SOME CONTRIBUTIONS TO THE THEORY AND APPLICATIONS OF NONPARAMETRIC SUBSET SELECTION PROCEDURES

by

SAJIDAH ALSAEED

A dissertation submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN APPLIED MATHEMATICAL SCIENCES

2024

Oakland University
Rochester, Michigan

Doctoral Advisory Committee:

Dorin Drignei, Ph.D., Chair
Gary C McDonald, Ph.D., Co-Chair
Hon Yiu (Henry) So, Ph.D.
Li Li, Ph.D.

To my family, whom I am deeply grateful for their unwavering support and encouragement, THANK YOU!

ACKNOWLEDGMENTS

First and foremost, I want to take a moment to express my heartfelt gratitude to my advisor and mentor, Gary C. McDonald, for his invaluable guidance, consistent encouragement and incredible support throughout my PhD journey. His enthusiasm and immense knowledge, and constructive feedback have been a great help to my research work. I truly appreciate the time and effort he dedicated to mentoring me and the confidence he instilled in me

I would also like to extend my sincere thanks to the rest of my committee members: Dorin Drignei, Hon Yiu So, and Li Li for their insightful comments and helpful instructions.

I thank Oakland University and the Department of Mathematics and Statistics for offering me the chance to seek higher education and for their support in my Teaching Assistantship.

Last but not least, special thanks to my parents and little family whom their warm love and endless support got me through challenging obstacles and difficulties, I will always be grateful.

Sajidah Alsaeed

ABSTRACT

SOME CONTRIBUTIONS TO THE THEORY AND APPLICATIONS OF NONPARAMETRIC SUBSET SELECTION PROCEDURES

by

Sajidah Alsaeed

Adviser: Dorin Drignei, Ph.D.

Choosing the best population (based on some parameter performance) among alternative populations can be challenging. The decision-making process depends on what factors are considered important by the person making the decision. On order to determine which populations are better among k of them, an experiment should be designed, and samples from each population are to be taken and the sample size needs to be determined. The target here is not to estimate any unknown parameter, but rather to select the population with the best parameter value.

The methodological topic of this dissertation is the nonparametric approach to subset selection of populations so as to contain the "best" population -to be defined- with a user-prescribed probability of a correct selection, where we will examine how to apply a versatile, nonparametric framework to efficiently and accurately choose subsets of populations while ensuring that users can control the reliability of the selections made. The common approach in this problem context has been to rank the data (as it is a well-established approach) and base the statistical inference on population rank sums. The process of ranking and calculating rank sums is relatively straightforward, facilitating

easier interpretation of results and it can be adapted to different contexts and objectives, whether for hypothesis testing or selecting the best population, enabling researchers to base their choices on relative performance as opposed to absolute measurements.

In this dissertation, alternative rank scoring methods are considered and shown, in some cases, to yield smaller selected subsets with the same assurance probability as with rank sums. The investigations herein considered are for a two-way experimental block design.

An application of the research developed here is made to state motor vehicle traffic fatality rates for the years 1994–2022 with the goal of selecting a subset of states to contain the best (worst) with a prescribed probability. The effects of alternate scoring rules are displayed and shown to support a practical conclusion for other applications.

TABLE OF CONTENTS

TABLE OF CONTENTS—Continued

TABLE OF CONTENTS—Continued

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Overview and history:

In the practical world of everyday life, full of complexities, we often face the baffling problem of making myriad decisions. Every different scenario requires of the problem under a suitable theoretical model followed by a cautious analysis of the evidence leading to a decision whose validity is based on reason. In the advancement of scientific reasoning, statistical inference provides the methodology developed to meet these requirements.

The decision-making process frequently involves comparing a range of alternatives and the goal is choosing the best one based on specific performance. Consequently, a selection process is necessary to determine the number of samples required from each alternative and subsequently identify the best alternative based on the sample data. During an experiment, one common question an experimenter must find the answer to is how to compare several categories or populations and how to choose the best one/s. That includes many categories to be grouped such as different treatments/ drugs for a certain disease, different given textbooks for a specific course, or even as simple as different varieties of grains, all are daily events we face that need the job of finding the "best" through the process of *ranking and selection* that gained recognition as an active area of research on which a great deal of work has been done by many authors in the last 70 years. Ranking and selecting procedures are methods used to organize and choose items, individuals, or entities based on specific criteria. These procedures are used as a

way of organizing things according to their importance, value, quality, or some other relevant factor and they are applied in various fields and contexts to make informed decisions.

Statistically speaking, ranking and selection methods are statistical tools used to compare the characteristics of multiple populations, with the assumption these characteristics are not identical. Simply put together, the problem of interest is concerned with selecting and ranking $k\ (\geq 2)$ populations where each one is characterized by the value of a parameter of interest, say θ. At the beginning, we assume that the populations do not share identical parameter values and that they can be arranged in a meaningful manner based on these parameter values. This arrangement could range from the least favorable to the most favorable population in a clearly defined manner, such as from the population with the smallest parameter value to the population with the largest one. A sample of size n is randomly selected from each population. Utilizing this sample data, a *statistical selection process* is employed to choose populations in a manner that allows us to confidently claim, with a specified level of confidence, that the selected populations have the best θ values. A *statistical ranking process* uses the sample data to make an order of populations in such a way that we can state, based on a specified level of confidence, that the ordering made is correct. The selection and ranking may be defined in terms of a parameter of the population. This may physically represent the mean, the variance, some quantile, or maybe a function of all these quantities.

Usually, in selection and ranking problems, populations with large (small) values of the parameters are considered desirable and accordingly, we define the population with the largest (smallest) of the unknown values of the k parameters to be the best. Therefore,

the best refers to the population with the largest (or smallest) parameter performance depending on the goal of the experimenter. The formulation considered is that of selecting a subset of k populations, which contains the "best" population, or all those populations which are better than a standard. Therefore, in many situations the experimenter is interested in selecting a subset, selecting as far as possible the best ones. In general, multiple decision (selection and ranking) rules may be defined which select a subset of a fixed or random size.

In the example of treatments/drugs, an appropriate measure of the effectiveness of a treatment/drug, or in the example of the textbooks, the number of proven theories or solved exercises it contains, could be the parameters on which the ranking and selection process helps to compare the given populations (treatments/ drugs, textbooks) with a goal to rank them and select the best among them, where the best refers to the category with the largest (or smallest) parameter performance depending on the goal of the experimenter. It is natural to think if these measures are the same among all categories which is clearly unrealistic and it would make more sense to assume at the outset the parameters are unequal and formulate the main problem as a ranking problem.

Historically then, the problems of statistical inference were formulated by the classical and conventional approaches as these of estimation or testing of hypotheses. The classical approach to such a problem is to test the homogeneity (null) hypothesis:

H_0: $\theta_1 = \theta_2 = \ldots = \theta_k$, where $\theta_1, \ldots, \theta_k$ are the values of the parameters for these k populations. In the case of normal populations with means $\theta_1, \ldots, \theta_k$ and a common variance σ^2, the test can be carried out using the F-ratio of the analysis of variance.

It should be recognized that a satisfactory solution to any statistical inference problem depends on the goal of the experiment. In many problems of practical interest, the classical approach of homogeneity is inadequate and unrealistic in the sense that it is not formulated to give satisfactory answers to the main question of how to identify the best population/category. As R. E. Bechhofer states, in his first pioneering paper relating to ranking and selection, as follows: “Thus in an agricultural problem the hypothesis that several essentially different varieties of grain have the same (population) mean yield is an unrealistic one since it is obvious that if the varieties actually are different, the (population) mean yields also be different, and a sufficiently large sample will establish this fact at any preassigned level of significance. Moreover, should a significant result be obtained, the experimenter's problem usually has just begun. For having established that the varieties are different he may now desire to select the one which is ‘best’. Here the best variety might be defined as the one having the largest (population) mean yield. Whenever the experimenter ultimately is faced by the prospect of having to choose a best variety, it seems reasonable that the experimenter should have been designed with this outcome in mind.”

In fact, the method of the least significant differences based on t-tests has been used in the past to detect differences between the average yields of different varieties and thereby choose the ‘best’ variety. However, this method is indirect, less efficient, and does not easily provide an overall probability of a correct selection. In recent years, some work has been done toward developing techniques which try to incorporate in the original statistical formulations of the problem the plans of the experimenter for further analysis after the hypothesis of homogeneity is tested. In cases where the experimenter tests for

homogeneity and regardless of the outcome ranks the populations on the basis of further analysis of the same data, it would clearly be more realistic to assume at the outset the parameters are unequal and formulate the main problem as a ranking problem. In fact, the multiple comparison techniques developed largely by Tukey (1949) and Scheffe´(1953) arose from the desire to draw inferences about the populations when the homogeneity hypothesis is rejected. Although the deficiencies of the classical approach of tests of homogeneity are evident, it should be borne in mind that these drawbacks are not in the design aspects of the procedure, but rather in the types of decisions that are based on the data. The continued use of the classical tests in instances where the formulation was inadequate for the problem was largely due to the lack of procedures under more realistic formulations. The origin of selection and ranking procedures lies in seeking formulations based on goals that are more realistic.

1.2 Thesis Outline

The focus of this research is on a nonparametric method for selecting subsets of populations, making fewer assumptions about the underlying data. This flexibility can be advantageous when the data does not fit standard distributions. The goal is to identify a subset of populations (or groups) from a larger set. This selection is aimed at identifying the “best” population according to a defined criterion, which could involve metrics like highest mean, lowest variance, or some other criterion relevant to the research context. The approach will incorporate a probabilistic element, allowing users to specify the desired likelihood of selecting the correct population. This means that the method not only aims for accuracy but also quantifies the certainty associated with the selection ensuring that users can control the reliability of the selections made.

In Chapter two, a description and comparison of the main two approaches used in the formulation of multiple decision methods in the framework of ranking and selection procedures and how different objectives may lead to various methodological approaches and algorithms that can be employed to tackle different ranking and selection problems.

In Chapter three, a short review and abstract of the scoring tests to be applied to the scoring rules used later on and how they set methods that combine theoretical statistics with practical ranking techniques to provide a robust framework for selection problems, emphasizing the importance of distributional characteristics in scoring and ranking processes.

In Chapter four, an overview and an assessment of the main problem regarding the non-parametric subset selection procedures/rules and the main characteristics to consider in application of these rules.

In Chapter five, the development of the scoring rules used in this research, based on the problem specification, is applied and the rules are established depending on the goal of the experimenter and whether the rules are used to find the best/worst populations. The choice of the scoring rule to be preferred is discussed with an application for two different datasets.

In Chapter six, it is shown how specifying such procedures requires determination of the parameter configuration that minimizes the probability of a correct selection and subsequent calculation of the constant required to implement the procedure so that the probabilistic inference is valid over the entire parameter space to get an idea on the performance of the selection procedure under two different configurations of the populations' parameters.

Finally, Chapter seven concludes with a summary of the contribution to the statistical science regarding R&S procedures discussed and covers potential ideas for improvement and future analysis.

With the outline established, the next step is to delve deeper into the context and significance of this research. The introduction provided a comprehensive overview of the history behind the study, leading to a clearer understanding of the research motivation and objectives answering some questions that may have been asked in the current literature. It will articulate the relevance of this research within the broader field and its potential implications for practice, ultimately guiding the reader through the rationale and significance of the work undertaken.

CHAPTER TWO

DESCRIPTION OF THE TWO APPROACHES: INDIFFERENT ZONE AND SUBSET SELECTION

2.1 Preliminary

Decision makers often face the problem of selecting the best from a finite set of alternatives, where the best refers to the alternative with the largest (or smallest) mean performance. For example, an inventory manager might choose an inventory policy to reduce the overall average cost, while a financial investor may seek an investment strategy that maximizes the average return. In practice, the mean performance of these alternatives are not explicitly available and can only be evaluated by running simulation experiments and often a decision procedure is required to find appropriate sample sizes for all alternatives and to select an appropriate alternative. There is a huge volume of literature on designing such procedures, among these procedures, there are frequentist and Bayesian approaches.

Frequentist procedures, e.g. Rinott (1978), Kim and Nelson (2001), which approve simulation observations to the various alternatives so as to attain a pre-specified probability of correct selection (PCS) for even the least favorable configuration of the means, are usually conservative, i.e., requiring more samples than necessary for an average case. Bayesian procedures (e.g., Chick and Inoue 2001, Chick and Frazier 2012) typically allocate a finite computing budget to different alternatives to either maximize the posterior (Bayesian) PCS or minimize the opportunity cost. They usually require fewer observations compared with the frequentist procedures but generally cannot

guarantee a fixed (frequentist) PCS. In this paper, we take the frequentist viewpoint, and our goal is to design procedures that deliver a user-specified PCS.

It is crucial to align the selection with the true objective at hand when trying to select the best population/ category to ensure a more nuanced and contextually relevant decision-making process. Moreover, it is essential to acknowledge that objectives may vary depending on the specific needs and priorities of the situation at hand. Various types of ranking and selection problems arise from the explicit representations of different objectives. These can be described as:

- Selection of the population based on the largest/ smallest performance measure which is a very natural objective. One formulation of this type is the indifferent zone formulation that will be discussing an upcoming paragraph.
- Comparison with a designated standard population, and only choose another population if it outperforms the standard significantly.
- Selection of the population is most likely to be the best in the form of likely performing the best in one single trial.
- Selection of the population with the largest probability of success. Success here is a binary outcome which means if a population performs above a certain threshold, then it outputs 1, otherwise it outputs 0 if it fails.

If the populations under consideration are very complex, they may be computationally expensive to draw from, we may have a fixed “budget” of computational effort available to expend on the ranking and selection. In this case, the objective becomes a decision on

how to allocate this effort to achieve the highest probability of a correct selection based on some kind of prior information about each population.

Different objectives are often appropriate in different situations. We will discuss two different formulations for choosing the best population.

2.2 *Description and comparison:*

The formulation of a $\boldsymbol{k}$-sample problem as a multiple-decision problem enables the experimenter to answer his natural questions regarding the best category. The formulation of multiple decision procedures in the framework of selection and ranking procedures has been generally accomplished either using the *indifference zone approach* or the *(random-sized) subset selection approach*. Briefly, an *indifference zone* in the parameter space is preassigned, the common number of observations needed is tabulated and the final decision is the selection of *single* population which is asserted *to be the best population*.

The former approach was introduced by Bechhofer (1954). Mainly, he worked on the problem of selecting *only one* population and guaranteeing with probability P^* that the selected population is the best provided some other condition on the parameters is satisfied. He considered that for the means of k *normal* populations with a *common and known variance*. In the formulation of this problem, he is interested in guaranteeing the probability, of selecting the best one, to be a specified number P^* whenever the standardized difference δ between the largest (best) mean and the second largest mean is at least equal to a specified value δ^*. He also solved the problem of explicitly determining the common sample size n to guarantee the probability P^* under the indifference zone $\delta \geq \delta^*$. This "indifference" zone approach thus requires that the experimenter specify two quantities, $P^*$and δ^*, where the decision rule is to choose the

population with the largest observed sample mean. Along with other investigators, such as Sobel and Dunnett (1954), he solved basically similar problems, with obvious modifications in the definition of indifference zone, where they discussed a two-sample multiple decision procedure for ranking means of normal populations with a common but unknown variance.

A second and different formulation is used primarily in decision-making processes where multiple alternatives or treatments are compared and evaluated based on some performance criteria offering a structured framework for choosing the best alternatives based on statistical analysis. This might involve hypothesis testing to determine if the differences between alternatives are statistically significant. It requires that the probability of selecting the best population in the *subset* (hopefully small) be guaranteed to be a specified number $\boldsymbol{P}^*$ regardless of the possible configurations of the parameters. Thus, there is no "indifference zone" and no "least favorable" configuration specified by the analyst. However, the size of the selected subset is a random variable. In the formulation of random-sized subset selection, the number of observations is given, constants needed for carrying out the procedure are tabulated, and the final decision is the selection of a *subset* of populations that is asserted *to contain the best population*.

Notice that in the indifference zone approach, an appropriate distance metric (such as the mean) is defined to measure the difference between the population we want to identify (typically the best) and the remaining populations. That provides a systematic way to evaluate and compare different populations. By focusing on measurable differences, it helps identify which populations are significantly different from the best and which can be considered similar enough to be treated as equivalent in decision-

making contexts. Distributional results in this case are derived under a least favorable parametric configuration and do not result in distribution-free procedures. While in the subset selection formulation, the least favorable configurations are obtained when populations are identically distributed, and the procedures can be distribution-free.

