RANKING AND SELECTION PROCEDURES AND THEIR APPLICATIONS TO THE EDUCATIONAL PERFORMANCE OF STUDENTS AT ADDIS ABABA UNIVERSITY

By

MEKONNEN TADESSE

A THESIS SUBMITTED IN (PART) FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE IN STATISTICS IN THE ADDIS ABABA UNIVERSITY

JUNE 1987

ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES

RANKING AND SELECTION PROCEDURES AND THEIR APPLICATIONS TO THE EDUCATIONAL PERFORMANCE OF STUDENTS AT ADDIS ABABA UNIVERSITY

By
Mekonnen Tadesse

Department of Statistics
Faculty of Science

Approved by Board of Examiners:

Prof. Dr. W. Winkler ______ External Examiner

Dr G. SADASIVAN ______ Internal Examiner

Dr. Abebe Tessera ______ Internal Examiner

Dr G. P. MEHTA ______ Advisor

CONTENTS

ACKNOWLEDGEMENT

I am very grateful to Dr. G.P. Mehta, Associate Professor in the Department of Statistics for his continuous advice and useful comments on the manuscript. Many thanks are also due to Ato Andargachew Ijigu, Associate Registrar for records of Addis Ababa University for allowing me to enter into the records of students to collect the necessary data for this study.

ABSTRACT

The statistical methodology of ranking and selecting procedures has been investigated. Some of the procedures considered are parametric in nature while some others are non-parametric procedures. These procedures have then been applied to problems concerning educational performance of students at Addis Ababa University.

Some of the problems considered are the following:

I. Compare students academic performance (with the same grade point average in the Ethiopian School Leaving Certificate Examination (ESLCE) and thereby determine the faculty/college that corresponds to the best students performance.

II. Compare the academic performance of three groups of students admitted to Addis Ababa University in 1975 E.C.

Group 1: Students who took ESLCE only in 1974 E.C.

Group 2: Students who took ESLCE in any one year other than 1974 E.C.

Group 3: Students who took ESLCE in two or more years.

III. Determine which ESLCE is a better predictor of success in a college/faculty.

The relevant primary data collected from the records available with the Registrar's Office at Addis Ababa

University have been used to give empirical solutions to the problems. The statistical answers to the ranking and selection problems are given in such a way that the probability of making a correct selection (PCS) is controlled. Based on the computed results, conclusions and important recommendations are made.

CHAPTER I

INTRODUCTION

1.1. General

There has been marked and steady growth in the literature on higher education in recent years. Individuals, Colleges and Universities, educational systems and organizations, foundations, and federal, state, and local political units are seeking answers to the varied and urgent problems of higher education in contemporary society. Such demands could produce significant research and great theoretical formulations, but significant research in education is rare.

Many Statistical Problems donot fit naturally into the classical mold of testing hypothesis or estimation, but should instead be treated as multiple decision problems. In this report a study of some Ranking and Selection Procedures and their applications to the Educational (Academic) Performance of Students at Addis Ababa University is carried out.

In Ethiopia at present, the grade point averages (GPA) of the Ethiopian School Leaving Certificate Examination (E.S.L.C.E.) are being used as the chief criteria for admission to universities and colleges. However the validity of placing students based on their GPA in the ESLCE alone as predictor of academic success in the university has not been thoroughly studied.

GPA in the ESLCE is calculated by simply looking into the grades of English, Mathematics and any other three best grades.

In this study a statistical analysis of the ESLCE GPA, Year of ESLC Examination, and individual ESLC Examination grades as predictors of academic success in several faculties and colleges of Addis Ababa University has been carried out. The academic performance of students at their undergraduate studies has also been analysed.

The main objective of this paper is to introduce the idea of ranking and selection procedures, a methodology which many of our statisticians are not aware of, and show its applications on finding solutions to problems that may not be answered by the classical statistical procedures. Since education for the professions involves achievement of skills, acquisition of knowledge, development of attitudes, and synthesis of these into an integrated performance, scholars in this field in collaboration with statisticians tend to be concerned with the design of new techniques and the testing of old ones. We need new approaches to research in education.

The analysis is based on the group of students who entered Addis Ababa University in 1975 E.C.. Students who dropped out before 1978 E.C. were discarded from this analysis. The analysis provides some basis for estimating the utility of ESLC examination grades and GPA in the ESLCE for forcasting university achievement at Addis Ababa University. If these are worthwhile indicators of university achievement, then they might be used to select Freshman Candidates. This is simply an attempt to see how the year of ESLC examinations, the results of individual subjects, and GPA in the ESLCE alone or a combination of these predict achievement in a particular college/faculty of Addis Ababa University.

Institutions with similar aims and with student bodies of comparable ability but with different value orientations cannot realize their aims in the same way. When applications for admission exceed admission quotas, many universities have used a combination of test scores and secondary school GPA to determine which applicant to select. Consideration of scores and grades together predicts a candidates college success better than either scores or grades alone. Scores and grades are then combined in a formula that provides the admissions office with a basis for estimating the likelihood that the candidate will succeed in college if he is admitted. From the over supply of applicants, those with the higher chances for success are admitted.

In Ethiopia at present, the condition of the grade record keeping systems among the many secondary schools is thought to be of uneven quality. So secondary school grades may be unreliable in some instances. But the ESLC Examination certainly must be regarded as a measure of secondary school achievement. Research is needed to learn, whether individual subject test grades and year of ESLCE besides GPA in the ESLCE forecasts university success in a particular faculty better than does GPA in the ESLCE alone.

The criteria of University Success, the standard against which different types of preadmission variables are used should be evaluated. Those responsible for defining selection policy naturally want to choose the best students, those who are likely to be better college students than those not selected. In most cases the earned degree has been taken as the true measure of a college students success.

The report also presents an analysis of the validity of the Ethiopian School Leaving Certificate Examinations. In this study we have used data from the Faculty of Science, Faculty of Technology, College of Social Sciences, School of Pharmacy, Institute of Language Studies, Faculty of Law and Faculty of Medicine.

1.2. PROBLEMS CONSIDERED IN THE ANALYSIS

Problem I: Compare students academic performance (with the same grade point average in the Ethiopian School Leaving Certificate Examination and thereby determine the Faculty/College that corresponds to the best student's performance.

Problem II: Compare the academic performance of three groups of students admitted to Addis Ababa

University in 1975 E.C., namely,

i) Students who took Ethiopian School Leaving Certificate Examination only in 1974 E.C.

ii) Students who took the Ethiopian School Leaving Certificate Examination in any one year other than 1974 E.C.

iii) Students who took the Ethiopian School Leaving Certificate Examinations in two or more years.

Problem III: Which Ethiopian School Leaving Certificate Examination is a better Predictor of Success in a College/Faculty?

Problem IV: Determine the degree of Correlation between the Ethiopian School Leaving Certificate Examination grades and grades on the same subject taken in the university and thereby determine the subject which has highest correlation and need least improvement in the Ethiopian School Leaving Certificate Examination.

Problem V: Compare various Faculties/Colleges towards female education and determine the one with the highest probability of success among females.

Problem VI: Determine the most popular subject or depart ment among bright students in a particular College/ Faculty, in Science Faculty, say.

Problem VII: Compare the academic performance of students at various Faculties/Colleges who entered the various Faculties/Colleges with the same grade point average in the Ethiopian School Leaving Certificate Examination, to determine a subset of faculties that contains the faculty that corresponds to the best student's performance.

Problem VIII: To select a subset of Faculties/Colleges that contains all "good" faculties.

Results/solutions of these problems will definitely help us to revise the university admission criteria and also to improve the Ethiopian School Leaving Certificate Examinations. Moreover, we will be able to frame our policies to improve the education level of females.

The present report is divided into four chapters, including chapter I, an introduction. In chapter II the methodology of ranking and selection problems is introduced and some such selection procedures are discussed as are relevant to the problems raised above in the present chapter. These procedures are then applied to the data collected on Addis Ababa University students who were admitted in the undergraduate program in 1975 E.C. and completed their four years education in this programme as scheduled in the academic year 1978 E.C..

The relevant data analysis forms the contents of chapter III. In chapter IV we give a brief summary of the present work and also give the recommendations.

CHAPTER II

METHODS USED IN THE STUDY

2.1. INTRODUCTION

In this chapter we introduce the methodology of ranking and selection problems and study some of the available ranking and selection procedures. These procedures will then be used to the data collected on the students of Addis Ababa University, which shall form parts of chapter III.

Ranking and selection procedures are modern statistical techniques for comparing two or more populations. These selection procedures are designed specifically to identify the 'best' single population, or the best subset of populations, or some subset that contains the best population, or the like. We will assume that the populations are not all the same and can be ordered in some meaningful way.

Ranking and selection procedures answer questions that are raised in many investigations but seldom answered by the more traditional methods of analysis. Infact the statistical answer to ranking and selection problems must be given in such a way that the probability of a correct selection is controlled. Before we get into specific ranking and selection procedures we have included a brief discussion of the classical procedures. This is because the classical procedures for comparing k populations are familiar to most persons acquainted with elementary statistical methods and these procedures are based on the same kind of data, it will be possible for us to point out the similarities and/or differences between the two types of procedures (techniques).

In the frame-work of testing hypothesis, the classical procedure attempts to determine whether k parameters all have a common value. Each of the parameters represents the same type of description, attribute or response for all populations, but the populations may differ. For example, if we want to compare the academic performance of students at k different Colleges/Faculties, and if performance is measured by some parameter θ_j for the j^{th} College/Faculty, the classical procedure permits us to decide about the following null hypothesis, sometimes called the homogeneity hypothesis.

$$H_0: \theta_1 = \theta_2 = \ldots = \theta_k$$

The laternative hypothesis is that the θ values are unequal for atleast one pair of Colleges/Faculties.

