the order statistics of a given sample; Epanechnikov (1968) found an exact expression for the distribution of the statistic, and Steck (1971) obtained a closed form of the distribution under the hypothesis H_1. When k is any finite integer greater than or equal to 1, we use the formulas of Section 1.3 to determine, under the null hypothesis H_1, the null distribution of the statistic

$$\bar{U} = \sup_z \left| F^i_{n_i}(z) - G(z) \right|$$

which is a special form of the statistic U_1 when $c_1 = \ldots = c_k = 1$. In fact, this is only a combination of Steck's (1971) result. However, we obtain further the conditional null distribution of the same statistic subject to the restriction on the ordering of the order statistics between the k samples.

Now we define the following statistics which constitute the $\bar{U}$ statistic.

$$\bar{D}^i_+(n_i) = \sup_z (G(z) - F^i_{n_i}(z))$$

$$\bar{D}^i_-(n_i) = \sup_z (F^i_{n_i}(z) - G(z))$$

and

$$\bar{D}^i(n_i) = \max \{\bar{D}^i_+(n_i), \bar{D}^i_-(n_i)\}$$

are usually called the one-sample Kolmogorov statistics. The first two are one-sided and the third one is two-sided. In the k-sample case, we

define

$$\bar{D}_+(n_1,\ldots,n_k) = \sup_{1\le i\le k} \{\bar{D}_+^i(n_i)\}$$

$$\bar{D}_-(n_1,\ldots,n_k) = \sup_{1\le i\le k} \{\bar{D}_-^i(n_i)\}$$

and

$$\bar{U}(n_1,\ldots,n_k) = \sup_{1\le j\le k} \{\bar{D}^i(n_i)\}$$

Since the statistic $\bar{U}(n_1,\ldots,n_k)$ is a function of the one-sample statistics $D^i(n_i)$, i=1, ..., k, the distribution of $U(n_1,\ldots,n_k)$ is based on the distributions of these one-sample statistics. Therefore, we need the following lemma and theorem which are originally due to van der Waerden (1969) and Epanechnikov (1968), respectively. For the sake of completeness, the proofs are also sketched here.

Lemma 2.3.1 $$\sup_z(F^i(z) - F^i_{n_i}(z)) = \sup_{1\le h\le n_i}\left(x^i_h - \frac{h-1}{n_i}\right)$$

$$\sup_z(F^i_{n_i}(z) - F^i(z)) = \sup_{1\le h\le n_i}\left(\frac{h}{n_i} - x^i_h\right)$$

where x^i_h is the sample point of the order statistic X^i_h, for every $h = 1, \ldots, n_i$ and $i = 1, \ldots, k$.

Proof: Without loss of generality, we may assume that $i = 1$. Since a continuous monotone transformation of the z axis leaves the differences $(F^1(z) - F^1_{n_i}(z))$ unchanged, we can replace z and x^1_h by the new variables $z' = F^1(z)$ and $(x^1_h) = F^1(x^1_h)$ without changing the maximal difference

$\sup_z (F^1(z) - F^1_{n_1}(z))$. Let us call the new variables z and x^1_h again, so the distribution function assumes the simple form

$$F^1(z) = z \qquad 0 < z < 1.$$

Since all of the x^1_h's lie between 0 and 1, we can set

$$F(z) = 0 \qquad z \leq 0$$

$$F(z) = 1 \qquad z \geq 1$$

Hence, the probability density function (p.d.f.) is

$$f(z) = \begin{cases} 1 & 0 < z < 1 \\ 0 & \text{otherwise.} \end{cases}$$

At the point x^1_h, the function $F^1_{n_1}(z)$ jumps from $\frac{h-1}{n_1}$ to $\frac{h}{n_1}$. It is clear that the maximum of the difference $F^1(z) - F^1_{n_1}(z)$ must occur at one of the points $x^1_1, x^1_2, \ldots, x^1_h, \ldots, x^1_{n_1}$. Thus, the results of the lemma follow.

Theorem 2.3.1 For any integer i, $1 \leq i \leq k$ and any real number r, $-1 \leq r \leq 1$, the following hold under the null hypothesis H_1.

(1) $\text{Prob}\{\bar{D}^{i}_{+}(n_i) \leq r\} = n_i! \int_{G^i_+(r)} \cdots \int dx^i_1 \ldots dx^i_{n_i}$,

where $G^i_+(r)$ is a n_i dimensional region of integration specified by the conditions $0 \leq x^i_1 \leq \ldots \leq x^i_{n_i} \leq 1$ and $0 \leq x^i_h \leq \min\left\{r + \frac{h-1}{n_i}, 1\right\}$, for $h = 1, \ldots, n_i$.

(2) Prob $\{\bar{D}^i_-(n_i) \leq r\} = n_i! \int_{G^i_-(r)} \cdots \int dx^i_1 \ldots dx^i_{n_i}$,

where $G^i_-(r)$ is a n_i dimensional region of integration specified by the conditions $0 \leq x^i_1 \leq \ldots \leq x^i_{n_i} \leq 1$ and $\max\ 0, \frac{h}{n_i} - r \leq x^i_h \leq 1$ for $h = 1, \ldots, n_i$.

(3) Prob $\{\bar{D}^i(n_i) \leq r\} = n_i! \int_{G^j(r)} \cdots \int dx^i_1 \ldots dx^i_{n_i}$,

where $G^i(r)$ is a n_i dimensional region of integration specified by the conditions $0 \leq x^i_1 \leq \ldots \leq x^i_{n_i} \leq 1$ and $\max \left\{0, \frac{h}{n_i} - r\right\} \leq x^i_h \leq \min\left\{r + \frac{h-1}{n_i}, 1\right\}$, for $h = 1, \ldots, n_i$.

Proof: Since the order statistics $X^i_1, \ldots, X^i_{n_i}$ are i.i.d., the sample points $x^i_1, \ldots, x^i_{n_i}$, after the transformation described in the proof of Lemma 2.3.1, are distributed with p.d.f.

$$f(x^i_h) = \begin{cases} 1 & 0 < x^i_{n_i} < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, the joint p.d.f. of the order statistics is $n_i!$ in the region specified by $0 \leq x^i_1 \leq \ldots \leq x^i_{n_i} \leq 1$. Under the null hypothesis H_1, we have $F^1 = \ldots = F^k = G$, therefore, Lemma 2.3.1 implies that the condition $D^i_+(n_i) \leq r$ is equivalent to

$$x^i_h - (h-1)/n_i \leq r \qquad h = 1, \ldots, n_i.$$

This proves (1) of the theorem. (2) and (3) follow analogously.

Given k populations, an experiment $\mathcal{E}_1$ for testing the null hypothesis H_1: $F^1 = F^2 = \ldots = F^k = G$ may be performed in the following steps:

(1) a sample of size n_i, namely $X_1^i, X_2^i, \ldots, X_{n_i}^i$ such that $X_1^i \leq X_2^i \leq \ldots \leq X_{n_i}^i$ is drawn from the i^{th} population, $i = 1, \ldots, k$.

(2) Compute the statistics defined at the beginning of this section by using Lemma 2.3.1.

In order to find the null distributions of the statistics, we recall that I(k, m, A, B, C, D) and IB(k, m, A, B, H, D) are functions whose values have been determined in Section 1.3 for given vectors A, B, C, D and H satisfying the conditions specified by (1.2.2) and (1.2.3) respectively.

The forthcoming theorems and corollary are the main results of this section.

Theorem 2.3.2 Under the conditions that $k \geq 1$, $n_1 = \ldots = n_k = n$ and $x_j^i \leq x^{i+1}_j$ for every $j = 1, \ldots, n$ and $i = 1, \ldots, k-1$, the following is true when the null hypothesis H_1 holds.

$$\text{Prob}\left\{\bar{D}_-^{-1}(n) \leq r_1,\ \bar{D}_+^{k}(n) \leq r_2\right\} = (n!)^k I(k, n, A, B, 0, 0)$$

where $A = (a_1, \ldots, a_n)$, $B = (b_1, \ldots, b_n)$ and $a_j = \min\left\{r_2 + \frac{j-1}{n}, 1\right\}$, $b_j = \max\left\{0, \frac{j}{n} - r_1\right\}$ for $j = 1, \ldots, n$.

Proof: From the proof of Theorem 2.3.1, we know that the joint p.d.f. of the i.i.d. order statistics $X_1^i, \ldots, X_{n_i}^i$ is $n_i!$, for every

$i = 1, \ldots, k$. Since the random variables between the samples are also independent of each other, the joint p.d.f. of the k samples of order statistics is equal to $(n!)^k$. Hence, Theorem 2.3.1 implies that the joint probability distribution of the statistics $\bar{D}_-^1(n)$ and $\bar{D}_+^k(n)$ considered in this theorem is the integration of the constant $(n!)^k$ over the region which is the same as the region of integration defined by the function $I(k, n, A, B, 0, 0)$. Thus, the result of the theorem follows.

Similarly, we can prove the following theorem.

Theorem 2.3.3 Under the conditions that $k \geq 1$, $n_1 = \ldots = n_k = n$ and $x_j^i \leq x_j^{i+1} + d_i$, $j = 1, \ldots, n$ and $i = 1, \ldots, k-1$, where d_i's are non-negative real numbers such that $0 \leq d_i \leq 1$ for $i = 1, \ldots, k-1$, and $\sum_{i=1}^{k-1} d_i \leq \max\left\{0, \frac{1}{n} - r_1\right\}$, the following is true when the null hypothesis H_1 holds.

$$\text{Prob}\left\{\bar{D}_-^1(n) \leq r_1, \bar{D}_+^k(n) \leq r_2\right\} = IB(k, n, A, B, 1, D)$$

where A and B are the same as those defined in Theorem 2.3.2, and $D = (d_1, \ldots, d_{k-1})$.

Corollary 2.3.1 Under the conditions that $k \geq 1$, $n_1 = \ldots = n_k = n$ and $x_j^i \leq x_j^{i+1}$ for $j = 1, \ldots, n$ and $i = 1, \ldots, k-1$, then the following holds under the null hypothesis H_1.

$$\text{Prob}\left\{\bar{D}_-(n_1, \ldots, n_k) \leq r_1, \bar{D}_+(n_1, \ldots, n_k) \leq r_2\right\} = (n!)^k I(k,n,A,B,0,0)$$

where A, B, and D are the same as those defined in Theorem 2.3.3.

Proof: This is true because of the fact that

$$\text{Prob}\left\{\bar{D}_-(n_i, \ldots, n_i) \underset{-}{\leq} r_1,\ \bar{D}_+(n_1, \ldots, n_k) \underset{-}{\leq} r_2\right\}$$

$$= \text{Prob}\left\{\bar{D}_-^{1}(n) \underset{-}{\leq} r_1,\ \bar{D}_+^{k}(n) \underset{-}{\leq} r_2\right\}$$

which can be verified from the given conditions and Lemma 2.3.1.

Remark 2.3.1 The conditional null distribution of the statistic $\bar{U}(n_1, n_2, \ldots, n_k)$ has been determined by Corollary 2.3.1. For the special case that k = 1, we obtain Steck's (1971) result. In general, the null distribution of the statistic $\bar{U}(n_1, n_2, \ldots, n_k)$ can be expressed as

$$\text{Prob}\left\{\bar{U}(n_1, \ldots, n_k) \underset{-}{\leq} r\right\} = \prod_{i=1}^{k} \text{Prob}\left\{\bar{D}^{i}(n_i) \underset{-}{\leq} r\right\},$$

since the k samples $X_1^i, \ldots, X_{n_i}^i$, i= 1,...,k are independently drawn in the experiment $\mathcal{E}_1$ and therefore the statistics $\bar{D}^i(n_i)$, i = 1,...,k are independent. This enables us to compute the statistic $\bar{U}(n_1,\ldots,n_k)$ based on the 1-sample result.

The conditions imposed on the order statistics when we find the null distributions in Theorem 2.3.2 and Theorem 2.3.3 arise due to the restrictions given on the matrices used to specify the regions of integrations of the k·m fold multiple integrals detetermined in Theorem 1.3.1 and Theorem 1.3.2. Therefore, if simple expressions can be found for the integrals defined in Remark 1.3.1, the corresponding results on the conditional null distributions obtained in this chapter can also be improved.

2.4 Procedure of Testing the Null Hypotheses

Here we describe precisely how the k-sample statistics defined in this chapter can be used to test either the null hypothesis H_2: $F^1=\ldots=F^k$ or the null hypothesis H_1: $F^1=\ldots=F^k = G$.

Let S represent any one of the statistics whose c.d.f. has been found in this chapter. Let

$$p = \text{Prob}\{S \leq r\} \tag{2.4.1}$$

then we are able to determine the value p for a given r, or to approximate the number r for a fixed p. Therefore, the following testing procedures can be used as an analogue to those given by Maag and Stephens (1968).

Suppose there are k populations with continuous distributions. Draw one sample of prescribed size from each population and calculate the value of the statistic S according to its definition. If we find that S = s, then the two test at significance level α can be performed in the following two directions:

(i) Compute p_s such that $p_s = \text{Prob}\{S \leq s\}$. If $1 - p_s \leq \alpha$, reject the null hypothesis at the significance level α.

(ii) Approximate the value of r for $p = 1 - \alpha$, such that (2.4.1) is satisfied. If $s > r$, reject the null hypothesis at the significance level α.

CHAPTER III

A STUDY OF ROOTED PLANE TREES

3.1 Introduction

In this chapter, we use the definition of a rooted plane tree given by Klarner (1970). Two representation theorems will be presented in Section 3.2, namely, the pseudo-search code representation and the matrix representation. This reformulates the results of Chorneyko and Mohanty (1972, 1975) who have identified both a rooted plane tree and a pseudo-search code with a lattice path. Then we are able to enumerate in Section 3.3 certain classes of rooted plane trees by using the generating function techniques or the formulas of Chapter I. Finally, construction of optimal alphabetic q-ary trees are studied in Section 3.4. This involves the investigation of the connections between rooted plane trees and codes in information theory. As a result, we are able to establish a necessary and sufficient condition on the path lengths of a q-ary tree (a refinement of Kraft's inequality) and provide an algorithm for constructing an optimal alphabetic q-ary tree in terms of pseudo-search codes by using a computer (a generalization of Schwartz's and Kallick's (1964) algorithm). An important application of the optimal q-ary trees can be found in the k-sample group testing problem of Chapter IV.

3.2 Definitions and Representation Theorems

The notion of a rooted plane tree considered here is the one used by Klarner (1969). For undefined terms see the book by Harary and Palmer (1973).

Let $T(V, E, v, \alpha)$ be a rooted plane tree where V is the vertex set, E the edge set (a set of 2-subsets of V), v a distinguished vertex called the root, and α a linear order relation on V possessing the following properties:

(i) For $x, y \in V$, if $\rho(x) < \rho(y)$, then $x \alpha y$, where $\rho(x)$ is the path length from v to x and is called the path length of x. In particular, $\rho(v) = 0$.

(ii) If $\{r, s\}, \{x, y\} \in E$, $\rho(r) = \rho(x) = \rho(s) - 1 = \rho(y) - 1$ and $r \alpha x$, then $s \alpha y$.

A rooted tree is called a planted tree if the degree of the root is 1.

Any vertex of degree 1 other than the root is called an end vertex. Any vertex with degree greater than 1 is called a branch vertex. A tree having all its branch vertices of degree $q + 1$ is called a q-ary tree.

Two rooted plane trees $T_1(V, E_1, v_1, \alpha_1)$ and $T_2(V, E_2, v_2, \alpha_2)$ are isomorphic if there exists a permutation ϕ of V such that $\phi(v_1) = v_2$, $E_2 = \{\{\phi(x), \phi(y)\}; \{x, y\} \in E_1\}$ and $x \alpha_1 y$ if and only if $\phi(x) \alpha_2 \phi(y)$. One can easily draw a diagram of a rooted plane tree by arranging the vertices in levels so that vertex x is in level $\rho(x)$ and then arranging the vertices in each level from left to right according to the order relation α.

We note that two trees are isomorphic when they have "the same" diagram in the plane.

Example 3.2.1 A diagram of a rooted plane tree $T(V, E, v, \alpha)$ where $V = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$, $E = \{\{1, 2\}, \{1, 3\}, \{2, 4\}, \{2, 5\}, \{3, 6\}, \{3, 7\}, \{6, 8\}, \{6, 9\}, \{6, 10\}\}$ is given in Figure 3.2.1. The linear ordering α on the vertices is $1 \alpha 2 \alpha 3 \alpha 4 \alpha 5 \alpha 6 \alpha 7 \alpha 8 \alpha 9 \alpha 10$.

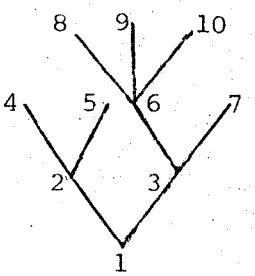


Figure 3.2.1 A Rooted Plane Tree

It is clear from the definition that even though the vertices of a rooted plane tree are not labelled, the ordering α imposed on the vertices has already implied a natural labelling.

One may ask, "Can we represent a rooted plane tree analytically without showing its diagram?". The answer is positive. In fact, many authors, for example, Klarner(1970), Chorneyko and Mohanty (1975) have tried to establish various representations of such trees. Here we provide two methods, namely, the pseudo-search code representation and the matrix representation. They can be considered as new versions of the results of Chorneyko and Mohanty (1972, 1975).

3.2.1 Pseudo-search Code Representation

Let us define the lexicographic labelling of a rooted plane tree as the following. The root is labelled by e. The vertices of path length 1 are labelled from left to right as 0, 1, ..., p, provided that there are p + 1 of them. The labelling of the remaining vertices are determined firstly by the labelling of the branch vertex adjacent to it (joined to it by an edge) but with a shorter path length, and, secondly by the left to right ordering among the vertices of same path length and adjacent to the same branch vertex. For example, if a branch vertex of degree q + 1 is labelled as x where x is a concatenation of natural numbers, then the q vertices adjacent to it but with longer path lengths are labelled from the left to the right as x0, x1, ..., xq, respectively, where xi is the concatenation of x and i, i = 1, ..., q.

It is easy to see that trees in the same isomorphism class have the same lexicographic labelling. Furthermore, the set of all labels at the end vertices of a rooted plane tree resulting from the lexicographic labelling determines uniquely the isomorphic class of the tree.

We define the lexicographic ordering β of the vertices of a rooted plane tree to be the lexicographic ordering of the corresponding labels obtained from the lexicographic labelling of the tree. Assume that the label e of the root is always first in order.

Example 3.2.2 When the tree T(V, E, v, α) of Example 3.2.1 is lexicographically labelled, we represent the tree by T(V, E, v, β). The diagram of such a tree is shown in Figure 3.2.3. The vertices in the lexicographic ordering is e β 0 β 1 β 00 β 01 β 1 β 10 β 100 β 101 β 102 β 11. The set C = {00, 01, 100, 101, 102, 11} determines the isomorphic class of the tree uniquely.

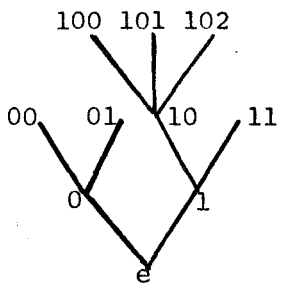


Figure 3.2.2 The Tree Represented by the Set C = {00, 01, 100, 101, 102, 11}

Later we shall see that sets like the set C of Example 3.2.2 are in fact a pseudo-search code defined by Chorneyko and Mohanty (1972) who modified Rényi's(1969) definition of a search code. For the sake of completeness, we list the necessary definitions:

Definition 3.2.1 A finite sequence of non-negative integers is called a codeword. The length of a codeword is the number of non-negative integers contained in it.

We denote the codewords by small Latin letters (a, b, c, ...), where each of the a, b, c, ... is a codeword of length $\ell \geq 1$. When no confusion arises, we omit the comma and bracket signs. The length of the codeword a is denoted by $\ell(a)$. It is convenient to consider the empty sequence, e, as a codeword. The set of all codewords is denoted by Z.

Definition 3.2.2 A codeword b is called a prefix of a codeword C if there exists a codeword d such that C = bd.

It is obvious that $\ell(bd) = \ell(b) + \ell(d)$.

Definition 3.2.3 A finite set C of different codewords is called a code.

The empty set is considered to be a code and is called the empty code. The code consisting of the empty codeword e only is called the trivial code.

Definition 3.2.4 If C is a code and a any codeword, then C_a is the set of all codewords $b \in Z$ such that $ab \in C$.

We denote by N(C) the number of codewords in the code C.

Definition 3.2.5 A code C is branched if one of the following occurs:

(1) C is the empty code.

(2) C is the trivial code.

(3) C does not contain e and there exists an integer $b(C) \geq 1$, such that for k, the codeword consisting of the single letter k, $k \neq 0, 1, 2, \ldots$, the code C_k is empty or non-empty according as $k \geq b(C)$ or $k < b(C)$.

We call b(C) the branching number of C. To complete the definition of branching number, it is convenient to put b(C) = 0 if C is the empty or the trivial code.

Definition 3.2.6 C is a pseudo-search code if C_a is branched for every $a \in Z$.

Definition 3.2.7 For C a pseudo-search code, we call those $a \in Z$ for which $b(C_a) \geq 1$ the branch points of C and $b(C_a)$ is the branching number of the branch point a.

Definition 3.2.8 A pseudo-search code is called regular of degree $q \geq 1$ if each $a \in Z$ such that $b(C_a) \geq 1$, $b(C_a) = q$.

It is convenient to say that the branching point a has $b(C_a)$ branches at the branch point a.

If a pseudo-search code C does not contain e, then C is a search code according to Rényi (1969) if $b(C_a) \geq 2$ for every branch point a of C. Also, Rényi defines a regular search code as a regular pseudo-search code of degree $q \geq 2$.

The following example may clarify the above definitions.

Example 3.2.3 Consider the codes:

$$C^1 = \{00, 01, 100, 101, 102, 11\}$$

$$C^2 = \{0, 11, 2\}$$

$$C^3 = \{0, 21\}$$

C^1 is a pseudo-search code. C^2 is branched but is not a pseudo-search code (since e.g. C^2_1 is not branched) and C^3 is not branched ($C^3_1 = \phi$ but $C^3_2 = \{1\}$)

Note that C^1 is the code representing the tree in Example 3.2.2 and Figure 3.2.2.

Given a rooted plane tree, it is easy to check that the set of all labels at the end vertices of the tree obtained from the lexicographic labelling is a pseudo-search code. Conversely, given a pseudo-search code, in order to show that there corresponds a unique isomorphism class of rooted plane trees with lexicographic labelling, we need to construct the sets defined below.

For C a pseudo-search code, i a non-negative integer, define C^i to be a code consisting of distinct codewords formed by the first i integers of each codeword with length greater than i contained in C, together with all those codewords in C having lengths less than or equal to i. Suppose the length of the longest codeword in C is h, then we construct C^1, C^2, ..., C^h. It is obvious that $C^h = C$.

The following lemmas concerning the sets C^1, C^2, ..., C^h are essential to the proof of our main results of this section. Verifications of the lemmas are elementary but lengthy. Therefore, we only outline the proofs.

Lemma 3.2.1 Let a be any codeword in C with length $\ell(a)$, then for any non-negative integer i such that $i \leq h$ and $i - \ell(a) \geq 1$, we have $(C^i)_a = (C_a)^{i-\ell(a)}$.

Proof: It can be shown from the definition that for each codeword $b \in (C^i)_a$, we have $b \in (C_a)^{i-\ell(a)}$. Thus, $(C^i)_a \subseteq (C_a)^{i-\ell(a)}$. Similarly we can also show that $(C_a)^{i-\ell(a)} \subseteq (C^i)_a$. This gives the result of the lemma.