Both subset selection and the concept of an indifferent zone are crucial for effective decision-making under uncertainty. Combining subset selection and the indifferent zone approach fosters a comprehensive strategy for decision-making under uncertainty. In statistical analysis, both approaches are essential for increasing model effectiveness and interpretability as they provide frameworks for navigating complex choices and ensuring that selections are not just statistically valid but also meaningful in practical terms. These frameworks reduce the cognitive burden of decision-making by clarifying which variables or choices truly matter, ultimately leading to more informed and confident decisions.

CHAPTER THREE

REVIEW OF SCORING TESTS

3.1 Overview and abstract:

The order of a series of random variables does not typically get paid particular attention to. However, it would be interesting to discover if we would do it. Suppose, for example, we needed to know the probability that the fourth largest value was less than 65. Or, we needed to know the 80th percentile for random cardiac rate data. Either way, we need to know more about how the order of information functions and behaves therefore we could use that knowledge to draw a proper conclusion about something we are experimenting on. That's to say, we need to know something about the ***order statistics*** defined by sorting the values in increasing order and the ***k***th order statistic of a statistical sample is equal to its *k*th-smallest value. Important special cases of the order statistics are the minimum and maximum value of a sample, and the sample mean or other sample quantiles as the sample median.

In the other hand, assigning ranks to data points in a dataset where each data point is replaced by its rank that indicates its position relative to other points in the dataset when arranged in order whether data is sorted in ascending or descending order, is known as the ***rank statistics*** that focus on the rankings of data rather than the raw data values themselves.

Order statistics, together with rank statistics, are among the most fundamental components in non-parametric statistics and inference. They provide powerful tools for estimating parameters, conducting hypothesis tests, and making inferences without the

stringent assumptions of traditional parametric methods. They do not rely on assumptions about the underlying distribution of the data, making them particularly useful in many real-world scenarios where data may not meet parametric assumptions.

Under the subset selection formulation, the main problem is to select a subset of k given populations that contains the 'best' population with a pre-assigned probability. The random variables associated with a fixed population are assumed to be independent and identically distributed with a continuous distribution function depending on a scalar parameter θ (usually, a location parameter belongs to some interval on the real line). This parameter is expected to stochastically determine the order of the k distribution functions, and the best population should be a stochastically largest (or smallest) population. The parameter θ maybe, for example, the mean, the variance, some quantile, or some function of these quantities. In case several populations possess the largest (or smallest) parameter value, one of them is tagged at random and called the best.

In the past few years, many papers have appeared on both formulations of the selection problem. As expected, most of this research has been devoted to rules that assume a specific distributional form of the underlying observations, uniform, normal, binomial, multinomial, etc. However, *Nonparametric Statistical* analysis that makes minimal assumptions about the underlying distribution of the data being studied is also wildly used. These distribution-free methods are useful for analyzing data that might not satisfy the distributional assumptions of parametric methods. In cases where the research hypothesis entails comparing subjects under different conditions or time points or comparing two subject samples on an outcome variable, nonparametric rank score tests can be invoked.

3.2 Scoring Tests: Rank sums, Normal Scores

About fifty years ago, McDonald (1972,1973) developed a class of nonparametric subset selection rules for *block (two-way) experimental data design*. These statistical methods are used where experimental units are grouped into homogeneous blocks based on a known nuisance variable, and then treatments are randomly assigned within each block, allowing for the analysis of both treatment effects and block effects simultaneously, effectively controlling for the variability introduced by the blocking factor. In a block design, there are essentially two factors being considered: the treatment factor (the variable being tested) and the blocking factor (the variable that creates the blocks). Blocking allows us to further reduce experimental error by grouping similar experimental units into blocks, thereby isolating the effect of the treatment from the influence of the blocking factor. It is particularly useful when there is a known nuisance variable that could significantly influence the outcome.

The selection procedures developed by McDonald (1972,1973) were based on scores or scoring rules, i.e., functions of the rank values of the data. The scoring rules discussed in this research are based on the expected values of order statistics of the uniform distribution (yielding rank values) and of normal distribution (yielding normal score values).

Among nonparametric tests, the Wilcoxon *rank sum* test is widely utilized and applicable to various research inquiries. It compares the ranks of two independent samples to assess whether their distributions differ by computing the sum of ranks for each group and evaluates the differences. When the data is not normally distributed or involves ordinal variables, the rank sum test, due to its non-parametric nature, can be

utilized to evaluate whether there is a significant difference between two independent groups. It works where all data points from both groups are ranked together, with the smallest values assigned to the lowest rank and the largest values assigned to the highest rank. After ranking, the test assesses the cumulative sum of ranks in each group. Assuming that both groups come from identical populations, the rank sums should be roughly the same. When one group consistently outperforms or underperforms the other, it suggests a substantial disparity between the groups.

The other scoring test that fits into our theoretical framework and is of interest for the application of this paper is the normal score test. The term *normal score* is used with two different meanings in statistics. One of them relates to creating a single value which can be treated as if it had arisen from a standard normal distribution (zero mean, unit variance). The second one relates to assigning alternative values to data points within a dataset, with the broad intention of creating data values that can be interpreted as being approximations for values that might have been observed had the data arisen from a standard normal distribution. It is associated with data values derived from the ranks of the observations within the dataset. A given data point is assigned a value that is either exactly, or an approximation, to the expectation of the order statistic of the same rank in a sample of standard normal random variables of the same size as the observed data set.

Scoring tests that utilize rank sums and normal scores are powerful tools in statistical analysis, particularly when the assumptions of parametric tests are not fulfilled. Together, they enhance the ability to make informed decisions based on statistical evidence in various research contexts as researchers may compare rank sums with normal scores to evaluate whether the transformations provide similar conclusions about group

differences. The ability to leverage both rank sums and normal scores enhances researchers' capacity to make informed decisions based on statistical evidence across various research contexts, including medical studies, social sciences, and market research. These methods are adaptable, making them suitable for diverse data types and research questions. Researchers can select the most appropriate method based on the characteristics of their data and the specific objectives of their analysis.

Constructing non-parametric subset selection rules based on scoring tests, such as rank sums and normal scores, involves developing criteria to identify the best-performing groups or treatments without relying on strict distributional assumptions. The formulation of non-parametric subset selection rules based on these scoring tests along with a comparison and applications using real data will be discussed next.

CHAPTER FOUR

NON-PARAMETRIC (DISTRIBUTION-FREE) SUBSET SELECTION RULES

4.1 Overview and assessment:

Non-parametric (distribution-free) subset selection rules are methods used to identify the best subset of options without assuming a specific form of probability distribution for the data making them robust to violations of these assumptions and thereby enhancing the reliability of their conclusions as they often involve ranking the options based on the performance metrics observed, such as means (or medians), without assuming specific distributional properties. These rules are especially handy when working with data that does not come from a standard statistical distribution. Besides the advantage and more flexibility of using simpler models, since they don't rely on specific distributional assumptions, nonparametric methods often involve less computational work and therefore are easier to apply than other statistical methods and much of the theory behind them may be developed rigorously.

Subset selection, i.e., finding a bunch of items from a collection to achieve specific goals, has wide applications in information retrieval, statistics, and machine learning. In an experiment involving several treatments, a natural statistical inference format (and maybe better) is to select a subset of these treatments such that the optimal treatment is included in the subset with a pre-specified confidence level. These procedures assume no prior order information about the treatment effects. However, in some applications partial order information about the treatment effects at increasing treatment levels is readily available. Such subset selection procedures have been

developed since Gupta (1956, 1965) (for a comprehensive review, see Bechhofer, Santner, and Goldsman, 1995). In the subset-selection approach to ranking and selection, a decision-maker seeks a subset of simulated systems that contains the best options/populations with high probability where the number of the populations in the selected subset is a random variable and later it will be specifically defined. Therefore, a single population is not necessarily chosen; rather a subset of the given k populations is selected which is guaranteed to contain the best population with probability at least P^*, the basic probability requirement in these procedures.

The problem of selecting the population with the largest (or smallest) parameter for distributions that belong to a location or a scale parameter family is considered using the subset selection as well as the indifference zone. The focus here is on ensuring that the selection process reliably identifies this best population of interest based on the preferred parameter. A correct selection (CS) is said to occur under the subset selection approach if and only if the best population is included in the selected subset. Using the indifference zone approach, the main goal is to select one of the k populations with the guarantee that the probability of a correct selection (PSC) is at least P^*. These approaches offer a comprehensive strategy to making effective decisions in statistical inference.

To effectively implement our selection rules, we first need to determine the worst-case scenario for the probability of making a correct selection, given any configuration of parameters. By finding the infimum of this probability, we can understand the limitations of our selection process and ensure that our rules are robust enough to perform well even in less favorable conditions.

4.2 Probability of a Correct Selection

Let $\pi_1, \dots, \pi_k$, $k(\geq 2)$ be independent populations. For each of these populations, the researchers are interested in a parameter of a specified characteristics. For example, the $\pi_1, \dots, \pi_k$ could be the 8 textbooks suggested for a math course and our interest could be the average number of proven theories they include. So the number of populations is given as 8 and the parameter of interest that will later help to rank and select the given textbooks is the average number of proved theories. The associated random variables X_{ij} , $i = 1, \dots, k$, $j = 1, \dots, n$, are independent samples of size n from the k populations. Assume the random variables X_{ij} have a continuous cumulative distribution function (CDF) $F_j(x, \theta_i)$, where θ_i's belong to some interval Θ on the real line. Suppose $F_j(x, \theta)$ is a stochastically increasing family of distributions in θ; i.e., if $\theta_1 < \theta_2$, then $F_j(x, \theta_1)$ and $F_j(x, \theta_2)$ are distinct and $F_j(x, \theta_1) \leq F_j(x, \theta_2)$ for all x. Examples of such families of distributions are (1) any location parameter family, i.e., $F_j(x, \theta) = F_j(x - \theta)$; (2) any scale parameter family, i.e., $F_j(x; \theta) = F_j\left(\frac{x}{\theta}\right), \theta > 0, x > 0$; (3) any family of distribution functions whose densities possess the monotone likelihood ratio property. If the observations are taken in n *blocks*, which are well specified in advance of the data analysis, the subscript *j* then indicates the particular block in which the observation x_{ij} occurs.

Let R_{ij} denote the rank of the observation x_{ij} among $x_{1j}, x_{2j}, \dots, x_{kj}$; i.e., if there are exactly r of the observations $x_{mj}, m = 1, \dots, k$ less than x_{ij} then $R_{ij} = r + 1$. These ranks are well-defined with probability one, since the random variables are assumed to have a continuous distribution and take integer values from $1 \; to \; k$ inclusive. Now let

$Z(1) \leq Z(2) \leq \cdots \leq Z(k)$ denote an ordered sample of size k from any continuous distribution G, such that:

$-\infty < a(r) \equiv E[Z(r)|G] < \infty, r = 1, \ldots, k.$ With each of the random variables X_{ij} associate the number $a(R_{ij})$ and define

$$H_i = \sum_{j=1}^{n} a(R_{ij}), i = 1, \ldots, k. \tag{4.2.1}$$

The quantity $a(R_{ij})$ is called the score of X_{ij}, and the quantities H_i will define the procedures for selecting a subset of the k populations. Letting θ_i denote the i^{th} smallest unknown parameter, it follows that

$$F_j(x; \theta_{[1]}) \geq F_j(x; \theta_{[2]}) \geq \cdots \geq F_j(x; \theta_{[k]}), \forall x. \tag{4.2.2}$$

The population whose associated random variables have the distribution $F_j(x; \theta_{[k]})$ is called the "best" population *(assuming that the parameter of interest is the smallest*). In a similar fashion, the "worst" population can be defined as that population characterized by the probability distribution $F_j(x; \theta_{[1]})$, where the assignment of "best" and "worst" is problem specified noted in the applications under study with the smallest (or largest) parameter performance depending on the goal of the experimenter.

A "Correct Selection" (CS) is said to occur if and only if the best population is included in the selected subset, i.e. the population with the largest $\theta_{[k]}$ of $\theta_1, \theta_2, \ldots, \theta_k$. Formally, for a given selection rule R, such that:

$$R: select\ the\ ith\ population\ iff\ x_i \geq x_{max} - d\ , d > 0. \tag{4.2.3}$$

where x_i is the observed value of the ith random variable X_i and x_{max} is the largest value of $x_1, x_2, \dots, x_k$. In the subset selection formulation, one wishes to select a subset such that the probability is at least equal to a preassigned constant P^* $(k^{-1} < P^* < 1)$ that the selected subset includes the best population. So

$$P\{CS\ |R\} = P\{\ X_{(k)} \geq X_{max} - d\} \tag{4.2.4}$$

where $X_{(k)}$ is the (unknown) variable that is associated with $\theta_{[k]}$. It should be noted that we do not know, in prior, the correct pairing of X_i and the ranked θ_i.

Now equation (4.2.4) can be written as:

(4.2.5)

$$P\{CS\ |R\} = P[\ X_{(k)} - \ \theta_{[k]} \geq \max\left(X_{(1)} - \theta_{[k]}, X_{(2)} - \theta_{[k]}, \dots, X_{(k)} - \theta_{[k]}\right) - d]$$

$$= P[X_{(j)} - \ \theta_{[i]} + \ \theta_{[i]} - \theta_{[k]} \leq X_{(k)} - \theta_{[k]} + d, j = 1,2, k-1]$$

$$= \int_{-\infty}^{\infty} \left[\prod_{j=1}^{k-1} F\left(t + d + \theta_{[k]} - \theta_{[j]}\right) \right] f(t)dt$$

From that, we see:

$$\underbrace{Inf}_{\Omega}\ P\ \{CS\ |R\} = \int_{-\infty}^{\infty} F^{k-1}\,(t+d) f(t)dt \tag{4.2.6}$$

Hence, if we choose d to satisfy $\int_{-\infty}^{\infty} F^{k-1}\,(t+d) f(t)dt = P^*$, then we have determined the smallest d for which

$$Inf_{\Omega} P(CS|R) \geq\ P^* \tag{4.2.7}$$

where $\Omega = \{\ \theta = (\theta_1, \dots, \theta_k) : \theta_i \epsilon\ \Theta\ , i = 1, \dots, k\}$, that is referred to as **the P^* condition.**

The choice of P^* is specified by the analyst and represents the confidence level that the resultant selected subset will contain the best population. For example, if $P^* = 0.95$, there is a 95% chance that the best population is included in the subset. The relationship between P^* and the size of the selected subset allows for flexibility and strategic decision-making, enabling analysts to balance confidence with the number of populations considered. Therefore, the number of populations in the selected subset is a non-decreasing function of P^*. This means that as the confidence level increases, the size of the subset does not decrease; it either stays the same or increases.

Selection procedures can analogously be defined with P^* requirements on the selected subset to contain the worst/best population. That is, these procedures can be structured to meet the P^* requirement, ensuring that the chosen subset satisfies the specified probability condition, and they may involve statistical tests, rank-based methods, or confidence intervals to assess which populations to include.

4.3 Expected Size of the Selected Subset

As mentioned before, the size of the (nonempty) selected subset is not determined in advance but the data themselves and the number of the populations in the selected subset is a random variable as a non-decreasing function of P^*, and from there we can rewrite the P^* condition in the form: $\Pr(S \ni \pi_{(k)}) \geq P^*$, given that S denotes the (random) selected subset and $\pi_{(k)}$ represents the best population. Also, as the value of P^* diminishes from 1, so that the requirement is less stringent, the expected size of the subset will become smaller. Consequently, S may take any of the values $1,2,\ldots,k$. When $S = 1$ the subset consists of only one element, which we designate as best. When $S = k$ the subset contains of all the populations which we know with certainty the best is

contained within. Of course, by taking $S = k$ we always satisfy the P^*condition, but the goal is to find the subset of the *smallest* (expected) size satisfying that requirement. Since the size of S is a random variable, we should consider its mean value or expectation $E(S)$ and we would like this mean to be small depending not only on P^* (as stated before), k, and the common sample size n, but also on the true configuration of $\theta_1, \theta_2, \ldots, \theta_k$. Therefore, for fixed values of P^*, k, and n, the expectation $\mathrm{E}(S|\theta_1, \theta_2, \ldots, \theta_k)$ is regarded as a criterion for measuring the efficiency of any selection procedure that satisfies the P^* condition. Thus, if we have two competitive procedures that satisfy the same P^* condition, we will use the one that yields a smaller expected subset size for any particular configuration of concern to the investigator since choosing the one that yields a smaller expected subset size enhances efficiency. This can lead to simpler interpretations, reduced costs in practical applications, and less complexity in subsequent analyses. In addition, in many real-world scenarios, having fewer options in the selected subset can facilitate decision-making, resource allocation, and implementation of findings. For instance, in clinical trials, selecting a smaller number of effective treatments based on a procedure with a smaller expected subset size can simplify patient management and treatment protocols.