If a test of homogeneity is the desired goal of an investigation or experiment, alternative methods of statistical analysis are not needed. However, there are many practical situations in which other kinds of information or other goals are of interest. For example, suppose that the null hypothesis of homogeneity is rejected. The investigator may want:

a) to determine which populations differ from which others, and in what direction; or

b) to see which population(s) can be considered 'best' in some well-defined sense of the term best.

Methods of multiple comparisons may provide information that are relevant for both cases, but as far as case (b) is considered ranking and selection procedures are more appropriate. One must not take the above to mean that ranking and selection procedures are simply additional answer to the problem of what to do next when the null hypothesis of homogeneity is rejected. Whenever an investigator wants to select the one best population, say, a test of homogeneity of all k populations is not pertinent to the primary problem of the investigator.

Now we will list down some among many different goals of ranking and selection procedures and then discuss the two approaches of ranking and selection before we get into specific procedures of ranking and selection which are applied to our problems.

2.2. POSSIBLE GOALS OF RANKING AND SELECTION

1. Selecting the one best population
2. Selecting a random number of populations such that all populations better than a control population or a standard are included in the selected subset.
3. Selecting the t best populations for $t \geq 2$,

a) in an ordered manner

b) in an unordered manner

4. Selecting a random number of populations, say r, that includes the t best populations.
5. Selecting a fixed number of populations, say r, that includes the t best populations.
6. Ordering all the k populations from best to worst (or vice versa).
7. Ordering a fixed size subset of the populations from best to worst (or vice versa).

In literature selection procedures are available to provide solutions to the problems of achieving these and many other goals. The available procedures are both for the parametric as well as for the non-parametric families of probability distributions. However we shall study only some of these.

2.3. INDIFFERENCE ZONE APPROACH

In the problem of selecting the best of k populations, a natural rule is to select the population corresponding to the largest (or at times even the smallest) sample value of an appropriate statistic. It is best to start our discussion with a simple instance of the application of the so-called indifference zone approach to the problem of ranking and selection.

Consider independent observations X_{ij} from each of k populations with cdf's $G(X-\theta_i)$ $(i = 1,2,...k,$

$j = 1,2,\ldots n$). The experimenter wishes to select the 'best' population associated with the largest parameter θ_i. For this purpose, we choose an appropriate statistic $Y_i = Y(X_{i1}, X_{i2}, \ldots, X_{in})$ with cdf $F_n\ (y-\theta_i)$ and use the natural selection rule that selects the population corresponding to the largest Y_i as the best. It is for this problem, that Bechhofer (1954) while dealing with normal distributions, introduced the indifference zone approach in which we determine the sample size n, prior to the experiment, to control the probability of a correct selection (PCS)

$$PCS = \int_{-\infty}^{\infty} \prod_{i=1}^{K-1} F_n\ (y + \theta_{/k/} - \theta_{/i/})\ dF_n(y) \quad 2.3.1)$$

Where $\theta_{/1/} \leq \theta_{/2/} \leq \cdots \leq \theta_{/k/}$ are the ordered values of the θ_i'S. Any selection of the population that leads to achieving the goal is called a correct selection. In controlling the PCS we need to specify a preference zone where the largest two parameters $\theta_{/k/}$ and $\theta_{/k-1/}$ are apart.

According to Bechhofer we have the following: Assume that two constants P^* ($\frac{1}{k} < P^* < 1$) and δ^* ($\delta^* > 0$) are specified by the experimenter. What is then needed is a statistical procedure for achieving the following goal.

Goal: Select a population and assert that it is the best with the guarantee that

$P(\pi_{(k)}$ is selected) $\geq P^*$ whenever $\theta_{/k/} - \theta_{/k-1/} \geq \delta^*$

Here $\Pi_{(i)}$ is the population with parameter value $\theta_{/i/}$, $i = 1, 2, \ldots, k$. δ^* is to be thought of as an indifference or tolerance constant with the understanding that the experimenter is indifferent as to the selection if $\theta_{/k-1/}$ lies within δ^* of $\theta_{/k/}$.

The subset Ω_{PZ}, of the parametric space, Ω, of vectors $\underline{\theta} = (\theta_1, \ldots, \theta_k)$ consisting of those $\underline{\theta}$ vectors with $\theta_{/k/} - \theta_{/k-1/} \geq \delta^*$ is referred to as the preference zone, and its complement being the indifference zone. The constant P* reflects the probability guarantee the experimenter demands. He would like to be atleast 100P*% sure having selected (the best population) correctly whenever $\underline{\theta}$ lies in the preference zone. A selection proceudre achieves the above goal if and only only if

$$\inf_{\theta \in \Omega_{pZ}} P\,(\Pi_{(k)} \text{ is selected}) \geq P^* \qquad (2.3.2)$$

By replacing the true parameteric configuration with the corresponding slippage configuration ($\theta_{/1/} = \theta_{/2/} = \ldots = \theta_{/k-1/}$, $\theta_{/k/} = \theta_{/k-1/} + \delta^*$) defined by its second best population, under some assumptions on the statistics, one can obtain the inequality

$$PCS \geq \int_{-\infty}^{\infty} F^{k-1}\,(y + \delta^*)\; dF(y) \qquad (2.3.3)$$

In all the above explanations one must not think that our goal is to estimate the value of $\theta_{/k/}$, nor to make a decision about the value of $\theta_{/k/}$. Our aim is only to select the population whose θ value is equal to $\theta_{/k/}$. Hence an error can occur only if the selection is not correct. The counterpart of the power of a test in a classical test of any null hypothesis is the probability of making a correct selection when $\theta_{/k/} > \theta_{/k-1/}$, and this probability provides an indication of the performance of the ranking and selection procedure for the particular model assumed and the sample size used.

.4. SUBSET SELECTION APPROACH

There is an alternative approach also available in the literature. This is the so called subset selection approach of Gupta (1956, 1965). In this approach, we select a random number of populations so as to guarantee that the selected subset contains the best population(s) with preassigned high probability. Moreover, in this approach there is no indifference zone involved and the selection rules are applicable on the entire parametric space. In case we restrict to the case of single best population, the goal with this approach is:

Goal: To select a subset, of the given k populations, that contains the best population with probability not less than a preassigned number P^* ($\frac{1}{k} < P^* < 1$).

In order to achieve the goal, we determine the values of some reasonable statistic for the independent samples and the populations which correspond to the observed statistics having values 'near' the largest (or at times smallest) statistic are then included in the selected subset. Such procedures are given in sections 2.9 and 2.10 where in we specifically deal with normal populations.

Remark 2.4.1. The following points clearly bring out the difference between the above two approaches to the ranking & selection problems.

Indifference Zone Approach	Subset Selection Approach
1. The parametric space is divided into two parts, namely, the preference zone & the indifference zone. The PCS is controlled only on the preference zone and experimenter remains indifferent to selection on the other zone.	1. No indifference zone is brought to bear and the PCS is controlled over the entire parametric space.
2. The sample size is determined prior to experimentation so as to control the PCS.	2. One works with the data already collected and the PCS is infact controlled by the size of the selected subset.
3. Number of populations to be selected is a constant and is known in advance.	3. The number of populations which one would actually select is a random variable.

4. One claims to have selected the best population(s).

4. One selects not only the best but also all those populations which seem to be 'near' the best.

5. The least favourable configuration comes out to be the slippage configuration $\theta_{[1]} = \dots = \theta_{[k-1]} = \theta_{[k]} - \delta^*$

5. The least favourable configuration is $\theta_{[1]} = \dots = \theta_{[k]}$

We shall next discuss some specific ranking and selection procedures which would then be used to the data on students performance of Addis Ababa University. Some of the procedures are for normal distributions and while using these to our data, in chapter III, we have assumed the GPA to be normally distributed. This assumption of normality seems reasonable in view of the following facts:

i) Most Instructors are grading students on the basis of normal distribution.

ii) Since the sample sizes used are large it is reasonable to assume normality by the central limit theorem.

iii) In independent studies of the distribution of GPA's for previous years it has been found that the GPA's are approximately normally distributed.

. SELECTING THE BEST AMONG NORMAL DISTRIBUTIONS WITH A COMMON VARIANCE

In this section we shall first study the selection procedure for selecting the single best normal population, the one having the largest mean, and then provide estimates (both point and interval estimates) for the true PCS.

2.5.1. SELECTING THE POPULATION WITH THE LARGEST MEAN:

Here we consider the case of k populations where the underlying characteristic in Π_i has normal distribution $N(\theta_i, \sigma^2)$, $i = 1,2,\ldots,k$. The common variance σ^2 may be either known or unknown. Let $\theta_{/1/} \leq \theta_{/2/} \leq \ldots \leq \theta_{/k/}$ denote the ordered θ values. The best population may be defined to be the one associated with the largest mean, $\theta_{/k/}$. In case there is a tie among some θ-values for the largest, one of them is arbitrarily labelled as the largest. Let $\delta^* > 0$ and P^* $(\frac{1}{k} < p^* < 1)$ be both specified by the experimenter in advance. The goal is:

Goal 1: To select the single best population such that the probability of correct selection is atleast P^* $(\frac{1}{k} < P^* < 1)$ whenever $\theta_{/k/} - \theta_{/k-1/} \geq \delta^*$

The goal obviously cannot be achieved by selecting one of the populations merely at random. We do need the observations from each of the k populations. Let $X_{i1}, X_{i2}, \ldots, X_{in}$ denote the n observations from population Π_i, $i=1,2,\ldots,k$, and let all the nk observations be independently drawn. Let $\bar{X}_i = \frac{1}{n}\sum_{j=1}^{n} X_{ij}$ denote the sample mean of the observations from Π_i, $i=1,2,\ldots,k$

and let $\bar{X}_{/1/} \leq \bar{X}_{/2/} \leq \cdots \leq \bar{X}_{/k/}$ denote the corresponding ordered values. Bechhofer's procedure is the following:

Rule R_1: Select the population that corresponds to the largest sample mean $\bar{X}_{/k/}$ as the best population.