Lemma 3.2.2 If C is a pseudo-search code and the length of the longest codeword in C is h, then C^1, C^2, ..., C^h are also pseudo-search codes.

Proof: For any fixed non-negative integer i such that $1 \leq i \leq h$, we obtain the following from Lemma 3.2.1:

$$(C^i)_a = \begin{cases} 0 & i < \ell(a) \\ \{e\} & i = \ell(a) \text{ and } C_a \neq \phi \\ \phi & i = \ell(a) \text{ and } C_a = \phi \\ (C_a)^{i-\ell(a)} & i > \ell(a) \end{cases}$$

Then we only need to show that in the non-trivial case, $(C_a)^{i-\ell(a)}$ is also a pseudo-search code. The verification is elementary.

For any pseudo-search code C, let $\mathbb{B}(C)$ be the set of all branch points of C. Then we have:

Lemma 3.2.3

(i) $\mathbb{B}(C^1) \subsetneq \mathbb{B}(C^2) \subsetneq \ldots \subsetneq \mathbb{B}(C^h) = \mathbb{B}(C)$.

If a is a branch point of C^i, then a is also a branch point of C^j with the same branch number for any $i < j$, $i,j = 1, \ldots, h$.

(ii) $\mathbb{B}(C^1) = \{e\}$

$$\mathbb{B}(C^k) = \bigcup_{i=1}^{k-1} C^i - C^k, \quad k = 2, 3, \ldots, h.$$

Proof: (i) For any $i < j$, $i,j = 1, 2, \ldots, h$, it follows from the definition that $C^i = (C^j)^i$. If a is a codeword of length $\ell(a) < i$, then Lemma 3.2.1 implies that

$$(C^i)_a = ((C^j)^i)_a = ((C^j)_a)^{i-\ell(a)}$$

Therefore, if a is a branch point of C^i, it must be a branch point of C^j with the same branch number.

The strict inclusions hold because of the fact that the longest codeword in C is of path length h.

(ii) It is easy to see that $\mathbb{B}(C^1) = \{e\}$ and $\mathbb{B}(C^k) \cup C^k = \bigcup_{i=1}^{k} C^i$ for $k = 1, \ldots, h$. But $\mathbb{B}(C^k) \cap C^k = \phi$, therefore, we conclude that

$$\mathbb{B}(C^k) = \bigcup_{i=1}^{k-1} C^i - C^k.$$

Theorem 3.2.1 Given a pseudo-search code C consisting of n codewords $a_1, a_2, \ldots, a_n$ arranged in lexicographic order, suppose $\max_{1\le i\le n} \{a_i\} = h$. There exists a rooted plane tree T with height h and with n end vertices such that the lexicographic labellings at the end vertices are $a_1, a_2, \ldots, a_n$ from the left to the right. Furthermore, the degree at the root of the tree T corresponds to the branching number of the pseudo-search code C. Any other branch vertices of degree q + 1 corresponds to a branch point with branch number q in the pseudo-search code C.

Proof: Construct the codes $C^1, C^2, \ldots, C^k$. Then they are pseudo-search codes by Lemma 3.2.2. If C^1 consists of q codewords, they must be 0, 1, ..., q-1 , then C^1 corresponds to a tree T^1 with height 1 and with q end vertices labelled from left to right as 0, 1, ..., q-1. Let m be any integer < h. Suppose there always corresponds a tree T^k to the pseudo-search code C^k, for any $k \le m < h$, then the tree T^{m+1} representing the code C^{m+1} can be obtained by adding the proper number of branches to the end vertices of the tree T_m which corresponds to a codeword in C^m but not in C^{m+1}. This is possible because of Lemma 3.2.3. Thus, the theorem is proved by induction.

Remark 3.2.1 Since we have already mentioned that the converse of Theorem 3.2.1 is also true, we can establish a one-to-one correspondence

between the set of all pseudo-search codes and the set of all isomorphism classes of rooted plane trees in the manner described in Theorem 3.2.1. This enables us to enumerate or construct trees by dealing analytically with the corresponding pseudo-search codes. The process of coding, that is, given a tree to determine the code; and decoding, that is, given a code to determine the tree (representing an isomorphism class) is relatively simple compared to Klarner's (1970) method of representing a planted plane tree by a sequence of integers or Chorneyko's and Mohanty's (1975) original method of representing a planted plane tree by a lattice path.

3.2.2 Matrix Representation

The definitions and results of this subsection apply to trees with vertices under either a linear ordering or a lexicographic ordering. Although different classes of trees are encountered when different orderings are specified in the matrix representation, we find that the one-to-one correspondence between these two classes naturally exists.

Definition 3.2.9 Given a planted plane tree with n end vertices and k branch vertices other than the root such that the i^{th} branch vertex is of degree $q_i + 1$, $i = 1, \ldots, k$, then the sequence $(q_1, \ldots, q_k)$ is called the degree sequence of the tree. Suppose there are x_i end vertices between the i^{th} and the $(i + 1)^{th}$ branch vertices, $i = 1, \ldots, k-1$, and that $x_k = q_k$, then the sequence $(x_1, \ldots, x_k)$ is called the end vertex sequence of the tree.

Let $Q_j = \sum_{i=1}^{j} q_i - j$ and $X_j = \sum_{i=1}^{j} x_i$ for $j = 1, \ldots, k$. The sequence $(Q_1, \ldots, Q_k)$ is called the cumulative degree sequence of the tree and the sequence $(X_1, \ldots, X_k)$ is called the cumulative end vertex sequence of the tree.

Remark 3.2.2 The following hold for any rooted plane tree:

(i) If $(Q_1, \ldots, Q_k)$ is the cumulative degree sequence, then $Q_i + 1$ denotes the total number of end vertices joined to any one of the first i branch vertices of the tree by an edge, $i = 1, \ldots, k$. On the other hand, if $(X_1, \ldots, X_k)$ is the cumulative end vertex sequence, then X_i denotes the total number of end vertices between the first and the $(i + 1)^{th}$ branch vertex of the tree, $i = 1, \ldots, k-1$ and X_k denotes the total number of end vertices of the tree.

(ii) It is obvious that a degree sequence (resp. end vertex sequence) of a tree is uniquely determined for a given cumulative degree sequence (resp. cumulative end vertex sequence), and the converse is also true.

Definition 3.2.10 Let $U = (U_1, \ldots, U_r)$ and $V = (V_1, \ldots, V_r)$ be two nondecreasing sequences of non-negative integers such that $U_i \geq V_i$, $i = 1, \ldots, r$. We say that the vector U dominates the vector V, or in other words, the vector V is dominated by the vector U.

Theorem 3.2.2 Let $(Q_1, \ldots, Q_k)$ be the cumulative degree sequence of a planted plane tree and $(X_1, \ldots, X_k)$ be the cumulative end vertex sequence of the same tree. Then we have $X_k = Q_k + 1$ and the vector $(Q_1, \ldots, Q_{k-1})$ dominates the vector $(X_1, \ldots, X_{k-1})$

Proof: By definition, $X_k = Q_k + 1$. The fact that $X_i \leq Q_i$ for $i = 1, \ldots, k-1$ follows from (1) of the remarks.

Theorem 3.2.3 Given any two nondecreasing sequences of non-negative integers $(Q_1, \ldots, Q_k)$ and $(X_1, \ldots, X_k)$ such that $X_k = Q_k + 1$ and $(Q_1, \ldots, Q_{k-1})$ dominates $(X_1, \ldots, X_{k-1})$, then there exists a unique

isomorphism class of planted trees with $(Q_1, \ldots, Q_k)$ as its cumulative degree sequence and $(X_1, \ldots, X_k)$ as its cumulative end vertex sequence.

Proof: Determine the vectors $(q_1, \ldots, q_k)$ and $(x_1, \ldots, x_k)$ by the systems of equations $\sum_{i=1}^{j} q_j - j = Q_j$ and $\sum_{i=1}^{j} x_i = X_j$ for $j = 1, \ldots, k$. Then we claim that $(q_1, \ldots, q_k)$ is the degree sequence and $(x_1, \ldots, x_k)$ is the end vertex sequence of a tree which represents a unique isomorphism class of rooted plane trees. This can be shown by drawing a tree with the i^{th} branch vertex of degree $q_i + 1$ and with x_i end vertices between the i^{th} and the $(i + 1)^{th}$ branch vertex, $i = 1, \ldots, k-1$ and with $Q_k + 1$ total number of end vertices. Such a tree can always be drawn because of the given conditions.

Remark 3.2.3 We conclude from the above two theorems that there is a one-to-one correspondence between the set of all isomorphism classes of rooted plane trees with given degree sequences and end vertex sequences, and the set of all matrices of the form

$$\begin{pmatrix} X_1 & \cdots & X_k \\ Q_1 & \cdots & Q_k \end{pmatrix}$$

where the vectors satisfy the conditions given in Theorem 3.2.2.

3.3 Enumeration Methods

Various techniques for counting labelled non-planar trees have been gathered by Moon (1970). However, not all the techniques can be used to enumerate unlabelled plane trees. The reason why labelling is unnecessary to be considered in the enumeration of the rooted plane trees is that the vertices of such trees have already been assigned an order relation. Therefore the ordering of the vertices can be taken as a natural labelling.

Klarner (1970) used generating functions to enumerate certain classes of planted plane trees with a fixed number of vertices and with specified degrees at the branch vertices. Chorneyko and Mohanty (1975) determined, via the enumeration of the lattice paths, the total number of planted plane trees with a specified number of branch vertices and some boundary conditions on the degree sequences and the end vertex sequences.

In this section, we first extend the generating function technique to the enumeration of certain classes of homeomorphically irreducible planted plane trees (planted plane trees with no branch vertex of degree 2) with a fixed number of end vertices. Further results are also obtained with additional restrictions on the degrees at the vertices or heights of the trees with a fixed number of end vertices. We then use the formulas of Chapter I to determine, via the enumeration of matrices, certain classes of k-tuples of planted plane trees with a specified number of branch vertices and some more generalized boundary conditions on the degree sequences and end vertex sequences compared to the results of Chorneyko and Mohanty (1975). It is also demonstrated that, under certain circumstances, the same results can be obtained by both methods.

3.3.1 Generating Function Techniques

A generating function is a function of the form $f(x) = \sum_{i=0}^{\infty} a_i x^i$

where $x^0 = 1$. The variable x is an indeterminate or a tag whose powers identify the coefficients which are numbers of various kinds of trees in question associated with the powers. The enumeration proceeds by finding relations for the generating functions and solving the function equations for the generating functions. A variety of techniques is required to solve these function equations so that an explicit solution or a recurrence relation for the coefficients of a generating function can be found.

Unless otherwise specified, the results in the following examples are original. They are given as a demonstration of the above mentioned techniques.

Example 3.3.1 Let t(n) be the total number of homeomorphically irreducible planted plane trees with n end vertices. Let the generating function be $T(x) = \sum_{n=1}^{\infty} t(n)\, x^n$. We can always combine k planted plane trees where $k \geq 2$, to form a new planted plane tree by joining them at their roots, and then adding an edge at the root. The new planted plane tree has a total number of end vertices equal to the sum of the end vertices at the k original trees. We therefore have the relation

$$T(x) = x + \sum_{i=2}^{\infty} (T(x))^i$$

where we note that t(1) = 1 and the term x arises for the case that the tree has only a single edge. Using Lagrange's formula which can be found in Chapter 5 of the book by Pólya and Szegő (1970)), we obtain

$$T(x) = \sum_{n=1}^{\infty} \frac{x^n}{n!} \left\{ \frac{d^{n-1}}{dT^{n-1}} \{\phi(T)\}^n \right\}_{T=0}$$

where $\phi(T) = \dfrac{1}{1 - \sum_{i=1}^{\infty} T^i}$. Consequently, by comparing the coefficients, we get

$$t(n) = \frac{1}{n!} \left\{ \frac{d^{n-1}}{dT^{n-1}} \{\phi(T)\}^n \right\}_{T=0}$$

In particular, $t(1) = 1$, $t(2) = 2$, $t(3) = 3$, $t(4) = 11$, and $t(5) = 45$, etc. Pictures of planted plane trees with n end vertices are given in Figure 3.3.1 for n = 1, 2, 3, and 4. For n = 5, only one tree is shown as a representation for all those trees which are the same as a non-plane tree. However, the total number of possible planted plane trees is illustrated in the bracket below each representation.

Example 3.3.2 Let $\bar{t}(n_i)$ be the total number of q-ary planted plane trees with k branch vertices and n_i end vertices where $n_i = qi - i + 1$. Let the generating function be $\phi(x) = \sum_{i=0}^{\infty} \bar{t}(n_i)\, x^i$. Then it is clear that $t(n_0) = 1$ because there is only one tree consisting of a single edge with no branch vertex at all.

If we combine q such trees at their root and add an edge to the root, then the resultant tree is again a q-ary planted plane tree with a total number of branch vertices equal to 1 plus the sum of the branch vertices at the q original trees. Therefore, we obtain the relation

$$\phi(x) = 1 + x(\phi(x))^q$$

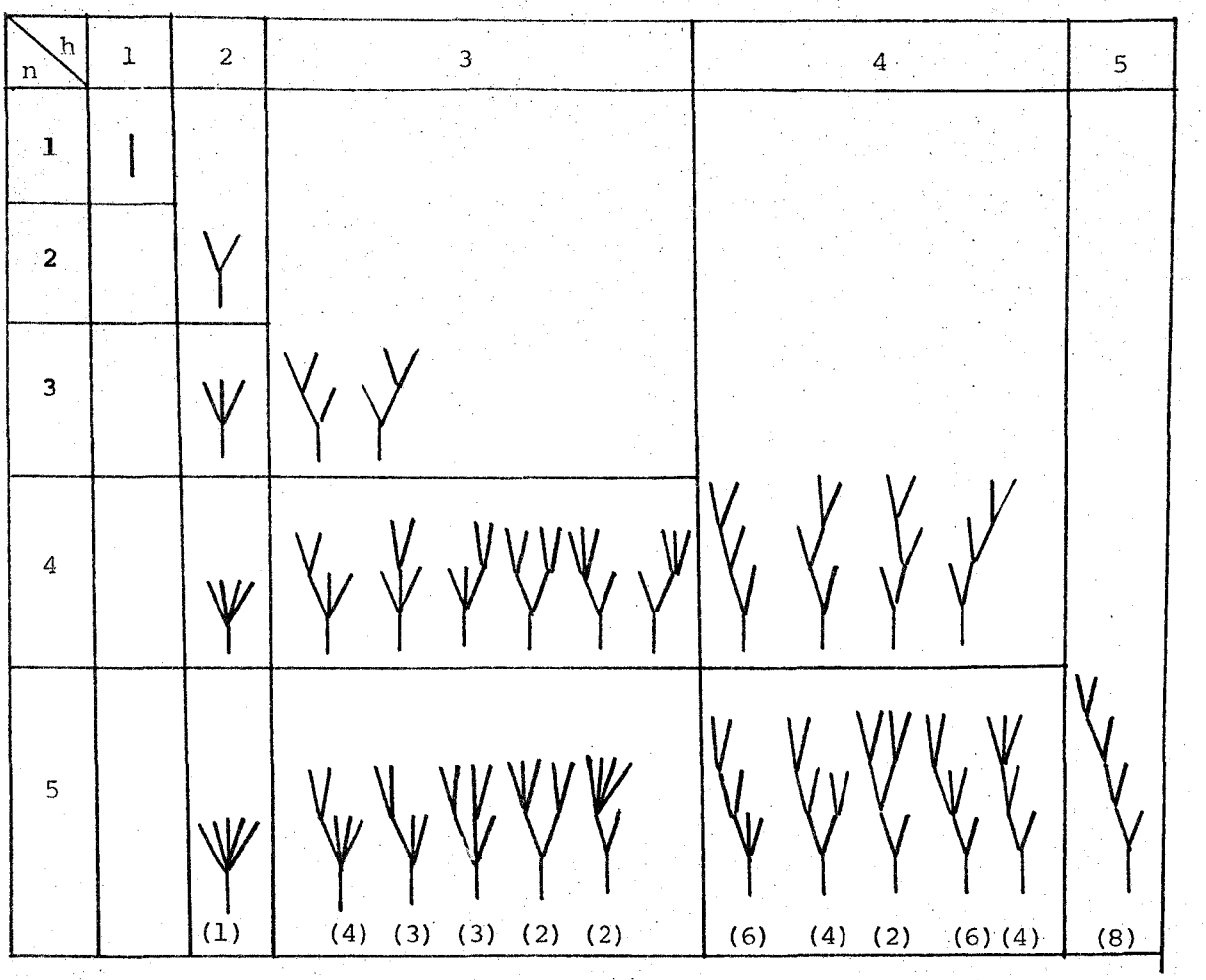


Figure 3.3.1 Planted plane trees with n end vertices and with height h. For n = 5, the numbers in the brackets below the trees denote the total number of planted plane trees of that kind which can be obtained by switching the branch vertices within the same level.

where the term 1 appears due to the trivial case when the tree consists of no branch vertex at all. Using the formula in Problem 211 (Chapter 5, Vol. Vol. 1 of the book by Pólya and Szegö (1970)), we get

$$\phi(x) = 1 + \sum_{i=1}^{\infty} \binom{qi}{i-1} \frac{x^i}{i}$$

By comparing the coefficients, we conclude that

$$\bar{t}\,(n_i) = \binom{qi}{i-1} \frac{1}{i} .$$

Note that $\bar{t}(n_i)$ also denotes the total number of q-ary planted plane trees with $n_i + i + 1 = qi + 2$ vertices. The same result has been obtained by Klarner (1970) who used a generating function of the form T(x) in Example 3.3.1. However, his function equation for T(x) is more difficult to solve than our relation for $\phi(x)$.

On the other hand, if we identify a q-ary planted plane tree with a pseudo-search code by using Theorem 3.2.1 and its converse, the same result has also been obtained by Chorneyko and Mohanty (1972) in terms of pseudo-search codes.

The similar result for labelled non-planar binary trees is due to Harding (1971).

Example 3.3.3 Let u(n) be the total number of homeomorphically irreducible planted plane trees with n end vertices and with branch vertices of degrees less than or equal to 4. Let $U(x) = \sum_{n=1}^{\infty} u(n)\, x^n$ be the generating function. Since we can only combine 2 or 3 such trees at their roots and add an edge at the root to form a new tree of the same type, we obtain

$$U(x) = x + (U(x))^2 + (U(x))^3$$

where the term x appears due to the trivial case that the tree consists of a single edge. To solve this function equation, we first find a transformation $U(x) = V(x) + c$ for some constant c such that the equation in terms of V(x) has the coefficient of the second degree term zero. By simplifying $V(x) + c = x + (V(x) + c)^2 + (V(x) + c)^3$ and setting the coefficient of the second degree term zero, we find that $c = -\frac{1}{3}$. Consequently, the equation becomes

$$(V(x))^3 - \frac{4}{3} V(x) + x + \frac{11}{27} = 0.$$

In order to reduce the power of V(x) in the equation, we move those terms with degree of V(x) less than 3 to the right hand side of the equality sign and take a ln transformation so that

$$3 \ln |V(x)| = \ln \left| \frac{4}{3} V(x) - x - \frac{11}{27} \right|.$$

Differentiation of both sides yields the result

$$24V(x)\, V'(x) - 11V'(x) - 27xV'(x) = -9V(x).$$

If we substitute V(x) by $U(x) + \frac{1}{3}$, then

$$8U(x)U'(x) - U'(x) - 9xU'(x) = -3U(x) - 1. \qquad (3.3.1)$$

Recall that $U(x) = \sum_{n=1}^{\infty} u(n)x^n$, hence $U'(x) = \sum_{i=1}^{\infty} nu(n)x^{n-1}$ and

$U(x)U'(x) = \sum_{n=1}^{\infty} \sum_{i=1}^{\infty} iu(i)u(n - i + 1)x^n$. Substituting the above terms into Equation (3.3.1), we have

$$8 \sum_{n=1}^{\infty} \sum_{i=1}^{\infty} iu(i)u(n-i+1)x^n - \sum_{n=1}^{\infty} nu(n)x^{n-1} - 9 \sum_{n=1}^{\infty} nu(n)x^n = -3 \sum_{n=1}^{\infty} u(n)x^n - 1.$$

Finally, the following recurrence relation is obtained by comparing the coefficients:

$$u(n+1) = \frac{1}{n+1}\left\{3(1-3n)u(n) + 8\sum_{i=1}^{n} iu(i)u(n-i+1)\right\}.$$

Since $u(1) = 1$, the above relation implies that $u(2) = 1$, $u(3) = 3$, $u(4) = 10$, $u(5) = 38$, $u(6) = 154$, and so on.

Generating functions can also be used to enumerate certain classes of planted plane trees with restrictions on the heights and the total number of end vertices. In the literature, Riordan (1960) enumerated unlabelled non-planar trees with given heights and total numbers of vertices. Gordon and Kennedy (1975) obtained recurrence formulas for counting unlabelled non-planar q-ary trees with given heights. In the following two examples, our approach is analogous to Riordan's (1960). The relations of the generating functions obtained seem to be relatively simple for planted plane trees. However, the solutions are still not in simple and explicit form.

Example 3.3.4 Let $s_h(x)$ denote the total number of homeomorphically irreducible planted plane trees with height less than or equal to h and with n end vertices. Let $S_h(x) = \sum_{n=1}^{\infty} s_h(n)x^n$ be the generating function. If we combine i trees of height less than or equal to h at their roots and add an edge at the root, we obtain a new tree of height less than or equal to h + 1 and with a total number of end vertices equal to the sum of the end vertices at these i original trees, for any i = 2, 3, ... Thus we have the relation

$$S_{h+1}(x) = x + \sum_{i=2}^{\infty} (S_h(x))^i$$

where the term x appears due to the trivial case when the tree consists of a single vertex.

Since $S_1(x) = \sum_{i=1}^{\infty} s_1(n)x^n$ and $s_1(1) = 1$, $s_1(2) = s_1(3) = \ldots = 0$,

therefore

$$S_1(x) = x$$

$$S_2(x) = x + \sum_{i=2}^{\infty} x^i = \frac{x}{1-x}$$

$$S_3(x) = x + \left(\frac{x}{1-x}\right)^2 \frac{1}{1 - \frac{x}{1-x}}$$

$$= x + x^2 + 3x^3 + 7x^4 + \ldots$$

and so on. We see for example, $s_3(4) = 7$ is the total number of homeomorphically irreducible planted plane trees with 4 end vertices and of height less than or equal to 3. These 7 trees can be found in Figure 3.3.1.

Example 3.3.5 Let $\bar{s}_h(n)$ denote the total number of q-ary planted plane trees with height less than or equal to h and consisting of i branch vertices and n_i end vertices where $n_i = ki - i + 1$. Let $\psi_h(x) = \sum_{i=0}^{\infty} \bar{s}_h(n_i)x^i$ be the generating function, where $x^0 = 1$ and $\bar{s}_h(n_0)$ denotes the trivial case when the tree consists of no branch vertex at all. Since we can only combine q-ary planted plane trees of height less than or equal to h at their root and then add an edge at the root to form a new q-ary planted plane tree of height less than or equal to h + 1, we obtain

$$\psi_{h+1}(x) = 1 + x(\psi_h(x))^q$$

where the term 1 appears due to the trivial case that the tree has only one edge.