Moreover, while P^* corresponds to the confidence coefficient, the size of S (whether it is fixed or random variable), denoted by $|S|$, corresponds to the 'length' of the confidence interval. Thus, any subset selection rule for 'constructing' S meets the criterion of validity by satisfying the P^*- condition mentioned above (whatever the configuration of the unknown parameters with no preference zone) and $|S|$ serves as a measure of the sensitivity or performance of the rule.

Now we derive an expression for $E(S)$, the expected size of the selected subset. Therefore,

$$E(S) = \sum_{i=1}^{k} P[Selecting\ the\ population\ with\ parameter\ \theta_{[i]}] \tag{4.3.1}$$

$$= \sum_{i=1}^{k} P[X_{(i)} \geq X_{max} - d]$$

$$= \sum_{i=1}^{k} \int_{-\infty}^{\infty} [\prod_{j=1, j\neq i}^{k} F(t + d + \theta_{[i]} - \theta_{[j]})]\ f(t)dt.$$

In using the rule R, the ranks of the selected populations are random variables, and one may want to evaluate the expected sum of ranks of the selected populations. Let the population with parameter $\theta_{[i]}$ be assigned rank i where $i = 1, 2, \ldots, k$. Then the expected sum of ranks is

$$E(Sum\ of\ ranks\ of\ selected\ populations) = \sum_{i=1}^{k} i \int_{-\infty}^{\infty} [\prod_{j=1, j\neq i}^{k} F(t + d + \theta_{[i]} - \theta_{[j]})]\ f(t)dt \tag{4.3.2}$$

As the expected size of the selected subset is preferred to be as small as possible, one may wonder what the maximum value of $E(S)$ could be. It will be shown that the maximum value of $E(S)$ takes place when all the parameters θ_i are equal. If we set the m smallest parameters $\theta_i (1 \leq m \leq k)$ equal to a common value, say θ, and define $\delta_i = \theta_{[i]} - \theta$ and $\delta_{ij} = \theta_{[i]} - \theta_{[j]}$ then writing $Q = E(S \mid \theta_{[1]} = \theta_{[2]} = \cdots = \theta_{[m]} = \theta)$, we obtain

$$Q = \sum_{i=m+1}^{k} \int_{-\infty}^{\infty} F^{m}(x+d+\delta_i)\,[\prod_{j=m+1, j\neq i}^{k} F(x+d+\delta_{ij})]\; f(x)dx \tag{4.3.3}$$

$$+ m \int_{-\infty}^{\infty} F^{m-1}(x+d)[\prod_{j=m+1}^{k} F(x+d-\delta_j)]\; f(x)dx.$$

The right-hand side of the equation (4.3.3) is a non-decreasing function of θ for values $\frac{1}{k} \leq P^* \leq 1$. Note that for values of $P^* \leq \frac{1}{k}$, there always exists a no-data decision rule. This proves that it is maximum where $\theta = \theta_{[m+1]}$ and since this holds for any integer $m < k$, the desired result will follow. To show that Q is monotone we differentiate it with respect to θ and show that the result is positive for $\frac{1}{k} \leq P^* \leq 1 \; and \; \theta < \theta_{[m+1]}$. That results in

$$\frac{dQ}{d\theta} = -m \sum_{i=m+1}^{k} [\int_{-\infty}^{\infty} F^{m-1}(x+d+\delta_i) f(x+d+\delta_i) \tag{4.3.4}$$

$$\cdot \left[\prod_{j=m+1, j\neq i}^{k} F(x+d+\delta_{ij})\right] f(x)\, dx - \int_{-\infty}^{\infty} F^{m-1}(x+d)$$

$$\cdot \left[\prod_{j=m+1, j\neq i}^{k} F(x+d-\delta_i)\right] f(x+d-\delta_i) f(x) dx].$$

For $\frac{1}{k} \leq P^* \leq 1$, it is clear that $d > 0$.

If $x = \dot{x} + \delta_i$ in the second integral and drop primes, then (4.3.4) becomes

$$\frac{dQ}{d\theta} = -m \sum_{i=m+1}^{k} \int_{-\infty}^{\infty} F^{m-1}(x+d+\delta_i) \left[\sum_{j=m+1, j\neq i}^{k} F(x+d+\delta_{ij}\right] \cdot \tag{4.3.5}$$

$$\{f(x+d+\delta_i) f(x) - f(x+d) f(x+\delta_i)\}\, dx$$

It suffices that for each x, the expression in braces above is negative. Clearly this is so if we assume that the density $f(x-\theta)$ has a monotone likelihood ratio in x provided that $d > 0 \; and \; \delta_i > 0 \; for \; i = m+1, m+2, \ldots, k$. Hence,

$$\underbrace{Max}_{\Omega} E(S) = k \int_{-\infty}^{\infty} F^{k-1}(x+d) f(x) dx = kP^* \tag{4.3.6}$$

Thus, subject to the basic probability requirement, $\inf_{\Omega} P(CS|R) \geq P^*$, the procedure satisfies the condition that the expected size of the selected subset is $\leq kP^*$ for all points in Ω.

It is worth mentioning that for some problems we may wish to have a subset of fixed size so that the number of populations in the subset is known ahead, rather than a varying number as in some instances the size of the subset is large or small. This is primarily a question of what the experimenter wants. If he wants a subset of fixed size, there is no point in trying to convince him of the desirability of obtaining any slight increase in efficiency by using a procedure with a random subset size and we would rather give him the best-fixed subset size procedure available. In addition to that, when it comes to studying single-sample procedures which provide more flexibility to the experimenter than does either the fixed subset size rule or the subset selection procedure by allowing to specify an upper bound – say m - on the number of populations included in the selected subset. Such procedures will be called restricted subset selection procedures highlighting their focus on managing how many populations are chosen based on the data. If the data clearly shows that one population is superior this rule shares the advantage of the subset selection procedure over the fixed size subset rule in permitting fewer than m populations to be selected. In contrast, if the data do not make

the choice of best populations as obvious, the latter rule selects a larger subset but guarantees that no more than m populations are selected.

Besides the problem of selecting the subset containing (with the smallest size) the best of the given k populations, another problem that has been investigated from the early period is that of comparing k experimental populations with a control population or a fixed standard say, θ_0. The goal is to select a subset that contains all populations *better than* the standard or the control. Although we say 'better than', we are actually interested in selecting populations *as good as or better than* the control. A correct selection with this goal is defined as including all populations with θ values that are as large or larger than θ_0, or equivalently, as eliminating only populations with θ values smaller than θ_0. This goal is more flexible than the goal of selecting the one best population in that the investigator chooses a subset of populations but does not specify any ordering with the subset. In this case, an empty subset indicates that the control is best and none of the contenders is better than it. Moreover, Traditional procedures, developed since Gupta (1956, 1965), assume no specific order of treatment effects. However, in some applications partial order information about the treatment effects at increasing treatment levels is readily available. For example, the treatment effect usually increases with the increase of dosage up to a certain unknown optimal level, then makes a downturn due to overwhelming side effects. This type of up-then-down response pattern is called an umbrella ordering. These procedures aim to ensure that the optimal treatment level is included in a selected subset with a specified confidence level. They help with addressing concerns about comparing two populations (e.g., males and females) that exhibit similar

ordering patterns. This type of additional order information is utilized explicitly in Pan, G. (1996).

Beyond just selecting a minimal subset, it's essential to consider how these selections are made, the robustness of the criteria used, and how to handle the inherent complexities of the problem. This way, you can ensure that your selection is both effective and efficient. The number of possible subsets grows exponentially with the size of the set. This makes it computationally challenging to evaluate all options. So the analyst needs to ensure that the selection process remains effective even under uncertainty or variations in the underlying data where the selected subset should ideally have a high probability of being the best or most appropriate choice according to some criteria. Ensuring the effectiveness of the selection process amidst uncertainty requires careful consideration of the methods employed and a focus on achieving high probabilities of success based on sound statistical criteria. This approach not only enhances decision-making but also increases the reliability of the outcomes derived from the analysis.

CHAPTER FIVE

FORMULATION OF THE SUBSET SELECTION RULES

5.1 Early developments (1950- 1965)

Early investigations of subset selection rules predictably centered around well-known parametric families of distributions, namely, normal, binomial, and gamma. Gupta (1956) considered a procedure for selecting the population with the largest mean from k *normal* populations with means $\mu_1, \mu_2, \dots, \mu_k$ and a common variance σ^2. He considered the case of known as well as unknown σ^2. Based on samples of size n from these populations, his rule in the case of *known* σ^2 is:

$$R_1: Select\ \pi_i\ iff\ \bar{X}_i \geq \max_{1 \leq j \leq k} \bar{X}_j - \frac{d\sigma}{\sqrt{n}}, \tag{5.1.1}$$

where $\bar{X}_i$ is the mean of the sample from the population $\pi_i, i = 1,2, \dots, n, and\ d > 0$ is the smallest constant such that the P^*- condition is met.

When σ^2 is *unknown***,** the rule R_2 of Gupta (1956) is the same as R_1 except that σ is replaced with s, given that s^2 is the usual pooled unbiased estimator of σ^2 and the constant d will have a different value here.

For selecting the population with the largest scale parameter from k *gamma* populations with (unknown) scale parameters $\theta_1, \theta_2, \dots, \theta_k$, and a common known shape parameter ν, Gupta (1963) proposed the rule:

$$R_3: select\ \pi_i\ iff\ \bar{X}_i \geq c \max_{1 \leq j \leq k} \bar{X}_j, \tag{5.1.2}$$

where $\bar{X}_i$ is the mean of the sample from the population $\pi_i, i = 1,2, \dots, n, and\ c \in (0,1)$ is to be determined to satisfy the P^*- condition.

The rules such as $R_1, R_2, and\ R_3$ are all referred to as *Gupta's maximum-type rules*. Of course, these have their counterparts for the problem of selecting the population with the smallest parameter of interest. These maximum-type rules have been investigated extensively in literature; their optimal properties have also been studied.

As opposed to the maximum type rules, average type rules were proposed by Seal (1955, 1957, 1958a). In the case of selecting the normal population with the largest mean when the common variance σ^2 is unknown, the average-type rule is:

$$R_4: select\ \pi_i\ iff\ \bar{X}_i \geq \sum_{r=1}^{k-1} c_r \bar{X}_{[r]}^{(i)} - st, \tag{5.1.3}$$

where $\bar{X}_{[1]}^{(i)} \leq \bar{X}_{[1]}^{(i)} \leq \cdots \leq \bar{X}_{[k-1]}^{(i)}$ denote the ordered sample means after deleting $\bar{X}_i (i = 1, \dots, k)$, s^2 is the usual pooled unbiased estimator or $\sigma^2, c_1, \dots, c_{k-1}$ are nonnegative constants subject to the constraint $\sum_{i=1}^{k-1} c_i = 1$, and $t = t(\ k, P^*, c_1, \dots, c_{k-1})$ is chosen to again satisfy the P^*-condition. However, the maximum type rules are found to be approximately Bayes optimal under reasonable loss functions. The additional simplicity in determining the constants associated with these rules makes them more appealing and useful.

The initial investigations of the rules for normal means, normal variances and gamma scale parameters were concerned with derivations of the properties of the rules such as monotonicity, and of results relating to the supremum of the expected size of the selected subset for these specific distributions. The first paper in the direction of a unified treatment was by Gupta (1965) who treated selection in terms of location and scale

parameters. It was assumed that the selection statistics used in the rule have distributions differing in a location or a scale parameter.

Let T_i be the statistic associated with the sample from π_i, $i = 1,\ldots,k$ defined as $T_i = \sum_{j=1}^{n} R_{ij}$, and R_{ij} denote the rank of the observations $x_{ij}'s$, then the distributions of the T_i's are $F(x-\theta_i), i = 1,\ldots,k, or\ F\left(\frac{x}{\theta_i}\right), where\ the\ \theta_i's$ are the parameters to be ranked. The rules investigated by Gupta (1956) are R_5 (location case) and R_6 (scale case) are given as:

$$R_5: select\ \pi_i\ iff\ T_i \geq \underbrace{max}_{1\leq j\leq k}\ T_J - d \tag{5.1.4}$$

and

$$R_6: select\ \pi_i\ iff\ T_i \geq c\ \underbrace{max}_{1\leq j\leq k}\ T_J \tag{5.1.5}$$

where $d > 0\ and\ c\ \epsilon\ (0,1)$ are to be determined to satisfy the P^*-condition.

The statistic T_i here represents the *rank sums* where the structure for determining the rank sums is outlined in Table 5.1 below:

Table 5.1: The Structure for Determining Rank Sums

Block/ π	$\boldsymbol{\pi_1}$	$\boldsymbol{\pi_2}$	$\boldsymbol{\pi_3}$	**…**	$\boldsymbol{\pi_k}$	**SUM**
Block 1	$X_{11} \approx R_{11}$	$X_{21} \approx R_{21}$	$X_{31} \approx R_{31}$	…	$X_{k1} \approx R_{k1}$	$k(k+1)/2$
Block 2	$X_{12} \approx R_{12}$	$X_{22} \approx R_{22}$	$X_{32} \approx R_{32}$	…	$X_{k2} \approx R_{k2}$	$k(k+1)/2$
.	.	.	.	…	.	.
.	.	.	.	…	.	.
.	.	.	.	…	.	.

Table 5.1 – Continued

Block/ π	π_1	π_2	π_3	…	π_k	**SUM**
Block *n*	$X_{1n} \approx R_{1n}$	$X_{2n} \approx R_{2n}$	$X_{3n} \approx R_{3n}$	…	$X_{kn} \approx R_{kn}$	$k(k+1)/2$
RK SUM	T_1	T_2	T_3	…	T_k	$nk(k+1)/2$

Gupta (1965) showed that the infimum of the PCS is attained in either case when the parameters are equal and this infimum is independent of their common value. He also established some important properties that are enjoyed by both procedures. These are:

1- The procedures are monotone, i.e., for $\theta_i > \theta_j$, the probability of including π_i in the selected subset is at least as large as that of including π_j.

2- The probability of selecting the best population in the selected subset of size $|S|$ (not known in advance) is maximum among all possible subsets of size $|S|$.

3- If the density $f(x, \theta)$ possesses a monotone likelihood ratio in x, then the $E(|S|)$ is maximized over all parametric configurations when the $\theta_i's$ are equal and this maximum is $kP*$.

For selecting the *binomial* population with the largest success probability, Gupta and Sobel (1960) proposed a *location-type* rule. Let X_i be the number of successes in n trials associated with $\pi_i, i = l, \ldots, k$. Their rule is:

$$R_7: select\ \pi_i\ iff\ X_i \geq \underbrace{max}_{1 \leq j \leq k}\ X_j - d, \tag{5.1.6}$$

where d is the smallest nonnegative integer for which is $P^* -$ condition is met.

An interesting aspect of this procedure R_7 is that the infimum of the PCS occurs when all the parameters are equal but it is not independent of their common value, say, p.

For $k = 2$, Gupta and Sobel (1960) showed that the infimum of PCS over p occurs when $p = \frac{1}{2}$. When $k > 2$, the common value p for which this infimum takes place is not known. However, it is known that this common value $p \to \frac{1}{2}\ as\ n \to \infty$. The investigations of these early periods were mainly under the assumption that the sample sizes were equal and that the nuisance parameters (such as σ^2 for the normal means problem) are equal.

5.2 Establishment of the Selection Rules

In the analysis of state motor vehicle traffic fatality rates (MVTFRs) as given in McDonald (2016), to identify subsets of states that contain the 'best' (lowest MVTFR) and 'worst' (highest MVTFR) states with a prescribed probability, four subset selection rules are considered.