The probability of a correct selection using Rule R_1, is easily seen to be

$$PCS = \int_{-\infty}^{\infty} \prod_{i=1}^{k-1} \Phi\,(x + \sqrt{n}(\theta_{/k/}-\theta_{/i/})/\sigma)d\Phi(X) \quad (2.5.1)$$

Where Φ is the standard normal cdf. Obviously this PCS depends on the parametric configuration $(\theta_1,\ldots\theta_k)$ through $\theta_{/k/}-\theta_{/i/}$, $i=1,\ldots K-1$ and is monotone increasing function of each of $\theta_{/k/}-\theta_{/i/}, i=1,2,\ldots,k-1$. So, the PCS attains its least value on the preference zone when $\theta_{/k/}-\theta_{/i/} = \delta^*$, $i=1,2\ldots k-1$, that is, when the configuration $\underline{\theta}$ is $\theta_{/1/} = \cdots = \theta_{/k-1/} = \theta_{/k/} - \delta^*$, the so called least favourable configuration. Hence, PCS satisfies, for all $\theta\varepsilon\Omega_{pz}$,

$$PCS \geq \int_{-\infty}^{\infty} \Phi^{k-1}(x+\tau)d\Phi(x) \quad (2.5.2)$$

where $\tau = \frac{\sqrt{n}\delta^*}{\sigma}$. In case σ is known, the right hand side is equated to p^* and thereby we determine n (for given δ^* and p^*). Such a choice of the common sample size n will ensure (in view of 2.5.2) that the probability requirement (2.5.1) is satisfied. Some such values are given in Table A-1 in the Appendix. Table A-1 infact gives us a value of τ for specified p^* and known k.

This in turn, using $n=(\frac{\sigma\tau}{\delta^*})^2$, would give the value of the required common sample size n.

Remark 2.5.1: Various tables, except Table C-1, in the Appendix are taken from Gibbons, Olkin and Sobel (1977), Table C-1 is taken from kim (1986). These tables have been included to make this report self contained.

Remark 2.5.2. An approximate solution for the selection problem with unequal sample sizes n_i, $i=1,2,\ldots,k$ can be obtained by computing a certain generalized average sample size, denoted by n_o, by the square mean-root formula, given by

$$n_o = \left(\frac{\sum_{i=1}^{k} \sqrt{n_i}}{k}\right)^2 \qquad (2.5.3)$$

and use n_o in place of n in order to determine the least PCS that can be guaranteed.

Remark 2.5.3: A more realistic problem will be for the case of common unknown σ. However, there doesnot exist a fixed sample size procedure based on indifference zone approach to achieve Goal 1 and also to have infimum of the PCS independent of unknown σ. One way out is to consider a modified goal and use some other approach (for example, subset selection approach). Another way out in case σ is unknown is to exploit the consistency of the sample variance and use, for large total sample size, the pooled sample variance as if it is the population variance. This will reduce the problem to the case of known variance. This is what has been done in Chapter III, sections 3.1, 3.2.

2.5.2. ESTIMATING THE TRUE PROBABILITY OF A CORRECT SELECTION

Let us assume that the sample size is common but its value was not determined by a preassigned pair (δ^*, P^*), as the investigator might have been unable to specify these numbers. In such a case also the selection rule R_1 is defined and we wish to estimate the PCS attained by this procedure. This true probability of correct selection depends on the true configuration $\underline{\theta}$ of population means. Let $\delta_j=\theta_{/k/}-\theta_{/j/}$, $j=1,2,\ldots,k-1$. So we can estimate δ_j by using the ordered sample means, i.e, by

$$\hat{\delta}_j=\bar{X}_{/k/}-\bar{X}_{/j/} \quad j=1,2,\ldots,k-1 \quad (2.5.4)$$

If we denote $\tau_j=\frac{\delta_j\sqrt{n}}{\sigma}$, then the estimates of τ_j are

$$\hat{\tau}_j=\frac{\hat{\delta}_j\sqrt{n}}{\sigma} , \; j=1,2,\ldots,k-1 \quad (2.5.5.)$$

We use the values of $\hat{\tau}_j$ for estimating the true PCS given by (2.5.1). We first provide a lower bound p_L and an upper bound p_U for the actual probability of correct selection. Then (p_L, p_U) is an interval estimate of the actual PCS. The point estimate P_E of the actual PCS is also obtained and this will obviously lie in the interval (p_L, p_U).

To determine p_L, we compute $\hat{\tau}_j$ and consider each $\hat{\tau}_j$ separately to find the corresponding probability p_j of a correct selection between the j^{th} and k^{th} ordered populations, for each $j=1,2,...,k-1$. Then

$$p_L = \prod_{j=1}^{k-1} p_j \qquad (2.5.6)$$

p_j for each j can be found by interpolating in table A-2 in Appendix with $k=2$ for the corresponding $\hat{\tau}_j$. P_L can also be determined from

$$p_L = \prod_{i=1}^{k-1} \Phi\left(\frac{\hat{\tau}_i}{\sqrt{2}}\right) \qquad (2.5.7)$$

where $p_j = \Phi\left(\frac{\hat{\tau}_j}{\sqrt{2}}\right)$

To determine p_U, we first compute

$$\bar{T} = \frac{\sum_{j=1}^{k-1} \hat{\tau}_j}{k-1}$$

and then p_U is the entry in Table A-2 (with the original value of k) that corresponds to have value $\bar{T}$. Finally to find p_E the first step is to partition the k-1 values of $\hat{\tau}$ into nonoverlaping subgroups or clusters that are approximately of equal size. Suppose we decide to form m clusters. Let $\bar{T}_j$ be the arithmetic mean of $\hat{\tau}$ values in the j^{th} cluster. Then the j^{th} cluster gives rise to a probability value p_j. The value of p_j is the entry in Table A-2, that corresponds to the values of $\bar{T}_j$ and k_j. Then p_E is given by

$$p_E = \prod_{j=1}^{m} p_j \qquad (2.5.8)$$

It may be noted that in the above procedure no statement about the probability of coverage of the true PCS is given. So possibly the above interval cannot be interpreted as a confidence interval for the true PCS. A better procedure is due to Kim(1986), who has provided one sided lower confidence bound for the true PCS. According to him, with 100(1 -α)% confidence.

$$PCS \geq \int_{-\infty}^{\infty} \Phi^{k-1} (x+\sqrt{2}\, h(\sqrt{n}(\bar{X}_{/k/}-\bar{X}_{/k-1/})/\sqrt{2}\sigma)d\Phi(x)$$
$$(2.5.9)$$

where the function h(t) satisfies

$$\Phi(h(t)-t)+\Phi(-h(t)-t)=\alpha \qquad (2.5.10)$$

and lies between $t-Z_{\alpha/2}$ and $t-Z_{\alpha}$ for $t \geq Z_{\alpha/2}$, that is

$$t-Z_{\alpha/2} < h(t) < t-Z_{\alpha}, \text{ for } t \geq Z_{\alpha/2} \qquad (2.5.11)$$

where Z_{α} denotes the upper α quantile of standard normal distribution. Using values of $n, \sigma, \bar{X}_{/k-1/}$ and $\bar{X}_{/k/}$ we first obtain

$$t = \sqrt{n}\,(\bar{X}_{/k/} - \bar{X}_{/k-1/})/\sqrt{2}\,\sigma$$

and determine h(t) from Table C-1 corresponding to $t - Z_{\alpha/2}$. The right hand side of (2.5.9) can then be obtained from table A-2 with $\tau = \sqrt{2}\,h(t)$. This value is infact the 100 $(1-\alpha)$% lower confidence bound for the true PCS.

Remark 2.5.4: In case the sample sizes are unequal, we determine average sample size n_o using (2.5.3). Estimating the true PCS can now be carried out exactly as before except that the average sample size n_o is substituted for n in any calculations.

Remark 2.5.5: The above selection rule R_1 has been used to provide a solution to problems I and II using data on students of Addis Ababa University. This has been done in section 3.1 for problem I and in section 3.2 for problem II. Apart from the solution to the selection problems, the above point and interval estimates for the true PCS have also been obtained for our data.

2.6. SELECTING THE TREATMENT (POPULATION) THAT HAS THE LARGEST CORRELATION WITH THE OVERALL EFFECT: NORMAL POPULATION CASE

Suppose $\Pi_1, \Pi_2, \ldots, \Pi_k$ are k different bivariate normal populations. Let $(X,Y_1),\ldots,(X,Y_k)$ be k bivariate variables where $Y_1\ldots, Y_k$ denote measurements on the k different populations, and X is a measure of the overall effect of all these populations. If $\rho_1,\rho_2,\ldots\rho_k$ denote the correlation coefficients and $\rho_{/1/}\leq\ldots\leq\rho_{/k/}$ the corresponding ordered values, then the best population may be defined to be the one associated with $\rho_{/k/}$, the largest ρ value. That is, we wish to select the population that has the highest correlation with the overall effect. So the goal is:

Goal 2: To select the single best population such that the probability of a correct selection is atleast p^* whenever the distance measure $\rho_{/k/}-\rho_{/k-1/}\geq\delta^*$.