Since $\psi_1(x) = \sum_{i=1}^{\infty} \bar{s}_1(n_i)x^i$ and we know that $\bar{s}_1(n_0) = 1$, $\bar{s}_1(n_1) = \bar{s}_1(n_2) = \ldots = 0$, therefore,

$$\psi_1(x) = 1$$

$$\psi_2(x) = 1 + x$$

$$\psi_3(x) = 1 + x(1 + x)^q$$

$$\psi_4(x) = 1 + x(1 + x(1 + x)^q)^q$$

$$\psi_5(x) = 1 + (1 + x(1 + (1 + x(1 + x)^q)^q)^q)^q$$

and so on. Thus, for example, if we assume that $q = 3$, then $s_1(n_0) = 1$ and $s_1(n_1) = s_1(n_2) = \ldots = 0$; $s_2(n_0) = s_2(n_1) = 1$ and $s_2(n_2) = s_2(n_3) = \ldots = 0$; $s_3(n_0) = s_3(n_1) = 1$, $s_3(n_2) = s_3(n_3) = 3$, $s_3(n_4) = 1$ and $s_3(n_5) = s_3(n_6) = \ldots = 0$. Here we find that there are three 3-ary trees with height less than or equal to 3, in fact, all three of them are of height 3 since $s_2(3) = 0$.

3.3.2 Matrix Enumeration Techniques

It has been shown in Subsection 3.2.2 that a planted plane tree T_i with m branch vertices can be represented by a unique matrix of integers of the form $\begin{pmatrix} X_{i1} & \cdots & X_{im} \\ Q_{i1} & \cdots & Q_{im} \end{pmatrix}$ such that $0 \leq Q_{i1} \leq \ldots \leq Q_{im}$ is the cumulative degree sequence and $0 \leq X_{i1} \leq \ldots \leq X_{im}$ where $X_{im} = Q_{im} + 1$ and $X_{ij} \leq Q_{ij}$ for $j = 1, \ldots, m-1$ is the **cumulative** end vertex sequence of the tree. Furthermore, the converse is also true (cf. Theorem 3.2.2 and Theorem 3.2.3). Let $T = (T_1, T_2, \ldots, T_i, \ldots, T_k)$ be a k-tuple of trees T_i, $i = 1,2,\ldots,i, \ldots, k$, where $k \geq 1$, defined as the above. Let $\{T\}$ be the set of all k-tuples of trees of the form T satisfying the following conditions:

$$\left.\begin{array}{lll}
\text{(a)} & Q_{ij} = a_j + \sum_{t=1}^{k-i} d_t & j = 1, \ldots, m \text{ and } i = 1, \ldots, k \\
\text{(b)} & c_i \leq X_{i1} \leq X_{i2} \leq \ldots \leq X_{im} & i = 1, \ldots, k \\
\text{(c)} & X_{ij} \leq X_{i+1,j} + d_i & i = 1, \ldots, k-1 \text{ and } j = 1, \ldots, m \\
\text{(d)} & b_j \leq X_{ij} & j = 1, \ldots, m \\
\text{(e)} & X_{kj} \leq a_j & j = 1, \ldots, m
\end{array}\right\} \quad (3.3.1)$$

where $a_1, \ldots, a_m$; $b_1, \ldots, b_m$; $c_1, \ldots, c_k$ and $d_1, \ldots, d_{k-1}$ are non-netative integers such that $a_1 \leq \ldots \leq a_m$, $b_1 \leq \ldots \leq b_m$ and $c_1 \leq \ldots \leq c_k$.

We define $A = (a_1, \ldots, a_{m-1})$, $B = (b_1, \ldots, b_{m-1})$, $C = (c_1, \ldots, c_k)$ and $D = (d_1, \ldots, d_{k-1})$. Then Theorem 3.2.2 and Theorem 3.2.3 imply that the cardinality of $\{T\}$ equals to $N(k,m-1,A,B,C,D)$ which **has** been defined and evaluated in Section 1.2 (cf. Theorem 1.2.1).

Similarly, if $b_1 \geq \sum_{t=1}^{k-1} d_t$, where $d_0 = 0$, we let $\{T'\}$ be the set of all k-tuples of trees of the form $T' = (T_1, T_2, \ldots, T_i, \ldots, T_k)$, where T_i, $i = 1, \ldots, k$ are the trees defined at the beginning of this subsection which satisfy the conditions (c), (d), (e) of (3.3.1) and the following:

$$(a') \quad Q_i = \min \left\{ a_j + \sum_{t=1}^{k-i} d_t, h_i \right\} \qquad j = 1, \ldots, m \text{ and } i = 1, \ldots, k$$

$$(b') \quad 0 \leq x_{i1} \leq \ldots \leq x_{im} \leq h_i, \qquad i = 1, \ldots, k$$

where $0 \leq h_1 \leq \ldots \leq h_k$ are non-negative integers. Then the cardinality of the set $\{T'\}$ equals to the number NB(k, m-1,A,B,H,D) which has been defined and evaluated in Section 1.2 (cf. Theorem 1.2.2).

Note that for the special case that k = 1 and C dominated by B (resp. A dominated by H) the determinant representing N(k, m-1,A,B,C,D) (resp. NB(k, m-1,A,B,H,D)) can be reduced to the determinant given in Corollary 1.2.1. This becomes the result which has been given by Chorneyko and Mohanty (1972, 1975) for both pseudo-search codes and planted plane trees of this kind. This is also a generalization of Klarner's (1970) result for binary trees with a given number of vertices (cf. Example 3.3.2 of Subsection 3.3.1).

3.4 Optimal Alphabetic q-ary Trees

Let T be a planted plane tree with n end vertices. Let $V=\{v_1,\ldots,v_n\}$ be the set of all end vertices of T. Let ℓ_i be the path length of the end vertex v_i and w_i be a non-negative weight associated to the end vertex v_i, $i = 1, \ldots, n$. We say that the tree T is a weighted tree. The cost of the tree T is defined to be the weighted path length of its end vertices, that is, $\sum_{i=1}^{n} \ell_i w_i$. A q-ary tree achieving the minimal cost for a given set of weights $W = \{w_1, \ldots, w_n\}$ is called an optimal q-ary tree for W. An optimal alphabetic q-ary tree for any given sequence of weights $\vec{W} = \{w_1, \ldots, w_n\}$ is defined to be a weighted q-ary tree T which achieves the minimal cost for the given sequence of weights $\vec{W}$ under the restriction that the left to right sequence of end vertices must follow the order $v_1, v_2, \ldots, v_n$ and with the weight w_i associated to the end vertex v_i, $i = 1, \ldots, n$. It is clear that the cost of an optimal alphabetic q-ary tree for a given sequence of weights is always greater than or equal to the cost of an optimal q-ary tree for the same set of weights.

For applications to the k-sample group testing problem of Chapter IV, we are required to construct an optimal alphabetic q-ary tree, where $q = k+1$ in this case, for a valley sequence of weights $\vec{W} = \{w_1, \ldots, w_n\}$, that is, for any w_i, $2 \le i \le n-1$, $w_i \le \max\{\min_{i<j} w_i, \min_{j<k} w_k\}$

Constructions of optimal binary trees and optimal alphabetic binary trees are well known. Huffman's (1952) algorithm for constructing an optimal binary tree was first given for coding purposes. Schwartz and Kallick (1964) provided a computer algorithm which transforms an optimal binary tree obtained from Huffman's algorithm to an optimal alphabetic binary tree of

the same cost for a monotone sequence of weights. Hu and Tucker (1971) proposed a T-C algorithm for constructing binary trees which can be converted into an optimal alphabetic q-ary tree for a valley sequence of weights. The contents of the above mentioned algorithms will be studied later in Subsection 3.4.2.

In this section, we first mention briefly the basic properties of a weighted q-ary tree observed by Huffman (1952) and Knuth (1968, 1971). Then we summarize how Huffman's algorithm can be generalized for constructing optimal q-ary trees and the T-C algorithm by Hu and Tucker (1971) can be generalized for constructing optimal q-ary trees which can be converted into optimal alphabetic q-ary trees for valley sequences of weights. The process of generalization is mainly straightforward since the principles of these original algorithms do not depend on the assumption that the trees have to be binary. In addition, we provide a computation algorithm based on pseudo-search code construction which enables us to use a computer to convert an optimal q-ary tree constructed by using the generalized T-C algorithm to an optimal alphabetic q-ary tree for a valley sequence of weights without increasing the cost. This can be considered as a non-trivial extension of the algorithm by Schwartz and Kallick (1964). Finally, the upper and lower bounds of the cost of an optimal q-ary tree are found as functions of the entropy. When $q = 2$, these bounds have been found by Hwang (1974).

3.4.1 Basic Properties of the Weighted Trees

Consider any weighted tree T with n end vertices, namely, v_1, v_2, ..., v_n, from the left to the right. Let w_i be the non-negative weight associated to the end vertex v_i, $i = 1, 2, \ldots, n$. To every branch vertex

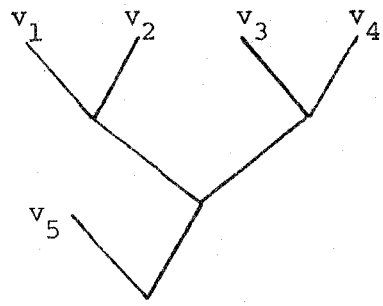
of the tree T starting from those with the longest path length to those with the shortest path length, assign a weight which equals to the sum of the weights at the vertices adjacent to it but with a longer path length. An example of such a weighted tree is given below to clarify the definitions.

Example 3.4.1 Let $W = \{w_1, w_2, w_3, w_4, w_5\}$ where $w_1 = 5$, $w_2 = 1$, $w_3 = 2$, $w_4 = 1$, $w_5 = 1$. If we assign the weight w_i to the vertex v_i of the tree in Figure 3.4.1(a), the resultant tree in Figure 3.4.1(b) is a weighted tree for W. In fact, this is also an optimal tree for W. We also say that the tree is an optimal alphabetic binary tree for the sequence of weights $\vec{W} = \{w_1, w_2, w_3, w_4, w_5\}$. However if we let $\vec{W}' = \{w_1, w_2, w_3, w_4, w_5\}$, the tree is not an optimal alphabetic tree for $\vec{W}'$. Weights can also be assigned to the branch vertices of the tree according to the previous paragraph so that the tree in Figure 3.4.1(c) is obtained. The cost of the tree is equal to:

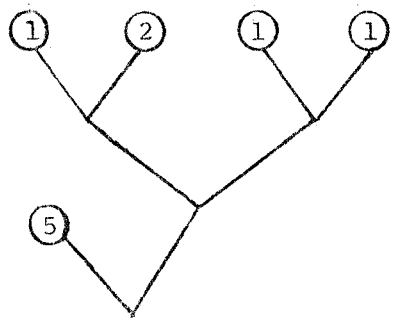
$$\sum_{i=1}^{5} \ell_i w_i = 1x5 + 3x1 + 3x2 + 3x1 + 3x1 = 20.$$

It is interesting to note that the sum of the weights at the branch vertices is also 20. This is explained in the following lemmas which were given as an exercise in the book by Knuth (1968). They can be easily proved by using induction on the total number of branch vertices of a tree.

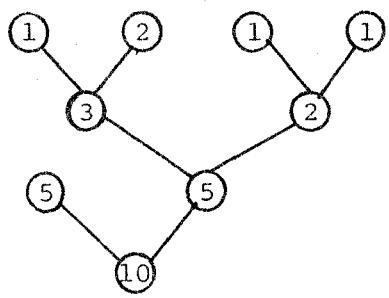
Lemma 3.4.1 The weight associated to any branch vertex v of a weighted tree T equals to the sum of the weights at the end vertices of the subtree of T rooted at v, that is, those end vertices of T that can be reached by a path from the root of T through the vertex v.



(a) A binary tree with n end vertices v_1, v_2, ..., v_5.



(b) A weighted tree for the set of weights
W = {5, 1, 2, 1, 1}



(c) A weighted tree for W with branch vertices labelled by the associated weights.

Figure 3.4.1 The assigment of weights to a binary tree with n end vertices.

Proof: It is obviously true for weighted trees with only one branch vertex. Suppose it is true for all trees with number of branch vertices less than or equal to k. For a weighted tree with k+1 branch vertices, the lemma holds for all the branch vertices except the root because of the induction hypothesis. However, the weight at the root is defined to be the sum of the weights at the branch vertices adjacent to it. Thus, the weight at the root equals to the sum of the weights at all the end vertices of the tree. Therefore, the lemma is also true for trees with k+1 branch vertices. This completes the proof by induction.

Lemma 3.4.2 The weighted path length of a weighted tree T equals to the sum of all the weights at the branch vertices of T.

Proof: It is obviously true for trees with only one branch vertex. Suppose it is true for all trees with total number of branch vertices less than or equal to k. For a tree T' with k+1 branch vertices, let there be m planted subtrees rooted at the root of T'. From the left to the right, let the i^{th} subtree be T_i' which has n_i end vertices with weights w_{i1}, w_{i2}, ..., w_{in_i} and path lengths ℓ_{i1}, ℓ_{i2}, ..., ℓ_{in_i} respectively, for i=1,...,m.

Then the weighted path length of T' is

$$\sum_{i=1}^{m}\sum_{j=1}^{n_i}(\ell_{ij}+1)w_{ij} = \sum_{i=1}^{m}\sum_{j=1}^{n_i}\ell_{ij}w_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{n_i}w_{ij}$$

Since the numbers of branch vertices at the subtrees T_1', ..., T_m' are less than k, by induction hypothesis, the first part of the expression equals the sum of the weights at all the branch vertices of the subtrees. But, the second part of the expression is known as the weight associated to the

root from Lemma 3.4.1. Hence, the lemma is also true for trees with k+1 branch vertices. This completes the proof by induction.

Before we study the properties of an optimal tree for a given set of weights $W = \{w_1, \ldots, w_n\}$, we note that such a tree always exists because there are only finitely many distinct planted plane trees with n end vertices (cf. Example 3.3.1). Note that if we want to find an optimal tree for W and allow the degrees at the branch vertices to be any number from $1, \ldots, q$, $1 \leq q \leq n$, then the tree with the most branch vertices of degree q is always the optimal one. Therefore, it suffices to consider optimal q-ary trees of W, provided that $n = 1 \bmod(q-1)$. The condition on n can always be fulfilled by adding some extra zero weights to the set W.

The following result for optimal q-ary trees is a generalization of the result on optimal binary trees by Huffman (1952).

Theorem 3.4.1 Let v_i and v_j be any two given vertices (each of which may be either a branch vertex or an end vertex, but they cannot be joined by the same path through the root) of an optimal q-ary tree with path lengths ℓ_i, ℓ_j and weights w_i, w_j respectively. Then the following hold true:

(i) $\ell_i > \ell_j$ implies $w_i \leq w_j$

(ii) there exists an optimal q-ary tree with the property that $w_i \leq w_j$ implies $\ell_i \geq \ell_j$.

Proof: Let T_o be a given optimal q-ary tree. Let v_i and v_j be the given vertices of T_o. Let T_1 and T_2 be the two q-ary subtrees rooted at v_1 and v_2 respectively. In the case when v_1 or v_2 or both are end vertices,

the corresponding subtrees become the vertices themselves. Let T_3 be the subtree rooted at the root of T_o such that the end vertices of T_3 include all those end vertices of T_o which do not belong to either T_1 or T_2 plus the vertices v_i and v_j. From Lemma 3.4.2, we know that the cost of T_o equals to the sum of the costs of T_1, T_2, and T_3. Now if the locations of T_1 and T_2 are interchanged such that v_i takes the place of v_j and vice versa, then the resultant tree T'_o is formed by the subtrees T_1, T_2, and T'_3, where T'_3 is the same as T_3 except that the end vertices v_i and v_j of T_3 are interchanged.

(i) Given that $\ell_i > \ell_j$, suppose it was true that $w_i > w_j$, then the cost of T'_3 is less than the cost of T_3. It follows from Lemma 3.4.2 that the cost of T'_o is less than the cost of T_o. This contradicts the optimality of T_o. Hence $w_i \leq w_j$.

(ii) Given that $w_j \geq w_i$, suppose that $\ell_i < \ell_j$, then the cost of T'_3 is no more than the cost of T_3. Therefore, T'_o is the optimal q-ary tree satisfying the required condition.

3.4.2 Algorithms of Construction

The constructions of an optimal q-ary tree for an arbitrary set of positive weights and an optimal alphabetic q-ary tree for a valley sequence of weights are studied in this subsection. They are the generalizations of Huffman's (1952) algorithm, Hu's and Tucker's (1971) T-C algorithm, and Schwartz's and Kallick's (1964) algorithm.

Generalized Huffman's Algorithm: Given a set of n positive integers $W = \{w_1, w_2, \ldots, w_n\}$, where $n = 1 \bmod (q-1)$ and $n > k$, an optimal q-ary

tree for W can be constructed as the following. The n weights are said to be the end vertices of the tree. The q end vertices with smallest weights, say $w_{1'}, w_{2'}, \ldots, w_{q'}$, are combined to form a new vertex of weight $w_{1'} + w_{2'} + \ldots + w_{q'}$, being the father of the q vertices. We say that a branch vertex of weight $w_{1'} + w_{2'} + \ldots + w_{q'}$ is created. Repeat the same procedure to the remaining n-q end vertices and the new branch vertex. The procedure continues until all the vertices are combined, that is, a branch vertex of weight $w_1 + w_2 + \ldots + w_n$ is created. This is possible because $n = 1 \bmod(q-1)$ and $n > k$. The resultant tree is called a Huffman's q-ary tree for W. An example will be given at the end of this subsection.

Theorem 3.4.2 A Huffman's q-ary tree for a set of weights $W = \{w_1, w_2, \ldots, w_n\}$ is an optimal q-ary tree for W.

Proof: The proof for q = 2 can be found on p. 403, Vol. 1 of the book by Knuth (1968) or the paper by Hu and Tucker (1971). The proof in general is a straightforward extension of the special case that q = 2. This can be shown by induction on the total number of branch vertices of the tree. Use the fact that if T_o is a q-ary weighted tree for W with the q smallest weights, say $w_{1'}, w_{2'}, \ldots, w_{q'}$, combined to form a branch vertex of the tree, then T_o is an optimal q-ary tree for W if and only if the subtree with the q end vertices $w_{1'}, w_{2'}, \ldots, w_{q'}$ excluded is also an optimal q-ary tree for the remaining n-q weights and the weight $w_{1'} + \ldots + w_{q'}$.

In order to obtain an optimal alphabetic q-ary tree for a given sequence of weights we have to use a more restrictive construction method which is a generalization of the T-C algorithm given by Hu and Tucker (1971) for binary trees.

The Generalized T-C Algorithm: Let $\vec{W} = \{w_1, w_2, \ldots, w_n\}$ be a sequence of weights arranged in required order. We call this an initial construction sequence of end vertices. Whenever q of the vertices are combined to form a new vertex, say w, having a weight which equals to the sum of the q weights, we say that a branch vertex w is created. The new construction sequence becomes the one without these q vertices but with w taking the place of the leftmost one of these q vertices. Therefore, in general, a construction sequence may contain end vertices or branch vertices or both.

Two vertices in a construction sequence are called tentative-connecting, abbreviated as T-C, if the sequence of vertices between them is either empty or consists of entirely branch vertices. Any q vertices in a construction sequence is called a T-C q-tuple if any two of its members are T-C whenever the sequence of vertices between them does not contain any one of its own members.

The generalized T-C algorithm asserts that a T-C q-tuple of minimum sum of weights should be combined in each construction sequence. In the case of a tie, combine the leftmost T-C q-tuple of minimum sum of weights. The procedure terminates when all the vertices are combined, that is $(n-1)/(q-1)$ branch vertices are created. The resultant tree is called the q-ary weighted tree for $\vec{W}$ constructed by using the generalized T-C algorithm. An example will be given at the end of this subsection.

Lemma 3.4.3 Let $\vec{W} = \{w_1, \ldots, w_n\}$ be a valley sequence of weights. A weighted q-ary tree for $\vec{W}$ constructed by using the generalized T-C algorithm is a Huffman's q-ary tree.

Proof: The proof for q = 2 can be found in the paper by Hu (1973). The proof in general is similar and based on the fact that in every step of the construction using the generalized T-C algorithm, a T-C q-tuple of minimum sum of weights always consists of the q smallest weights. This is true because the given sequence of weights is a valley sequence.

Theorem 3.4.3 Let $\vec{W} = \{w_1, \ldots, w_n\}$ be a valley sequence of weights. A q-ary weighted tree T_c constructed by using the generalized T-C algorithm can be converted into an optimal alphabetic q-ary tree for $\vec{W}$ with the same cost. Furthermore, T_c is also an optimal q-ary tree for $\vec{W}$.

Proof: The proof for q = 2 is due to Hu (1973). The proof in general can be obtained by considering the minimum sum T-C q-tuples instead of the minimum sum T-C pairs in Hu's (1973) proof. Because $\vec{W}$ is a valley sequence, we can always reassign the weights of the end vertices at the same level of T_c to obtain a tree with the weights at the end vertices arranged from left to right in the required order without increasing the cost. The resultant tree is optimal because of Lemma 3.4.3 and Theorem 3.4.2.

Before we introduce pseudo-search code construction algorithm, we define and study the following terms which are due to Hu (1973) in the binary case.

Definition 3.4.1 A sequence of n positive integers $A = \{a_1, a_2, \ldots, a_n\}$ is called a q-ary feasible sequence if there exists a q-ary tree with n end vertices having path lengths corresponding to the integers from the left to the right.

In addition, suppose $\max\{a_1, a_2, \ldots, a_n\} = h$, we say that A is a q-ary feasible sequence of height h.

The following lemma is a straightforward extension of Hu's (1973) result.

Lemma 3.4.4 A finite sequence of positive integers is a q-ary feasible sequence if and only if the following conditions are satisfied.

(i) If the largest integer in the sequence is h, then the number of h's in the sequence must be a multiple of q and such h's always occur in consecutive sets of length q.

(ii) If we form a reduced sequence from the original sequence by successively replacing (from left to right) every q consecutive h's by one occurrence of the integer h - 1, then the reduced sequence again satisfies (i).

(iii) If the process of (ii) is replaced by considering the reduced sequence as the original sequence, (i) is still satisfied until a reduced sequence of q 1's is found.

Proof: The proof is a straightforward extension of the one given by Hu and Tucker (1971) for q = 2. It can be verified from the definition.

Theorem 3.4.4 Given a q-ary feasible sequence of height h, a q-ary rooted plane tree of height h is uniquely determined.

Proof: Based on the properties of a q-ary feasible sequence of height h stated in Lemma 3.4.4, it can be shown by induction on the height h.

Given a feasible sequence $(a_1, a_2, \ldots, a_n)$, Theorem 3.4.4 implies that a unique rooted q-ary tree can be determined. Therefore, by Theorem 3.2.1, we know that a unique pseudo-search code which corresponds to the rooted q-ary tree can also be determined. Such a pseudo-search code can be constructed by using the following algorithm which can be considered as a

non-trivial extension of the algorithm by Schwartz and Kallick (1964).

Pseudo-search Code Construction Algorithm:

(i) the first codeword is formed by a_1 0's.

(ii) the i^{th} codeword is formed by a q-ary addition* of 1 to the $(i-1)^{th}$ codeword (in lexicographic order) and affixing or removing zeros at the end so that the resultant codeword is of length a_i, $i = 2, 3, \ldots, n$.

*By a q-ary addition, we mean an operation (+) defined on the set of integers $\{0, 1, \ldots, q-1\}$ such that

$$r(+)s = \begin{cases} r + s & \text{if } r + s < q \\ 10 & \text{if } r + s = q \\ 1a & \text{if } r + s > q \text{ and } a = (r+s) \bmod q,\ a < q \end{cases}$$

Theorem 3.4.5 Given a q-ary feasible sequence, the code obtained by using pseudo-search code construction algorithm is a pseudo-search code.

Proof: Since q-ary addition is employed in the construction, the resultant code is branched with branching number q at every branch point. Since the existence of the corresponding q-ary tree is known due to Theorem 3.4.4, it is always possible for us to affix or remove zeros at the end of the i^{th} codeword obtained in step (ii) of the algorithm, so that the resultant one is of desired length representing the i^{th} end vertex of the q-ary tree, $i = 2, \ldots, n$. Therefore, the code is a pseudo-search code.