The two selection rules for choosing a subset containing the *worst* population are given by:

$$R_1: Select\ \pi_i\ iff\ H_i \geq \underbrace{max}_{1 \leq j \leq k} H_j - b_1 \tag{5.2.1}$$

$$R_2: Select\ \pi_i\ iff\ H_i \geq b_2 \tag{5.2.2}$$

While the two selection rules for choosing a subset containing the *best* population are given by:

$$R_3: Select\ \pi_i\ iff\ H_i \leq \underbrace{min}_{1 \leq j \leq k} H_j + b_3 \tag{5.2.3}$$

$$R_4: Select\ \pi_i\ iff\ H_i < b_4 \tag{5.2.4}$$

where $H_i = \sum_{j=1}^{n} a(R_{ij}), i = 1, \ldots, k$. The quantity $a(R_{ij})$ is called the score of X_{ij}.

In cases considered here, the constants b_i's are calculated assuming the population parameters are equal and, thus, the distribution of the H statistics are distribution-free. If $k = 2$, the two selection rules $R_1 and\ R_2$ are equivalent, as are $R_3 and\ R_4$, since $H_1 + H_2$ is a constant. In the next section, the calculation of selection rule constants (based on the above rules) for two distributional choices will be derived.

5.3 Calculation of Selection Rule Constants

5.3.1 Calculation of selection rule constants, $G = Uniform\ Distribution\ (0, 1)$

With the choice of G to be the uniform distribution on the interval $(0, 1)$, the expected value of the order statistics $a(r) \equiv E[Z(r)] = r / (n + 1)$. The selection procedures can then be stated in terms of ranks and rank sums. Thus, the two selection rules for choosing a subset containing the *worst* population are given by:

$$R_1: Select\ \pi_i\ iff\ T_i \geq \underbrace{max}_{1 \leq j \leq k} T_j - b_1 \tag{5.3.1}$$

$$R_2: Select\ \pi_i\ iff\ T_i > b_2 \tag{5.3.2}$$

Similarly, the two selection rules for choosing a subset containing the *best* population are given by:

$$R_3: Select\ \pi_i\ iff\ T_i \leq \underbrace{min}_{1 \leq j \leq k} T_j + b_3 \tag{5.3.3}$$

$$R_4: Select\ \pi_i\ iff\ T_i < b_4 \tag{5.3.4}$$

where $T_i = \sum_{j=1}^{n} R_{ij}\ , i = 1, \dots, k$, represent the rank sums as defined before.

Note that the rules $R_1\ and\ R_2$ could be written in the form select $\pi_i\ iff\ T_i \geq b$, where b is a stochastic quantity for the rule R_1 and a deterministic quantity for the rule R_2. A similar statement can be made for the rules $R_3\ and\ R_4$.

To choose the constants $b_i, i = 1, \ldots, 4$ so that the basic probability requirement is satisfied, it is sufficient to determine the parameter configuration at which the infimum (either over Ω or an appropriately defined subspace) of the probability of a correct selection is attained. Then this minimum value is set equal to P^* and constants $b_i's$ determined. Since the selection rules $R_1 \; and \; R_2$ are based on the sum of ranks, the probability of a correct selection assumes only a finite number of values for any given k and n. Hence, when a P^* value is specified in advance, the minimum probability may not assume this value. The experimenter may then in fact be working with a slightly larger P^* condition than the nominal value. Selection rules $R_1 \; and \; R_2$ are representative members of two classes of selection procedures investigated by McDonald (1972, 1973). These rules were used by McDonald (1979) and by Lorenzen and McDonald (1984) in an analysis of state traffic fatality rates over a period of nine years.

In some instances, the infimum of the probability of a correct selection may be guaranteed to be less than P^* only on a well-defined subspace $\Omega' \; of \; \Omega$. This will be illustrated in a later analysis when Ω'represents a *slippage configuration, i.e.* $\Omega' = \{ \theta \in \Omega : \; \theta_{[1]} = \cdots = \; \theta_{[k-1]} \}$. In order to justify the nonparametric nature of these rules, the constants $b_i's, i = 1, \ldots, 4$ are calculated assuming the $\theta_i's$ are all equal. As developed by McDonald (1972, 1973, 1985), $R_1 \; and \; R_3$ are justified over a slippage space Ω', with the possible exception of $\theta_{[k]}$in case of rule R_1 or $\theta_{[1]}$ in case of rule R_3; and $R_2 \; and \; R_4$ are applicable over the entire parameter space. That means that These rules $R_1 \; and \; R_3$ are valid or applicable within a specific subset of parameters called the slippage space, Ω'. This implies that in this context, the rules operate under certain conditions or constraints. In contrast, rules $R_2 \; and \; R_4$ can be applied universally across

all possible parameter values in the broader parameter space, Ω. This indicates that these rules are more general and not limited by the same conditions as R_1 *and* R_3.

The non-negative constants $b_1, b_3, and\ b_4$ are chosen as small as possible and b_2 is chosen as large as possible preserving the probability P^*- condition. The asymptotic value of b_1 to meet the P^*- condition is the solution to:

$$\int_{-\infty}^{\infty} [\emptyset(x + cb_1)]^{k-1}\ \varphi(x)dx = P^* \tag{5.3.5}$$

where $\emptyset(x) and\ \varphi(x)$ are the CDF and density, respectively, of a standard normal random variable, and

$$c = c(n,k) = [\frac{12}{nk(k+1)}]^{1/2} \tag{5.3.6}$$

The value of $b_3 = b_1$. The values of $b_2\ and\ b_4$ are given by:

$$b_2 = [\frac{n(k^2-1)}{12}]^{1/2}\ \emptyset^{-1}(1-P^*) + \frac{n(k+1)}{2} \tag{5.3.7}$$

$$b_4 = n(k+1) - b_2 \tag{5.3.8}$$

where $\emptyset^{-1}(\cdot)$ is the inverse standard normal CDF. The selection rules defined above are based on rank sums. These arise from the expected values of the order statistics from a standard uniform distribution. The question is: how does the choice of the distribution of the order statistics affect the performance of the subset selection procedures by choosing an alternate distribution for G?

5.3.2 Calculation of selection rule constants, $G = Standard\ Normal\ Distribution$

With the choice of G to be the normal distribution with mean 0 and standard deviation 1, the expected value of the *k*th largest observation in a sample of size *n* from G

is given by the equation:

$$E(x_{k|n}) = \frac{n!}{(n-k)!\,(k-1)!}\int_{-\infty}^{\infty} x\left[\frac{1}{2} - \emptyset(x)\right]^{k-1}\left[\frac{1}{2} + \emptyset(x)\right]^{n-k}\varphi(x)dx \tag{5.3.9}$$

where $\varphi(x) = (2\pi)^{-\frac{1}{2}}\, e^{-\frac{1}{2}x^2}$ and $\emptyset(x) = \int_0^x \varphi(x)dx$. The selection procedures can then be stated in terms of normal scores and score sums. Thus, the two selection rules for choosing a subset containing the *worst* population are given by:

$$Q_1: Select\ \pi_i\ iff\ S_i \geq \underbrace{max}_{1\leq j\leq k} S_j - d_1 \tag{5.3.10}$$

$$Q_2: Select\ \pi_i\ iff\ S_i > d_2 \tag{5.3.11}$$

Similarly, the two selection rules for choosing a subset containing the *best* population are given by:

$$Q_3: Select\ \pi_i\ iff\ S_i \leq \underbrace{min}_{1\leq j\leq k} S_j + d_3 \tag{5.3.12}$$

$$Q_4: Select\ \pi_i\ iff\ S_i < d_4 \tag{5.3.13}$$

where $S_i = \sum_{j=1}^{n} a(R_{ij})\,, i = 1, \ldots, k$, represent the normal scores sums.

The calculation of the constants $d_i's, i = 1,\ldots,4$ follow the same lines of derivation as for the respective constants used with the uniform distribution order statistics previously.

The asymptotic value of the value of d_1 to meet the P^* requirement is the solution to:

$$\int_{-\infty}^{\infty} [\emptyset(x + hd_1)]^{k-1}\ \varphi(x)dx = P^* \tag{5.3.14}$$

where $\emptyset(x) and\ \varphi(x)$ are the CDF and density, respectively, of a standard normal random variable, and

$$h = h(n,k) = [\frac{(k-1)}{(n \cdot ssq)}]^{1/2} \quad (5.3.15)$$

and $ssq = \sum_{i=1}^{k}[a(R_{ij})]^2$.

The value of $d_3 = d_1$. The values of d_2 *and* d_4 are given by:

$$d_2 = [\frac{n \cdot ssq}{k}]^{1/2} \emptyset^{-1}(1 - P^*) \quad (5.3.16)$$

$$d_4 = -d_2 \quad (5.3.17)$$

5.4 State Motor Vehicle Traffic Fatality Rates

5.4.1 Data Description (MVTFRs)

The first data of interest in the upcoming analysis is the motor vehicle traffic fatality rates (MVTFRs) for each of the 51 states (including the District of Columbia) for the years 1994 through 2012. By definition the motor vehicle traffic fatality rate is the number of motor vehicle deaths per 100,000,000 (10^8) vehicle miles of travel (VMT). The motor vehicle fatalities are recorded for the state in which the accident occurs and the number of vehicle miles are estimated by the individual states. The National Highway Traffic Safety Administration (NHTSA) publishes the MVTFR for all U.S. states each year in the Fatality Analysis Reporting System (FARS). This data can be obtained from the government website: www-fars.nhtsa.dot.gov.

The data is used here to illustrate the impact that the two rank scores described earlier in section 5.3 have on the selected subsets using the selection procedures $R_1, \dots, R_4, Q_1, \dots, Q_4$. The data is given in Appendix A (to 2 dp) where the two letter abbreviation for states is given as the variable "State" and the fatality rates for the respective years are given "y1994",…, "y2012" in the order of states specified in "State".

Thus $k = 51$ populations (states) and $n = 19$ blocks (years) comprise the data set. The nonparametric subset selection based on ranks can be easily described by assigning a rank "1" to the state with the lowest MVTFR, "2" to the second lowest, and so on, until a rank of "k" ($k = 51$ in our case here) is assigned to that which has the highest MVTFR for each of $n = 19$.

The model (for the observation X_{ij}) to be used is of the additive form

$$X_{ij} = \mu + \theta_i + \beta_j + \varepsilon_{ij} \tag{5.4.1}$$

where, μ is the overall mean, θ_i is the state (population) effect, β_j is the year (block) effect, and ε_{ij} is the error term that may have any (but not necessarily normal) continuous distribution function $F_j(x)$ of mean 0. The observations are taken in n blocks (time periods). The j indicates the particular block to which the observation X_{ij} corresponds and i indicates the population, and $F_j(x, \theta)$ is a stochastically increasing family of distributions in θ.

The above two-way model is used in earlier studies of MVTFRs by McDonald (1979), by Lorenzen and McDonald (1984), and by Green, McDonald and Rao (2006). It is discussed extensively by Neter, Kutner, Wasserman and Nachtsheim (1996), Kuehl (2000), Winter (1971), Christensen (1998), and others.

Since there is only one observation for each state for each year, there is no general test for additivity, i.e., lack of interaction between states and years. Tukey (1949) developed a one degree-of-freedom test for non-additivity when there is a single observation per cell, as given here. Thus, a transformation of the MVTFRs will allow reasonable use of the additive model desired. Green, McDonald and Rao (2006) and Green and McDonald

(2009) used this test to establish the plausibility of the model (5.4.1) for a power transformation (raised to 0.3 power) of the MVTFRs.

That is, the two-way additive model for the transformed rates given as

$$X_{ij}^{0.3} = \mu + \theta_i + \beta_j + \varepsilon_{ij} \qquad (5.4.2)$$

is plausible as it indicates no significant evidence of interaction. For the purpose of the nonparametric analysis to follow, the original MVTFR data will be used because ranks are invariant to monotone increasing transformation. That is, since the power transformation is a monotone transformation, the ranks of the transformed data are identical to the ranks of the original fatality rates to be used here. The cited references provide a more detailed discussion of the data and the form of the assumed additive model.

5.4.2 Data Application

The goal is now to choose a subset of the 51 states that can be asserted, with a specified confidence, to contain the state with the highest MVTFR (worst population), and similarly a state with the lowest MVTFR (best population) using nonparametric ranking and selection procedures. The best population is defined to be that unknown state which has the lowest rate and, similarly, the worst population is the one with the highest rate. Based on these ranks, the selection procedure for choosing a subset of the 51 states asserts that the best state (or worst state) is contained with a specified confidence level P^*.

Similar to the structure as outlined in section 4.1 with the assumption being held, where each random variable $X_{ij}, j = 1, \dots, n \; and \; i = 1, \dots, k$ is assumed to be independent

samples of size n from k populations with a continuous CDF $F_j(x;\theta)$ that is stochastically increasing family distributions of θ, where θ_i's belong to interval Θ on the real line.

The selection procedures $R_1, \ldots, R_4, Q_1, \ldots, Q_4$ in section 5.3 can now be applied to the MVTFRs data after finding the rank sums ($T_i = \sum_{j=1}^{n} R_{ij}$) and normal scores sums ($S_i = \sum_{j=1}^{n} a(R_{ij})$) for each state.

Execution of the R-code in Appendix A yields the state rank sums and the state normal score sums listed in Table 5.2 form McDonald and Alsaeed (2024)

Table 5.2: Rank Sums and Normal Score Sums for MVTFRs, *k* = 51, *n* = 19

State	RK Sum	State	RK Sum	State	NS Sum	State	NS sum
MA	23	PA	288	MA	-41.30654	PA	-0.28389
CT	88	IA	492	RI	-28.60640	IA	-0.13781
RI	93	GA	499	CT	−28.16906	GA	0.25425
NJ	104	KS	605	NJ	−24.72987	KS	5.63425
MN	124	MO	606	MN	−24.61870	MO	5.64446
NH	135	TX	612	NH	−23.45682	TX	6.00957
WA	169	NC	614	WA	−18.93026	NC	6.02964
NY	181	OK	649	NY	−17.93579	OK	8.26094
MD	221	AK	652	VT	−17.14453	AK	8.68151
VT	226	FL	699	MD	−15.01171	FL	10.89390
VA	230	ID	714	VA	−14.56338	ID	11.77507

Table 5.2 – Continued

State	RK Sum	State	RK Sum	State	NS Sum	State	NS Sum
CA	258	NV	721	CA	−12.79833	NV	12.90143
OH	261	TN	744	OH	−12.38324	TN	13.44741
MI	304	AL	760	DC	−11.42801	AL	14.72405
IL	305	KY	769	MI	−10.20329	KY	15.43561
IN	313	WY	772	IL	−10.02509	NM	15.80883
WI	324	NM	773	IN	−9.65541	WY	16.41764
DC	325	SD	786	WI	−8.87739	SD	17.81267
ME	330	AZ	823	ME	−8.52842	AZ	20.20069
UT	350	WV	824	UT	−7.78518	WV	20.55222
OR	397	AR	887	OR	−5.02838	AR	25.49252
HI	445	LA	899	HI	−2.53294	LA	27.43415
ND	449	SC	910	ND	−2.33842	SC	28.77247
DE	453	MT	927	DE	−2.30543	MS	34.07669
NE	462	MS	930	NE	−1.60001	MT	34.48870
CO	469			CO	-0.36437		

However, as the rules are defined, the constants $b_1, \dots, b_4 \; and \; d_1, \dots, d_4$ need to be obtained first. Since the values of $k \; and \; n$ are large for this application, ($k = 51 \, , n = 19$), the selection rules are determined by the asymptotic formulas of the constants. Recall that the values for $b_1 \; and \; d_1$ are based on determining the values of the

products $c \cdot b_1 \; and \; h \cdot d_1$ based on (5.3.5),(5.3.6) and (5.3.14),(5.3.15) for given values of $k, n, and \; P^*$.These two products are equal and share a common value call it $w \, , so, \; c \cdot b_1 = w = h \cdot d_1$ and therefore the values of both $b_1 \; and \; d_1$ are calculated by dividing the product by $c \; and \; h$ respectively. Notice that w can be easily obtained by noting that the integral in (5.3.5) is an increasing function of it and by using the following short R-code such as

w<3.5

*func<-function(x){(pnorm(x+w))^50*dnorm(x)}*

integrate(func,lower = -Inf,upper = Inf)

and successive integral halving to converge on $w = 3.6666$ for $k = 51, n = 19$ and $P^* = 0.90$.