To achieve this goal we first need to take a sample of size n on the (k+1) dimensional random variable $(Y_1, Y_2,\ldots, Y_k,X)$ and then compute the sample correlation coefficient r_j for the j^{th} population using the n paired observations on (Y_j,X), i.e.

$$r_j = \frac{\Sigma(\chi_i-\bar{\chi})\,(y_{ij}-\bar{y}_j)}{\Sigma(\chi_i-\bar{\chi})^2 \quad \Sigma(y_{ij}-\bar{y}_j)^2} \tag{2.6.1}$$

where $(y_{i1},\ldots,y_{ik},\chi_i)$ denotes the i^{th} observation on $(Y_1,Y_2,\ldots,Y_k,X)$ and $\bar{\chi}$, $\bar{y}$, denotes the respective sample means for the x-observations and for the observations from the j^{th} population, $j=1,2,\ldots,k$. The above k sample correlation coefficients are then ordered as $r_{/1/}\leq\ldots\leq r_{/k/}$.

The selection rule is (see Gibbons et al. 1977):

Rule R_2: select the population that corresponds to the largest sample correlation coefficient $r_{/k/}$ as the best population.

Since the determination of the exact least favourable configuration is difficult in this problem we have not included it.

In order to determine the least probability of correct selection that can be guaranteed we proceed as follows. Using common sample size n and the pre specified value of δ^*, we determine τ_t from the relation $\tau_t=\sqrt{n}\delta^*$. With this value of τ_t and known value of k we enter table A-2 and determine p^*. This value of p^* is then the least PCS that can be guaranteed. This procedure R_2 has been used to handle problems III and IV dealing with the data obtained for students of Addis Ababa University and the solution is provided in section 3.3. for problem III and section 3.4. for problem IV.

2.7. SELECTING THE ONE BEST POPULATION FOR BINOMIAL (OR BERNOULLI) DISTRIBUTIONS

In this section we consider the problem of selecting the one best population out of k populations where each element in each population is classified into exactly one of two mutually exclusive (nonoverlapping) and exhaustive categories, such as male and female.

Let Π_1, ..., Π_k be k populations where the underlying characteristic in Π_i has binomial distribution

$b(n_i, p_i)$, $i=1,\ldots,k$; where p_j denotes the probability of success for the j^{th} population.

Let $p_{/1/} \leq \cdots \leq p_{/k/}$ denote the ordered p-values. The best population may be defined to be the one associated with $p_{/k/}$, the largest probability of success. Let p^* $(\frac{1}{k} < p^* < 1)$ and $\delta^* (o < \delta^* < 1)$ be pre-specified constants. The goal is:

<u>Goal 3</u> To select the single best population such that the PCS is atleast p^* **whenever** $p_{/k/} - p_{/k-1/} \geq \delta^*$

Since the parameters of interest are the probabilities of success in each population, it is natural to base our decision on the proportion of successes in each sample.

Let $\bar{X}_j$ denote the proportion of successes observed in the j^{th} sample and let $\bar{X}_{/1/} \leq \cdots \leq \bar{X}_{/k/}$ denote the corresponding ordered values. The rule R_3 is proposed by Sobel and Huyett (1957):

<u>Rule R_3</u> Select the largest sample proportion and assert that the population that produced that sample is the best.

The generalized least favourable configuration in the binomial case for all k and n is $p_{/1/} = p_{/2/} = \cdots = p_{/k-1/}$ and $p_{/k/} = p_{/k-1/} + \delta^*$. The PCS under this generalized least favourable configuration depends not only on the value of δ^* but also on the value of $p_{/k/}$. A theoretical investigation has shown that the least favourable configuration for practical purposes, and particularly for large sample sizes, is given by the single configuration:

$$p_{/1/}=\ldots=p_{/k-1/}=0.5-\frac{\delta^*}{2} \text{ and } p_{/k/}=0.5+\frac{\delta^*}{2}$$

In section 3.5, the above selection rule has been used to provide an answer to problem V, using the data on the students of Addis Ababa University. For specified values of p^* and δ^*, the smallest common sample size n required to meet the p^* condition is determined from figure D-1 given in the appendix. Such a choice of n would automatically guarantee that the probability requirement is met.

2.8. SELECTING THE ONE POPULATION WITH THE LARGEST q^{th} QUANTILE:

Let $\Pi_1, \Pi_2, \ldots, \Pi_k$ be k populations with associated distribution functions $F_1, F_2, \ldots, F_k$ and let ξ_{qj} denote the q^{th} quantile of the cdf F_j, where q $(0\leq q\leq 1)$ is pre-specified. The bestness is defined interms of the q^{th} quantiles. Let the ordered ξ's be denoted by

$$\xi_{q_{/1/}}\leq\ldots\leq\xi_{q_{/k/}}$$

and let the population having $\xi_{q_{/k/}}$ as the q^{th} quantile be defined to be the best population. Let $F_{/j/}$ denote the cdf of the population having $\xi_{q_{/j/}}$ as the q^{th} quantile, for j=1,2,...,k. Let d* be a specified positive quantity such that $d^*\leq \min(q,1-q)$. Look at the points q-d* and q+d* on the ordinate, and the corresponding quantiles ξ_{q-d^*} and ξ_{q+d^*} on the abcissa for the distribution $F_{/k/}$. Let I denote the interval with these two quantiles as end points. The distance measure d between $F_{/k/}$ and other distributions

$F_{/1/},\ldots,F_{/k-1/}$ is then defined to be

$d=\min(d_{1k},d_{2k},\ldots,d_{k-1},k)$

where $d_{ik}=\min_{X\varepsilon I}|F_{/k/}(x)-F_{/i/}(x)|$ denotes the distance between $F_{/k/}$ and $F_{/i/}$, $i=1,2,\ldots,k-1$. Let $p^*(\frac{1}{k}<p^*<1)$ be prespecified.

The Goal is:

Goal 4: To select the one population with the largest q^{th} quantile such that the PCS is atleast p* whenever $d \geq d^*$.

Inorder to achieve this goal we start by first difining a sample quantile of arbitrary order q, which we denote by y_q. If q (n+1) is an integer, then the sample quantile of order q is defined as the order statistic $Y_{/q(n+1)/}=X_{/r/}$. If q (n+1) is not an integer, the sample quantile lies between two order statistics, $X_{/r/}$ and $X_{/r+1/}$, where r is the largest integer contained in q(n+1), and then

$$y_q=\left[r+1-q(n+1)\right]X_{/r/}+\left[q(n+1)-r\right]X_{/r+1/}.$$

In our problem we have k populations. Assuming that we have n observations from each population the sample data from the j^{th} population are denoted by $X_{1j},\ldots,X_{nj}$, for $j=1,2,\ldots,k$. The n observations within the j^{th} sample are ordered as $X_{/1/_j}\leq\cdots\leq X_{/n/_j}$, and the sample estimate of the q^{th} quantile for the j^{th} population, denoted by

y_{qj}, is

$y_{qj}=X_{/r/_j}$ if $q(n+1)$ is an integer for $r=q(n+1)$, (2.8.1)

and $y_{qj}=\left[r+1-q(n+1)\right] X_{/r/_j}+\left[q(n+1)-r\right] X_{/r+1/_j}$ if $q(n+1)$ is not an integer (2.8.2.)

Let $y_{q/1/} \leq \ldots \leq y_{q/k/}$ be the ordered y_{qj}'s. The selection rule R_4, proposed by Sobel (1976), is

<u>Rule R_4</u> Assert that the population that gives rise to the largest sample quantile, $y_{q_{/k/}}$ is the best population.

For specified values of d^* and p^* such that $0<d^*<\min(q,1-q)$ and $\frac{1}{k}<p^*<1$, the smallest common sample size n required to meet the p^*-condition is then determined from table B-2 given in the Appendix. Such a choice of n would automatically guarantee that the probability requirement is met. The rule R_4 has been applied to the data on the students of Addis Ababa University to provide a solution to problem VI. This is done in section 3.6.

Remark 2.3.1. In practice one may face the problem of ties This problem is handled as follows: within the j^{th} sample if two or more observations are equal to $X_{/r/}$ or $X_{/r+1/}$, then y_{qj} is still uniquely defined. If there are ties for first place between samples, i.e., among the y_{qj}'s, $j=1,\ldots k$, then we will be content to select one of the contenders by some random mechanism and assert that it has the largest population quantile.

Remark 2.8.2. If the sample sizes are unequal, say the sample from the j^{th} population is of size n_j, the sample estimate y_{q_j} is computed from (2.8.1.) and (2.8.2.) with n replaced by n_j, for each $j=1,\ldots k$.

2.9. SELECTING A SUBSET CONTAINING THE BEST POPULATION NORMALMEANS CASE:

In this section we consider the goal of selecting a subset of random size that contains the best of k normal populations having common known variance. The bestness is again defined in terms of population means and the best population is the one having largest mean.