Remark 3.4.1 Let $\vec{W} = \{w_1, \ldots, w_n\}$ be a valley sequence of weights. Theorem 3.4.4 asserts that the q-ary weighted tree for $\vec{W}$ constructed by using the generalized T-C algorithm can be converted into an optimal

alphabetic q-ary tree without changing the corresponding path lengths of the weights. However, the process of converting was carried out graphically from level to level. Here we outline how it can be done analytically by employing pseudo-search code construction algorithm.

(1) Let ℓ_i be the path length of the weight w_i at the q-ary weighted tree constructed by using the generalized T-C algorithm. By Theorem 3.4.3, $(\ell_1, \ell_2, \ldots, \ell_n)$ must be a feasible q-ary sequence representing the left to right path lengths of the end vertices of an optimal alphabetic q-ary tree for W.

(2) Theorem 3.4.4 and Theorem 3.4.5 imply that the optimal alphabetic q-ary tree for $\vec{W}$ can be obtained in the form of its corresponding pseudo-search code by using pseudo-search code construction algorithm.

The following example may clarify the concept.

Example 3.4.1 Given that $\vec{W} = \{w_1, w_2, \ldots, w_n\}$, where $w_1 = 25$, $w_2 = 6$, $w_3 = 4$, $w_4 = 1$, $w_5 = 1$, $w_6 = 1$, $w_7 = 1$, $w_8 = 2$, $w_9 = 2$, $w_{10} = 2$, $w_{11} = 5$, then we know by definition that W is a valley sequence. We illustrate the construction of an optimal alphabetic q-ary tree for $\vec{W}$ in the following, where we let $q = 3$.

(a) A Huffman's 3-ary tree for $\vec{W}$ can be constructed by using the generalized T-C algorithm (cf. Lemma 3.4.3). The resultant tree is shown in (a) of Figure 3.4.1.

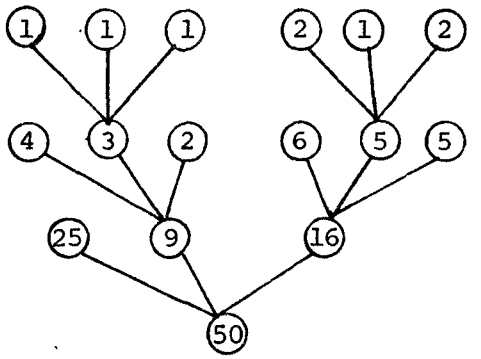
(b) An optimal alphabetic 3-ary tree for $\vec{W}$ converted from the tree in (a) by the procedure described in the proof of Theorem 3.4.3 is shown in (b) of Figure 3.4.1.

(c) From (a), we know that $\ell_1 = 1$, $\ell_2 = 2$, $\ell_3 = 2$, $\ell_4 = 3$, $\ell_5 = 3$, $\ell_6 = 3$, $\ell_7 = 3$, $\ell_8 = 3$, $\ell_9 = 3$, $\ell_{10} = 2$, $\ell_{11} = 2$. Pseudo-search code

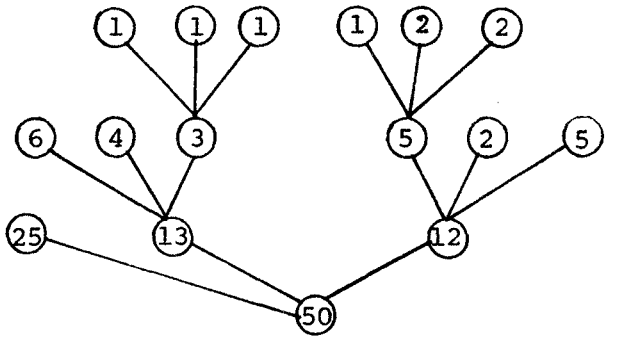
algorithm can be employed to determine the corresponding pseudo-search code representing the optimal alphabetic 3-ary tree for W instead of finding the tree in (b) directly.

codeword number (in lexicographic order)	length	3-ary addition	resultant codeword
i	ℓ_i	(+)	a_i
1	1	0	0
2	2	0+1=1	10
3	2	10+1=11	11
4	3	11+1=12	120
5	3	120+1=121	121
6	3	121+1=122	122
7	3	122+1=200	200
8	3	200+1=201	201
9	3	201+1=202	202
10	2	202+1=210	210
11	2	21+1=22	22

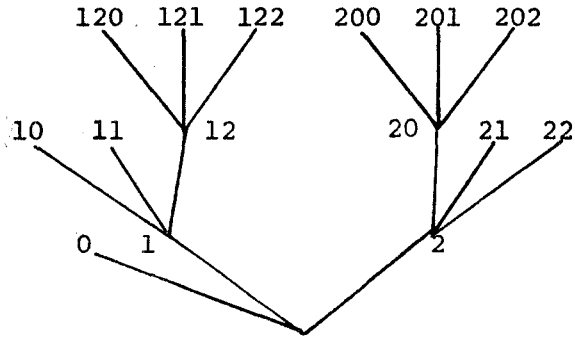
The code $A = \{a_1, a_2, \ldots, a_{10}\}$ is a pseudo-search code due to Theorem 3.4.5. The tree corresponding to A is shown in (c) of Figure 3.4.1. This tree is in the same isomorphism class as the tree in (b).



(a) An optimal (Huffman's) 3-ary tree for the set of weights W ={ 25,6,4,1,1,1,2,1,2,2,5 } constructed by using the generalized T-C algorithm.



(b) An optimal alphabetic 3-ary tree the the sequence of weights $\vec{W}$ = { 25,6,4,1,1,1,2,1,2,2,5} constructed by reassigning the weights at the vertices of same path length on the tree obtained from (b).



(c) The tree corresponding to the pseudo-search code A = { 0,10,11 120,121,122,200,201,202,21,22 }. This tree is isomorphic to the tree obtained from (b) but it is constructed by using the pseudo-search code algorithm.

Figure 3.4.2 Illustrations of the construction procedures.

3.4.3 Entropy Bounds for the Costs of the Optimal q-ary Trees

In this section, we would like to find the upper and lower bounds of an optimal q-ary tree in terms of the functions of entropy. Before we can apply theorems in information theory and coding, we study the following definition and a necessary and sufficient condition for a sequence of n positive integers to be the path lengths of a q-ary tree with n end vertices.

Definition 3.4.2 A full q-ary tree of height h is a q-ary tree with a number of q^h end vertices of path length h.

The following theorem is a refinement of Kraft's inequality (cf. Ash (1965) or Abramson (1963)).

Theorem 3.4.6 A necessary and sufficient condition for a sequence of n integers $\ell_1, \ell_2, \ldots, \ell_n$ to be the path lengths of the end vertices of a q-ary tree with n end vertices is that

$$\sum_{i=1}^{n} q^{-\ell_i} = 1.$$

Proof: Observe that any q-ary tree T which has height h and n end vertices can be obtained by excluding some vertices from a full q-ary tree T_F of height h. Thus, a vertex v of T_F is an end vertex of T if and only if all the vertices on the subtree rooted at v are excluded except the root. As a result, every end vertex of T_F is either left as an end vertex of T or excluded. Therefore, we have

$$\sum_{i=1}^{n} a^{h-\ell} = q^h$$

where by definition, $h = \max_{1 \le i \le n} \{\ell_1, \ell_2, \ldots, \ell_n\}$. Dividing both sides of the

above equality by q^h gives the condition of the theorem.

Conversely, given that $\sum_{i=1}^{n} a^{-\ell_i} = 1$, a q-ary tree with n end vertices of path lengths $\ell_1, \ell_2, \ldots, \ell_n$ can be obtained by reversing the arguments we have already used. (The details can be found as an analogue to the proof of Kraft's inequality for binary codes, cf. p. 59 of Abramson's (1963) book or p. 34 of Ash's (1965) book).

Similarly, we can prove the following theorem referred to as Kraft's inequality in terms of trees.

Theorem 3.4.7 The necessary and sufficient condition for a sequence of n positive integers $\ell_1, \ell_2, \ldots, \ell_n$ to be the path lengths of a q-ary tree with n end vertices and with branch vertices of degrees $\leq$ q+1 is that

$$\sum_{i=1}^{n} q^{-\ell_i} \leq 1.$$

The following is a very useful lemma in information theory.

Lemma 3.4.5 Let $p_1, p_2, \ldots, p_M$ and $q_1, q_2, \ldots, q_M$ be arbitrary positive numbers with $\sum_{i=1}^{M} p_i = \sum_{i=1}^{M} q_i = 1$. Then $-\sum_{i=1}^{M} p_i \log_e p_i \leq -\sum_{i=1}^{M} p_i \log_e q_i$ with equality if and only if $p_i = q_i$ for all i.

Proof: See p. 16 of Abramson's (1963) book or p. 16 of Ash's (1965) book for details. We sketch the proof in the following.

For any real number $x \geq 1$, we have

$$\int_1^x \left(1 - \frac{1}{x}\right) dx = x - 1 - \log x \geq 0$$

and for $x \leq 1$, we have

$$\int_x^1 \left(\frac{1}{x} - 1 \right) dx = -\log x - 1 + x \geq 0.$$

Therefore, $x - 1 \geq \log x$ with equality if and only if $x = 1$. Let $x = q_i/p_i$, we have

$$\sum_{i=1}^{M} p_i \log(q_i/p_i) \leq \sum_{i=1}^{M} p_i \frac{q_i}{p_i} - 1 = 0$$

with equality if and only if $q_i = p_i$ for all i. This proves the lemma.

We can now prove our main results of this subsection. They are known for binary codes (cf. Ash (1965) and Abramson (1963)). Here we state and prove the theorems for q-ary weighted trees.

Theorem 3.4.8 Let $p_1, p_2, \ldots, p_n$ be arbitrary positive numbers such that $\sum_{i=1}^{n} p_i = 1$. Let $\ell_1, \ell_2, \ldots, \ell_n$ be the path lengths of the end vertices of a tree with branch vertices of degree less than or equal to $q + 1$ and with n end vertices.

Let p_i be the weight associated with the end vertex of path length ℓ_i, $i = 1, \ldots, n$. Then we have

$$\sum_{i=1}^{n} p_i \ell_i \geq - \sum_{i=1}^{n} p_i \log_q p_i$$

with equality if and only if $p_i = q^{-\ell_i} / \sum_{i=1}^{n} q^{-\ell_i}$ for all i.

Proof: Let $q_i = q^{-\ell_i} / \sum_{i=1}^{n} q^{-\ell_i}$. From Lemma 3.4.5, we have

$$- \sum_{i=1}^{n} p_i \log p_i \leq - \sum_{i=1}^{n} p_i \log(q^{-\ell_i} / \sum_{i=1}^{n} q^{-\ell_i})$$

$$= \sum_{i=1}^{n} p_i \ell_i \log q + \sum_{i=1}^{n} p_i \log(\sum_{i=1}^{n} q^{-\ell_i})$$

$$\leq \sum_{i=1}^{n} p_i \ell_i \log q$$

since $\log(\sum_{i=1}^{n} q^{-\ell_i}) \leq 0$. The equality holds if and only if $p_i = q^{-\ell_i} / \sum_{i=1}^{n} q^{-\ell_i}$ for all i.

Theorem 3.4.9 Given that $p_1, p_2, \ldots, p_n$ are n positive numbers such that $\sum_{i=1}^{n} p_i = 1$, there exists a tree which has its branch vertices of degrees less than or equal to q + 1 and n end vertices of path lengths $\ell_1, \ell_2, \ldots, \ell_n$ such that $- \sum_{i=1}^{n} p_i \log_q p_i \leq \sum_{i=1}^{n} p_i \ell_i \leq 1 - \sum_{i=1}^{n} p_i \log_q p_i$.

Proof: (cf. p. 38 of Ash's (1965) book). Select ℓ_i such that $-\log_q p_i \leq \ell_i \leq -\log_q p_i + 1$ for all i = 1, ..., n, then the required condition is satisfied. Observe that this is possible because $-\log_q p_i \leq \ell_i$ implies that $-\log_q p_i \leq \ell_i \log q$ and $p_i \geq q^{-\ell_i}$.

Therefore, $\sum_{i=1}^{n} q^{-\ell_i} \leq 1$ and Theorem 3.4.8 asserts that the required tree exists.

We note that the tree in Theorem 3.4.9 can always be considered as a qpary tree by adding vertices of weight zero adjacent to those branch vertices of degree less than q + 1. Thus, combining the above two theorems, we obtain the following corollary for the bounds of an optimal q-ary tree.

Corollary 3.4.1 Let $W = \{w_1, w_2, \ldots, w_n\}$ be a set of n weights such that $\sum_{i=1}^{n} w_i = 1$. Let T be an optimal q-ary tree for W such that the end vertex associated with the weight w_i is of path length ℓ_i, $i = 1, \ldots, n$. Then the cost of T is bounded in the following fashion:

$$-\sum_{i=1}^{n} w_i \log_q w_i \leq \sum_{i=1}^{n} w_i \ell_i \leq 1 - \sum_{i=1}^{n} w_i \log_q w_i.$$

Proof: The lower bound is obtained from Theorem 3.4.8. Since T is an optimal q-ary tree for w, its cost must be less than or equal to the cost of the tree determined in the proof of Theorem 3.4.9. Thus we obtain the upper bound also.

Remark 3.4.2 Let $p_1, \ldots, p_n$ be n positive numbers with $\sum_{i=1}^{n} p_i = 1$. The quantity $\sum_{i=1}^{n} p_i \log_q p_i$, where q is any positive real number, is called entropy in the information theory.

CHAPTER IV

K-SAMPLE OPTIMAL NESTED BINOMIAL GROUP TESTING

4.1 Introduction

Consider a population P of N units, each with a nonzero probability p of being defective, and a probability q = 1 - p of being good. Thus, the units in any sample X of size n $\leq$ N chosen from P have a joint distribution of binomial type with parameters (p, n). We say that X is binomial. In particular, when X = P, we say that P is a binomial population. A group test is a simultaneous test on a sample X of arbitrary size chosen from P with two possible outcomes: X is identified as good if all the units in it are good, and identified as defective if otherwise. The usual purpose is to find a certain number of defectives or all the defectives from P, or to determine that P contains no defective units. A sequence of group tests used to attain the purpose is called a group testing procedure. The cost of a group testing procedure is defined to be the expected number of tests required to attain the purpose. An important criterion of evaluating a group testing procedure is called the optimality criterion, that is, a procedure is said to be optimal if it is of minimal cost.

The concept of group testing originated from Dorfman's (1943) procedure for a blood testing problem. Suppose there are N blood samples subject to a test which reveals the presence or absence of "syphilitic antigen". A blood sample which contains syphilitic antigen is said to be defective. Dorfman (1943) suggested that n blood samples, where n < N, should be pooled and tested simultaneously as a unit at each step.

If none of the n blood samples contributed to the pool contains syphilitic antigen, then the pool will not contain it either and will be classified as good. Otherwise, the pool will be classified as defective and the individual samples making up the pool should be retested to determine which of the members are defective. It is not necessary to draw a new blood sample for this purpose since sufficient blood for both the test and the retest can be taken at once. Dorfman (1943) also formulated the cost of his procedure as a function of n and p, where p is the probability that a blood sample is defective. Hence, the value of n which minimizes the cost can be computed.

Sterrett (1959) modified Dorfman's procedure by proposing that individual testing of the units in a defective pool should cease once a defective is found and the remaining units should again be pooled and tested simultaneously in one group test.

Watson (1961) applied Dorfman's method in group screening problems where a large number of variables are screened by group testing to identify the important ones.

Sobel and Groll (1959, 1960, 1966) found many industrial applications of group testing. For example, the testing of electrical devices and the elimination of defectives from manufactured products. They also improved Dorfman's procedure by allowing that the sizes of each sample tested in a group test may vary in order to save the cost. Furthermore, they required that the tests be nested, that is, once a defective sample is found,the next k samples to be tested must be chosen from the same defective sample. Kumar and Sobel (1971) pointed out that although such a procedure is not optimal for classifying all the defectives, it is optimal for finding a single defective

from the given population P. Furthermore, the exact cost functions for such an optimal nested group testing procedure were also conjectured for the cases when P is either finite or infinite.

Hwang (1974) proved the above conjectures. His proof is based on the construction of an optimal binary tree (Huffman's tree). Later, Hwang (1976) further improved the procedure for classifying all the defectives from a given binomial population P. Garey and Hwang (1974) also generalized the problem by allowing each unit in the given population to have a distinct nonzero probability of being defective. They referred to such a population as the generalized binomial population. For this problem, Hwang (1975) also gave a dynamic programming algorithm to obtain an optimal Dorfman's procedure for a generalized binomial population of finite size.

Recently, Moon and Sobel (1977) found a formula in terms of Catalan number, for enumerating a class of group testing procedures which classify all the defectives from a given population of N units. Hwang (1978) considered hypergeometric group testing procedures for the cases when the given population was known to contain either exactly d or at most d defective units.

In this chapter, we extend the group testing method to the case when two or more experimenters are working on a single population of N units by carrying out simultaneous group tests (each of which takes the same fixed time) and cooperating so as to minimize the time required to attain the purpose. By a k-sample group testing procedure,we mean that at each step, k samples are chosen from the given population so that each is tested by an experimenter using a group test at a same fixed time period. When k=1, this is the group testing procedure studied by various authors mentioned

in the previous paragraphs. The terminology used in the literature of one-sample group testing can readily be carried over to k-sample group testing in general by regarding every k-sample group test as a unit. However, we note that in k-sample testing, more than one defective sample or good sample may be obtained from each k-sample test. Thus, in a nested k-sample testing procedure, we require that the defective samples be kept separate since the next k-sample test should be performed on a subset of one of the defective samples. It is evident that when k increases, the total expected number of individual tests required to find the defectives cannot be decreased. However, the total expected number of stages (units) of k-sample tests can be siginificantly reduced. Therefore, the purpose of using a k-sample testing procedure is mainly concerned with the saving of time required to find defectives.

Here we outline the main steps for finding all the defectives from a certain sample S which may be a portion or the whole of a population P:

Step (i): to find a single defective from S.

If there is no defective in S, we conclude that S is a good sample. If a single defective is found from S, there may be three types of samples resulting from the tests. They are good samples, defective samples and binomial samples. The good samples can be excluded from consideration. The binomial samples can be combined to form one binomial sample. But the defective samples have to be kept separate since our k-sample testing prcedure is nested.

Step (ii): to find a single defective from a defective sample.

The new defective samples resulting from step (ii) are again kept separate. But the binomial samples can be combined with those binomial

samples obtained after step (i).

Repeat step (ii) until all the defective samples are exhausted. Then go back to step (i) with the remaining binomial sample. The procedure ends when all the defectives of S are classified.

The main results of this chapter include the design of k-sample optimal nested binomial group testing procedures required to attain the purposes of step (i) and step (ii) respectively. The most interesting part is the optimal allocation for finding a single defective from an infinite binomial population.

4.2 (K+1)-ary Tree Representation

Unlike the 1-sample group testing procedures, not every k-sample group testing procedure can be represented by a (k+1)-ary tree when $k > 1$. The type of k-sample group testing procedures which can be represented by a (k+1)-ary tree will be defined later in this section. An example will also be given to clarify the above statements.

Let I be a population of n units. Suppose the i^{th} unit I_i has a probability p_i of being defective and a probability $q_i = 1 - p_i$ of being good. Without loss of generality, we may assume that $p_1 \geq p_2 \geq \ldots \geq p_n$ because this can always be achieved by properly permuting the indices of the units in I. Denote by O_i the event that the unit I_i is defective, $i = 1, \ldots, n$ and O_{n+1} the event that all the n units in I are good. Let f be a k-sample testing procedure for finding a single defective from I. When I is a binomial population, the set of all possible outputs of f is $\{O_1, \ldots, O_{n+1}\}$. When I is a defective population, the set of all possible outputs of f is $\{O_1, \ldots, O_n\}$.

Let each k-sample group test be performed on a k-tuple of samples of the form $J = (J_1, \ldots, J_k)$, where $J_1, \ldots, J_k$ are k samples chosen from I. Then each of the k samples is tested separately at a same fixed time period. Consider the type of k-sample group testing procedure which has the property that after a k-sample group test is performed on a k-tuple of samples, say, $J = (J_1, \ldots, J_k)$, the next k-sample group test is completely determined either when (i) the least index defective sample is found, or (ii) when all the k samples $J_1, \ldots, J_k$ are good. A procedure of this type can be represented by a (k+1)-ary tree in the following steps:

(i) The root of the tree is labelled by a k-tuple of samples $J^o = (J_1^o, \ldots, J_k^o)$ to be tested in the first k-sample test. The degrees and labellings of the k vertices joined to the root (by an edge) are determined according to the following criteria:

(a) if a single defective can be found when J_i^o is the least index defective sample, then the i^{th} vertex from the left is an end vertex and is labelled by that single defective unit, i=1,...,k.

(b) if the size of J_i^o is greater than one when J_i^o is the least index defective sample, then the i^{th} vertex from the left is a branch vertex and is labelled by the k-tuple of samples $J^i = (J_1^i, \ldots, J_k^i)$ to be tested in the next k-sample group test, i = 1, ..., k.

(c) if I can be determined as a good population when all $J_1, \ldots, J_k$ are found to be good, then the $(k+1)^{th}$ vertex from the left is an end vertex and is labelled by the empty set I_{n+1}, denoting that all the units of I are good.

(d) if I cannot be determined as a good population when all the samples $J_1, \ldots, J_k$ are found to be good, then the $(k+1)^{th}$ vertex from the left is a branch vertex and is labelled by the k-tuple of samples $J^{k+1} = (J_1^{k+1}, \ldots, J_k^{k+1})$ to be tested in the next k-sample group test.

(ii) Every branch vertex of the tree can be regarded as the root of a subtree so that step (i) can be applied. This continues until all the branch vertices of the tree are encountered.

It is clear that a (k+1)-ary tree representing a k-sample group testing procedure for finding the defectives of I has the property that

every end vertex of the tree should either be labelled by a unit of I or labelled by an empty set I_{n+1}, denoting that I is good. Furthermore, every unit of I must be the label of at least one of the end vertices of the tree.

Whenever a procedure is determined by its testing result to reach a stage described by a vertex of the corresponding tree, we say that the procedure reaches that vertex. In particular, when an end vertex is reached, then the procedure is terminated since an output is obtained.

Example 4.2.1 Suppose that $I = \{I_1, \ldots, I_{10}\}$. Let I_{11} be an empty set denoting the event that all the units in I are good. The testing procedure represented by the tree in Figure 4.2.1 can be interpreted as the following:

(i) the first 2-sample test is to be performed on the 2-tuple of samples $((I_1, I_2, I_4), (I_3, I_5, I_6, I_7, I_8))$.

(ii) the second 2-sample test will be performed on one of the following three possible 2-tuples of samples:

(a) $((I_1, I_3, I_5), (I_2))$ if (I_1, I_2, I_4) is defective. Note that the decision is made independent of whether or not the sample $(I_3, I_5, I_6, I_7, I_8)$ is defective.

(b) $((I_3, I_5, I_6), (I_7))$ if (I_1, I_2, I_4) is good but the sample $(I_3, I_5, I_6, I_7, I_8)$ is defective. For in this case, the latter sample is the least index defective sample.

(c) $((I_9), (I_{10}))$ if both of the samples (I_1, I_2, I_4) and $(I_3, I_5, I_6, I_7, I_8)$ are good.

(iii) the third 2-sample test will be performed on $((I_1), (I_3))$ if (I_1, I_3, I_5) is found to be defective, and on $((I_3), (I_5))$ if (I_3, I_5, I_6)

is found to be defective.