The resultant constants for implementing the eight subset selection procedures $R_1, \ldots, R_4, Q_1, \ldots, Q_4$ are given in the Table 5.3 (to 2 dp) given in McDonald and Alsaeed (2024).

Table 5.3: Selection rules constants for the MVTFRs ($k = 51, n = 19 \; and \; P^* = 0.90$)

R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
$b_1 =$ 237.53	$b_2 =$ 411.77	$b_3 =$ 237.53	$b_4 =$ 576.23	$d_1 =$ 15.72	$d_2 =$ -5.44	$d_3 =$ 15.72	$d_4 =$ 5.44

The code uses the R function “rank” to order the state MVTFRs for each of the nineteen years. This function provides six methods for ranking. The one used here is the

"random" option. If two states have the same fatality rate and are thus tied for, say, ranks r1 and r2, the allocation of those two ranks to the tied states would be done randomly, i.e., each state would have the same probability of assignment of r1 and r2. Consequently, for each of the years the 51 ranks are the whole numbers 1, 2, …, 51. The "average" option would assign to each of the tied states the average of r1 and r2. With averaging, not all of the states would have whole numbers assigned.

Using $P^* = 0.90$, it can be asserted using rules (5.3.1) and (5.3.2) that the following subset of states contain the one characterized as the worst population (state):

R_1: $Select\ \pi_i\ iff\ T_i \geq \underbrace{max}_{1 \leq j \leq k} T_j - b_1 = 930 - 237.53 = 692.47$. That is, all states with rank sums greater than or equal to 692.47 which are 16 states are chosen as worst states. *This means the 16 states chosen using R_1 can be asserted, with 90% confidence, to contain the state with the highest MVTFR, if 50 of the states have the same MVTFR and one (unknown) state has a rate at least as large as this common value, given that R_1 is applicable over a slippage parameter space and indicated before.*

R_2: $Select\ \pi_i\ iff\ T_i > b_2 = 411.77$. That is, all states with rank sums greater than 411.77 which are 30 states are chosen as worst states. Similarly, the same confidence statement can be made here. *The 30 states chosen using R_2 can be asserted, with 90% confidence, to contain the state with the highest MVTFR, given that R_2 is not limited to any constraints and can be applied over the whole parameter space.*

Similarly, with $P^* = 0.90$, it can be asserted using rules (5.3.3) and (5.3.4) that the following subset of states contain the one characterized as the best population (state):

R_3: $Select\ \pi_i\ iff\ T_i \leq \underbrace{min}_{1 \leq j \leq k} T_j + b_3 = 23 + 237.53 = 260.53$. That is, all states with rank sums less than or equal to 260.53 which are 12 states are chosen as best states. *This means the 12 states chosen using R_3 can be asserted, with 90% confidence, to contain the state with the lowest MVTFR, if 50 of the states have the same MVTFR and one (unknown) state has a rate at least as large as this common value, given that R_3 is applicable over a slippage parameter space and indicated before.*

R_4: $Select\ \pi_i\ iff\ T_i < b_4 = 576.23$. That is, all states with rank sums less than 576.23 which are 29 states are chosen as best states. Similarly, the same confidence statement can be made here. *The 29 states chosen using R_4 can be asserted, with 90% confidence, to contain the state with the lowest MVTFR, given that R_4 is not limited to any constraints and can be applied over the whole parameter space.*

Following the same methodology for the NORMAL SCORS SUMS selection rules, Table 5.4, given in McDonald and Alsaeed (2024) provides the number of selected states in the subsets chosen by the four rules using rank sums and the four rules using normal scores sums.

Table 5.4: Number of states chosen be selection rules with

$(k = 51, n = 19\ and\ P^* = 0.90)$

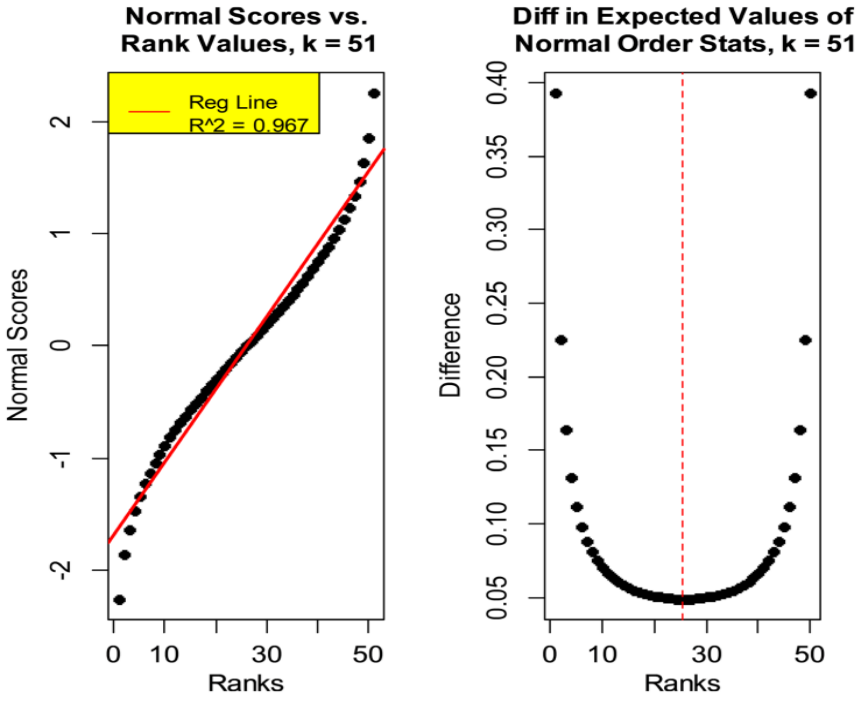
R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
16	30	12	29	7	31	3	29

Clearly, the number of populations chosen using the rank sums vs. the normal scores sums makes a substantial difference. The subset size using Q_1 is slightly less than

half of that using R_1 (7 vs.16). The subset size using Q_3 is a quarter of that using R_3 (3 vs. 12). However, the subset sizes produced from Q_2 *and* R_2 (Q_4 *and* R_4) are within one of each other (are equal), respectively.

The correlation between rank values and normal scores is very high (0.983), indicating a strong linear relationship between them. This suggests that as one variable increases, the other does as well implying that normal scores can be a good approximation of rank scores as they can be well explained by the normal scores and that the underlying distribution of the data is not too far from normality, or the transformation done maintains the relative ordering of the ranks effectively. The R-code of Appendix B produces the following figure that illustrates this relationship with a comparison of how the two scoring rules are similar/ different for $k = 51$.

Figure 5.1: Comparison of Normal Scores and Ranks for $k = 51$



Comparison of normal scores and ranks for $k = 51$.

The left plot displays the normal scores (which are often used to assess how well data fits a normal distribution) vs. the rank scores (where each value is assigned based on its position in the ordered list) along with the least regression line (Reg Line) fitted to the data. The closeness of the data points to the regression line can be quantified using measures like $R^2 = 0.967$, which indicates how well the regression line explains the variability in the data and suggests a good fit overall with the exception of the two or three end points on both sides of the line. The combination of normal scores and rank scores in a scatter plot with a least squares regression line provides a comprehensive view of the relationship between these two sets of data. It allows for both visual and quantitative assessments of how well they align, enabling better interpretations and predictions based on this relationship

However, the right plot highlights the deviation with a notable difference in the rank values and the normal score values is the spacing between the successive values. For rank values, the spacing between any two successive ranks of the uniform order statistics is $1/k + 1$ with k populations, and the values will increase by one for each successive value creating a consistent difference of 1 (which is a constant) between the rank values. In contrast, normal score values (derived from the standard normal distribution) have variable spacing. This is especially noticeable at the extreme values (tail ends) of the distribution, where the scores tend to spread out more than in the center and the spacing is much larger than the other spacing. Thus, the more extreme values carry a substantial difference weight than the more moderate values, and the spacing are symmetric about the midpoint 25.5 as indicated by the vertical line.

5.4.3 Extended Data Application

In this study, we expand upon our research by incorporating an extended dataset of the MVTFRs described in the previous section with a goal to significantly enhance the breadth and depth of our analysis. This new dataset includes additional observations and variables, where the fatality rates are given for 51 states (taking the District of Columbia as a state) for the years 1994, …, 2022, allowing for a more comprehensive examination of the underlying patterns and relationships. Thus, $k = 51$ populations (states) and $n = 29$ blocks (years).

By leveraging this enriched data, we aim to refine our conclusions and provide more robust insights into the phenomena under investigation. The integration of this extended dataset not only improves the statistical power of our analyses but also facilitates a nuanced understanding of the complexities inherent in the subject matter.

The analysis done is to be assumed carried out in a similar way done on the original dataset that has a different input (where $k = 51$ populations (states) and $n = 19$ blocks (years)) in accordance with the framework established in section 4.1, that each random variable $X_{ij}, j = 1, \dots, n \; and \; i = 1, \dots, k$ consists of independent samples of size n drawn from k populations and characterized by a continuous CDF $F_j(x; \theta)$ where these distributions are part of a stochastically increasing family with respect to the parameter $\theta \, \epsilon \, \Theta$ on the real line.

One point to just keep in mind in analyzing that extended data ($n = 29$) with the nonparametric selection rules is, we need to be reasonably assured of no interaction between states and years (as we did with the original data set with $n = 19$). The Tukey one degree of freedom test allows testing for this when there is just one observation per

cell as is the case here. The test with the raw rates rejects the hypothesis of no interaction. But testing with the rates raised to the 0.1 power gives a p-value greater than 0.9, and hence fails to reject the hypothesis of no interaction. Thus, we can proceed with the nonparametric selection rules since the rankings of the data raised to the 0.1 power are the same as the rankings of the original rates which is a similar logic followed before.

The selection procedures $R_1, \ldots, R_4, Q_1, \ldots, Q_4$ can now be applied to the expanded MVTFRs data after finding the rank sums for each state.

The R-codes in Appendix P and Appendix Q yield the state rank sums and the state normal score sums along with the constants needed to apply the selection rules, respectively. They can be viewed in Table 5.5 and Table 5.6 below

Table 5.5: Rank Sums and Normal Score Sums for MVTFRs, $k = 51$, $n = 29$

State	**RK Sum**	**State**	**RK Sum**	**State**	**NS Sum**	**State**	**NS sum**
MA	37	PA	746	MA	-41.30654	IA	-0.28714
RI	149	DE	772	RI	-28.64992	PA	-0.23157
MN	152	GA	803	CT	-27.74082	GA	0.30321
NJ	154	MO	900	NJ	-24.55634	KS	5.62678
CT	206	NC	927	MN	-24.55324	TX	5.73501
NH	241	KS	933	NH	-23.62102	MO	5.74621
NY	283	TX	990	WA	-18.69788	NC	6.01958
WA	293	AK	1005	NY	-18.07752	OK	8.22282
MD	311	NV	1015	VT	-17.02779	AK	8.73345

Table 5.5 - Continued

State	RK Sum	State	RK Sum	State	NS Sum	State	NS Sum
VT	324	ID	1038	MD	-15.43363	FL	10.93145
VA	382	OK	1058	VA	-14.48558	ID	11.77347
DC	404	WY	1079	CA	-12.89841	NV	12.91360
OH	417	FL	1102	OH	-12.66343	TN	13.51545
WI	423	AL	1116	DC	-11.39293	AL	14.72909
UT	442	TN	1117	MI	-10.29806	KY	15.35518
MI	491	SD	1145	IL	-9.94410	NM	15.93390
IL	492	NM	1195	IN	-9.70501	WY	16.39790
IN	493	KY	1211	WI	-8.87626	SD	17.91957
CA	502	AZ	1244	ME	-8.39963	AZ	20.08083
ME	534	WV	1261	UT	-7.69293	WV	20.38802
HI	601	AR	1337	OR	-4.85427	AR	25.49252
IA	666	LA	1357	HI	-2.58399	LA	27.43415
NE	691	MT	1393	DE	-2.46654	SC	28.80465
ND	708	SC	1417	ND	-2.27741	MS	33.96465
CO	735	MS	1424	NE	-1.64933	MT	34.71417
OR	738			CO	-0.36437		

The constants $b_1, \ldots, b_4$ & $d_1, \ldots, d_4$ needed to apply the selection rules are displayed in the table 5.6 below:

Table 5.6: Selection rules constants for the MVTFRs ($k = 51, n = 29\ and\ P^* = 0.90$)

R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
$b_1 =$ 293.49	$b_2 =$ 652.41	$b_3 =$ 293.49	$b_4 =$ 855.59	$d_1 =$ 19.42	$d_2 =$ -6.72	$d_3 =$ 19.42	$d_4 =$ 6.72

Now to apply rules $R_1, \dots, R_4, Q_1, \dots, Q_4$, with the information given in Tables 5.5 & 5.6, all needed is simple algebraic calculations that will give the number of states in the subset selected by a particular rule. See Table 5.7

Table 5.7: Number of states chosen be selection rules with

($k = 51, n = 29\ and\ P^* = 0.90$)

R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
10	30	10	29	11	31	6	33

In comparison to previous findings, one can clearly observe that the change of the number of blocks changes the outcomes. We can see, in the data with $n = 29$, that the number of states chosen by R_1 (10 worst states) is *less* than those chosen by Q_1 (11 best states) indicating that the rank sums are of a better choice as a scoring test to be applied here. That contradicts our findings before with the data where $n = 19$ and the results favored normal scores as the scoring test (R_1 chose 16 worst states in comparison to the 7 best states chosen by Q_1).

So, what scoring tests/rules should be carried out in practice with real-world applications? That remains to the analyst decide what rule to apply, yet analyses should

be conducted using multiple scoring rules. By applying various scoring rules, you can assess different aspects of the data, providing a more thorough understanding of the underlying patterns and relationships which helps to verify the consistency of the findings and can help identify the best-performing populations more effectively, as each rule may emphasize different strengths or weaknesses and if different methods yield similar conclusions, it increases confidence in the results.

Moreover, different scoring rules can show how responsive the findings are to modifications in the assumptions or properties of the data and how sensitive the results are to changes in assumptions or data characteristics. Carrying out the results of the R-code in Appendix Q listed in Table 5.8, we can examine how changes in the input assumptions or characteristics of the data affect the results of a model or analysis.

Table 5.8: Number of States Chosen by Selection Rules with Different P^*

$(k = 51, n = 29)$

P^*/ **Rule**	R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
0.99	16	31	14	32	16	38	9	37
0.95	12	30	10	29	13	33	6	34
0.90	10	30	10	29	11	31	6	33
0.75	9	28	6	29	7	30	3	29

The table shows that adjusting the value of the preassigned probability $P^* = 0.75, 0.90, 0.95, 0.99$ changes the outcomes, and that might suggest that the changes done on that character may have significant impact on the outcomes and may be critical to understanding the data being studied. This can help researchers see that their conclusions might depend heavily on the assumptions they make on particular characteristics of their data.

It demonstrates the relationship between P^* and the size of the selected subset. That is the number of populations included in the selected subset is a non-decreasing function of P^*. This means that as the confidence level increases, the size of the subset does not decrease; it either stays the same or increases. The justification behind it is that a higher confidence level generally requires a broader subset to ensure that the best population is included. For example, if a higher value of P^* is assigned (like 99% instead of 95%), you might need to include more populations to meet that requirement, as it is displayed in the table. Overall, the choice of P^* directly impacts the statistical power of the selection procedure. By specifying P^* and understanding its implications, analysts can better manage risks associated with making incorrect selections, ensuring that the decision-making process is statistically valid, after all, P^* represents the probability of a correct selection as defined.

5.5 State Homicide Rates

Data Description (SHRs) and Application

Homicide rates can vary significantly by state and overtime. They are typically expressed as the number of homicides per 100,000 residents. For the most current and

specific homicide rates by state, you would need to refer to the latest data from the CDC's National Center for Health Statistics in the website:

https://www.cdc.gov/nchs/pressroom/sosmap/homicide_mortality/homicide.htm

The data used here, as analyzed by Wang and McDonald 2022, can be found (to 2 dp) in Appendix C for the years 2005, 2014 to 2020 of 50 states. *(There is an existing gap from the year 2005 – 2013 whiles addressing the data as the CDC website does not contain the data through those years).* Thus, $k = 50 \; and \; n = 8$. An indicated rate of 0.00 is not actually zero since only 2 decimal points were retained for the data.