Let $\Pi_1,\ldots,\Pi_k$ be k normal populations having means $\theta_1,\ldots,\theta_k$ and common known variance σ^2. Let $\theta_{/1/}\leq\ldots\leq\theta_{/k/}$ denote the ordered θ's. The goal is:

Goal $_5$: To select a subset of random size that contains the population with the largest mean $\theta_{/k/}$

We have a correct selection in case the selected subset happens to include the best population. For a pre-assigned p^* $(\frac{1}{2^k}<p^*<1)$ and using the available observations, we need to have a selection rule that achieves the above Goal with probability not less than p^*. Let $X_{i1},\ldots,X_{in}$ denote the observations from the

i^{th} population and let $\bar{X}_i$ denote their mean, i=1,2,...k. Let $\bar{X}_{/1/} \leq \ldots \leq \bar{X}_{/k/}$ be the ordered sample means. We use the following selection rule of Gputa (1965) (see, Gibbons et al. (1977)

Rule R_5 For any fixed i ($1 \leq i \leq k$), include Π_i in the selected subset if

$$\bar{X}_i \geq \bar{X}_{/k/} - \frac{d\sigma}{\sqrt{n}} \qquad (2.9.1)$$

Here d>0 is to be chosen so as to meet the probability requirement. For known k and preassigned p*, the choice of d is infact the entry τ in Table A-1. In case the common variance is unknown and we have large samples, we use the pooled estimate of the common unknown variance instead of the true variance. This is done keeping inview the consistency of the sample variance.

The above procedure is used to the data on students of Addis Ababa University to provide an empirical solution to the selection problem VII and this is done in section 3.7.

2.10. SELECTING A SUBSET CONTAINING ALL 'GOOD'POPULATIONS: CASE OF NORMAL POPULATIONS

In this section we consider the problem of selecting a random sized subset that contains all good populations; where a population is defined to be good if it is 'near' the best population. The number of good populations is also unknown.

Let $\Pi_1,\ldots,\Pi_k$ be k normal populations with means $\theta_1,\ldots,\theta_k$ respectively and with common known variance σ^2. Let $\theta_{/1/} \leq \cdots \leq \cdots \leq \theta_{/k/}$ denote the ordered θ-values. A population Π_i is defined to be a good population if $\theta_i \geq \theta_{/k/} - \varepsilon$, where $\varepsilon > 0$ is pre-specified by the experimenter. Any population having $\theta_i < \theta_{/k/} - \varepsilon$ is labelled as a 'bad' population. The goal is:

Goal 6: To select a random sized subset of the given populations such that the selected subset includes all the good populations with probability not less than a preassigned number p*.

Let $X_{i1},\ldots,X_{in}$ be a random sample available from Π_i, $i=1,\ldots,k$, and let these k random samples be independently drawn. Let $\bar{X}_{/1/} \leq \cdots \leq \bar{X}_{/k/}$ denote the ordered sample means. The selection rule of Lam (1986) is used to achieve the above goal. The rule is

<u>Rule R_6</u> For any $i(1 \leq i \leq k)$, include Π_i in the selected subset if $\bar{X}_i \geq \bar{X}_{/k/} - \varepsilon - q\frac{\sigma}{\sqrt{n}}$

Here $q=q(k,p^*)$ is the p*-quantile of the range distribution of k standard normal random variables.

In view of Theorem 1 of Lam(1986), with the above selection rule we will make a correct selection with probability not less than p*. Even the infimum of the probability of a correct selection may still be higher than p*.

The tabulated values of q are available in literature. In place of q one may use $q(k,\nu,p^*)$, the p^*-quantile of the studentized range distribution on k treatments with $\nu=\infty$ degrees of freedom (see for example, Pearson and Hartley (1970)). This is what has been done in our case, because the sample sizes are large.

The above procedure is used on the data on Addis Ababa University students' to provide a solution to problem VIII and this is explicitly done in section 3.8.

CHAPTER III

EMPIRICAL SOLUTIONS TO VARIOUS SELECTION PROBLEMS

In this chapter, we make use of various selection rules, as introduced in the preceeding chapter, to data on the students of Addis Ababa University and thereby provide answers to various academic problems formulated in chapter 1.

3.1. PROBLEM I AND ITS EMPIRICAL SOLUTION

We wish to compare the performance of students in the various faculties of Addis Ababa University with the same GPA in the ESLCE and thereby determine the faculty that corresponds to the best students performance.

Inorder to find a solution to this problem, we have applied the procedure and/or rule R_1 of selecting the best normal population with the largest population mean. This has been discussed in sub-section 2.5.1. of the preceding chapter.

The following results have been observed for the data on students who took their fourth year under-graduate examination in 1978 E.C.

MEANS AND STANDARD DEVIATIONS OF CUMMULATIVE GPA(CGPA)

College/Faculty	Sample size (n_j)	Mean of CGPA ($\bar{x}_j$)	Standard dev (S_j)
1. Language studies	26	2.37	0.24
2. Law	12	2.45	0.31
3. Medicine	26	2.64	0.31
4. Pharmacy	25	2.29	0.28
5. Science	81	2.41	0.30
6. Social Science	101	2.48	0.30
7. Technology	33	2.50	0.32

The selection rule R_1 dictates that the population corresponding to the largest sample mean, i.e, the population with label 3 be asserted to be the best population. This implies that among the students who joined Addis Ababa University with the same grade point average in the ESLCE, the students of the faculty of Medicine have performed better than the students at other faculties or colleges. Since the total sample size $\Sigma n_j=304$ is large, we can get reasonably good approximate solutions by estimating σ^2 by S^2 given by

$$S^2=\frac{1}{n-k}\Sigma(n_j-1)S_j^2$$

For our data this becomes $S^2 = \frac{26.258}{297} = 0.0884108$,

from which we get $S=0.2973395$

Since the sample sizes are unequal, we must find, the average sample size from the square mean root formula given by (2.5.3.):

$n_o=38.540291.$

We need n_o to make probability statments about the above selection procedure. Here we are comparing k=7 faculties/colleges and have decided about the 'best' faculty/college interms of the students performance. We will now determine the infimum of the probability of correct selection (PCS). Let the prespecified value of δ* be 0.10. In order to get the infimum of the PCS from tables we first determine τ_c:

$$\tau_c = \frac{\delta^* \sqrt{n_o}}{s} = 2.0878769$$

Therefore from table A-2 the least probability of correct selection that can be guaranteed is 0.759.

<u>ESTIMATION OF THE TRUE PCS</u>: If we assume that we have no idea about the value of δ*, we have by (2.5.4) of section 2.5. that

$$\hat{\delta}_j = \bar{X}_{/7/} - \bar{X}_{/j/} \quad j=1,2,3,4,5,6$$

This resulted in $\hat{\delta}_1$=0.35, $\hat{\delta}_2$=0.27, $\hat{\delta}_3$=0.23, $\hat{\delta}_4$=0.19, $\hat{\delta}_5$=0.16, $\hat{\delta}_6$=0.14, and from (2.5.5.) the τ_j values are found to be $\hat{\tau}_1$=7.303, $\hat{\tau}_2$=5.637, $\hat{\tau}_3$=4.802, $\hat{\tau}_4$=3.967, $\hat{\tau}_5$=3.341, $\hat{\tau}_6$=2.923. To find p_j we use table A-2 and each of the above $\hat{\tau}_i$ values, this gives p_1=1 p_2=1 p_3=0.9997 p_4=0.9978 p_5=0.001 p_6=0.9796. Therefore by (2.5.6), p_L=0.968.

Inorder to compute P_U we need to compute $\bar{T} = \frac{\sum_{i:1} \hat{\tau}_i}{6}$

We have $\bar{T}$=4.663 and this inturn gives P_U =0.9974. An interval estimate of the true PCS is then

$$0.968 \le PCS \le 0.997$$

To determine the point estimate P_E we use the clustering procedure by choosing the three clusters shown below:

Cluster j	Cluster size	Cluster mean	P_j
(7.308, 5.637)	2	6.4725	1
(4.802, 3.967)	2	4.3845	0.9979
(3.341, 2.923)	2	3.132	0.982

The overall point estimate using (2.5.8) is then $P_E=0.9799$.

Using Kim's (1986) procedure discussed in sub-section 2.5.2., let us now obtain a 90 percent lower confidence interval for thetrue PCS. So $\alpha=.1$ and we have

$$t=\sqrt{n}_o(\bar{X}_{/7/}-\bar{X}_{/6/})/\sqrt{2}S=2.067.$$

From tables of standard normal distribution we have upper $\alpha/2$ quantile to be $Z_{.05}=1.645$. This gives $t-Z_{\alpha/2}=2.067-1.645=.422$. Using this value of $t-Z_{\alpha/2}$ and linear interpolation in Table C-1, gives $h(t)=0.7717$ and $\sqrt{2}h(t)=1.09$. Entering table A-1 with $\tau=1.09$ and using linear interpolation, we see that the lower bound to the true PCS, with 90 percent confidence, is 0.444.

Remark 3.1.1: In our case, although the actual value of the common variance was unknown, yet we have taken the pooled sample variance as if it is the known population variance (exploiting the consistency of sample variance). The degrees of freedom of the sample variance are 296, which is large enough for the sample variance to be fairly close to the actual population variance with a high probability.

In case, the sample sizes are small, we cannot exploit consistency of sample variance and we also know that there doesnot exist a fixed sample size procedure based on indifference zone approach whose infimum of the PCS is independent of the unknown σ. However, in such a case, the lower confidence interval bound for the true PCS (that depends on σ as well) can be determined.

3.2. PROBLEM II AND ITS EMPIRICAL SOLUTION

The problem is tocompare the academic performance of the following three groups of students who were admitted as Freshman in 1975 E.C.

Group 1: Students who took ESLCE only in 1974 E.C.

Group 2: Students who took ESLCE in any one year other than 1974 E.C.

Group 3: Students who took ESLCE in two or more years.

For this problem also, we shall use the procedure R_1 discussed in section 2.5. of chapter II. The following results have been observed.