The rest of the procedure follows analogously. For example, if $((I_9), (I_{10}))$ is tested and found to be good, then the vertex labelled by I_{11} is reached and we conclude that all the units of I are good.

Note that the test described in Figure 4.2.1 is not a nested test. The reason is that when (I_1, I_2, I_4) is defective, the next k-tuple of samples to be tested is $((I_1, I_3, I_5), (I_2))$, but (I_1, I_3, I_5, I_2) is not a subset of (I_1, I_2, I_4).

It is also interesting to note that if the procedure described in Example 4.2.1 is slightly changed such that (a) of (ii) becomes -

(a') $((I_1, I_3, I_5), (I_2))$ if both (I_1, I_2, I_4) and $(I_3, I_5, I_6, I_7, I_8)$ are defective.

$((I_3, I_5, I_6), (I_2))$ if (I_1, I_2, I_4) is defective, but $(I_3, I_5, I_6, I_7, I_8)$ is good.

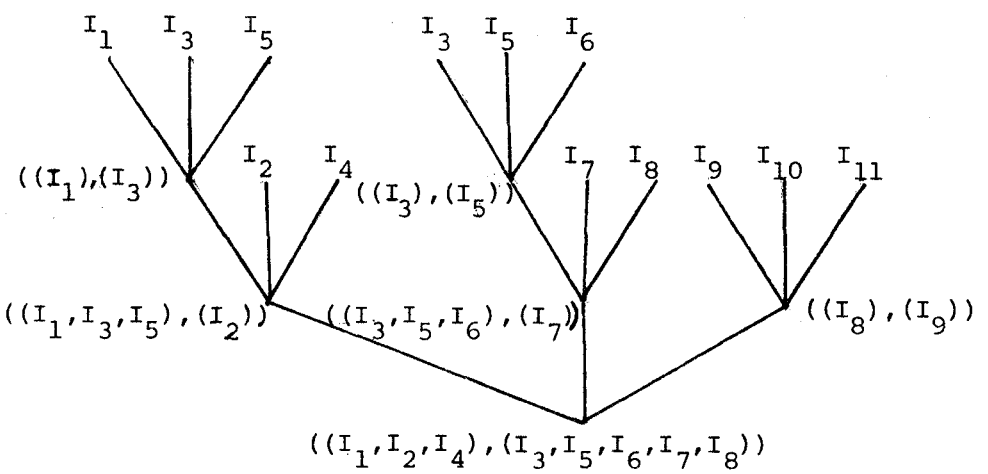


Figure 4.2.1 A 3-ary tree representing a 2-sample testing procedure for finding defectives from $I = \{I_1, \ldots, I_{10}\}$. I_{11} is an empty set denoting the event that I is good.

Let f be a k-sample testing procedure that can be represented by a (k+1)-ary tree, say T_f. Denote by $E(f)$ the cost of f, that is, the expected number of k-sample tests required to achieve an output.

Let $\{v_1, \ldots, v_m\}$ be the set of all end vertices of T_f. Let ℓ_i be the path length of v_i, and w_i the probability that the procedure f reaches v_i. Let $W = \{w_1, \ldots, w_m\}$ be a set of these m weights. Let the weight w_i be associated with the end vertex v_i, $i = 1, \ldots, m$. Define $C(T_f)$ to be the weighted path length of the end vertices of the tree T_f, which equals $\sum_{i=1}^{m} \ell_i w_i$. We say that $C(T_f)$ is the cost of the tree T_f.

The following lemma has been proved by Hwang (1973) for 1-sample optimal group testing procedures. Extension to the k-sample case is straightforward since the arguments in Hwang's proof do not depend on k.

Lemma 4.2.1 $E(f) = C(T_f)$

Proof: The procedure f reaches an end vertex of T_f if and only if it achieves an output. The number of k-sample tests required for f to reach an end vertex, say v_i, is ℓ_i and the probability that f reaches v_i is w_i'. Thus

$$E(f) = \sum_{i=1}^{m} \ell_i w_i' = C(T_f).$$

4.3 On Defective Populations of Small Sizes

Suppose it is known in advance that I contains at least one defective unit, then I is a defective sample. Recall that $p_1 \geq p_2 \geq \ldots \geq p_n$. Without loss of generality, we may assume that $n \equiv 1 \bmod k$. For otherwise, we can always include extra units I_{n+j} with probability 0 of being defective for $j = 1, \ldots, a$ where a is an integer such that $0 \leq a < k$ and $n + a \equiv 1 \bmod k$. Then we consider $n + a$ instead of n. Let $W = \{w_1, \ldots, w_n\}$ be a sequence of weights such that

$$w_i = \alpha^{-1} p_i \prod_{j=1}^{i-1} q_j \qquad i = 1, \ldots, n \tag{4.3.1}$$

where $\alpha = 1 - \prod_{j=1}^{n} q_j$. Let A_w be an optimal alphabetic (k+1)-ary tree for W. Now we label the vertices of A_w according to the following rules:

(i) the end vertex with weight w_i is labelled by I_i.

(ii) every internal vertex is labelled by a k-tuple of subsets of I such that the i^{th} coordinate consists of the subset of units of I labelling the end vertices which are reachable from the i^{th} leftmost edge joined to that internal vertex, $i = 1, \ldots, k$.

Hwang (1973) has proved the following theorem for $k = 1$. The proof, in general, is entirely similar since the argument of the original proof does not depend on the restriction that $k = 1$.

Theorem 4.3.1 A_w defines a k-sample testing procedure f_w for finding the least index defective in I.

Let D_i be the event that I_i is the least index defective in I. Then $\{D_1, \ldots, D_n\}$ is the set of all possible outputs of f_w. Since w_i is the probability that I_i is the least index defective of I, w_i is also the

probability that f_w reaches the end vertex I_i of A_w, $i = 1, \ldots, n$.

Let $\mathcal{J}_w$ be the family of all k-sample testing procedures that can be represented by a (k+1)-ary tree and have $\{D_1, \ldots, D_n\}$ as the set of all possible outputs.

The following lemma and theorem have been proved by Garey and Hwang (1974), for the case that $k = 1$. The proofs in general are similar but depend on the results established in Chapter III for (k+1)-ary weighted trees. This will be illustrated in the proof of the theorem.

Lemma 4.3.1 Let $\{v_1, v_2, \ldots, v_n\}$ be the set of all end vertices of a (k+1)-ary tree. Suppose v_i is of path length ℓ_i, $i = 1, 2, \ldots, n$ and $\ell_1 \leq \ell_2 \leq \ldots \leq \ell_n$. Let $w_1^*, w_3^*, \ldots, w_n^*$ and $w_1', w_2', \ldots, w_n'$ be two sequences of weights such that

$$\sum_{i=h}^{n} w_i^* \leq \sum_{i=h}^{n} w_i'$$

for any integer $h = 1, 2, \ldots, n$, then

$$\sum_{i=1}^{n} \ell_i w_i^* \leq \sum_{i=1}^{n} \ell_i w'_i.$$

Theorem 4.3.2 $E(f_w) \leq E(f^*)$ for every $f^* \in \mathcal{J}_w$.

Proof: Consider any testing procedure f_s in $\mathcal{J}_w$. Let T_s be the (k+1)-ary testing tree representing f_s. Suppose $s_1, \ldots, s_m$ is the left-to-right sequence of all the end vertices of T_s. Since each unit of I must label at least one end vertex of T_s, we must have $m \geq n$. Let $W' = \{w_1', \ldots, w_m'\}$, where w_i' is the probability that the procedure f_s reaches the end vertex s_i, $i = 1, \ldots, m$. Now we order W' to obtain a new sequence $W_r' = \{w'_{r(1)}, \ldots, w'_{r(m)}\}$ such that $w'_{r(1)} \geq \ldots \geq w'_{r(m)}$.

Denote by L the set of all labels of $s_{r(1)}, \ldots, s_{r(h)}$. Let $\ell = |L|$. Then necessarily $1 \leq \ell \leq \min\{h,n\}$. Furthermore, the event that f_s reaches one of the end vertices $s_{r(1)}, \ldots, s_{r(h)}$ is disjoint with the event that all the ℓ units in L are good. Hence

$$\sum_{i=1}^{h} w'_{r(i)} \leq 1 - \alpha^{-1} \prod_{j\in\{k:\ I_k\in L\}} q_j \leq \alpha^{-1}\left(1 - \prod_{j\in\{k:\ I_k\in L\}} q_j\right).$$

But $q_1 \leq \ldots \leq q_n$ and $\ell \leq \min\{h,n\}$ implies that

$$\prod_{j\in\{k:\ I_k\in L\}} q_j \geq \prod_{j=1}^{\ell} q_j \geq \prod_{j=1}^{\min\{h,n\}} q_j$$

Thus when $1 \leq h \leq n$, we have

$$\sum_{i=1}^{h} w'_{r(i)} \leq \alpha^{-1}\left(1 - \prod_{j=1}^{h} q_j\right) = \sum_{i=1}^{h} w_i,$$

where $w_1, \ldots, w_n$ are defined by (4.3.1). Define $w_i = 0$ for $n < i \leq m$. Then

$$\sum_{i=1}^{h} w'_{r(i)} \leq 1 = \sum_{i=1}^{h} w_i$$

for all h, $n < h \leq m$.

Let $H_{w'_r}$ be an optimal (k+1)-ary tree for W'_r. Let ℓ'_i be the path length of the end vertex associated with weight $W'_{r(i)}$, $i = 1, \ldots, m$. By Theorem 3.4.1 (ii), we may assume that $\ell'_1 \leq \ldots \leq \ell'_m$, since there always exists such an optimal (k+1)-ary tree for W'_r. Then we can apply Lemma 4.3.1 and obtain

$$\sum_{i=1}^{m} w_i \ell'_i \leq \sum_{i=1}^{m} w_{r(i)} \ell'_i.$$

Lemma 4.2.1 implies that $E(f_w) = C(A_w)$ which also equals to the cost of an optimal (k+1)-ary tree for W by Theorem 3.4.3, since the sequence of

weights in W is monotone. Therefore, $E(f_w) \leq \sum_{i=1}^{m} w_i \ell'_i$.

On the other hand, $E(f_s) = C(T_s)$, which is no less than the cost of an optimal (k+1)-ary tree for W'. Therefore, $E(f_s) \geq \sum_{i=1}^{m} w_i \ell'_i$.

Hence, we conclude that $E(f_s) \geq E(f_w)$ and the proof is complete.

If follows from Theorem 4.3.2 that the procedure f_w is a k-sample optimal nested group testing procedure in the family $\mathcal{F}_w$. The upper and lower bounds of $E(f_w)$ are given by the following theorem which is due to Garey and Hwang (1974) when k = 1.

Theorem 4.3.3 $\sum_{i=1}^{n} w_i \log_{(k+1)} w_i^{-1} \leq E(f_w) \leq 1 + \sum_{i=1}^{n} w_i \log_{(k+1)} w_i^{-1}$

where $w_i = \alpha^{-1} p_i \sum_{j=1}^{n} q_j$ and $\alpha = 1 - \sum_{j=1}^{n} q_j$.

Proof: Since $\sum_{i=1}^{n} w_i = 1$, the result follows directly from Corollary 3.4.1.

4.4 On Binomial Populations of Small Sizes

Suppose it is not known whether I contains a defective or not, then I is a binomial population. Recall that $p_1 \geq p_2 \geq \ldots \geq p_n$. Without loss of generality, we may assume that $n \equiv 0 \bmod k$. For otherwise, we can always include extra units I_{n+j} with probability 0 of being defective for $j = 1, \ldots, a$ where a is an integer such that $0 \leq a < k$ and $n + a \equiv 0 \bmod k$. Then we consider $n + a$ instead of n. Let $U = \{u_1, \ldots, u_{n+1}\}$ be a sequence of weights such that

$$u_i = \begin{cases} p_i \prod_{j=1}^{i-1} q_j & i = 1, \ldots, n \\ q^n & k = n+1. \end{cases} \tag{4.4.1}$$

Let A_U be an optimal alphabetic (k+1)-ary tree for U. We label the vertices of A_U according to the following rules:

(i) the end vertex with weight u_i is labelled by the units I_i, $i = 1, \ldots, n$ and the rightmost end vertex with weight u_{n+1} is labelled by the empty set I_{n+1}.

(ii) every internal vertex is labelled by a k-tuple of samples such that the i^{th} coordinate consists of the sample of units of I labelling the end vertices which are reachable from the i^{th} leftmost edge joined to that internal vertex, $i = 1, \ldots, k$.

We state the following theorem which has been proved by Hwang (1973) for k = 1. The proof in general is entirely similar.

Theorem 4.4.1 A_U defines a k-sample testing procedure f_U for either finding the least index defective from the binomial population I or determining that I is good.

Let D_i be the event that I_i is the least index defective in I, $i = 1, \ldots, n$ and let D_{n+1} be the event that all the units in I are good. Then $\{D_1, \ldots, D_n, D_{n+1}\}$ is the set of all possible outputs of f_U. Furthermore u_i is the probability that f_U reaches the end vertex I_i of A_U, $i = 1, \ldots, n+1$.

Let $\mathcal{J}_U$ be the family of all the k-sample testing procedures that can be represented by a (k+1)-ary tree and have $\{D_1, \ldots, D_n, D_{n+1}\}$ as the set of all possible outputs.

By using Theorem 3.4.1, Theorem 3.4.3 and Lemma 4.3.1, Hwang's (1973) proof can be extended to obtain the following result for k-sample testing procedures with any $k \geq 1$.

Theorem 4.4.2 $E(f_U) \leq E(f^*)$ for every $f^* \in \mathcal{J}_U$.

The procedure f_U is therefore a k-sample optimal nested group testing procedure for either finding a least index defective or determining that the binomial population I is good. The upper and lower bounds of f_U can again be obtained by using Corollary 3.4.1. Thus we have

Theorem 4.4.3 $$\sum_{i=1}^{n+1} u_i \log_{k+1} u_i^{-1} \leq E(f_U) \leq 1 + \sum_{i=1}^{n+1} u_i \log_{k+1} u_i^{-1}$$

where u_i, $i = 1, \ldots, n+1$ are defined in (4.4.1).

Proof: Note that $\sum_{i=1}^{n+1} u_i = 1$ and therefore Corollary 3.4.1 applies.

4.5 On Defective Populations with Units from a Unique Binomial Distribution

Given a population I of n units each with the same probability p of being defective and the probability q = 1 - p of being good, suppose it is known that I contains at least one defective, then I is a defective population. The problem of finding a single defective from I is a special case of the one discussed in Section 4.3.

Under the present conditions, the cost $E(f_w)$ of the k-sample optimal nested group testing procedure f_w belonging to the family $\mathcal{F}_w$ (see Section 4.3) can be formulated as a function of n, p and q only. For fixed p and q, we denote $E(f_w)$ by $F^{k+1}(n)$.

In the procedure f_w, after a k-sample test has been performed on a k-tuple of samples $J = (J_1, \ldots, J_k)$, where J_i, $i = 1, \ldots, k$ are the disjoint subsets of I, the next k-sample test is required if no single defective has been found. The units to be tested in the next k-sample test must be a subset of the sample chosen according to the following criteria:

(i) if one or more of the samples $J_1, \ldots, J_k$ are found to be defective, the least index defective sample should be chosen.

(ii) if none of the samples $J_1, \ldots, J_k$ are defective, the next k-tuple of samples should be chosen from the units of I which have never been tested.

The testing procedure continues until a single defective is found.

In order to find an expression for $F^{k+1}(n)$, we let $(J_1^o, \ldots, J_k^o)$ be the first k-tuple of samples tested in the proecdure f_w. Let y_i be the size of the sample J_i^o, $i = 1, \ldots, k$. Since I is known to have at least one defective, the probability that J_i^o is the least index defective sample

equals $q^{y_0+\ldots+y_{i-1}}(1 - q^{y_i})/(1 - q^n)$, where $y_0 \equiv 0$. The minimal expected cost of finding a defective from J_i when it is the least index defective sample equals $F^{k+1}(y_i)$. Therefore, the cost of f_w can be written as

$$F^{k+1}(n) = \min_{R_n^{k+1}} \left\{ 1 + \frac{\sum_{i=1}^{k+1} q^{y_0+\ldots+y_{i-1}}(1 - q^{y_i})\, F^{k+1}(y_i)}{1 - q^n} \right\}$$

for $n \equiv 1 \bmod k$, where the minimum is taken over the region

$$R_n^{k+1} = \{(y_1, \ldots, y_{k+1}): \sum_{i=1}^{k+1} y_i = n \text{ and } y_i \equiv 1 \bmod k,\ i = 1, \ldots, k\}.$$

The term 1 arises due to the cost of testing the samples $J^o = (J_1^o, \ldots, J_{k+1}^o)$. It is trivial that $F^{k+1}(k+1) = 1$. The reason why we impose the conditions that $y_i \equiv 1 \bmod k$, $i = 1, \ldots, k+1$ is that the procedure f_w is primarily designed for finding a single defective from a defective sample of size n with $n \equiv 1 \bmod k$.

It will be verified in Theorem 4.5.1 that $F^{k+1}(n)$ is in fact the cost of an optimal alphabetic (k+1)-ary tree for the sequence of weights $W = \{w_1, \ldots, w_n\}$, where $w_i = pq^{i-1}/(1 - q^n)$, $i = 1, \ldots, n$. Similar tree interpretations can also be found for the following functions:

$$F^{t,k+1}(n) = \min_{R_n^t} \left\{ 1 + \frac{\sum_{i=1}^{t} q^{y_0+\ldots+y_{i-1}}(1 - q^{y_i})\, F^{k+1}(y_i)}{1 - q^n} \right\},$$

for $n \equiv t \bmod k$, where the minimum is taken over the region

$$R_n^t = \{(y_1, \ldots, y_t): \sum_{i=1}^{t} y_i = n \text{ and } y_i \equiv 1 \bmod k,\ i = 1, \ldots, t\}, \text{ for } t = 1, \ldots, k+1.$$

It is clear that when $t = k+1$, we have $F^{k+1,k+1}(n) = F^{k+1}(n)$.

For simplicity in notation, we write

$$F^{t,k+1}(n) = F^t(n), \quad \text{for } t = 1, \ldots, k+1. \tag{4.5.1}$$

The function $F^k(n)$ can be found useful when one wants to determine the cost of an optimal nested k-sample group testing procedure for finding a single defective from a binomial population (see Section 4.6 and Section 4.7). In order to find the tree interpretations of the functions $F^t(n)$, $t = 1,\ldots,k+1$, we introduce here the notion of a t-sum (k+1)-ary tree which is trivial when $k = 1$.

Definition 4.5.1 Let $n = n_1 + \ldots + n_t$, where $n_i \equiv 1 \bmod k$, $i = 1,\ldots,t$ and $1 \leq t \leq k+1$ are positive integers. A t-sum (k+1)ary tree with n end vertices is a tree formed by combining t planted (k+1)-ary trees $T_1,\ldots,T_t$ at their roots, where T_i has n_i end vertices, $i = 1,\ldots,t$ and $1 \leq t \leq k+1$.

Thus when $t = k+1$, a t-sum (k+1)-ary tree is again a (k+1)-ary tree.

Definitions concerning the optimality and alphabetic optimality of a t-sum (k+1)-ary tree can be given analoguous to those defined for (k+1)-ary trees. Analogues of generalized Huffman's algorithm, generalized T-C algorithm and the related theorems can also be established since the only difference between a t-sum (k+1)-ary tree and a (k+1)-ary tree is the degree at the root when $t < k+1$.

Let $\{w_1^*, w_2^*,\ldots, w_n^*\}$ be a sequence of n weights. Denote by

$$C^t\{w_i^*\}_j^r = C^t\{w_j^*, w_{j+1}^*,\ldots,w_r^*\}$$

the cost of an optimal t-sum (k+1)-ary tree for the sequence of weights $\{w_j^*, w_{j+1}^*,\ldots, w_r^*\}$, where $1 \leq j \leq r \leq n$. In particular, when $w_i^* = w_i = pq^{i-1}/(1 - q^n)$, $i + 1,\ldots, n$, we write

$$C^t\{w_i\}_1^r = C^t(r), \qquad 1 \leq r \leq n.$$

Similarly, denote by

$$A^t\{w_i^*\}_j^r = A^t\{w_j^*, w_{j+1}^*, \ldots, w_r^*\}$$

the cost of an optimal alphabetic (k+1)-ary tree for the sequence of weights $w_j^*, w_{j+1}^*, \ldots, w_r^*$, where $1 \le j \le r \le n$. In particular, when $w_i^* = w_i = pq^{i-1}/(1-q^n)$, $i = 1, \ldots, n$, we write

$$A^t\{w_i\}_1^r = A^t(r), \qquad 1 \le r \le n. \tag{4.5.2}$$

The following results have been proved by Hwang(1974) when $k = 1$ and $t = k+1$.

Lemma 4.5.1 If $w_i' = a w_i^*$, $i = 1, \ldots, n$, for some given non-negative constant a, then we have $C^t\{w_i'\}_1^n = a C^t\{w_i^*\}_1^n$ and $A^t\{w_i'\}_1^n = a A^t\{w_i^*\}$

Proof: This is a result of Lemma 3.4.2..

Theorem 4.5.1 $F^t(n) = A^t(n) = C^t(n)$, for $t = 1, \ldots, k+1$.

Proof: Let $y_o \equiv 0$ and T be the optimal alphabetic t-sum (k+1)-ary tree under consideration. Let $T_0, T_1, \ldots, T_t$ be (k+1)-ary subtrees of T such that T_0 is the one consisting of the root and the k+1 vertices of path length 1 as its vertex set and T_i is the one rooted at the i^{th} leftmost end vertex of T_0, $i = 1, \ldots, t$. As a result of Lemma 3.4.2, we can write

$$A^t(n) = \min_{R_n^t} \left\{ A^t\{ \sum_{i=1}^{y_1} w_i, \sum_{i=y_1+1}^{y_1+y_2} w_i, \ldots, \sum_{i=y_1+\ldots+y_{t-1}}^{y_1+\ldots+y_t} w_i \} + \sum_{i=1}^{t} A^{k+1}\{w_i\}_{1+y_o+\ldots+y_{i-1}}^{y_1+\ldots+y_i} \right\}$$

for $n \equiv t \bmod k$, where the minimum is taken over the region

$$R_n^t = \{ (y_1,\ldots, y_t): \sum_{i=1}^{t} y_i = n \text{ and } y_i \equiv 1 \bmod k,\ i = 1,\ldots,t\},\ \text{for } t = 1,\ldots,k+1.$$

Now Lemma 4.5.1 implies that

$$A^t\{w_i\}_{1+y_1+\ldots+y_{i-1}}^{y_1+\ldots+y_i} = \frac{q^{y_0+y_1+\ldots+y_{i-1}}(1-q^{y_i})}{1-q^n} A^{k+1}\frac{pq^{\ell-1}}{1-q^{y_i}}{}_1^{y_i}$$

$$= \frac{q^{y_0+y_1+\ldots+y_{i-1}}}{1-q^n}(1-q^{y_i})\, A^{k+1}(y_i)$$

for $i = 1,\ldots, t$. By definition,

$$A^t\{ \sum_{i=1}^{y_1} w_i,\ldots, \sum_{i=y_1+\ldots+y_{i-1}}^{y_1+\ldots+y_t} w_i \} = 1.$$

thus we have

$$A^t(n) = \min_{R_n^t}\left\{1 + \frac{\sum_{i=1}^{t} q^{y_0+\ldots+y_{i-1}}(1-q^{y_i})\, A^{k+1}(y_i)}{1-q^n}\right\} \tag{4.5.3}$$

which is exactly of the same form as $F^t(n)$. So we conclude that $A^t(n) = F^t(n)$. The fact that $A^t(n) = C^t(n)$ follows from Theorem 3.4.3 and its analogue for t-sum (k+1)-ary trees, since the weights are monotone. This completes the proof.