As in the MVTFR data described in the previous section, we would assume that the assumptions from section 4.1 are held, and the power transformation, $x^{0.4}$, applied to the SHRs results in a two-way additive model which, plausibly, lacks interaction between the categorical variables 'state' and 'year' based on the Tukey one degree-of-freedom test as shown in the cited reference. Since this transformation is monotone, analysis can be done directly with the original rates without the power transformation as it is a monotone increasing transformation. The rank sums and the normal scores sums, calculated with the R-code in Appendix C, are given in the following table.

Table 5.9: Rank Sums and Normal Score Sums for SHRs $(k = 50, n = 8)$

State	**RK Sum**	**State**	**RK Sum**	**State**	**NS Sum**	**State**	**NS Sum**
VT	24	WV	217	NH	−14.25764	WV	0.66234
NH	25	TX	230	VT	−13.84980	TX	1.31531
ME	27	PA	239	ME	−13.11673	PA	1.77072

Table 5.9 – Continued

State	RK Sum	State	RK Sum	State	NS Sum	State	NS Sum
UT	58	IN	244	ND	−10.29260	IN	1.83492
ID	61	KY	245	ID	-9.15845	KY	2.12983
MA	61	AZ	249	RI	−9.09363	AZ	2.64468
ND	61	FL	256	UT	−8.85947	FL	2.66931
MN	69	MI	266	MN	−7.98909	MI	3.22598
HI	72	NC	274	HI	−7.72306	NC	3.66404
IA	84	NV	282	WY	−7.07594	NV	4.18924
OR	98	AK	285	IA	−6.88568	DE	4.33158
WY	99	DE	285	OR	−5.78677	AK	4.78824
NE	103	OK	307	NE	−5.69598	OK	5.72751
CT	116	GA	314	CT	−4.66888	GA	6.03714
WA	128	IL	316	WA	−3.96127	IL	6.27973
NY	133	TN	333	NY	−3.72165	TN	7.46995
SD	149	AR	346	SD	−2.85342	AR	8.60699
MT	154	NM	347	MT	−2.57816	NM	8.72040
NJ	156	SC	356	NJ	−2.46596	SC	9.40882
CO	157	MO	361	CO	−2.39117	MO	10.12647
WI	158	MD	362	WI	−2.34701	MD	10.37644
KS	196	AL	384	KS	−0.40414	AL	13.09039
VA	199	MS	390	VA	−0.24508	MS	14.87734

Table 5.9 – Continued

State	RK Sum	State	RK Sum	State	NS Sum	State	NS Sum
CA	202	LA	398	CA	−0.06694	LA	17.20416

Applying the selection rules $R_1, \ldots, R_4, Q_1, \ldots, Q_4$ to the SHRs data set with $P^* = 0.90$, the constants $b_1, \ldots, b_4$ *and* $d_1, \ldots, d_4$ need to be obtained as in subsection 5.4.2 and are given in Table 5.6 (to 2 dp).

Table 5.10: Selection rules constants for the SHRs ($k = 50, n = 8$ *and* $P^* = 0.90$)

R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
$b_2 = 148$	$b_2 = 151$	$b_3 = 148$	$b_4 = 257$	d_1 =10.13	d_2 =-3.53	$d_3 = 10.13$	$d_4 = 3.54$
150.85	*151.7*	*150.85*	*256.3*	*10.18*	*-3.53*	*10.18*	*3.53*

For this data set, $k = 50$ states and $n = 8$ years. The b_1 (and b_3) are obtained from the R-code given in Appendix D based on 50,000 simulations. The b_2 *and* b_4 values are obtained using the Appendix E R-code. Similarly, the d_1 (and d_3) are obtained from Appendix F, and d_2 *and* d_4 from Appendix G. A simulation approach seems preferable to the asymptotic approach used in subsection 5.4.2 since n is relatively small. For comparison, the asymptotic values of the selection constants are given in the last row of Table 5.10 in *italics* and are seen to be quite close to the simulated values in the first row. The sum of squares for the normal scores, ssq for

$k = 50\ and\ n = 8$, is 47.4217 and is used in the calculations for the asymptotic values, where $ssq = \sum_{i=1}^{k}\left[a\left(R_{ij}\right)\right]^2$ as defined before. Table 5.11 provides the number of selected states in the subset chosen by the four rank sums rules and by the four normal scores sums rules.

Table 5.11: Number of states chosen be selection rules with

$(k = 50, n = 8\ and\ P^* = 0.90)$

R_1	R_2	R_3	R_4	Q_1	Q_2	Q_3	Q_4
19	32	22	32	9	33	15	34

As noted for the MVTFRs before in Table 5.4, clearly here the number of populations chosen using the rank sum vs. the normal score sum makes a substantial difference. The subset size using Q_1 is less than half that using R_1 (9 vs. 19). The subset size using Q_3 is substantially less in comparison to that of R_3 (15 vs. 22). However, the subset sizes $Q_2\ and\ R_2\ (Q_4\ and\ R_4)$ are within one (two) of each other. The results of the comparative analyses of the MVTFRs and the SHRs are very similar.

5.6 Key Findings and Related Remarks

Comparing the rules R_1 and R_2, it is observed that, R_2 chooses substantially more populations in the selected subset than does rule R_1. This might be expected since R_2 guarantees a probability of correct selection to be no less than P^* for any configuration of the population θ-parameters, while that guarantee for rule R_1 is proven for slippage configurations of the θ-parameters. However, limited simulation studies do suggest that

the stronger unconstrained P^* guarantee for R_1 may hold for some classes of distributions (e.g., see Lorenzen and McDonald 1984). In general, for the rank sums,

$$\frac{n(k+1)}{2} \leq \max\left(T_j, j = 1, \ldots, k\right) \leq n \cdot k \tag{5.6.1}$$

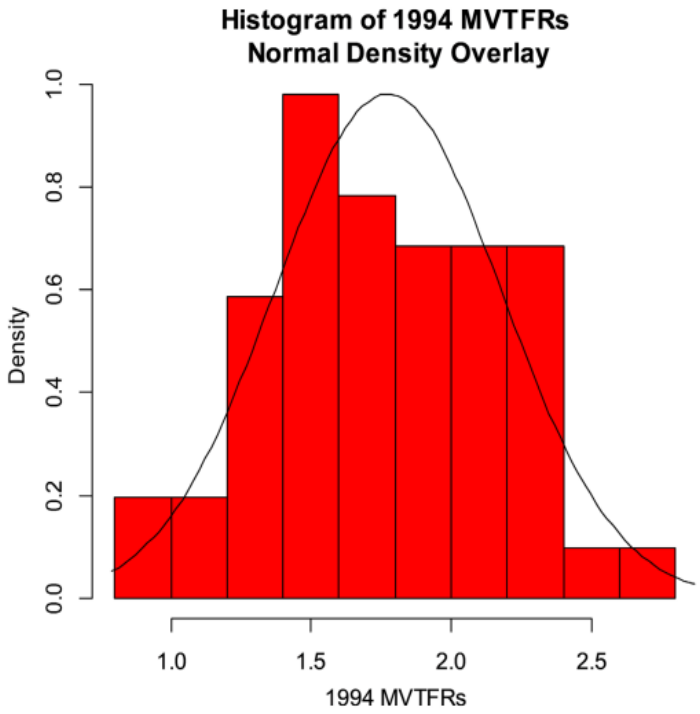
So, for $k = 51\ and\ n = 19$, the two-sided inequality becomes: $494 \leq \max\left(T_j\right) \leq 969$. For $P^* = 0.90$, $b_1 = 237.53$, so $494 - 237.53 \leq \max\left(T_j\right) - b_1 \leq 969 - 237.53$ yields to $256.47 \leq \max\left(T_j\right) - b_1 \leq 731.47$. With $b_1 = 237.53$ and $\max\left(T_j\right) = 930$, then we have $256{,}47 \leq 930 - 237.53 \leq 731.47$ or equivalently $256.47 \leq 692.47 \leq 731.47$ and thus rule R_1 selects all states such that $T_i \geq 692.47$. (Recall, $R_1: select\ \pi_i\ iff\ T_i \geq \max T_j - b_1$).

For rule R_2 the determination of constant b_2 does not depend on the ranks, it depends only on $k, n\ and\ P^*$. So with $b_2 = 411.77$ the rule R_2 chooses all states such that $T_i > 411.77$. (Recall, $R_2: select\ \pi_i\ iff\ T_i > b_2$). So which rule places more population in the selected subset depends on $\max T_j$. If its value is relatively close to the upper bound $n \cdot k$, then R_1 chooses fewer populations than R_2. If its value is relatively close to the lower bound $n(k + 1)/2$, then R_1 chooses more populations than R_2.

It is clear that in both datasets used, the rank sums scores and the normal score sums produced different numbers of the selected subsets. With the traffic fatality rates considered earlier, rule Q_1 placed seven states in the selected subset and rule R_1 placed sixteen states in the selected subset. So, it is natural to wonder which of these two rules to use in practice? From Figure 5.1, it appears that Q_1 would be the appropriate choice when it is desired to place relatively greater weight on the extreme three or four observations and the underlying distribution of the data is approximately symmetric. Figure 5.2 shows

the values of the MVTFRs for the year 1994 to be approximately symmetric and normal, a characteristic shared by most of the years. Such a pattern seems to favor the choice of normal scores over the rank scores.

Figure 5.2: Distribution of 1994 MVTFRs for states



Distribution of 1994 motor vehicle traffic fatality rates for states.

The same statements would apply to the choice between Q3 and R3. Clearly there is more work to be done in this area of statistical inference. Sections 5.4 and 5.5 compared only two scoring rules based on the expected values of order statistics from two distributions, the uniform distribution and the normal distribution. Substantial differences in the size of selected subsets result from the application of these nonparametric subset selection rules to a study of state motor vehicle traffic fatality rates (state homicide rates)

over a nineteen (eight) year period. So is it possible for R_1 to place fewer populations in the selected subset than rule Q_1?

While the applications given in sections 5.4 and 5.5 demonstrate that the selected subset size using rule Q_1 (normal scores) can be smaller than that using rule R_1 (rank scores), it should be noted that this is not always so.

Consider the case where the probability distributions for each of the populations share support over the same interval. Then it's possible that within each block any rank order of the population observations can occur. Using the R-code in Appendix H with $k = 7, n = 2, and\ P^* = 0.75$, it's determined that $b_1 = 6\ and\ d_1 = 2.70436$. Assuming the observations yield the ranked values given in Table 5.12 as follows:

Table 5.12: Ranked data for ($k = 7\ , n = 2\ and\ P^* = 0.75$)

Popula-tion	π_1	π_2	π_3	π_4	π_5	π_6	π_7
Block1 ranks	1	2	3	4	5	6	7
Block2 ranks	1	3	2	7	5	6	4
T_i	2	5	5	11	10	12	11
S_i	-2.7043	-1.1100	-1.1100	1.3521	0.7054	1.5147	1.3521

Then using the selection rules given in (5.3.1) and (5.3.10) we have:

$R_1: select\ \pi_i\ iff\ T_i \geq \max T_j - b_1 = 12 - 6 = 6$. So, rule R_1 chooses **four** populations, and

$Q_1: select\ \pi_i\ iff\ S_i \geq \max S_j - d_1 = 1.51474 - 2.70436 = -1.18962$. So rule Q_1 chooses **six** populations (*a larger number*).

With $k = 7$, there are $7! = 5040$ permutations of possible rank orders for a given block. Thus, with $k = 7\ and\ n = 2$ there are $5040^2 = 25{,}401{,}600$ possible rank order configurations for this experimental design with two blocks. The number of these configurations yielding specific differences in number of populations chosen by the two ranking procedures is calculated with the R-code in Appendix H and is given in Table 5.13.

Table 5.13: the Number of Configurations (N) for Which the Number of Populations Chosen by R_1 less than the Number Chosen by Q_1 is Equal to Δ for ($k = 7, n = 2\ and\ P^* = 0.75$)

Δ	**-2**	**-1**	**0**	**1**
N	40,320	665,280	23,486,400	1,209,600

With these closing findings, what ought to be done in real life? The state of theoretical development along with observance of outcomes using differing scoring rules, suggests analyses should be carried out with several scoring rules, such as ranks and normal scores used here, including but not limited to Cauchy, logistic, and lognormal scoring. This approach allows for a more comprehensive evaluation of data, enabling decision-makers to identify the most appropriate scoring method for their specific

context. For a fixed value of P^*, analyses should focus on determining the results that yield the smallest subset size. This practice ensures efficiency in decision-making, allowing stakeholders to make well-informed choices without unnecessary complexity. Then use the results that yield the smaller subset size for the given probability of correct selection criteria. Incorporate these methods into existing decision-making frameworks to enhance the interpretability and effectiveness of results. This integration will facilitate a clearer understanding of how statistical outcomes translate into practical actions.

CHAPTER SIX

PERFORMANCE OF THE SELECTION PROCEDURES UNDER DIFFERENT CONFIGURATION OF POPULATIONS' PARAMETERS

6.1 Introduction to the Least Favorable Configuration (LFC)

The subset selection procedure selects a subset of the populations with the goal of including within the chosen subset the population which is "best" (e.g., that characterized with the largest mean), with a prescribed probability no less than a user prescribed probability, P^* . Specifying such procedures requires determination of the parameter configuration that *minimizes* the probability of a correct selection and subsequent calculation of the constant required to implement the procedure so that the probabilistic inference is valid over the entire parameter space. This minimizing configuration is referred to as the "least favorable configuration". Generally, the term least favorable configuration, LFC, usually refers to the least advantageous or optimal where the performance across the options under study is particularly close, or statistically indistinguishable. In statistics, the LFC might be the distribution of data that maximizes the variance or error, making it hard to accurately estimate parameters as this can complicate identifying the best option or subset of options. The concept of LFC is important, especially for research containing decision making because it helps to clarify boundaries of performance and efficiency and in developing algorithms or methods that work well in the absolute worst case since the occurrence of the LFC impacts the traditional ranking procedures may turn out to be indecisive, which, in turn, increases uncertainty during the choice-making process. In general, the problem of determining the

LFC on population parameters is an open question, for selection procedures based on ranks.

Recall that the P^*- condition is defined as:

$\underbrace{Inf}_{\Omega} P\{CS|R\} = P^*$ where $\Omega = \{\theta = (\theta_1, \ldots, \theta_k) : \theta_i \epsilon \Theta$ some interval on the real line, $i = 1, \ldots, k\}$. The LFC is represented when all populations are **identically distributed** and the population parameters are the same, i.e., the LFC, Ω_0 , is given by the Equi parameter configuration and so, $\Omega_0 = \{\theta \epsilon \Omega: \theta_{[1]} = \theta_{[2]} = \ldots = \theta_{[k]}\}$.

However, Rizvi and Rizvi and Woodworth (1970) and McDonald (1972) show that for an unrestricted parameter space, the LFC occurs when all populations are identically distributed, and they present a class of distributions for which this condition ***does not hold*** for rules R_1 and R_3 which are justified over a slippage space, Ω' as derived in McDonald (1972). Various discussions have been made to get the LFC of some procedures. Their investigation involved two different population parameter configurations: slippage and equal spaced configurations.

6.1.1 Slippage Configuration

The first step in the direction of a somewhat more realistic approach came in the form of a test proposed for deciding whether one of populations has *slipped* to the right of the rest, under the null hypothesis that all populations are continuous and identical. So, suppose we have k samples of size n each. It is desired to test the alternative hypothesis that one of the populations, from which the samples were drawn, has been rigidly translated to the right relative to the remaining populations. The null hypothesis is that all the populations have the same location parameter. In other words, we test the hypotheses

mentioned above against one or more of the alternatives, which implies slippage to the right, or the left.

Statistically speaking and with respect to experimental design and analysis, the word slippage refers to a case when some kind of misalignment or error occurs in the intended structure of an experiment or data collection process. It typically refers to configurations or settings that result in inefficiencies or discrepancies from expected outcomes. The primary goal is to improve model accuracy and reliability. By adjusting parameters, researchers or analysts can better reflect the conditions or dynamics of the system being studied. This parameter configuration is described where all parameters θ_i are equal with the possible exception of $\theta_{[k]}(\theta_{[1]})$ in case of rule $R_1(R_3)$. Thus, $\Omega' = \{\boldsymbol{\theta} \in \Omega : \theta_{[1]} = \cdots = \theta_{[k-1]}\}$.