MEANS AND STANDARD DEVIATIONS OF CGPA

Group	Sample size	Mean of CGPA	Standard deviation
1	185	2.86	0.31
2	27	2.24	0.35
3	101	2.09	0.21
	313		

from this we get the pooled sample variance to be

$$S^2 = \frac{1}{N-k} \Sigma(n_j-1)S_j^2 = 0.08154$$

This gives an estimate of the common standard deviation to be

$$S=0.2855521$$

The selection rule R_1 dictates that the population corresponding to the largest mean, i.e. group 1 be asserted to be the best population. Since students who took ESLCE only in 1974 E.C. produced a larger mean we conclude that those students who took ESLCE only in 1974 E.C. have performed better than both the other groups of students. This means that students who did come to the university directly from secondary schools are performing better than the others.

The average sample size n_o by (2.5.3.) is n_o=92.464241. To determine the infimum of the PCS let us take $\delta^*=0.10$. Inorder to get the infimum of the PCS we first determine τ_c:

$$\tau_c = \frac{\delta^* \sqrt{n_o}}{S} = 3.367453$$

From table A-2 by interpolation, the least PCS that can be guaranteed is .984.

Let us now determine lower confidence interval for the true PCS. To find the $100(1-\alpha)\%$ lower confidence bound for the true PCS we use the method of Kim(1986), which was discussed in section 2.5. of chapter II. We have $t=\sqrt{n_o}(\bar{X}_{/3/}-\bar{X}_{/2/})/\sqrt{2}S = 10.477$. Let us choose $\alpha=.1$.

So $Z_{\alpha/2} = 1.645$ and $t - Z_{\alpha/2} = 8.832$.

Since table C-1 contains values of h(t) for $t - Z_{\alpha/2} \leq 1.5$, we have approximated h(t) by $t - Z_\alpha$ (using the least upper bound in (2.5.11). This gives $h(t) = t - Z_\alpha = 9.195$ and $\sqrt{2}\, h(t) = 13.0$ approximately. We enter table A-2 with $\tau = 13$ and find that the 90 percent lower confidence bound for the true PCS is 1.0. Infact, table A-2 gives values of p* (lower confidence bound in this case) for τ upto 5.0. The value at $\tau = 5$ is 0.999. Since lower confidence bound is increasing with τ we have taken this to be unity. So we can conclude that the true PCS is unity with 90 percent confidence.

Remark 3.2.1: The lower confidence bound has come out to be unity in view of the fact that not only there is big difference between the two largest sample means, the variance is small and the sample sizes are also large to make a correct selection with very high probability.

3.3. PROBLEM III AND ITS EMPIRICAL SOLUTION

Here the problem is to determine which ESLC Examinations are better predictors of success in a particular college/faculty. We have restricted our study, in this case, only to the Faculties of Science, Social Sciences and Technology. However, such a study can be carried over in other faculties as well.

To find a solution to this problem we apply the procedure R_2 of selecting the population with the largest correlation coefficient, which was discussed in section 2.6. The correlations of grades of the individual ESLC Examinations with cumulative GPA have been calculated. The subject with the highest correlation with CGPA is considered as the subject that predicts success better than the remaining subjects in that particular college or faculty.

The results observed are the following:

CORRELATIONS OF ESLCE SUBJECT GRADES WITH UNIVERSITY CGPA

College/Faculty	Subject	Correlation Coefficient
Science	English	$r = 0.29$
	Mathematics	$r = 0.63$
	Biology	$r = 0.28$
	Chemistry	$r = 0.29$
	Physics	$r = 0.48$
	History	$r = 0.11$
	Geography	$r = 0.13$
	Economics	$r = 0.17$
Social Science	English	$r = 0.47$
	Mathematics	$r = 0.21$
	History	$r = 0.42$
	Geography	$r = 0.50$
	Economics	$r = 0.31$

College/Faculty	Subject	Correlation Coefficient
Technology	English	r = 0.27
	Mathematics	r = 0.49
	Chemistry	r = 0.14
	Physics	r = 0.49
	History	r = 0.09
	Geography	r = 0.03

The selection rule R_2 dictates that the population corresponding to the largest sample correlation coefficient shall be asserted as the best population.

Accordingly we can conclude that grade in mathematics is a better predictor of success in the faculty of science, geography in college of social sciences, and physics and/or mathematics in the faculty of technology. We want the probability of a correct selection of the most useful predictor subject to be atleast $p^*=0.75$ when the true difference $\delta=\rho_{/k/}-\rho_{/k-1/}$ of correlation coefficients satisfies $\delta>\delta^*=0.15$ we enter table A-1 with the values of k and p* to obtain a quantity τ_t. Then n is found by substituting τ_t and the specified δ^* values into the expression

$$n=\tau_t^2/\delta^{*2}$$

Since the value of k is different for different faculties/colleges, the values of the common sample sizes used over different faculties/colleges is different. The sample sizes used are as given below.

1. Science 204
2. Social Science 152
3. Technology 172

3.4. PROBLEM IV AND ITS SOLUTION

The problem is to determine how well the grades in the various ESLC Examinations correlate with subsequent college grades in the same subject and determine the subject that needs least improvement in the ESLCE Examination. The table below shows the sample correlation coefficients between the ESLCE grades and the grades in the same subject taken in the university at the Freshman level. Each correlation coefficient is based on 620 observations.

CORRELATION OF ESLCE GRADES WITH CORRESPONDING FIRST YEAR UNIVERSITY SUBJECT GRADES

Subject	College/Faculty	Year of ESLCE	Correlation Coefficient
English	All	any	r =0.31
		1974 E.C.	r =0.47
Mathematics	where taught	any	r =0.33
		1974	r =0.49
Chemistry	where taught	any	r =0.51
		1974	r =0.54
Physics	where taught	any	r =0.26
		1974	r =0.29
History	where taught	any	r =0.31
		1974	r =0.31

Subject	College/Faculty	Year of ESLCE	Correlation Coefficient
Geography	where taught	any	r=0.34
		1974	r=0.37
Economics	where taught	any	r=0.19
		1974	r=0.15
Biology	where taught	any	r=0.27
		1974	r=0.27

These correlations differ from faculty to faculty. However, the above results are for the whole university, without restricting to any college/faculty as such. The selection rule R_2 dictates that the population corresponding to the largest sample correlation coefficient shall be asserted as the best population. We therefore conclude that the Chemistry question paper in the ESLCE is a good measure of the level of knowledge of the students and so it needs least improvement among various question papers.

We have used n=620 observations to determine various correlation coefficients. Let us take the prespecified value of δ^* to be $\delta^*=0.1$. This gives $\tau=\sqrt{n}\,\delta^*=2.489$. With $\tau=2.489$ and k=8 we enter Table A-2 and using interpolation we get $p^*=0.832$. Thus the least PCS that can be guaranteed is 0.832.

The need for an effort to improve the validity of these subject matter examinations, in most cases, is obvious. We feel that these subject matter correlations should be computed each year, so that progress in improving the examinations can be checked. A good subject examination should not have a very low correlation coefficient, say not less than 0.50, when correlated with grades in the same subject in college, although this depends on how similar the college subject matter is to the secondary school subject matter.

3.5. PROBLEM V AND ITS SOLUTION

The problem is to compare various faculties or colleges towards female education and determine the faculty/college corresponding to the largest proportion of successful female students, successful as scheduled in the four years of undergraduate program.

In order to provide an empirical solution to this problem we have used procedure R_3 of selecting the one best population for binomial distributions, which was discussed in section 2.7. of the proceding chapter. The following results have been observed:

PROPORTION OF SUCCESSFUL FEMALE STUDENTS BY COLLEGE/FACULTY

College/Faculty	Observed Proportion of Successes
1. Language studies	0.86
2. Law	0.50
3. Medicine	0.61
4. Pharmacy	0.86
5. Science	0.24
6. Social Sciences	0.27
7. Technology	0.80

According to rule R_3 we assert that the population that produced the largest proportion of female successes is the best population. As we can see from the above results the proportions of females in the Institute of Language Studies and School of Pharmacy happened to be the same we therefore choose one of them randomly and conclude that one of them is the best towards female education. If we prespecify $p^*=0.75$ and $\delta^*=.15$ then n is found from Figure D-1 in the Appendix. The Figure has been taken from Gibbons, Olkin and Sobel (1977). The value of n, as determined using this figure has been used. This value is infact n=25.

3.6. PROBLEM VI AND ITS EMPIRICAL SOLUTION

In this section we shall provide an empirical solution to the problem of determining the most popular subject/department among students in the Faculty of Science. Since the problem is that of choosing the popular area or field of study, we define the best area to be the one for which the median GPA on the first year examination is higher than that for any other areas. We are restricting to the comparisons in terms of median and not any other quantile merely for want of tables. For this we will follow the procedure discussed in section 2.8. Here we have compared the six departments of the faculty of science, i.e, k=6. Let the specified values of p^* and d^* be $p^*=0.75$ and

$d^*=0.20$. The size of the sample we should take from each department is n=23. This has been read from tableB- 1. The results observed are as given below.

MEDIAN GPA OF STUDENTS BY DEPARTMENT

Department	Median GPA
Biology	2.00
Chemistry	2.22
Geology	1.83
Mathematics	2.00
Physics	2.22
Statistics	2.42

As we can see from the above table, Department of Statistics happened to be the one with the largest median GPA. So we assert with probability atleast 0.75 that the Department of Statistics is most popular among the students in the Faculty of Science.