Let x be a positive integer which satisfies the inequalities

$$q^{x-1} + q^x + \ldots + q^{x+k-1} \geq 1 \geq q^x + q^{x+1} + \ldots + q^{x+k} \tag{4.5.4}$$

for any fixed q such that $0 \leq q < 1$. A closed form expression for $F^t(n)$, where $0 \leq n \leq 2x+k-1$, can be derived by constructing a perfect (k+1)-ary tree defined below.

Definition 4.5.2 A t-sum (k+1)-ary tree with n end vertices of path length $\ell_1,\ldots, \ell_n$ is called a perfect t-sum (k+1)-ary tree if and

only if $|\ell_i - \ell_j| \leq 1$, for all i and j such that $1 \leq i,j \leq n$.

Remark 4.5.1 Let n be the total number of end vertices of a t-sum (k+1)-ary tree. Suppose that (α, β) is a pair of non-negative integers satisfying

$$n = t(k+1)^{\alpha} + \beta, \qquad 0 \leq \beta \leq tk(k+1)^{\alpha}, \tag{4.5.5}$$

then the following are true:

(i) $\beta \equiv 0 \bmod k$, since $n \equiv t \bmod k$ implies that $t(k+1)^{\alpha} + \beta - t \equiv 0 \bmod k$, the result follows from the fact that $(k+1)^{\alpha} - 1 \equiv 0 \bmod k$.

(ii) $\frac{\beta}{k}(k+1)$ must be the total number of end vertices of path length $\alpha + 2$, since there is a total of $(k+1)^{\alpha}$ vertices of path length $\alpha + 1$ and an increase of k+1 vertices of path length $\alpha + 2$ implies a decrease of 1 end vertex of path length $\alpha + 1$. Or equivalently, there is a total of β/k branch vertices of path length $\alpha + 1$ and a total of $n - \frac{\beta}{k}(k+1)$ end vertices of path length $\alpha + 1$.

(iii) When n is of the form $t(k+1)^{\alpha}$, there are two pairs (α_1, β_1) and (α_2, β_2) satisfying (4.5.5), but they represent the same perfect tree.

Lemma 4.5.2 Let $w_1^*, w_2^*, \ldots, w_n^*$ be a sequence of n weights such that $n \equiv t \bmod k$ and $w_1^* \geq w_2^* \geq \ldots \geq w_n^*$. Then $w_{n-t}^* + \ldots + w_n^* \geq w_1^*$ implies that there exists an optimal alphabetic t-sum (k+1)-ary tree for these n weights which is also a perfect t-sum (k+1)-ary tree.

Proof: Using generalized Huffman's algorithm for constructing

an optimal alphabetic t-sum (k+1)-ary tree, the first step is to combine the k+1 vertices of smallest weights to form a new vertex of weight $w^*_{n+1} = w^*_{n-k} + \ldots + w^*_n \geq w^*_1$. Theorem 3.4.1 asserts that there exists an optimal (k+1)-ary tree such that

$$\ell_1 \geq \ell_{n+1} = \ell_n - 1$$

where ℓ_i is the path lenght of the vertex with weight w_i, $i = 1, \ldots, n+1$. Therefore,

$$| \ell_i - \ell_j | \leq \ell_n - \ell_1 \leq 1 , \quad \text{for all } 1 \leq i, j \leq n.$$

Thus it is a perfect t-sum (k+1)-ary tree. Furthermore, Theorem 3.4.3 and its analogue imply that such a tree can be converted into an optimal alphabetic t-sum (k+1)-ary tree of the same cost, since the sequence of weights $w^*_1, w^*_2, \ldots, w^*_n$ is monotone decreasing. This completes the proof.

Theorem 4.5.2 Let t be any fixed integer such that $1 \leq t \leq k+1$. For any integer n which satisfies the conditions $n \equiv t \bmod k$ and $n \leq x + k$, we have

$$F^t(n) = 1 + \alpha + \frac{q^{n - \frac{\beta}{k}(k+1)} - q^n}{1 - q^n}$$

Proof: Since

$$w_{n-k} + w_{n-k+1} + \ldots + w_n = \frac{p}{1-q^n} \left(q^{n-k-1} + q^{n-k} + \ldots + q^{n-1} \right)$$

$$= \frac{pq^{n-k-x}}{1-q^n} \left(q^{x-1} + q^x + \ldots + q^{x+k-1} \right)$$

$$\geq \frac{pq^{n-k-x}}{1-q^n} \geq \frac{p}{1-q^n}$$

when $n \leq x+k$ Lemma 4.5.2 implies that there exists an optimal alphabetic

t-sum (k+1)-ary tree for $w_1, \ldots, w_n$, which is a perfect t-sum (k+1)-ary tree. Suppose (α, β) is defined by the expression (4.5.5) in Remark 4.5.1 which implies

$$F^t(n) = (\alpha + 1) \sum_{i=1}^{h} w_i + (\alpha + 2) \sum_{i=h+1}^{n} w_i$$

$$= (\alpha + 1) + \frac{q^h - q^n}{1-q^n},$$

where $h = n - \frac{\beta}{k}(k+1)$. This completes the proof.

Theorem 4.5.3 Let t be any fixed integer such that $1 \leq t \leq k+1$. For any integer n which satisfies the conditions $n \equiv t \bmod k$, $n = x + h$ and $0 \leq h \leq x + k - 1$, we have

$$F^t(n) = 1 + \alpha + \frac{1}{1-q^n}\left(q^{m-\frac{\beta}{k}(k+1)-r} + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r} - 2q^n \right)$$

where m, r, α, β are defined by (4.5.9), (4.5.10) and (4.5.11).

Proof: Let w_{n+i} be the weight of the new vertex created from the i^{th} step of the generalized Huffman's algorithm, $i = 1, 2, \ldots$. Let i_o be the largest integer satisfying the following condition: for any $i \leq i_o$, in the i^{th} step of the algorithm, k + 1 weights from the original sequence $w_1, \ldots, w_n$ are combined. It is obvious that $i_o \geq 1$. For any integer $i \leq i_o$, we have

$$w_{n+i} = w_{(x+h)+i}$$

$$= w_{x+h-(k+1)i+1} + w_{x+h-(k+1)i+2} + \ldots + w_{x+h-(k+1)i+k+1}$$

$$= \frac{pq^{x+h-(k+1)i}}{1-q^n}(1 + q + \ldots + q^k)$$

From the definition of x given by (4.5.4), we find that

$$w_{n+i} \geq \frac{pq^{h-(k+1)i+1}}{1-q^n} \tag{4.5.6}$$

and

$$w_{n+i} \leq \frac{pq^{h-(k+1)i}}{1-q^n} \tag{4.5.7}$$

In particular, when $i = 1$ and $h \geq k+1$, we have

$$w_{h-k+1} \leq w_{n+1} \leq w_{h-k}. \tag{4.5.8}$$

Thus the weights of the new vertices created are always greater than or equal to $pq^{h-k}/(1-q^n)$. Therefore, we know that

$$i_o = \left\lfloor \frac{n - (h-k)}{k+1} \right\rfloor = \left\lfloor \frac{x+k}{k+1} \right\rfloor .$$

Let a be an integer such that $h \equiv a \bmod (k+1)$ and $0 \leq a \leq k$. Let $i_1 = \left\lfloor \frac{h}{k+1} \right\rfloor$ whenever $h \geq k+1$, then we have $i_1 = \frac{h-a}{k+1}$. It follows from the definitions that $i_1 \leq i_o$ since $h \leq x+k-1$. Inequalities (4.5.6) and (4.5.7) imply that

$$w_{a+2} \leq w_{n+i_1} \leq w_{a+1}, \quad \text{whenever } h \geq k+1.$$

Now we claim that $w_{n+i_1+1} \geq w_1$. This can be shown in the following two separate cases:

(i) if $i_1 + 1 \leq i_o$, then inequality (4.5.6) implies that

$$w_{n+i_1+1} \geq \frac{pq^{a-k}}{1-q^n} \geq w_i \text{ since } a \leq k.$$

(ii) if $i_1 + 1 > i_o$, then (4.5.8) implies that

$$w_{n+i_1+1} = (w_{x+a-k+1} + \ldots + w_{x+a}) + w_{n+1}$$

$$\geq (w_{x+a-k+1} + \ldots + w_{x+a}) + w_{h-k+1}$$

$$\geq (w_{x+a-k+1} + \ldots + w_{x+a}) + w_{x+a+1}$$

since $h \leq x+k-1$. From the definition of x given by (4.5.4), we have

$$w_{n+i_1+1} \geq \frac{pq^{a-k}}{1-q^n} \geq w_1 \text{ since } a \leq k.$$

Hence, after the i_1 steps of Huffman's algorithm, we obtain a sequence S' of weights in which the sum of the k+1 smallest weights is greater than or equal to w_1. Let m be the number of weights in S'. Then

$$m = x + h - \left\lfloor \frac{h}{k+1} \right\rfloor k \qquad (4.5.9)$$

Note that $n = x+h \equiv 1 \bmod k$ implies that $m \equiv 1 \bmod k$. By Lemma 4.5.2, there exists an optimal alphabetic t-sum (k+1)-ary tree $T_{S'}$ for the sequence S', which is a perfect t-sum (k+1)-ary tree. If (α, β) satisfies

$$m = t(k+1)^{\alpha} + \beta \text{ where } 0 \leq \beta \leq tk(k+1)^{\alpha} \qquad (4.5.10)$$

we can count exactly how many of the $m - \frac{\beta}{k}(k+1)$ end vertices of $T_{S'}$ with path length $\alpha+1$ are vertices formed by the combination of (k+1) end vertices in the original sequence $\{w_1, \ldots, w_n\}$. Let this number be r. Let the weights of these r vertices be w_{n+i_1-t+1}, $t = 1, \ldots, r$ whenever $r \geq 1$. The largest one is w_{n+i_1} which lies between w_{a+1} and w_{a+2}, provided that $h \geq k+1$, or equivalently $i_1 \geq 1$. There are k+1 weights from the original sequence between the weights w_{n+i_1+t} and w_{n+i_1+t+1}, $t = 0, 1, \ldots, r-1$, whenever $r \geq 2$. Therefore,

$$r = \min\left\{\left\lfloor\frac{h}{k+1}\right\rfloor, \left\lfloor\frac{m - \frac{\beta}{k}(k+1) + (k+1) - (a+1)}{k+2}\right\rfloor\right\}$$

$$= \min\left\{\left\lfloor\frac{h}{k+1}\right\rfloor, \left\lfloor\frac{m - \frac{\beta}{k}(k+1) + k - a}{k+2}\right\rfloor\right\} \qquad (4.5.11)$$

Hence, for the vertices in the original sequence, the first $m - \frac{\beta}{k}(k+1) - r$ vertices will have path length $\alpha + 1$, the last $\left(\left\lfloor\frac{h}{k+1}\right\rfloor - r\right)(k+1)$ vertices will have path length $\alpha + 3$ and the remaining

$$n - \left(\left\lfloor\frac{h}{k+1}\right\rfloor - r\right)(k+1) - \left(m - \frac{\beta}{k}(k+1) - r\right)$$

$$= (x+h) - \left\lfloor\frac{h}{k+1}\right\rfloor(k+1) + r(k+1) - \left(x+h - \left\lfloor\frac{h}{k+1}\right\rfloor k - \frac{\beta}{k}(k+1) - r\right)$$

$$= -\left\lfloor\frac{h}{k+1}\right\rfloor + (k+2)r + \frac{\beta}{k}(k+1)$$

vertices at the middle will have path length $\alpha + 2$. Now we can compute

$$F^t(n) = (\alpha+1)\sum_{i=1}^{H} w_i + (\alpha+2)\sum_{i=H+1}^{H+H^*} w_i + (\alpha+3)\sum_{i=H+H^*+1}^{n} w_i$$

where $H = m - \frac{\beta}{k}(k+1) - r$ and $H^* = \left\lfloor\frac{h}{k+1}\right\rfloor + (k+2)r + \frac{\beta}{k}(k+1)$. Consequently,

$$F^t(n) = 1 + \alpha + \sum_{i=H+1}^{n} w_i + \sum_{i=H+H^*+1}^{n} w_i$$

$$= 1 + \alpha + \frac{q^H(1 - q^{n-H})}{1-q^n} + \frac{q^{H+H^*}(1 - q^{n-H-H^*})}{1-q^n}$$

$$= 1 + \alpha + \frac{1}{1-q^n}\left\{q^H + q^{H+H^*} - 2q^n\right\},$$

then the result of the theorem follows by substituting the values of H and H^* into the above equation.

4.6 On Binomial Populations with Units from a Unique Binomial Distribution

Suppose I is a binomial population of n units, each with the same probability p of being defective and the probability q = 1 - p of being good. The problem of finding a single defective from I is a special case of the one discussed in Section 4.4.

Under the present conditions, the cost $E(f_U)$ of the k-sample optimal nested group testing procedure f_U belonging to the family $\mathcal{F}_U$ (see Section 4.4) can be formulated as a function of n, p and q only. For fixed p and q, we denote $E(f_U)$ by K(n).

It can be seen from the definition that the procedure f_U is the same as the procedure f_w described precisely in Section 4.5 except that f_U terminates in either one of the following cases:

(i) a single defective is found,

(ii) the population I is found to be good.

Let $J^o = (J_1^o, \ldots, J_k^o)$ be the first k-tuple of samples tested in the procedure f_U. Let y_i be the size of the sample J_i, i = 1, ..., k, $y_o \equiv 0$, and $y = \sum_{i=1}^{k} y_i$. The probability that J_i^o is the least index defective sample equals $q^{y_o+\ldots+y_{i-1}}(1 - q^{y_i})$, and the minimal expected cost of finding a single defective from J_i^o when J_i^o is the least index defective sample equals $F^{k+1}(y_i)$. The probability that all the k samples $J_1^o, \ldots, J_k^o$ are good equals q^y and the minimal expected cost of finding a single defective from the remaining n-y units when all these k samples are good equals K(n-y). Therefore we have

$$K(n) = \min_{R^{k+1}_{n+1,y}} \{1 + q^{y}K(n-y) + \sum_{i=1}^{k} q^{y_o+\ldots+y_{i-1}}(1 - q^{y_i})\, F^{k+1}(y_i)\} \qquad (4.6.1)$$

for $n \equiv 0 \bmod k$, where the minimum is taken over the region

$$R^{k+1}_{n+1,y} = \{(y_1,\ldots,y_{k+1}) : \sum_{i=1}^{k+1} y_i = n+1,\ y = n+1-y_{k+1} \text{ and } y_i \equiv 1 \bmod k,\ i=1,\ldots,k+1\}$$

Note that the term 1 arises due to the cost of testing the first k-tuple of samples $J^{o} = (J_1^{o}, \ldots, J_k^{o})$. The reason why we impose the condition that $\sum_{i=1}^{k+1} y_i = n+1$ is that the procedure f_U has n+1 possible outputs (cf. Section 4.4).

If we express K(n) in terms of the function $F^{k}(n)$ defined by (4.5.1), we find that

$$K(n) = \min_{y=1,\ldots,n} \{q^{y} + q^{y}K(n-y) + (1-q^{y})F^{k}(y)\}, \qquad (4.6.2)$$

where y is the total number of units to be tested in the first k-sample group test.

Theorem 4.6.1 $K(n) = A^{k+1}\{u_i\}_1^{n+1}$,

where $A^{k+1}\{u_i\}_1^{n+1} = A^{k+1}\{u_1, \ldots, u_{n+1}\}$ is the cost of an optimal alphabetic (k+1)-ary tree for the sequence of weights $u_1, \ldots, u_{n+1}$, such that $u_i = pq^{i-1}$, $i = 1, \ldots, n$ and $u_{n+1} = q^{n}$.

Proof: Define $y_o \equiv 0$. Then

$$A^{k+1}\{u_i\}_1^{n+1} = \min_{R^{k+1}_{n+1}} \left\{ \begin{array}{l} A^{k+1}\{\sum_{i=1}^{y_1} u_i,\ \sum_{i=y_1+1}^{y_1+y_2} u_i,\ \ldots,\ \sum_{i=y_1+\ldots+y_{k+1}}^{n+1} u_i\} \\ + \sum_{i=1}^{k+1} A^{k+1}\{u_i\}_{y_o+\ldots+y_{i-1}+1}^{y_1+\ldots+y_i} \end{array} \right\}$$

for $n \equiv o \bmod k$, where the minimum is taken over the region

$$R_{n+1}^{k+1} = \{ (y_1,\ldots,y_{k+1}) : \sum_{i=1}^{k+1} y_i = n+1 \text{ and } y_i \equiv 1 \bmod k,\ i = 1,\ldots,k+1 \}.$$

Now Lemma 4.5.1 implies that

$$A^{k+1}\{u_i\}_{y_o+\ldots+y_{i-1}+1}^{y_1+\ldots+y_i} = q^{y_o+\ldots+y_{i-1}}(1-q^{y_i})\, A^{k+1}\left\{\frac{pq^{\ell-1}}{1-q^{y_i}}\right\}_1^{y_i}$$

$$= q^{y_o+\ldots+y_i}(1-q^{y_i})\, A^{k+1}(y_i), \quad \text{for } i = 1,\ldots,k+1.$$

If we let $y = \sum_{i=1}^{k} y_i$, we obtain from the definition of $F^{k+1}(n)$ that

$$\min_{R_y^k}\left\{ \sum_{i=1}^{k} A^{k+1}\{u_i\}_{y_o+\ldots+y_{i-1}+1}^{y_1+\ldots+y_i} \right\} = (1-q^y)\, F^k(n) - 1 + q^y,$$

where the minimun is taken over the region

$$R_y^k = \{ (y_1,\ldots,y_k) : \sum_{i=1}^{k} y_i = y \text{ and } y_i \equiv 1 \bmod k,\ i = 1,\ldots,k \}.$$

Also, from Lemma 4.5.1, we know that

$$A^{k+1}\{u_i\}_{y_1+\ldots+y_k+1}^{n+1} = q^y\, A^{k+1}\{u_i\}_1^{n-y+1}.$$

Therefore,

$$A^{k+1}\{u_i\}_1^{n+1} = \min_{y=1,\ldots,n} \{ q^y + (1-q^y)\, F^k + q^y\, A^{k+1}\{u_i\}_1^{n-y+1} \}, \qquad (4.6.3)$$

which is of exactly the same form as K(n). This proves the theorem.

Remark 4.6.1 Theorem 4.6.1 agrees with Lemma 4.2.1 which asserts that the cost of the k-sample testing procedure f_U is equal to the cost of the corresponding (k+1)-ary tree representing it.

The exact value of K(n) can be found by using the recursive expression (4.6.2), provided that the values $F(\ell)$, $\ell = 1,\ldots,n$ are known.

However, Theorem 4.5.2 and Theorem 4.5.3 only give the values of $F(\ell)$ for $0 \leq \ell \leq 2x+k-1$, where x is defined by the inequalities (4.5.4). Therefore, in order to compute K(n), it is essential for us to find an upper bound, say h, of the value y in the expression (4.6.2) such that $h \leq 2x+k-1$. The following lemma is established for this purpose.

Lemma 4.6.1 Let q be a real number such that $0 \leq q \leq 1$. If x is a positive integer satisfying the inequality

$$1 \geq q^{x} + \ldots + q^{x+k-1} + q^{x+k} ,$$

then we have

$$1 \geq (k+1)^{2} q^{2x+k} .$$

Proof: The lemma can be proved by two independent approaches, namely, the combinatorial approach and the analytical approach. The former is rather lengthy but straightforward. However, the later one requires the knowledge of defining a monotone decreasing function, namely $\phi(q)$, whose lower bound is of our interest.

(1) The combinatorial approach: for any integer r, we have $(1-q^{r})^{2} \geq 0$ which implies that

$$1 + q^{2r} \geq 2q^{r} . \tag{4.6.4}$$

The expression

$$(q^{x} + q^{x+1} + \ldots + q^{x+k-1} + q^{x+k})^{2}$$

can be written as a sum of the squares which equals

$$\sum_{i=1}^{k+1} \sum_{i=1}^{k+1} q^{a_{ij}}$$

where

$$a_{ij} = (x+i-1) + (x+j-1) = 2x-2+i+j, \text{ for } i,j = 1,\ldots,k+1. \tag{4.6.5}$$

Let $A = \{a_{ij}\}$ be a matrix of size $(k+1)\times(k+1)$ with a_{ij} defined by (4.6.5). Table 4.6.1 illustrates how the entries of A can be obtained by first assigning the value $(x + i - 1)$ at the i^{th} row and the i^{th} column, then obtaining a_{ij} by adding the values at the i^{th} row and the j^{th} column, for $i, j = 1,\ldots, k+1$.

r_i \ c_j	x	x+1	x+2	.....	x+k-1	x+k
x	2x	2x+1	2x+2	.....	2x+k-1	2x+k
x + 1	2x+1	2x+2	2x+3	.....	2x+k	2x+k+1
x + 2	2x+2	2x+3	2x+4	.....	2x+k+1	2x+k+2
....	....	....	....	.....	.....	.....
x + k	2x+k	2x+k+1	2x+k+2	.....	2x+2k-1	2x+k

Table 4.6.1 The entries of the matrix $A = \{a_{ij}\}$ defined by (4.6.5). Note that $a_{ij} = r_i + c_j$, for $i,j = 1,\ldots, k+1$.

Observe from Table 4.6.1 that when $i + j = k+2$, we have

$$a_{i,k+2-i} = 2x + 2 + (k + 2) = 2x + k, \text{ for } 1 \leq i \leq k+1 \tag{4.6.6}$$

On the other hand, inequality (4.6.4) implies that

$$q^{a_{ij}} + q^{a_{k+2-j,k+2-i}} = q^{2x-2+i+j}(1 - q^{2k+4-2i-2j})$$

$$\geq q^{2x-2+i+j}(q^{k+2-i-j}) = 2q^{2x+k} \tag{4.6.7}$$

for $i = 1,\ldots,k+2-j$ and $j = 1,\ldots,k+1$.

Since a_{ij} and $a_{k+2-j,k+2-i}$ are symmetric entries with respect to the diagonal which divides the matrix A into upper left and lower right triangles, that is, the diagonal with the number 2x+k in every entry. We thus obtain:

$$1 \geq \sum_{i=1}^{k+1}\sum_{i=1}^{k+1} q^{a_{ij}} = \sum_{j=1}^{k}\sum_{i=1}^{k+1-j} \left(q^{a_{ij}} + q^{a_{k+2-j,k+2-i}} \right) + \sum_{j=1}^{k} q^{a_{k+2-j,j}}$$

$$\geq \left(\frac{(k+1)^2}{2} - \frac{k+1}{2} \right) \left(2q^{2x+k} \right) + (k+1)q^{2x+k}$$

$$= (k+1)^2 q^{2x+k}$$

as a result of (4.6.6) and (4.6.7). This completes the proof by using the combinatorial approach.