6.1.2 Equi-spaced Configuration

Equi-spaced configuration simply means a set of all points or elements which are evenly spaced or distributed. Evenly spaced parameters can help ensure that the entire range of parameter space is explored, aiding in identifying trends and behaviors. It is commonly used to assess how changes in parameters affect outcomes, providing a clear view of parameter influence. Equi-spaced configurations form the basis of a lot of mathematical and statistical methods since they ensure homogeneity and systematic covering of a domain under study for more precise and reliable results. This parameter configuration is described as $\Omega'' = \{\boldsymbol{\theta} \in \Omega : |\theta_{[i]} - \theta_{[i+1]}| = \delta, \delta > 0\}$.

6.2 Two-Way Randomized Block Design Model

In many statistical procedures, especially those involving probability and inference, constants are used to calibrate the method. This could be related to confidence

levels, error rates, or other statistical measures that ensure the method's validity. Computing the constants required to implement a specific nonparametric subset selection procedure based on ranks of data is widely studied in Statistical Randomized Block Experimental Designs. Randomized and Small block designs, which are commonly employed to control for variability within the blocks when experimental subjects vary in natural heterogeneity, are experimental designs in which treatments are organized into blocks smaller than the total number of treatments and every treatment is applied in every block. *However, it is sometimes impractical or impossible for all the treatments to be applied to each block, especially when the number of treatments is large, and the block size is limited. For example, if 20 different foods (treatments) are to be tasted, each judge (block) may find it quite difficult to rank accurately all 20 foods in order of preference. But if each judge tastes only 5 foods and then four times as many judges are used, the judging may be easier and more accurate. Thses experimental designs in which not all treatments are applied to each block are called incomplete block designs. This is not the focus of our application here, but it is worth mentioning.*

A model of three populations and two blocks (to get exact calculations) is used to compute the probability distribution of the relevant statistic, the maximum of the population rank sums minus the rank sum of the “best” population. Calculations are done for populations following a normal distribution, and for populations following a bi-uniform distribution.

Nonparametric (distribution-free) subset selection procedures are based on random data assumed to follow a continuous probability distribution stochastically ordered by the parameter of interest (e.g., a location or a scale parameter). The subset

selection procedure selects a subset of the populations with the goal of including within the chosen subset the population which is "best" (e.g., that characterized with the largest mean), with a prescribed probability no less than a user prescribed probability, P^*, that is the LFC. In the next section, the LFC for selection rule R_1 described earlier, will be thoroughly examined under two cases: normal populations differing in their mean values and bi-uniform counterexample distributions also differing in location parameters. The examples will be for a small model of $k = 3 \; and \; n = 2$ so as to facilitate exact calculations.

6.2.1 Normal Populations

The normal distribution is chosen as an example of a widely used distribution in practice that is unimodal and symmetric. We begin by deriving a general expression for $P_{123} = \Pr(X_1 \leq X_2 \leq X_3)$, where $X_i \sim Normal\ (\mu_i, \sigma_i)\ ,\ \ i = 1,2,3$, and the three variates are independent.

$$P_{123} = Pr[(X_1 - \mu_2)/\sigma_2 \leq (X_2 - \mu_2)/\sigma_2 \leq (X_3 - \mu_2)/\sigma_2] \tag{6.2.1}$$

$$= \int_{-\infty}^{\infty} \Pr\left(\frac{X_1 - \mu_2}{\sigma_2} \leq x\right) \cdot \Pr\left(x \leq \frac{X_3 - \mu_2}{\sigma_2}\right) \varphi(x)dx$$

$$= \int_{-\infty}^{\infty} \emptyset\left((\mu_2 - \mu_1 + x \cdot \sigma_2)/\sigma_1\right) \cdot \emptyset\left((\mu_3 - \mu_2 + x \cdot \sigma_2)/\sigma_3\right) \varphi(x)dx$$

where $\emptyset(x) \; and \; \varphi(x)$ are the cumulative distribution function (cdf) and probability density function (pdf), respectively, of a standard normal random variable (mean = 0, standard deviation = 1).

While Equation (6.2.1) and Appendix I is structured for three independent normal distributions, they can be easily adjusted to handle other location-scale families of distributions such as the logistic, t-distributions, etc.

There are six permutations of the digits 1,2,3 and hence there are six orderings of X_1, X_2, X_3 given as follows: Order1 $(X_1 \leq X_2 \leq X_3)$, Order2 $(X_1 \leq X_3 \leq X_2)$, Order3 $(X_2 \leq X_1 \leq X_3)$, Order4 $(X_2 \leq X_3 \leq X_1)$, Order5 $(X_3 \leq X_1 \leq X_2)$, Order6 $(X_3 \leq X_2 \leq X_1)$. The probabilities of these orders are designated by: Prob1, Prob2, …, Prob6, respectively. These six probabilities are calculated with the R-code in Appendix I using (6.2.1) matching the specific population means and standard deviations with the appropriate permutation of the population indices (1, 2, 3). Appendix , as herein stated, is set for the three means equal to 0 and the three standard deviations set equal to 1. These probabilities are the probabilities of the possible six orderings of the random draws from the three populations in each of the two blocks. There are thus 36 possible outcomes in the ordering of the data in a randomized block design for three populations and two blocks. The probability of permutation 1 for block one and permutation 2 for block two is, by block independence, Prob1 times Prob2. Similar calculations hold for the other thirty-five joint occurrences for the permutations. The probabilities of these 36 joint permutations events are given as the *df1 data.frame output.* The column under Prob1 are the probabilities of the six joint occurrences of permutation 1 with each of the six permutations 1 through 6 in that order. Similarly, the same type probabilities occur under the columns Prob2 through Prob6. The *df1 data.frame* provides the probabilities for the sample space of all possible rank orderings of a randomized block designed experiment with three normal populations and two blocks. The rank sums of the three populations can now be calculated along with the corresponding probabilities and from these, the probability mass function (pmf) and cumulative distribution function (cdf) of the statistic

$S = \underbrace{max}_{1 \le j \le 3} T_j - T_3$ can be determined.

Note that a CS, using the rule (5.3.1) $R_1: select\ \pi_i\ iff\ T_i \geq \underbrace{max}_{1 \le j \le 3} T_j - b_1$, occurs when

$T_k \geq \underbrace{max}_{1 \le j \le 3} T_j - b_1, or\ T_3 \geq \underbrace{max}_{1 \le j \le 3} T_j - b_1, given\ k = 3$. Which is equivalent to

$b_1 \geq \underbrace{max}_{1 \le j \le 3} T_j - T_3\ , or\ S \leq b_1$. Thus, the cdf of the statistic S is the relevant quantity

linking the value of b_1 with the preassigned P^*.

To illustrate the appropriate computation, suppose both blocks 1 and 2 yield Order1$(X_1 \leq X_2 \leq X_3)$. Then table 6.1 gives an explanation to compute the value of S as follows:

Table 6.1: Value of S when both blocks 1 and 2 yield Order1$(X_1 \leq X_2 \leq X_3)$

Block/ π	π_1	π_2	π_3
Block 1	$X_{11} \approx R_{11} = 1$	$X_{21} \approx R_{21} = 2$	$X_{31} \approx R_{31} = 3$
Block 2	$X_{12} \approx R_{12} = 1$	$X_{22} \approx R_{22} = 2$	$X_{32} \approx R_{32} = 3$
RK Sum T_i	$T_1 = 2$	$T_2 = 4$	$T_3 = 6$

And so, $S = \underbrace{max}_{1 \le j \le 3} T_j - T_3 = 6 - 6 = 0$. Applying same logic with the 36 possible

outcomes in the ordering of the data, therefore, the statistic S can assume only the values $d = 0,1,2,3,4$. The subsets of the 36 permutation configurations that result in values of 0, 1, 2, 3, 4 are given in Appendix I and coupled with the probabilities of the configurations, yield the pmf of S (denoted by Pr0, …, Pr4) and the cdf of S (denoted by cumdf0, …, cumdf4) displayed in Table 6.2 below.

Table 6.2: Pmf and cdf for the statistic $S = \underbrace{max}_{1 \le j \le 3} T_j - T_3$ with

$$\mu = c(0,0,0) \; and \; \sigma \; = c(1,1,1)$$

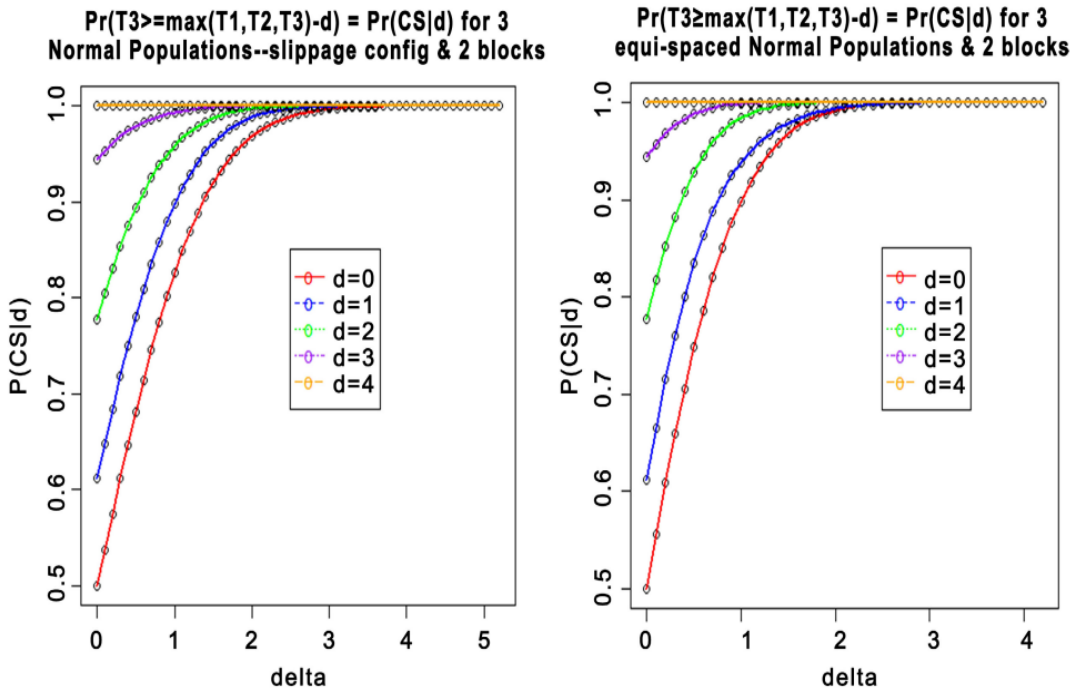
d	**Pr(*S*=*d*)**	**Pr(*S* ≤ *d*)**
0	0.5	0.5
1	0.11111	0.61111
2	0.16667	0.77778
3	0.16667	0.94444
4	0.05556	1

Implementing the subset selection rule R_1 requires the calculation of b_1 so that the P^*-condition is satisfied, i.e., that the probability of a correct selection is no less than a prescribed P^* for any underlying population parameter configurations. The analysis explores the LFC when the populations are normally distributed differing only in their mean values under the slippage configuration where two populations have an equal mean value of 0 and the third has a mean value of $\delta \geq 0$. The R-code in Appendix I is run with standard deviations equal to 1 and the mean values set at $c(0,0,\delta)$ with $\delta \; = \; 0 \; (0.1) \; 5.2$. The values of the cumcdf0, …, cumcdf4 are entered into the R-code of Appendix J as Pr0, …, Pr4 which then generates Figure 6.1 left.

The second is the equi-spaced configuration, where the first population has mean value of 0, the second population has mean value δ, and the third mean value 2δ. As

before, the R-code of Appendix I is run with standard deviations equal to 1 and the mean values set at $c(0, \delta, 2\delta)$ with $\delta = 0\ (0.1)\ 4.2$. The values of the cumcdf0, …, cumcdf4 are entered into the R-code of Appendix K as Pr0, …, Pr4 which then generates Figure 6.1 right.

Figure 6.1: Pr(CS) for Normal populations with equal variances and slippage configuration (left) and equal spaced configuration (right), $k = 3, n = 2$



As shown, in both parameter configurations the $Pr(CS|d)$ is a nondecreasing function of δ for $d = 0\ (1)\ 4$ and is minimized at $\delta = 0$, i.e., when the three populations have identical probability distributions, providing that the LFC occurs indeed when all population parameters are equal.

6.2.2 Bi-Uniform Distributions

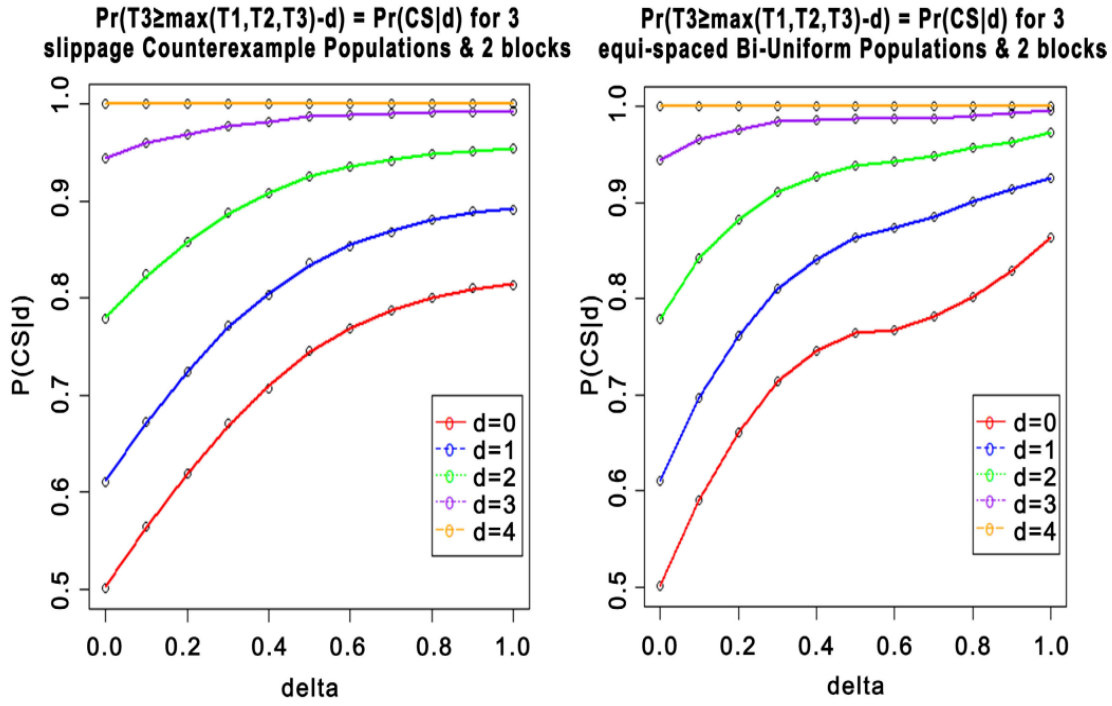
McDonald (1972) gives a class of continuous distribution functions for which the $Pr(CS)$ for nonparametric selection rule R_1 does not occur when the populations are

identically distributed, and hence yields a counterexample to that hypothesis. The bi-uniform distribution (counterexample distribution) is chosen since it is the singular example of a family of distributions which does not possess the "usual" property for the LFC and hence limits an unqualified endorsement for the nonparametric subset selection procedures (R_1 and R_3). For $0 < a < 1 < c$, let $F(x)$ be defined as:

$$F(x) = \begin{cases} 0 & x \le -c \\ c + x/2a & -c < x \le 0 \\ 1/2 & 0 < x \le 1 \\ \dfrac{a + x - 1}{2a} & 1 < x \le 1 + a \\ 1 & 1 + a < x \end{cases} \qquad (6.2.2)$$

$F(x)$ is composed of two non-overlapping uniform distributions: one on the interval $(-c, 0)$ and the other on the interval $(1, 1 + a)$ and each scaled to have mass 0.5. The counterexample is constructed for $k \geq 3$, the population location parameters are equal with the exception of the two largest which are themselves equal, and n is asymptotically large. A sufficiently large value of "c" can be chosen to establish the counterexample. In the analysis done here, the case $n = 2$ is considered. The R-code of Appendix L generates the cdf of the statistic S based on a large sample of bi-uniform variates as defined in (6.2.2). Based on input values of c, a, δ_1, δ_2, and nsim. The values δ_1, δ_2 are location parameters of the second and third population, and nsim is the number of random draws from the three populations. The nsim draws are structured in pairs so as to provide nsim/2 random block experimental designs with three populations and two blocks. The R-codes of Appendix M and Appendix N generate output similar to that of Appendix J and Appendix K respectively. The output is displayed in Figure 6.2. The cdf of the statistic S is based on nsim = 200,000 in Appendix L.