3.7. PROBLEM VII AND ITS EMPIRICAL SOLUTION

Now we shall compare the academic performance of students at various faculties of Addis Ababa University, and among those who entered the various colleges/faculties with the same GPA in the ESLCE, to select the subset that contains the faculty/college corresponding to the best students performance. We have applied the method (procedure) discussed in section 2.9. Since the rule R_5 works only for equal sample sizes and literature, at least locally, is not available for unequal sample sizes we have used equal sample size. The results observed are as given below.

MEANS AND STANDARD DEVIATIONS OF CGPA

College/Faculty	Sample size	Mean of CGPA	Standard deviation
1. Language Studies	30	2.35	0.27
2. Law	30	2.45	0.33
3. Medicine	30	2.65	0.36
4. Pharmacy	30	2.30	0.30
5. Science	30	2.39	0.33
6. Social Sciences	30	2.50	0.33
7. Technology	30	2.54	0.35

From table A-1, with k=7 and $p^*=0.95$, we obtain $\tau_t=3.2417$ and this value of τ_t is infact the value of d to be used in rule R_5. Using the values of d=3.2417, S=0.329 in place of σ and n=30 we get $\frac{d\sigma}{\sqrt{n}}=0.19235$. So the selection rule R_5 becomes:

Rule R_5: Include $\Pi_i(i=1,2,\ldots k)$ in the selected subset if $\bar{X}_i \geq \bar{X}_{/7/}-0.19235$

Thus the subset includes those populations with sample means smaller than $\bar{X}_{/7/}=2.65$ by atmost $\frac{ds}{\sqrt{n}}=0.19235$, that is, with sample mean not less than 2.45765. This implies that the selected subset consists of the populations that correspond to the sample means $\bar{X}_{/7/}=2.65$, $\bar{X}_{/6/}=2.54$ and $\bar{X}_{/5/}=2.50$. Hence we conclude with probability atleast 95 percent that the group consisting of tha faculties of Medicine, Technology and Social Sciences contains the best faculty in terms of students performance.

3.8. PROBLEM VIII AND ITS EMPIRICAL SOLUTION

In this section we deal with the problem of selecting a subset that includes all faculties/colleges with 'good' students performance. We shall apply the method of selecting a subset that contains all 'good' populations as discussed in section 2.10. and use the results obtained in section 3.7.

Let the pre-sepcified values of p^* and ε be $p^*=0.95$ and $\varepsilon=0.10$. From Section 3.7 it is known that k=7, n=30. The degrees of freedom of the sample variance are $\upsilon=k(n-1)=203$. So we may take the sample variance S^2 as if it is the known population variance σ^2. From tables we get $q=4.17$.

According to rule R_6 the population with sample mean $\geq \bar{X}_{/k/}-\varepsilon-q\frac{\sigma}{\sqrt{n}}$ will be included in the selected subset. Here $\varepsilon+q\frac{\sigma}{\sqrt{n}} = 0.147$. So the i^{th} population will be included in the selected subset if $\bar{X}_i \geq \bar{X}_{/k/}-0.147=2.503$. So we select the populations that correspond to $\bar{X}_{/6/}=2.65$ and $\bar{X}_{/6/}=2.54$ to be included in the selected subset. Thus we conclude, with probability atleast 0.95 , that the faculties of Medicine and Technology are the two 'good' faculties, interms of students performance.

CHAPTER IV

CONCLUSIONS AND RECOMMENDATIONS

We have already tried to express some of our opinions and tentative conclusions about the results found in the study. We would like to sum them up now.

The grade point average in the Ethiopian School Leaving Certificate Examination, while it may be somewhat useful as predictor of academic success in atleast some units of the university, is hardly valid enough to stand alone as the sole criterion for admissions to the University. Additional criteria, such as the ones based on grades of individual subject tests, year of the Ethiopian School Leaving Certificate Examination and the use of Secondary School ranking, must be developed. It has been shown in section 3.2. that students who took ESLCE only in 1974 E.C. performed better than both the other groups of students, namely, those who took ESLCE in any one year other than 1974 E.C. and those who took ESLCE in two or more years. This shows the importance of year of ESLCE. Aslo in section 3.3. it is shown which ESLCE is a better predictor of success in a particular College/ Faculty. For example, mathematics being the best predictor of success in the Faculty of Science, the grades in mathematics in addition to the minimum grade point average required must be looked into to place students in the faculty of science or a modified way of calculating grade point averages must be developed such as twice counting the grades of the subject(s) that best predicts success in the faculty and then take the average. The present practice is to calculate grade point average by simply looking into the grades of English, Mathematics and any other three best grades. This occasionally leads to undesirable admissions in various faculties. For example, a student with his individual subject grades A in Amharic, A in English, A in French, A in History and F in Mathematics may get admitted to the Faculty of Science with a high GPA (inface with GPA 3.2). Such a student might have got low grades in various science subjects like Biology, Physics, Chemistry, or even worse, he might not have eve taken these science subjects in his ESLCE. Such a student will infact be sooner than later dismissed due to poor academic performance. But if he is placed to some other faculty he might be successful.

In section 3.4. the correlations of ESLCE grades with corresponding first year University subject grades are determined. The ones with high correlation coefficient must be taken as good measure of the level of knowledge of the students and the subjects with low correlation coefficients need an improvement at the ESLCE level. For example, the Chemistry question paper in the ESLCE needs least improvement among various question papers. So the validity of the Ethiopian School Leaving Certificate Examinations must be strengthened by improving the individual subject examination.

In section 3.1. it is shown that there is a difference between the performance of students, with the same GPA, of different faculties. Thus, allowances must be made for differences among the faculties and colleges of the University, both as to the reliance placed on the grade point average in the Ethiopian School Leaving Certificate Examination as admission criteria, and as to the standard required for admission to each particular college or faculty.

Some adjustment should be made for the Ethiopian School Leaving Certificate Examination grades acquired in two or more different years, since it is clear from the analysis of section 3.2. that these grades are not comparable to those obtained in a single year. There it is shown that students who took ESLCE in two or more years had performed less than those who took ESLCE in a single year.

The comparison of theperformance of female students at various faculties was carried out in section 3.5. It was only based on one characteristic, namely the probability of success. To look for more general aspects would lead to a multivariate situation. Moreover, at some faculties it was found that the proportion of successful females is very low as compared to the others, the proportion of successful females being.24 in the Faculty of Science and .27 in the College of Social Sciences. This could be due to some other reasons. For example, some students might have been misplaced in different faculties due to poor selection criterion.

These recommendations for broadening and improving the university selection criteria and changing the admission standard must be interpreted as an attempt to upgrade the standards of the university. It is hoped that through better selection techniques we can admit students who are more able than some that we have now, but who have been excluded by the poor selection techniques used heretofore.

A fixed standard based on a variable, inaccurate and only semi-valid measuring instrument should be avoided. The best hope for upholding and improving the standards of the university in our admission is to develop new and more valid criteria (may be different criteria for different faculties) for choosing the best candidates,

Placement in the different departments of the university must also be based on the student's strength in individual subjects. Allocating students to various departments based on their GPA alone must be avoided. Most students with higher GPA, who are given their first choice, are choosing areas by merely looking into the job opportunities and future prospects but not by looking into their talents. Good orientation is needed to avoid this problem. The Authorities concerned must not just follow students preferences. They should look for their talents as well.

Table A.1

Values of τ (or $\tilde{\tau}$) for fixed P (or P^*)

K	P (or P^*) .750	.900	.950	.975	.990	.999
2	0.9539	1.8124	2.7718	3.2900	3.2900	4.3702
3	1.433	2.2302	2.7101	3.1284	3.6173	4.6450
4	1.6822	2.4516	2.9162	3.2220	3.7970	4.7987
5	1.8463	2.5997	3.0552	3.4532	3.9196	4.9048
6	1.9674	2.7100	3.1591	3.5517	4.0121	4.9855
7	2.0626	2.7972	3.2417	3.6303	4.0860	5.0504
8	2.1407	2.8691	3.3099	3.6953	4.1475	5.1046
9	2.2067	2.9301	3.3679	3.7507	4.1999	5.1511
10	2.2637	2.9829	3.4182	3.7989	4.2456	5.1916
15	2.4678	3.1734	3.6004	3.9738	4.4121	5.3407
20	2.6009	3.2986	3.7207	4.0899	4.5230	5.4409
25	2.6987	3.3911	3.8099	4.1761	4.6057	5.5161

Table A. 2

Values of P (or P*) for fixed τ (or [illegible])

K	0.0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4
2	.500	.556	.611	.664	.714	.760	.802	.839	.871	.898	.921	.940	.955
3	.333	.391	.452	.513	.574	.634	.690	.742	.789	.830	.866	.896	.921
4	.250	.304	.363	.425	.488	.552	.614	.674	.729	.779	.823	.861	.893
5	.200	.250	.305	.365	.429	.494	.559	.622	.682	.738	.788	.832	.869
6	.167	.212	.264	.322	.384	.449	.516	.581	.645	.704	.758	.807	.848
7	.143	.185	.234	.289	.350	.414	.481	.548	.613	.676	.733	.785	.830
8	.125	.164	.210	.263	.322	.385	.452	.520	.587	.651	.711	.766	.814
9	.111	.148	.191	.242	.299	.361	.427	.495	.563	.629	.691	.748	.799
10	.100	.134	.176	.224	.280	.341	.406	.474	.543	.610	.674	.732	.785
15	.067	.093	.126	.167	.215	.271	.332	.398	.467	.537	.606	.671	.731
20	.050	.072	.100	.135	.178	.228	.286	.349	.417	.488	.558	.626	.691
25	.040	.058	.083	.114	.153	.200	.254	.315	.381	.451	.522	.592	.659
50	.020	.042	.064	.086	.108	.130	.172	.223	.282	.347	.416	.488	.560