(2) The analytical approach: consider the function

$$\phi(q) = q^{-k/2} \left(1 + q + q^2 + \ldots + q^k \right), \quad 0 \leq q \leq 1$$

It suffices to show that $\phi(q) \geq k+1$ for all q such that $0 \leq q \leq 1$. By gathering the i^{th} and the $(k+2-i)^{th}$ term together in a common bracket, for $i = 1, 2, \ldots,$ we can write

$$\phi(q) = \left(q^{-\frac{k}{2}} - q^{\frac{k}{2}} \right) + \left(q^{-\frac{k}{2}+1} + q^{\frac{k}{2}-1} \right) + \left(q^{-\frac{k}{2}+2} + q^{\frac{k}{2}-2} \right) + \ldots$$

$$+ \left(q^{-\frac{k}{2}+\left\lfloor\frac{k}{2}\right\rfloor} + q^{\frac{k}{2}-\left\lfloor\frac{k}{2}\right\rfloor} \right) + R$$

where $R = \begin{cases} 0 & \text{when k is even} \\ q^{1/2} & \text{when k is odd.} \end{cases}$

Taking the derivative of $\phi(q)$ with respect to q, we find that

$$\phi'(q) = \frac{k}{2} q^{-1} \left(-q^{-\frac{k}{2}} + q^{\frac{k}{2}}\right) + \left(\frac{k}{2}+1\right) q^{-1} \left(-q^{-\frac{k}{2}+1} + q^{\frac{k}{2}-1}\right)$$

$$+ \left(-\frac{k}{2}+2\right) q^{-1} \left(q^{\frac{k}{2}+2} + q^{\frac{k}{2}-2} \right) + \ldots + \left(-\frac{k}{2}+\left\lfloor\frac{k}{2}\right\rfloor\right) q^{-1} \left(q^{-\frac{k}{2}+\left\lfloor\frac{k}{2}\right\rfloor} + q^{\frac{k}{2}-\left\lfloor\frac{k}{2}\right\rfloor}\right) + R'$$

where $R' = \begin{cases} 0 & \text{when k is even} \\ \frac{1}{2} q^{-1/2} & \text{when k is odd.} \end{cases}$

Therefore, $\phi'(q) \leq 0$ for any q in the closed interval [0, 1]. Thus we conclude that $\phi(q)$ is a decreasing function of q in [0, 1] and $\phi(q) \geq \phi(1)$ for any q such that $0 \leq q \leq 1$. Since $\phi(1) = k+1$, the result of the lemma follows.

Theorem 4.6.2 If x is the integer defined by the inequalities (4.5.4), then

$$K(n) = \min_{1 \leq y \leq \min\{2x+k-1,n\}} \{q^y + q^y K(n-y) + (1-q^y)F^k(y)\}. \qquad (4.6.8)$$

Proof: It suffices to prove for the case that $n > 2x+k-1$. Theorem 4.6.1 asserts that K(n) equals the cost of an optimal alphabetic (k+1)-ary tree T_U for the sequence of weights $u_1,\ldots,u_{n+1}$, where $u_i = pq^{i-1}$ for $i = 1,\ldots,n$ and $u_{n+1} = q^n$. It follows from Theorem 3.4.3 that the tree T_U is also an optimal tree for the sequence of weights $u_1,\ldots, u_{n+1}$ since it is a valley sequence of weights.

Let $v_1,\ldots, v_{k+1}$ be the left to right sequence of vertices of path length one on T_U. Let $w(v_i)$ be the weight associated to the vertex v_i, for $i = 1,\ldots, k+1$. By definition, we have

$$\sum_{i=1}^{k+1} w(v_i) = 1 . \qquad (4.6.9)$$

From the proof of Theorem 4.6.1, we know that the number y in (4.6.8) represents the total number of end vertices at the k leftmost planted subtrees rooted at the root of the tree T_U. Therefore,

$$w(v_{k+1}) = 1 - \sum_{i=1}^{k} w(v_i) = 1 - \sum_{i=1}^{y} pq^{i-1} = q^y . \qquad (4.6.10)$$

Let ℓ be the total number of end vertices of path length one on T_U. Then $0 \leq \ell \leq k+1$ since we assume that $n > 2x+k-1$. Now we claim that the value

of ℓ must be one of the following three values: 0, k, and k+1.

When $0 < \ell < k+1$, $v_1, \ldots, v_\ell$ must be the end vertices of T_U since the weights $u_1, \ldots, u_\ell$ are monotone decreasing when $\ell \leq k$; on the other hand, $v_{\ell+1}, \ldots, v_{k+1}$ must be the branch vertices. By Theorem 3.4.1, the optimality of T_U implies that each of the k+1 vertices of path length two joined to the vertices v_i, $i=\ell+1, \ldots, k$ is of weight less that or equal to $w(v_1) = p$. Thus we have

$$w(v_i) \leq (k+1)\, p\ , \qquad i = \ell+1, \ldots, k. \tag{4.6.11}$$

Now

$$\sum_{i=1}^{\ell} w(v_i) = \sum_{i=1}^{\ell} pq^{i-1} = 1 - q^{\ell}\ . \tag{4.6.12}$$

As a result of (4.6.9), (4.6.10), (4.6.11) and (4.6.12), we have

$$1 - q^{\ell} + (k-\ell)(k+1)\, p + q^{y} \geq 1\ ,$$

or equivalently,

$$q^{y} \geq q^{\ell} - (k-\ell)(k+1)\, p \geq q^{\ell}\ .$$

This is impossible unless $\ell = k$ since we know that $y \geq k$.

For the cases that $\ell = k$ and $\ell = k+1$, the theorem holds obviously and the proof is trivial.

For the case that $\ell = 0$, then $v_1, \ldots, v_{k+1}$ are all branch vertices. The optimality of T_U implies that

$$w(v_i) \leq (k+1)\, w(v_{k+1})\ , \qquad i = 1, \ldots, k, \tag{4.6.13}$$

since each of the k+1 vertices of path length two joined to v_i is of weight less that or equal to $w(v_{k+1})$. Now (4.6.9), (4.6.10) and (4.6.13) imply that

$$k(k+1)\, q^{y} + q^{y} \geq 1\ ,$$

or equivalently,

$$q^{y} \geq \frac{1}{k(k+1)+1} > \frac{1}{(k+1)^2} \geq q^{2x+k},$$

where the last inequality is obtained from Lemma 4.6.1. Thus we conclude that $y \leq 2x+k-1$. This completes the proof.

From the proof of Theorem 4.6.2, we obtain

Corollary 4.6.1 Let T_U be an optimal alphabetic (k+1)-ary tree for the sequence of weights $u_1, \ldots, u_{n+1}$, where $u_i = pq^{i-1}$ for $i = 1, \ldots, n$ and $u_{n+1} = q^n$. Let y be the total number of end vertices at the k leftmost planted subtrees rooted at the root of T_U. Then we have $y \leq 2x+k-1$.

Remark 4.6.2 The above results can be found in the paper by Hwang (1976) for the case that k = 1. Since the value of $F^k(y)$ is given by Theorem 4.5.2 when $y \leq x+k$ and given by Theorem 4.5.3 when $x \leq y \leq 2x+k-1$, Theorem 4.6.2 enables us to compute K(n) recursively within 2x+k-1 steps.

Let y^* be the value of y in the expression (4.6.8) such that

$$K(n) = q^{y^*} + q^{y^*} K(n-y^*) + (1 - q^{y^*}) F^k(y^*).$$

Then y^* denotes the total sum of units tested in the first k-sample test of the procedure f_U. The exact sizes of the first k samples tested can be found by considering the optimal alphabetic (k+1)-ary tree T_U which represents f_U. This is because the leftmost y^* end vertices of the tree T_U are also the end vertices of the optimal alphabetic k-sum (k+1)-ary tree for the sequence of weights $u_i = pq^{i-1}$, $i = 1, \ldots, y^*$, which can always be constructed.

4.7 Asymptotic Properties of the Cost Function

Suppose I is a binomial population with infinitely many units, each with a probability p of being defective and a probability $q = 1 - p$ of being good. The problem of finding a single defective from I is similar to the case when I is a finite population discussed in Section 4.6. In this case, we can even find a closed form solution for the cost $E(f_\infty)$ of a k-sample optimal nested group testing procedure f_∞ for solving the above stated problem. The procedure f_∞ is in fact an extension of the procedure f_U when the given population is infinite in size.

Let $J^o = (J_1^o, \ldots, J_k^o)$ be the first k-tuple of samples tested in the procedure f_∞. Let y_i be the size of the sample J_i^o, $i = 1, \ldots, k$, $y_o \equiv 0$ and $y = \sum_{i=1}^{k} y_i$. The probability that J_i^o is the least index defective sample equals $q^{y_o+\ldots+y_{i-1}}(1 - q^{y_i})$ and the minimal expected cost of finding a single defective from J_i^o when J_i^o is the least index defective sample equals $F^{k+1}(y_i)$, defined by (4.5.1), for $i = 1, \ldots, k$. The probability the all the k samples are good equals q^y and the minimal expected cost of finding a single defective from the rest of the units in I when all these y units are good equals $E(f_\infty)$ again, since I contains infinitely many units. Therefore, if we define the function

$$E_y = \min_{R_y^k} \{ 1 + \sum_{i=1}^{k} q^{y_o+\ldots+y_{i-1}}(1 - q^{y_i}) F(y_i) \} + q^y E_y \qquad (4.7.1)$$

where the minimum is taken over ther region

$$R_y^k = \{ (y_1, \ldots, y_k): \sum_{i=1}^{k} y_i = y \text{ and } y \equiv 1 \bmod k, \ i = 1, \ldots, k \}, \qquad (4.7.2)$$

then we have

$$E(f_\infty) = \min_{y=1,2,\ldots} E_y \ .$$

In terms of the function $F^k(y)$ defined by (4.5.2), we write,

$$E_y = \frac{q^y}{1-q^y} + F^k(y) \qquad (4.7.3)$$

and

$$E(f_\infty) = \min_{y=1,2,\ldots} \left\{ \frac{q^y}{1-q^y} + F^k(y) \right\}. \qquad (4.7.4)$$

The following theorems are identical to those obtained by Hwang (1974) for the case that k = 1.

Theorem 4.7.1 $E(f_\infty)$ equals to the cost of an optimal alphabetic (k+1)-ary tree for the sequence of weights $u_i = pq^{i-1}$, i = 1,2,... .

Proof: Let T_∞ be an optimal alphabetic (k+1)-ary tree for the infinte sequence of weights u_i, i = 1,2,... . Then there are k+1 planted (k+1)-ary subtrees rooted at the root of T_∞, namely, T_1, T_2,..., T_{k+1} from the right to the left. Let there be y_i end vertices at T_i, i=1,...,k and $y = \sum_{i=1}^{k} y_i$. Lemma 3.4.2 implies that these k subtrees T_1,..., T_k form an optimal alphabetic k-sum (k+1)-ary tree for the sequence of weights $u_i = pq^{i-1}$, for i = 1,..., y. On the other hand, the tree T_{k+1} is an optimal alphabetic (k+1)-ary tree for the infinite sequence of weights $u_i = pq^{i-1}$, i=y+1,y+2,... . Now, if we define the function

$$A_y = \min_{R_y^k} \{ 1 + \sum_{i=1}^{k} q^{y_0+\ldots+y_{i-1}} (1 - q^{y_i}) A^{k+1}(y_i) \} + q^y A_y \ ,$$

where the minimum is taken over the region R_y^k defined by (4.7.2) and the the function $A^{k+1}(y_i)$ is defined by (4.5.2), then by a similar argument as the one used in the proof of Theorem 4.6.1, we are able to conclude that the cost of the tree T_∞ equals

$$A\{u_i\}_1^\infty = \min_{y=1,2,\ldots} \{A_y\}.$$

But the expression for A_y can be simplified by means of the relation (4.5.3), so that

$$A_y = (1-q^y)A^k(y) + q^y A_y + q^y$$

and therefore,

$$A\{u_i\}_1^\infty = \min_{y=1,2,\ldots} \left\{ \frac{q^y}{1-q^y} + A^k(y) \right\},$$

which is of the same form as the function $E(f_\infty)$, since Theorem 4.5.1 implies that $A^k(y) = F^k(y)$. This completes the proof.

Theorem 4.7.2 $E(f_\infty) = \min_{1 \le y \le 2x+k-1} \{E_y\}$, where the function E_y is defined by (4.7.1) and x is the integer defined by (4.5.4).

Proof: This is a result of Theorem 4.7.1 and Theorem 4.6.2, since the upper bound of y in expression (4.6.8) is independent of n when n approaches infinity.

Remark 4.7.1 Theorem 4.7.2 enables us to find the exact value of $E(f_\infty)$, since the values of E_y, for $o \le y \le 2x+k-1$, can be computed by using Theorem 4.5.2 and Theorem 4.5.3. Here we give the closed form solutions for E_y when $0 \le y \le x+k$ or $x \le y \le 2x+k-1$. They will be frequently used in the proofs of our main results.

(i) When $y \leq x + k$ and $y \equiv 0 \bmod k$, we have

$$E_y = \frac{q^y}{1-q^y} + 1 + \alpha + \frac{q^{y-\frac{\beta}{k}(k+1)} - q^y}{1 - q^y}$$

$$= 1 + \alpha + \frac{q^{y-\frac{\beta}{k}(k+1)}}{1 - q^y} \qquad (4.7.5)$$

where (α, β) is a pair of non-negative integers satisfying the equation

$$y = k(k+1)^{\alpha} + \beta \quad , \qquad 0 \leq \beta \leq k^2(k+1)^{\alpha} \qquad (4.7.6)$$

(ii) When $x \leq y \leq 2x+k-1$ and $y \equiv 0 \bmod k$, we have

$$E_y = \frac{q^y}{1-q^y} + 1 + \alpha + \frac{1}{1-q^y}\left\{ q^{m-\frac{\beta}{k}(k+1)-r} + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r} - 2q^y \right\}$$

$$= 1 + \alpha + \frac{1}{1-q^y}\left\{ q^{m-\frac{\beta}{k}(k+1)-r} + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r} - q^y \right\} \qquad (4.7.7)$$

where $h = y - x$, $m = x + h - \left\lfloor \frac{h}{k+1} \right\rfloor k$, (α, β) is a pair of non-negative integers satisfying the equation

$$m = k(k+1)^{\alpha} + \beta \qquad 0 \leq \beta \leq k^2(k+1)^{\alpha} \qquad (4.7.8)$$

and

$$r = \min\left\{ \left\lfloor \frac{h}{k+1} \right\rfloor, \left\lfloor \frac{m - \frac{\beta}{k}(k+1) + k - a}{k+2} \right\rfloor \right\}, \text{ where } h \equiv a \bmod (k+1) \text{ and } a \leq k. \qquad (4.7.9)$$

It is easy to verify that the expression (4.7.7) can be reduced to the expression (4.7.5) when $x \leq y \leq x+k$.

Now we can prove the following two lemmas which lead to the main result. The proof of the second one seem to be more complicated than the proof when $k = 1$, given by Hwang (1974).

Lemma 4.7.1 $$\min_{\substack{1 \le y \le x+k-1 \\ y \equiv 0 \bmod k}} E_y = E_{x_o}, \qquad (4.7.10)$$

where $x_o = \left\lfloor \frac{x+k-1}{k} \right\rfloor k$ and x is the integer defined by the inequalities (4.5.4).

Proof: It suffices to show that E_y is a monotone decreasing function of y for $1 \le y \le x+k-1$. Let $y' = y + k$ for any integer y such that $1 \le y \le x - 1$ and $y \equiv 0 \bmod k$. There always exist pairs of non-negative integers (α, β) and (α', β') satisfying

$$\left.\begin{aligned} y &= k(k+1)^{\alpha} + \beta \qquad 0 \le \beta \le k^2(k+1)^{\alpha} \\ y &= k(k+1)^{\alpha'} + \beta' \qquad 0 \le \beta' \le k^2(k+1)^{\alpha'} \\ \alpha' &= \alpha \qquad \beta' = \beta + k. \end{aligned}\right\} \qquad (4.7.11)$$

Writing E_y and $E_{y'}$ in the form given by (4.7.5) of Remark 4.7.1, we find that

$$E_{y'} - E_y = \left\{1 + \alpha + \frac{q^{y' - \frac{\beta'}{k}(k+1)}}{1 - q^{y'}}\right\} - \left\{1 + \alpha + \frac{q^{y - \frac{\beta}{k}(k+1)}}{1 - q^{y}}\right\}$$

$$= \frac{q^{y - \frac{\beta}{k}(k+1) - 1} (1 - q)\left[1 - q^{y}(1 + q + \ldots + q^{k})\right]}{(1 - q^{y+k})(1 - q^{y})}$$

$$\le 0$$

since $q^{y}(1 + q + \ldots + q^{k}) \ge 1$ when $y \le x - 1$. (Cf. inequalities (4.5.4)). Hence E_y is a monotone decreasing function of y for $1 \le y \le x+k-1$. However, because of the restriction that $y \equiv 0 \bmod k$, x_o is the smallest value of y such that (4.7.10) is satisfied. This completes the proof.

Lemma 4.7.2 $$\min_{\substack{x \le y \le 2x+k-1 \\ y \equiv 0 \bmod k}} E_y = E_{x_o}, \qquad (4.7.12)$$

where x and x_o are the same as those of Lemma 4.7.1.

Proof: It suffices to show that E_y is a monotone increasing function of y for $x \le y \le 2x+k-1$. Let $y' = y + k$. We are required to show that $E_{y'} - E_y \ge 0$ for any integer y such that $x \le y \le 2x - 1$. In order to express $E_{y'}$ and E_y in the form of (4.7.7), we define, for a given y. the parameters h, m, (α, β), r and a according to the equations (4.7.8) and (4.7.9) of Remark 4.7.1. Similarly, we define, for a given y', the parameters h', m', (α', β'), r' and a'.

Since $h \equiv a \bmod(k+1)$ and $0 \le a \le k$, consider the two possible cases: (i) $a = 0$, or (ii) $0 < a \le k$. We claim that the quantity

$$d_k = (1 - q^y)(1 - q^{y'})(E_{y'} - E_y). \qquad (4.7.13)$$

is always non-negative in each of the above two cases.

Case (i) When $a = o$, then $a' = k$. We have

$h' = y' - x = y + k - x = (x + h) + k - x = h + k$ and

$\left\lfloor \frac{h'}{k+1} \right\rfloor = \left\lfloor \frac{h+k}{k+1} \right\rfloor = \left\lfloor \frac{h}{k+1} \right\rfloor = \frac{h}{k+1}$. Hence

$$m' = x + h' - \left\lfloor \frac{h'}{k+1} \right\rfloor k = x + (h+k) - \left\lfloor \frac{h}{k+1} \right\rfloor k = \left(x + h - \left(\frac{h}{k+1}\right)k\right) + k = m + k.$$

We can choose (α', β') satisfying $m' = k(k+1)^{\alpha'} + \beta'$, $0 \le \beta' \le k^2(k+1)^{\alpha'}$ and and also $\alpha' = \alpha$ and $\beta' = \beta + k$. Now

$$r' = \min\left\{\left\lfloor \frac{h'}{k+1} \right\rfloor, \left\lfloor \frac{m' - \frac{\beta'}{k}(k+1) + k - a}{k+2} \right\rfloor\right\} = \min\left\{\frac{h}{k+1}, \left\lfloor \frac{m + \frac{\beta}{k}(k+1) - 1}{k+2} \right\rfloor\right\}$$

$$= r \text{ or } r - 1. \qquad (4.7.14)$$

(a) If $r' = r$, (4.7.13) and (4.7.7) imply that, for $0 \le y \le 2x-1$, we have

$$d_k = (1 - q^y)\left(q^{m+k-\frac{\beta+k}{K}(k+1)-r} - q^{m+k-\frac{h}{k+1}+(k+1)r} - q^{y+k} \right)$$

$$- (1 - q^{y+k})\left(q^{m-\frac{\beta}{k}(k+1)-r} - q^{m-\frac{h}{k+1}+(k+1)r} - q^{y} \right)$$

$$= q^y(1 - q^k) + q^{m-\frac{\beta}{k}(k+1)-r-1}(1 - q) - q^{y+m-\frac{\beta}{k}(k+1)-r-1}(1 - q^{k+1})$$

$$- q^{m-\frac{h}{k+1}+(k+1)r}(1 - q^k).$$

Since $y = x + h$ and $q^y(1 - q^{k+1}) = q^{x+h}(1 - q)(1+q+\ldots+q^k) \le q^h(1 - q)$, so

$$d_k \ge q^{x+h}(1 - q^k) + q^{m-\frac{\beta}{k}(k+1)-r-1}(1 - q) - q^{h+m-\frac{\beta}{k}(k+1)-r-1}(1 - q)$$

$$- q^{m-\frac{h}{k+1}+(k+1)r}(1 - q^k). \qquad (4.7.15)$$

First suppose $r = \left\lfloor \frac{h}{k+1} \right\rfloor = \frac{h}{k+1}$, then $(k+1)r = h$ implies that

$$q^{m-\frac{h}{k+1}+(k+1)r} = q^{x+h-\left(\frac{h}{k+1}\right)k-\frac{h}{k+1}+h} = q^{x+h}.$$

Hence

$$d_k \ge q^{m-\frac{\beta}{k}(k+1)-r-1}(1 - q)(1 - q^h) \ge 0.$$

Next suppose $r = \left\lfloor \frac{m-\frac{\beta}{k}(k+1)+k}{k+2} \right\rfloor$. From (4.7.14), we know that $r' = \left\lfloor \frac{m-\frac{\beta}{k}(k+1)-1}{k+2} \right\rfloor$. Then $r' = r$ implies that $(k+2)r = m - \frac{\beta}{k}(k+1)-1$ must be true.

Besides, $m - \frac{h}{k+1} = x + h - \frac{h}{k+1}k - \left(\frac{h}{k+1}\right) = x$. Hence (4.7.15) implies that

$$\begin{aligned} d_k &\geq q^{x+h}(1 - q^k) + q^{(k+1)r}(1 - q) - q^{h+(k+1)r}(1 - q) - q^{x+(k+1)r}(1 - q^k) \\ &\geq q^{(k+1)r}(1 - q)(1 - q^h)\{-q^x(1 + q + \ldots + q^k) + q^{x+k} + 1\} \\ &\geq q^{(k+1)r}(1 - q)(1 - q^h)\{-1 + q^{x+k} + 1\} \\ &\geq 0. \end{aligned}$$

(b) If $r' = r - 1$, (4.7.13) and (4.7.7) imply that, for $0 \leq y \leq 2x-1$, we have

$$\begin{aligned} d_k &= (1 - q^y)\left(q^{m+k-\frac{\beta+k}{k}(k+1)-(r-1)} + q^{m+k-\frac{h}{k+1}+(k+1)(r-1)} - q^{y+k}\right) \\ &\quad - (1 - q^{y+k})\left(q^{m-\frac{\beta}{k}(k+1)-r} + q^{m-\frac{h}{k+1}+(k+1)r} - q^y\right) \\ &= q^y(1 - q^k)(1 - q^{m-\frac{\beta}{k}(k+1)-r}) + q^{m-\frac{h}{k+1}+(k+1)r-1}(1-q)\{1 - q^y(1+q+\ldots+q^k)\}. \end{aligned}$$

But $q^y(1 + q + \ldots + q^k) = q^{x+h}(1 + q + \ldots + q^k) \leq q^h$. Hence

$$\begin{aligned} d_k &\geq q^y(1 - q^k)(1 - q^{m-\frac{\beta}{k}(k+1)-r}) + q^{m-\frac{h}{k+1}+(k+1)r-1}(1 - q)(1 - q^h) \\ &\geq 0, \end{aligned}$$

since $m - \frac{\beta}{k}(k+1) - r \geq 0$ for it is the number of end vertices with path length $\alpha + 1$ on the k-sum (k+1)-ary tree constructed in the proof of Theorem 4.5.3.