Table A.2 (Continued)

K	2.6	2.8	3.0	3.2	3.4	3.6	3.8	4.0	4.2	4.4	4.6	4.8	5.0
2	.957	.976	.983	.988	.992	.995	.996	.998	.999	.999	.999	.9997	.9998
3	.941	.957	.969	.978	.985	.990	.993	.996	.997	.998	.999	.9993	.9996
4	.919	.940	.956	.969	.978	.985	.990	.993	.996	.997	.998	.999	.999
5	.900	.925	.945	.961	.972	.981	.987	.992	.995	.99	.998	.999	.999
6	.883	.912	.935	.953	.967	.977	.985	.990	.993	.996	[illegible]	.998	.999
7	.869	.900	.926	.946	.962	.974	.982	.988	.992	.995	.997	.998	.999
8	.855	.890	.918	.940	.957	.970	.980	.986	.991	.994	.996	.998	.999
9	.843	.880	.910	.934	.953	.967	.977	.985	.990	.994	.996	.998	.999
10	.831	.870	.902	.928	.948	.964	.975	.983	.989	.993	.996	.997	.998
15	.785	.832	.871	.904	.930	.950	.965	.976	.984	.990	.993	.996	.998
20	.750	.802	.847	.884	.915	.938	.957	.970	.980	.987	.992	.995	.997
25	.721	.777	.826	.867	.901	.928	.949	.965	.976	.984	.990	.994	.996
50	.630	.696	.756	.809	.853	.891	.920	.943	.961	.974	.983	.989	.993

Table B.1

K=6

d*	p* .750	.900	.950	.975	.990
05	387	733	997	1259	1607
10	97	183	249	313	401
15	43	81	109	139	177
20	23	45	61	77	99

Table C. 1

Values of $h_v(t)$ for $\alpha = .10$

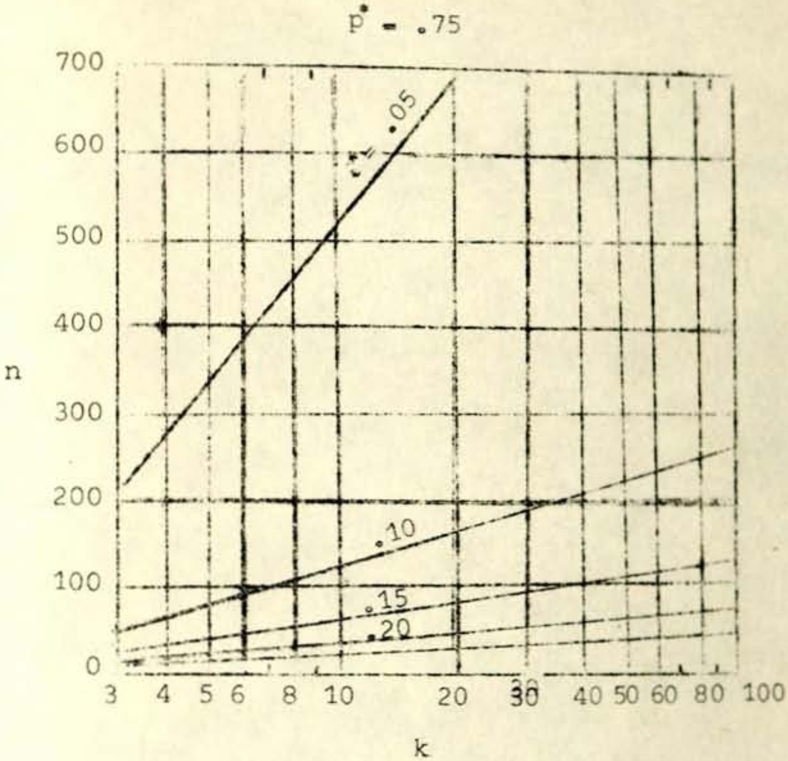
	$t - t_{\alpha/2}(v)$									
v	.05	.10	.15	.20	.25	.30	.35	.40	.45	.50
5	.224	.318	.392	.455	.512	.564	.613	.660	.704	.747
6	.228	.324	.400	.464	.522	.576	.626	.674	.720	.765
7	.231	.329	.405	.471	.530	.584	.636	.685	.732	.777
8	.233	.332	.409	.475	.535	.591	.643	.693	.740	.787
9	.235	.334	.412	.479	.540	.596	.648	.699	.747	.794
10	.236	.336	.415	.482	.543	.599	.653	.704	.753	.800
11	.237	.338	.417	.485	.546	.603	.656	.708	.757	.805
12	.238	.339	.418	.487	.548	.605	.659	.711	.761	.809
13	.239	.340	.420	.488	.550	.608	.662	.714	.764	.813
14	.240	.341	.421	.490	.552	.610	.664	.716	.767	.316
15	.240	.342	.422	.491	.553	.611	.666	.718	.769	.818
16	.241	.343	.423	.492	.555	.613	.668	.720	.771	.820
17	.241	.343	.424	.493	.556	.614	.669	.722	.773	.823
18	.242	.344	.424	.494	.557	.615	.670	.723	.774	.824
19	.242	.345	.425	.495	.558	.616	.671	.725	.776	.826
20	.242	.345	.426	.495	.558	.617	.673	.726	.777	.827
30	.244	.348	.429	.500	.563	.623	.679	.733	.786	.837
60	.246	.351	.433	.504	.569	.629	.686	.741	.794	.846
120	.247	.352	.435	.506	.571	.632	.689	.744	.798	.851
	.248	.354	.436	.508	.574	.635	.693	.748	.802	.856

Table C.1

Values of $h_\nu(t)$ for $\alpha = .10$

	$t - t_{\alpha 2}(\nu)$									
v	.60	.70	.80	.90	1.0	1.1	1.2	1.3	1.4	1.5
5	.830	.908	.984	1.058	1.131	1.202	1.272	1.342	1.410	1.479
6	.850	.932	1.011	1.089	1.165	1.240	1.314	1.387	1.460	1.532
7	.865	.949	1.031	1.111	1.190	1.268	1.345	1.421	1.497	1.572
8	.876	.962	1.046	1.128	1.209	1.289	1.369	1.448	1.526	1.604
9	.885	.972	1.058	1.142	1.225	1.307	1.388	1.469	1.549	1.629
10	.892	.981	1.068	1.153	1.237	1.321	1.404	1.486	1.568	1.650
11	.898	.988	1.076	1.162	1.248	1.333	1.417	1.501	1.584	1.667
12	.903	.994	1.083	1.170	1.257	1.343	1.428	1.513	1.598	1.682
13	.907	.999	1.088	1.177	1.264	1.351	1.438	1.524	1.609	1.694
14	.911	1.003	1.093	1.183	1.271	1.359	1.446	1.533	1.619	1.705
15	.914	1.007	1.098	1.188	1.277	1.365	1.453	1.541	1.628	1.715
16	.917	1.010	1.102	1.192	1.192	1.371	1.460	1.548	1.636	1.724
17	.919	1.013	1.105	1.196	1.286	1.376	1.465	1.554	1.643	1.731
18	.921	1.015	1.108	1.200	1.290	1.381	1.470	1.560	1.649	1.738
19	.923	1.018	1.111	1.203	1.294	1.385	1.475	1.565	1.655	1.744
20	.925	1.020	1.113	1.206	1.297	1.388	1.479	1.570	1.660	1.750
30	.936	1.033	1.129	1.224	1.319	1.413	1.506	1.600	1.693	1.786
60	.948	1.047	1.146	1.243	1.340	1.437	1.534	1.631	1.727	1.824
120	.954	1.054	1.154	1.253	1.352	1.450	1.549	1.647	1.745	1.843
	.959	1.061	1.162	1.263	1.363	1.463	1.563	1.663	1.763	1.863

Figure D.1


$P^* = .75$
700
600
500
400
300
200
100
0
n
.05
.10
.15
.20
3
4
5
6
8
10
20
30
40
50
60
80
100
k

BIBLIOGRAPHY

1. Bechhofer, R.E. (1954). A single sample multiple decision procedure for ranking means of normal populations with known variances. The Annals of Mathematical Statistics, 25, 16-39.

2. Gibbons, J.D., Olkin, I and Sobel, M. (1977). Selecting and ordering populations. A new statistical methodology, John wiley.

3. Gupta,S.S. (1956). On a decision rule for problem in ranking means. Ph.D. Dissertation, University of North Carolina, Chapel Hill, North Carolina.

4. Gupta, S.S. (1965). On some multiple decision (selection and ranking) rules. Technometrics, 7,225-245.

5. GUPTA, S.S. and PANCHAPAKESAN, S. (1979). Multiple Decision Procedures: Theory and Methodology of Selecting and Ranking Populations. John wiley.

6. Kim, Woo-Chul (1986). A lower confidence bound on the probability of a correct selection. JASA, 81, 1012-1017.

7. K.Lam (1986). A new procedure for selecting good populations. Biometrika, 73, 201-206.

8. Pearsin E.S. and Hartley, H.O.(1970) Biometrika Tables for statistician volume I.

9. Sobel,M. (1967). Non parametric procedures for selecting the population with the largest α-qunatiles. The annals of Mathematical Statistics, 38, 1804-1816.

10. Sobel,M. and M. Huyett(1957). Selecting the best one of several binomial populations. Bell system Technical Journal

DECLARATION

The thesis is my original work and has not been presented for a degree in any other university.

[signature]

MEKONNEN TADESSE

This thesis has been submitted for examination with my approval as university advisor.

[signature]

Dr G P MEHTA