Case (ii) When $0 < a \le k$, then $a' = a - 1$. We have

$h' = y' - x = y + k - x = (x+h) + k - x = h + k$ and

$\left\lfloor \frac{h'}{k+1} \right\rfloor = \left\lfloor \frac{h+k}{k+1} \right\rfloor = 1 + \left\lfloor \frac{h}{k+1} \right\rfloor$. Hence

$$m' = x + h' - \left\lfloor \frac{h'}{k+1} \right\rfloor k = x + (h+k) - \left(1 + \left\lfloor \frac{h}{k+1} \right\rfloor \right) k = x + h - \left\lfloor \frac{h}{k+1} \right\rfloor k = m.$$

Now we can choose (α', β') satisfying $m' = k(k+1)^{\alpha'} + \beta'$, $0 \le \beta' \le k^2(k+1)^{\alpha'}$ and also $\alpha' = \alpha$ and $\beta' = \beta$. Now

$$r' = \min\left\{ \left\lfloor \frac{h'}{k+1} \right\rfloor, \left\lfloor \frac{m' - \frac{\beta'}{k}(k+1) + k - a'}{k+2} \right\rfloor \right\} = \min\left\{ 1 + \left\lfloor \frac{h}{k+1} \right\rfloor, \left\lfloor \frac{m - \frac{\beta}{k}(k+1) + k - a + 1}{k+2} \right\rfloor \right\}$$

$$= r \text{ or } r+1. \qquad (4.7.16)$$

(a) If $r' = r + 1$, (4.7.13) and (4.7.7) imply that, for $0 \le y \le 2x-1$, we have

$$d_k = (1 - q^y)\left(q^{m - \frac{\beta}{k}(k+1) - (r+1)} + q^{m - \left(1 + \left\lfloor \frac{h}{k+1} \right\rfloor\right) + (k+1)(r+1)} - q^{y+k} \right)$$

$$- (1 - q^{y+k})\left(q^{m - \frac{\beta}{k}(k+1) - r} + q^{m - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r} - q^y \right)$$

$$= q^y(1 - q^k) + q^{m - \frac{\beta}{k}(k+1) - r - 1}(1 - q) - q^{y + m - \frac{\beta}{k}(k+1) - r - 1}(1 - q^{k+1})$$

$$- q^{m - \frac{h}{k+1} + (k+1)r}(1 - q^k)$$

$$\ge q^y(1 - q^k) + q^{m - \frac{\beta}{k}(k+1) - r - 1}(1 - q) - q^{h + m - \frac{\beta}{k}(k+1) - r - 1}(1 - q)$$

$$- q^{m - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r}(1 - q^k), \qquad (4.7.17)$$

since $q^y(1 - q^{k+1}) = q^{y-x}(1 - q)q^x(1+q+\ldots+q^k) \le q^{y-x}(1 - q) = q^h(1 - q)$.

First suppose $r = \left\lfloor \frac{h}{k+1} \right\rfloor$, then $(k+1)r = (k+1)\left\lfloor \frac{h}{k+1} \right\rfloor$. Hence

$$d_k \geq (1 - q^k)\left(q^{y} - q^{m - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)\left\lfloor \frac{h}{k+1} \right\rfloor} \right) + q^{m - \frac{\beta}{k}(k+1) - r - 1}(1 - q^h)(1 - q)$$

$$\geq 0 ,$$

since $m - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)\left\lfloor \frac{h}{k+1} \right\rfloor = x + h - \left\lfloor \frac{h}{k+1} \right\rfloor k - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)\left\lfloor \frac{h}{k+1} \right\rfloor = x+h = y \geq 0.$

Next suppose $r = \left\lfloor \frac{m - \frac{\beta}{k}(k+1) + k - a}{k+2} \right\rfloor$. From (4.7.16), we know that

$r' = \left\lfloor \frac{m - \frac{\beta}{k}(k+1) + k - a + 1}{k+2} \right\rfloor$. Then $r' = r + 1$ implies that

$(k+2)r = m - \frac{\beta}{k}(k+1) = k - a - (k-1) = m - \frac{\beta}{k}(k+1) - a - 1$, or

$m - \frac{\beta}{k}(k+1) - r - 1 = (k+1)r + a$. Besides,

$$m - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r = x + h - \left\lfloor \frac{h}{k+1} \right\rfloor k - \left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r = y - (k+1)\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r.$$

Hence, (4.7.16) implies that

$$d_k \geq q^y(1 - q^k) + q^{(k+1)r+a}(1 - q) - q^{h+(k+1)r+a}$$

$$- q^{y-(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r}(1 - q^k)$$

$$= - q^{y-(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r}(1 - q^k)\left(1 - q^{(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor - (k+1)r}\right)$$

$$+ q^{(k+1)r+a}(1 - q^h)(1 - q)$$

$$\geq - q^{y-(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor + (k+1)r}(1 - q^k)(1 - q^h) + q^{(k+1)r+a}(1 - q^h)(1 - q)$$

since $q^{(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor - (k+1)r} \geq q^{h-(k+1)r} \geq q^h$.

Now we have

$$d_k \geq (1-q)(1-q^h)\, q^{(k+1)r} \{ q^a - q^{y-(k+1)r\left\lfloor \frac{h}{k+1} \right\rfloor} (1+q+\ldots+q^k) + q^{y-(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor + k} \}$$

$$\geq (1-q)(1-q^h)\, q^{(k+1)r} \{ q^a - q^{h-(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor} + q^{y-(k+1)\left\lfloor \frac{h}{k+1} \right\rfloor + k} \}$$

$$\geq (1-q)(1-q^h)\, q^{(k+1)r} \{ q^{y-(k+1)r\left\lfloor \frac{h}{k+1} \right\rfloor + k} \}$$

$$\geq 0,$$

since by definition, $h - (k+1)\left\lfloor \frac{h}{k+1} \right\rfloor = h - (k+1)\left(\frac{h-a}{k+1}\right) = a$.

(b) If $r' = r$, (4.7.13) and (4.7.7) imply that, for $0 \leq y \leq 2x-1$, we have

$$d_k = (1-q^y)\left(q^{m-\frac{\beta}{k}(k+1)-r} + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor -1+(k+1)r} - q^{y+k} \right)$$

$$- (1-q^{y+k})\left(q^{m-\frac{\beta}{k}(k+1)-r} + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor +(k+1)r} - q^{y} \right)$$

$$= q^y(1-q^k)(1-q^{m-\frac{\beta}{k}(k+1)-r}) + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor -1+(k+1)r}(1-q)\{1-(1+q+\ldots+q^k)q^y\}$$

$$\geq q^y(1-q^k)(1-q^{m-\frac{\beta}{k}(k+1)-r}) + q^{m-\left\lfloor \frac{h}{k+1} \right\rfloor -1+(k+1)r}(1-q)(1-q^{h-x})$$

$$\geq 0,$$

since $y - x = h \geq 0$ and $m - \frac{\beta}{k}(k+1) - r \geq 0$ for it is the number of end vertices with path lenth $\alpha+1$ on the k-sum (k+1)-ary tree constructed in the proof of Theorem 4.5.3.

Therefore, we can conclude that E_y is a monotone increasing function of y for $x \leq y \leq 2x+k-1$. However, because of the restriction

that $x \leq y \leq 2x+k-1$, x_o is the smallest value of y such that (4.7.12) is satisfied. This completes the proof.

The following theorem generalizes the result of Hwang's (1974).

Theorem 4.7.3 $E(\ _\infty) = E_{x_o} = 1 + \alpha + \dfrac{x_o - \frac{\beta}{k}(k+1)}{1 - q^{x_o}}$,

where $x_o = \dfrac{x + k - 1}{k} k$, x is the non-negative integer defined by the inequalities (4.5.4) and (α, β) is a pair of non-negative integers satisfying

$$x_o = k(k+1)^{\alpha} + \beta, \quad 0 \leq \beta \leq k^2(k+1)^{\alpha} \tag{4.7.18}$$

Proof: This is a result of Theorem 4.7.2, Lemma 4.7.1 and Lemma 4.7.2. We note that the existence of (α, β) has been discussed in Remark 4.5.1.

Now we reach a conclusion that if we are given an infinite population with units from a binomial distribution such that each unit has the same probability p of being defective and the probability $q = 1 - p$ of being good, then the optimal nested k-sample group testing procedure for finding a single defective can be represented by an optimal alphabetic (k+1)-ary tree T_∞ for the infinite sequence of weights $u_i = pq^{i-1}$, $i = 1, 2, \ldots$. The total number of end vertices at the k leftmost planted subtrees rooted at the root of T_∞ is $x_o = \left\lfloor \dfrac{x+k-1}{k} \right\rfloor k$, where x is the non-negative integer defined by the inequalities (4.5.4). Furthermore, these k subtrees form an optimal alphabetic k-sum (k+1)-ary tree T_{x_o} for the sequence of weights

$u_i = pq^{i-1}$, $i = 1, \ldots, x_o$. Since $x_o \leq x + k - 1$, from the proof of Theorem 4.5.2, we see that the tree T_{x_o} is also a perfect k-sum (k+1)-ary tree. Suppose (α, β) is defined by (4.7.18). Then the first $x_o - \frac{\beta}{k}(k+1)$ end vertices of the tree T_{x_o} is of path length $\alpha + 1$ and the last $\frac{\beta}{k}(k+1)$ end vertices of T_{x_o} is of path length $\alpha + 2$.

Practically, the above statement means that the total number of units tested in the first k-sample group test is x_o. More precisely, let n_i be the total number of units tested in the i^{th} sample among the first k samples tested, $i = 1, \ldots, k$, then

$$\sum_{i=1}^{k} n_i = x_o$$

and

$$n_i = \begin{cases} (k+1)^{\alpha} \text{ for } i = 1, \ldots, \dfrac{x_o - \frac{\beta}{k}(k+1)}{(k+1)^{\alpha}} & \text{if } \dfrac{x_o - \frac{\beta}{k}(k+1)}{(k+1)^{\alpha}} \geq 1, \\ x_o - \dfrac{x_o - \frac{\beta}{k}(k+1)}{(k+1)^{\alpha}}(k+1)^{\alpha} - \dfrac{\beta}{k(k+1)^{\alpha}}(k+1)^{\alpha} \text{ for } i = \dfrac{x_o - \frac{\beta}{k}(k+1)}{(k+1)^{\alpha}} + 1 & \\ \qquad \text{if } \dfrac{x_o - \frac{\beta}{k}(k+1)}{(k+1)^{\alpha}} < k - \dfrac{\beta}{k(k+1)^{\alpha}}, & \\ (k+1)^{\alpha+1} \text{ for } i = k - \dfrac{\beta}{k(k+1)^{\alpha}} + 1, \ldots, k & \text{if } \beta \geq k(k+1)^{\alpha}. \end{cases}$$

CHAPTER V

CONCLUSIONS AND SUGGESTIONS FOR FUTURE STUDY

5.1 Contributions of the Thesis

In this thesis an attempt has been made to study two types of k-sample testing problems, namely, the testing of goodness-of-fit or homogeneity between k populations by using k-sample analogues of Kolmogorov-Smirnov statistics; and the binomial group testing for eliminating defectives by using an optimal nested k-sample group testing procedure. Several enumeration methods and results about the tree structures are included as a foundation of the study. The major contributions of the the thesis may be summarized as follows:

(1) A method of enumeration is used so that simple expressions can be obtained in the following two cases:

(a) a certain k×m fold summation of the number 1 is expressed as a determinant of size $(m+k)\times(m+k)$.

(b) a certain k×m fold definite integral of the number 1 is expressed as a determinant of size $(m+k)\times(m+k)$.

(2) Based on the formulae of (1), the null distributions and the conditional null distributions of the following statistics have been found:

(a) a k-sample analogue of the two-sample Kolmogorov-Smirnov statistics for testing the homogeneity hypothesis.

(b) a k-sample analogue of the one-sample Kolmogorov-Smirnov statistics for testing the goodness-of-fit hypothesis.

The testing procedure of using the above statistics has been proposed and the consistency property of the tests has also been mentioned.

(3) Rooted plane trees have been identified as matrices or pseudo-search codes. Using generating functions or matrix enumeration methods, the total number of distinct planted plane trees has been enumerated for each of the following cases:

(a) given n end vertices

(b) given height h and n end vertices

(c) given all branch vertices of degree q+1, height h and n end vertices.

(d) given all branches of degrees either 3 or 4, and n end vertices.

(e) with certain restrictions on the end vertex sequences and degree sequences.

(4) Further exploratory results have also been obtained for the case of q-ary rooted plane trees. For example, we have

(a) proposed a pseudo-search code construction algorithm for constructing an optimal alphabetic q-ary tree for a valley sequence of weights.

(b) determined in terms of entropy, the lower and upper bounds of the cost of an optimal q-ary tree.

(5) As an application of (3) and (4), a k-sample group testing procedure has been defined and the following problems have been solved:

(a) eliminating a single defective from a binomial population P by using an optimal nested k-sample group testing procedure.

(b) finding all the defectives from the population P.

5.2 Suggestions for Future Study

There are numerous directions in which the present study can be extended. A few of them are listed below:

(1) The enumeration methods used in Chapter I should be improved so that more specified classes of matrices satisfying the conditions given by Remark 1.2.1 or Remark 1.3.1 can be enumerated without using iteration methods. Such a problem is of importance because its solution will enable us to improve the conditional null distributions obtained in Chapter II. (cf. Remark 2.2.3 and Remark 2.3.1)

(2) The k-sample nested group testng procedure proposed in Chapter IV has been shown to be optimal for finding a single defective from a given binomial population P. However, this procedure is not optimal for classifying all the defectives from P. (See Section 4.1 for the case k = 1) It does not seem easy to find such an optimal procedure because it does not have a tree representation even when k = 1. However, one can consider k-sample group testing on a hypergeometric population, that is, a binomial population with a known number of defectives. Also, it is interesting to enumerate a class of nested k-sample group testing procedures for classifying all the defectives from a given population.

BIBLIOGRAPHY

[1] Abramson, N. (1963). *Information theory and coding.* McGraw-Hill, New York.

[2] Ahmad R. (1977). *On the multivariate k-sample problem and generalization of Kolmogorov-Smirnov test.* Ann. Inst. Statist. 48 (2), pp. 259-265.

[3] Anderson, T.W. (1962). *On the distribution of the two-sample Cramér-von Misses criterion.* Ann. Math. Statist. 33, pp. 1148-1159.

[4] Ash, R. (1965). *Information theory.* Interscience Tracts in Pure and Applied Mathematics, No. 19, New York.

[5] Bickel, P. J. (1969). *A distribution free version of the Smirnov two-sample test in the p-variate case.* Ann. Math. Statist. 40, pp. 1-23.

[6] Chorneyko, I.Z. and Mohanty, S.G. (1972). *On the enumeration of pseudo-search codes.* Studia Scientiarum Mathematicarum Hungarica 7, pp. 47-54.

[7] Chorneyko, I.Z. and Mohanty, S.G. (1975). *On the enumeration of certain sets of planted plane trees.* Journal of Combinatorial Theory 18, pp. 209-221.

[8] Conover, W.J. (1965). *Several k-sample Kolmogorov-Smirnov tests.* Ann. Math. Statist. 36, pp. 1019-1026.

[9] Conover, W.J. (1967). *A k-sample extension of the one-sided two-sample Smirnov test statistic.* Ann. Math. Statist. 38, pp. 1726-1730.

[10] Conover, W.J. (1971). *Practical Nonparametric Statistics.* Wiley, New York.

[11] David, H.T. (1958). *A three-sample Kolmogorov-Smirnov test.* Ann. Math. Statist. 29, pp. 842-851.

[12] Dorfman, R. (1943). *The detection of defective members of large populations.* Ann. Math. Statist. 14, pp. 436-440.

[13] Drion, E.F. (1952). *Some distribution free tests for the difference between two empirical cumulative distribution functions.* Ann. Math. Statist. 23, pp. 563-574.

[14] Dwass, M. (1960). *Some k-sample rank order tests.* Contri. to prob. and statistics, Essays in honor of H. Hotelling, pp. 198-202. Stanford University Press, Stanford, California.

[15] Epanechnikov, V.A. (1968). *The significance level and power of the two-sided Kolmogorov test in the case of small sample sizes.* Theor. Probability Appl. 13, pp. 686-690 (English translation).

[16] Garey, M.R. and Hwang, F.K. (1974). *Isolating a single defective using group testing.* J. Amer. Statist. Assoc. 69, pp. 151-153.

[17] Gnedenko, B.V. and Korolyuk, V.S. (1951). *On the maximum discrepancy between two empirical distributions* (in Russian). Doklady Akad. Nauk SSSR 80, pp. 525-528.

[18] Gordon, M. and Kennedy, J.W. (1975). *The counting and coding of trees of fixed diameter.* SIAM J. Applied Math. 28, pp. 376-398.

[19] Govindarajulu, Z., Alter, R. and Gragg, L.E. (1975). *The exact distribution of the one-sample Kolmogorov statistic.* Consejo Superior de Investigaciones Cientificas XXVI.

[20] Hájek, J. and Šidák, Z. (1967). *Theory of rank tests.* Academic Press, New York.

[21] Harding, E.F. (1971). *The probability of rooted tree-shapes generated by random-bifurcation.* Applied Probability Trust, pp. 44-77.

[22] Harary , F. and Palmer, E. (1973). *Graphical Enumeration.* Academic Press, New York.

[23] Hwang, F.K. (1973). *A group testing problem.* Proc. 4th S-E Conf. Combinatorics, Graph Theory and Computing, pp. 345-355.

[24] Hwang, F.K. (1974). *On finding a single defective in a binomial group testing.* J. Amer. Statist. Assoc. 69, pp. 146-150.

[25] Hwang, F.K. (1975). *A generalized binomial group testing problem.* J. Amer. Statist. Assoc. 70, pp.923-926.

[26] Hwang, F.K. (1976). *An optimal nested procedure in binomial group testing.* Biometrics 32, pp. 939-943.

[27] Hwang, F.K. (1978). *A note on hypergeometric group testing procedures.* SIAM J. Applied Math. 34, pp. 371-375.

[28] Hu, T.C. and Tucker, A.C. (1971). *Optimal computer search trees and variable-length alphabetical codes.* SIAM J. Applied. Math. 21, pp. 514-532.

[29] Hu, T.C. (1973). *A new proof of the T-C algorithm.* SIAM J. Applied Math. 25 pp. 83-94.

[30] Huffman, D.A. (1952). *A method for the construction of minimum-redundancy codes.* Proc. I.R.E. 40, pp. 1098-1101.

[31] Kaucký, J. (1975). *Kombinatorické identity.* Veda, Vydavatelstvo Slovenskej Akadémie Vied, Bratislava.

[32] Kiefer, J. (1955). *Distance tests with good power for the nonparametric k-sample problem (Abst.).* Ann. Math. Statist. 26, p. 775.

[33] Kiefer, J. (1959). *K-sample analogues of the Kolmogorov-Smirnov and Cramér-von Misses tests.* Ann. Math. Statist. 30, pp. 420-447.

[34] Klarner, D.A. (1970). *Correspondence between plane trees and binary sequences.* Journal of Combinatorial Theory 9, pp. 401-411.

[35] Knuth, D.E. (1968). *The art of computer programming, Vol. I.* Addison-Wesley, Reading, Massachusetts.

[36] Knuth, D.E. (1971). *Optimal binary search trees.* Acta Informatica 1, pp. 14-25.

[37] Kolmogorov, A.N. (1933). *Sulla determinazione emperica di una legge di distribuzione.* Giorn. dell'Instituto Ital. degli Attuari 4, pp. 83-91.

[38] Kolmogorov, A.N.(1941). *Confidence limits for an unknown distribution function.* Ann. Math. Statist. 12, pp.461-463.

[39] Kreweras, G. (1965). *Sur une classe de problemes de denombrement liés aû treillis des partitions des entiers.* Cahiers du Bur. Univ. de Opér. 6, pp. 5-105.

[40] Kumar, S. and Sobel, M. (1971). *Finding a single defective in binomial group testing.* J. Amer. Statist. Assoc. 66, pp. 824-828.

[41] Maag, U.R. and Stephens, M.A. (1968). *The V_{NM} two-sample test.* Ann. Math. Statist. 39, pp. 923-935.

[42] Massey, F.J., Jr. (1951). *The distribution of the maximum deviation between two sample cumulative step functions.* Ann. Math. Statist. 22, pp.125-128.

[43] Mohanty, S.G. (1967). *Restricted compositions.* The Fabonacci Quarterly 5, pp. 223-234.

[44] Mohanty, S.G. (1971). *A short proof of Steck's result on two-sample Smirnov statistics.* Ann. Math. Statist. 42, pp. 413-414.

[45] Mohanty, S.G. (1973). *On more combinatorial methods in the theory of queues.* Mathematical methods in queueing theory, Lecture Notes in Economics and Mathematical Systems 98, Springer-verlag, Berlin pp. 365-371.

[46] Mohanty, S.G. (1977). *Lattice path counting and applications.* McMaster Univ. Math. report no. 95, pp. 59-60.

[47] Moon, J.W. (1970). *Counting Labelled Trees.* Canadian Math. Monographs no.1, William Clowes and Sons, London and Beccles.

[48] Moon, J.M. and Sobel, M. (1977). *Enumerating a class of nested group testing procedures.* J. Combinatorial Theory, Series B 23,pp.184-188.

[49] Narayana, T.V. (1955). *A combinatorial problem and its application to probability theory I.* J. Ind. Agricult. Statist. 7, pp.169-178.

[50] Narayana, T.V. (to appear). *Lattice path combinatorics with statistical applications.* Univ. of Toronto Press.

[51] Pólya, G. and Szegö, G. (1970). *Aufgaden und Lebrsätze aus der Analysis I.* Springer-Verlag, Berlin.

[52] Putter, J. (1955). *The treatment of ties in some nonparametric tests.* Ann. Math. Statist. 26, pp. 368-386.

[53] Rényi, A. (1969). *Lectures on the theory of search.* Institute of Statistics, Univ. of North Carolina at Chapel Hill, mimeo series no. 6007.

[54] Rényi, A. (1970). *On the enumeration of search codes.* Acta Math., Acad. Sci. Hung. pp. 27-33.

[55] Riordan, J. (1958). *An introduction to combinatorial analysis.* Wiley, New York.

[56] Riordan, J. (1960). *The enumeration of trees by height and diameter.* IBM Journal of Research and Development 4, pp. 473-478.

[57] Schwartz, E.S. and Kallick, B. (1964). *Generating a canonical prefix encoding.* Communications of the Assoc. for Computing Machinery 7, pp. 166-169.

[58] Smirnov, N.V. (1939). *Estimate of deviation between empirical distribution functions in two independent samples* (in Russian). Bulletin Moscow Univ. 2, pp.3-16.

[59] Smirnov, N.V. (1948). *Tables for estimating the goodness of fit of empirical distributions.* Ann. Math. Statist. 19, pp.279-281.

[60] Sobel, M. and Groll, P.A. (1959). *Group testing to eliminate efficiently all defectives in a binomial sample.* The Bell System Technical Journal 38, pp. 1179-1252.

[61] Sobel, M. and Groll, P.A. (1960). *Group testing to classify efficiently all defectives in a binomial sample* (in R.E. Machol ed.) Information and Decision Process, McGraw-Hill Book Bo., pp. 127-161.

[62] Sobel, M. and Groll, P.A. (1966). *Binomial group testing with an unknown proportion of defectives.* Technometrics 8, pp. 631-656.

[63] Steck, G.P. (1969). *The Smirnov two-sample tests as rank tests.* Ann. Math. Statist. 40, pp. 1449-1466.

[64] Steck, G.P. (1971). *Rectangle probabilities for uniform order statistics and the probability that the empirical distribution function lies between two distribution functions.* Ann. Math. Statist. 42, pp. 1-11.

[65] Sterrett, A. (1957). *On the detection of defective members of large population.* Ann. Math. Statist. 28, pp. 1033-1036.

[66] Van der Waerden, B.L. (1971). *Mathematische Statistik.* Springer-verlag, Berlin.

[67] Wolf, E.H. and Naus, J.I. (1973). *Tables of critical values for a k-sample Kolmogorov-Smirnov test statistic.* J. Amer. Statist. Assoc. 68, pp. 994-997.

[68] Watson, G.S. (1961). *A study of the group screening methods.* Technometrico 3, pp. 371-388.