Figure 6.2: $Pr(CS)$ for Bi-Uniform populations with $c = a = 1$, slippage configuration (left) and equal spaced configuration (right), $k = 3, n = 2$



The patterns in Figure 6.2 are qualitatively the same as those in Figure 6.1. That is, in both parameter configurations the $Pr(CS|d)$ is a nondecreasing function of δ and is minimized at $\delta = 0$, i.e., when the three populations have identical probability distributions, providing that the LFC in this small model occurs indeed when all population parameters are equal.

6.3 Key Findings and Related Remarks

The analysis done in the previous section was key to get an idea on the performance of the selection procedure and searching for cases when LFC should occur

minimizing Pr (CS) under the two different configurations that lead to be when all the populations are identically distributed. As for the Bi-Uniform (as a counterexample leading to rejecting that hypothesis), it should be stated that to establish that to establish such counterexample, a critical inequality must be held. That is for any $\varepsilon > 0$, there exists $0.5 < P_\varepsilon < 1$ such that

$$A = A(P_\varepsilon, k) \leq (1 + \varepsilon)2^{1/2}\, \emptyset^{-1}(P_\varepsilon) \tag{6.3.1}$$

and A is given by:

$$\int_{-\infty}^{\infty} \emptyset^{k-1}(x + A)\, \varphi(x) dx = P_\varepsilon \tag{6.3.2}$$

For $k = 3$, the value of ε for the counterexample can be shown to be (to 5 dp) 0.06066. Using the R-code in Appendix O yields the values of A and P_ε to be 5.0 and 0.99960 (to 5 dp), respectively, to establish the counterexample for large sample sizes (blocks) n. The value of P_ε is excessively large for most practical applications making it become less of a big constraint.

In contrast, revisiting the Bi-Uniform distribution but for small model ($k = 3, n = 2$) was to get exact calculations and demonstrate when $Inf\ P(CS|R)$ occurs, reinforcing our belief (*not as a proof*) that for practical applications $Inf\ P(CS|R)$ occurs where it should.

The exact small model results given here, along with larger model simulation results, support the use of identical population distributions for the LFC and computation of P^* for the lower bound of the $Pr(CS)$. The results found indicate that the use of identical population distributions is justified, as they lead to consistent outcomes in both

small model analyses and larger simulations. This reinforces the validity of statistical methods that rely on this assumption.

One last point to add is for a given value of P^* and unrestricted parameter space, the decision rule R_2 (or R_4) is the only one which can be used with the assurance that the probability of a correct selection is at least as large as P^*. When in fact, the parameter space represents a slippage configuration, both $R_1 \& R_2$ ($R_3 \& R_4$) can be used with the guarantee that the basic probability requirement is satisfied. The choice of which rule to actually implement should then be based on additional operating characteristics of the selection procedures. A reasonable measure of "goodness" of a selection procedure is the ration of the expected number of populations in the selected subset to the probability of a CS. A rule *R* is said to be "better" than rule *R** if the ratio for *R* is less than the corresponding ration for *R**.

CHAPTER SEVEN

CONCLUSION

This research explored ranking and selection procedures through the eye of a nonparametric statistical process. Primary advantages in the nonparametric process are computational ease and valid inference based on weak assumptions. Since there are many practical situations where it may be better to select more than one population, the focus of this was the approach to subset selection of the best population (defined by the analyst depending on certain performance of the parameter of interest) from a set of candidates without relying on specific distributional assumptions. Various methodologies were highlighted throughout the study, including rank-based scoring tests on the expected values of order statistics that demonstrated how applicable they are across different scenarios and under distinct characteristics, as specific conditions or rules used to choose items can vary widely.

First, the main elements of the analysis that are to be known prior to the analysis were presented and it was shown how notably they can affect the decision-making process on the final results and how they can make a big impact on the size of the selected subset which is a crucial aspect that influences the robustness and reliability of the conclusions drawn from statistical analyses.. The number of populations, number of blocks along with incorporating a probabilistic element, allowing users to specify the desired likelihood of selecting the correct population. This means that the method not only aims for accuracy but also quantifies the certainty associated with the selection.

Four block design selection rules to provide a screening technique to that purpose were illustrated, for which two are aiming to choose the 'best' population and the other to choose the 'worst' one. With respect to nonparametric methods, the development of easy ways (asymptotically or by simulation) to compute the necessary constants used to relate a selected subset size to the probability of making a correct selection was applied using either calculator based or R-coding methods.

The calculations of such contents were done through scoring rules that are based on the expected values of order statistics from uniform distribution (yielding rank sums) and normal distribution (yielding normal score sums) and it was suggested that the ranking of items correlates very closely with their transformed normal scores. This means that higher ranks tend to correspond to higher normal scores, indicating a strong consistent relationship. A comparison was made via the datasets of MVTFRs over 19 years and SHRs over 8 years and a substantial difference in the size of selected subsets resulted that using the rule Q_1 (normal scores) is smaller than using the rule R_1. Which raised the question of whether the normal score sums will be always preferred in practical situations. It was noted, however, that it is not always so and a case of 8 populations and 2 block was considered, and it was shown that there are possible configurations yield smaller subset sizes by R_1 compared to that of Q_1. The status of these theoretical developments and findings and the utilizing various scoring rules indicates that many scoring rules should be used for studies. In addition, this result got the inspiration for future studies to discover and examine how other scoring rules can be applied, like Cauchy and lognormal scores then use the results that yield the smaller subset size for the given probability of correct selection criteria.

Extended data for MVTFRs was used and the same scoring rules implemented on the original data, implemented here as well with different values of P*. It was revealed that a higher confidence level generally requires a broader subset to ensure that the best population is captured. By allowing for flexibility and strategic decision-making, the relationship between P* and the size of the chosen subset allows analysts to strike a compromise between the number of populations taken into consideration and confidence improving the statistical inference's resilience. Ultimately, the value P* is essential for determining which populations are the best or worse.

Moreover, the configuration of population parameters for which P* is minimized known as the least favorable configuration was investigated and the done for populations following normal distribution, and bi-uniform distribution revealing that, indeed for a smaller model of 3 populations and 2 blocks, the LFC occurs when all populations are identically distributed, and the parameters are equal.

In conclusion, this research contributes to the understanding of nonparametric subset selection procedures by demonstrating their utility and effectiveness in practical applications. The results highlight how crucial it is to choose suitable approaches depending on the properties of the data and the objectives of the research. Future research ought to concentrate on improving these processes and extending their application to more complex data structures. In the end, this research emphasizes the necessity of ongoing innovation in subset selection techniques to improve decision-making in a variety of fields.

APPENDIX

CODES USED

Appendix A

```
#MVTFR Ranks 1994_2012
#Rank sums for the MVTFRs
#WIREs Comput Stat 2016, 8:222-237, doi: 10.1002/wics.1385
#k is the number of populations (e.g., states); n is the number of blocks (e.g.,years)
k=51;n=19
State=c("AL","AK","AZ","AR","CA","CO","CT","DE","DC","FL","GA","HI","ID","IL"
,"IN","IA","KS","KY","LA","ME","MD","MA","MI","MN","MS","MO","MT","NE","N
V","NH","NJ","NM","NY","NC","ND","OH","OK","OR","PA","RI","SC","SD","TN","
TX","UT","VT","VA","WA","WV","WI","WY")
y1994=c(2.21,2.05,2.33,2.44,1.56,1.74,1.14,1.59,2.00,2.2,1.72,1.54,2.15,1.68,1.59,1.86,1
.79,
1.95,2.25,1.51,1.47,0.94,1.67,1.49,2.77,1.9,2.22,1.75,2.26,1.13,1.26,2.18,1.49,1.99,1.39,
1.4,1.93,1.68,1.56,0.89,2.27,2.02,2.23,1.79,1.9,1.25,1.38,1.35,2.08,1.42,2.15)
y1995=c(2.2,2.11,2.61,2.37,1.52,1.84,1.13,1.61,1.67,2.19,1.74,1.64,2.13,1.68,1.49,2.03,1
.76,
2.07,2.31,1.49,1.5,0.92,1.79,1.35,2.94,1.87,2.28,1.61,2.24,1.11,1.27,2.29,1.46,1.9,1.13,
1.35,1.74,1.91,1.57,1.00,2.28,2.06,2.24,1.76,1.73,1.71,1.29,1.33,2.16,1.45,2.41)
y1996=c(2.23,1.97,2.36,2.21,1.43,1.71,1.1,1.51,1.59,2.12,1.76,1.84,1.99,1.53,1.49,1.73,1
.89,
1.98,2.37,1.32,1.32,0.83,1.67,1.3,2.65,1.88,2.12,1.8,2.18,1.22,1.31,2.25,1.34,1.89,1.26,
1.35,1.96,1.73,1.52,0.97,2.34,2.24,2.12,2.02,1.64,1.38,1.23,1.44,1.97,1.44,1.94)
y1997=c(2.23,1.76,2.19,2.35,1.32,1.62,1.19,1.79,1.8,2.08,1.69,1.65,2.01,1.41,1.36,1.67,1
.82,
1.97,2.44,1.45,1.31,0.87,1.58,1.22,2.73,1.89,2.82,1.77,2.13,1.12,1.23,2.21,1.37,1.81,1.47,
1.39,2.02,1.62,1.59,1.06,2.18,1.86,2.02,1.77,1.79,1.48,1.4,1.32,2.08,1.33,1.81)
y1998=c(1.94,1.55,2.17,2.2,1.2,1.6,1.12,1.4,1.63,2.05,1.63,1.5,1.97,1.38,1.42,1.55,1.82,1
.91,
2.3,1.42,1.25,0.78,1.46,1.31,2.77,1.81,2.47,1.79,2.19,1.11,1.15,1.91,1.23,1.87,1.25,1.36,1
.8, 1.61,1.48,0.93,2.34,2.04,1.94,1.74,1.65,1.58,1.29,1.27,1.9,1.26,1.92)
y1999=c(2.03,1.74,2.18,2.07,1.19,1.54,1.01,1.18,1.18,2.06,1.52,1.21,1.99,1.42,1.46,1.68,
1.95,
1.75,2.28,1.28,1.2,0.8,1.44,1.22,2.66,1.64,2.24,1.64,2.01,1.18,1.11,2.05,1.26,1.71,1.64,1.
36, 1.74,1.19,1.52,1.06,2.41,1.82,2.01,1.67,1.63,1.38,1.19,1.21,2.08,1.31,2.42)
y2000=c(1.76,2.3,2.11,2.24,1.22,1.63,1.11,1.49,1.37,1.99,1.47,1.55,2.04,1.38,1.25,1.51,1
.64,
1.75,2.3,1.19,1.17,0.82,1.41,1.19,2.67,1.72,2.4,1.53,1.83,1.05,1.08,1.9,1.13,1.74,1.19,1.2
9, 1.5,1.33,1.49,0.96,2.34,2.05,1.99,1.72,1.65,1.12,1.24,1.18,2.14,1.4,1.88)
y2001=c(1.75,1.89,2.12,2.08,1.27,1.73,1.03,1.58,1.81,1.77,1.53,1.61,1.84,1.37,1.27,1.49,
1.75,
```

1.83,2.2,1.33,1.27,0.9,1.34,1.06,2.18,1.62,2.3,1.36,1.72,1.15,1.08,2.00,1.2,1.67,1.45,1.29,
1.57,1.42,1.49,1.01,2.27,2.00,1.85,1.73,1.24,1.17,1.27,1.21,1.91,1.33,2.16)
y2002=c(1.8,1.82,2.18,2.13,1.27,1.71,1.04,1.4,1.33,1.76,1.41,1.34,1.86,1.35,1.09,1.31,1.
78,
1.95,2.09,1.47,1.23,0.86,1.28,1.2,2.43,1.77,2.59,1.64,2.12,1.01,1.1,1.97,1.15,1.7,1.32,1.3
1, 1.62,1.26,1.54,1.03,2.23,2.12,1.73,1.73,1.34,0.98,1.18,1.2,2.19,1.37,1.95)
y2003=c(1.71,1.98,2.07,2.09,1.31,1.48,0.95,1.57,1.87,1.71,1.47,1.43,2.05,1.36,1.15,1.42,
1.64,
1.99,2.13,1.39,1.19,0.86,1.27,1.18,2.33,1.81,2.41,1.54,1.91,0.98,1.05,1.92,1.11,1.66,1.41,
1.17, 1.47,1.46,1.48,1.24,2.01,2.38,1.73,1.71,1.29,0.83,1.23,1.09,1.96,1.42,1.79)
y2004=c(1.95,2.02,2.01,2.22,1.25,1.45,0.93,1.44,1.15,1.65,1.44,1.46,1.77,1.24,1.3,1.23,1
.57,
2.04,2.08,1.3,1.16,0.87,1.12,1.00,2.28,1.64,2.04,1.32,1.95,1.26,0.99,2.18,1.08,1.64,1.32,1
.15, 1.67,1.28,1.38,0.98,2.11,2.24,1.89,1.6,1.2,1.25,1.17,1.02,2.02,1.31,1.77)
y2005=c(1.92,1.45,1.97,2.05,1.32,1.26,0.88,1.4,1.29,1.75,1.52,1.39,1.85,1.27,1.31,1.45,1
.44,
2.08,2.14,1.13,1.09,0.8,1.09,0.98,2.32,1.83,2.26,1.43,2.06,1.24,1.01,2.04,1.03,1.53,1.62,
1.2,1.71,1.38,1.5,1.05,2.21,2.22,1.79,1.5,1.12,0.95,1.18,1.17,1.82,1.36,1.88)
y2006=c(1.99,1.49,2.07,2.01,1.29,1.1,0.98,1.57,1.02,1.65,1.49,1.58,1.76,1.17,1.27,1.4,1.
55,
1.91,2.17,1.25,1.16,0.78,1.04,0.87,2.2,1.59,2.34,1.39,1.97,0.93,1.02,1.88,1.03,1.53,1.41,
1.11,1.57,1.35,1.41,0.98,2.08,2.08,1.82,1.48,1.11,1.11,1.19,1.12,1.96,1.22,2.07)
y2007=c(1.81,1.59,1.7,1.96,1.22,1.14,0.92,1.23,1.22,1.56,1.46,1.33,1.6,1.16,1.23,1.43,1.
38,
1.8,2.19,1.22,1.09,0.79,1.04,0.89,2.04,1.43,2.45,1.32,1.68,0.96,0.95,1.54,0.97,1.62,1.42,
1.13,1.61,1.31,1.37,0.8,2.11,1.62,1.7,1.42,1.11,0.86,1.25,1.0,2.1,1.27,1.6)
y2008=c(1.63,1.27,1.52,1.81,1.05,1.15,0.95,1.35,0.94,1.5,1.37,1.04,1.52,0.98,1.11,1.34,1
.29,
1.74,2.03,1.06,1.07,0.67,0.96,0.78,1.79,1.41,2.12,1.09,1.56,1.06,0.8,1.39,0.92,1.4,1.33,1.
1, 1.55,1.24,1.36,0.79,1.86,1.35,1.5,1.48,1.06,1.0,1.0,0.94,1.82,1.05,1.68)
y2009=c(1.38,1.3,1.31,1.8,0.95,1.01,0.71,1.28,0.8,1.3,1.18,1.09,1.46,0.86,0.9,1.19,1.31,1
.67,
1.84,1.1,0.99,0.62,0.9,0.74,1.73,1.27,2.01,1.15,1.19,0.85,0.8,1.39,0.87,1.28,1.72,0.92,1.5
7, 1.11,1.22,1.01,1.82,1.48,1.4,1.35,0.93,0.97,0.94,0.87,1.82,0.96,1.4)
y2010=c(1.34,1.17,1.27,1.7,0.84,1.96,1.02,1.13,0.67,1.25,1.12,1.13,1.32,0.88,1.0,1.24,1.
44,
1.58,1.59,1.11,0.88,0.64,0.97,0.73,1.61,1.16,1.69,0.98,1.16,0.98,0.76,1.38,0.92,1.29,1.27,
0.97,1.4,0.94,1.32,0.81,1.65,1.58,1.47,1.29,0.95,0.98,0.9,0.8,1.64,0.96,1.66)
y2011=c(1.38,1.57,1.39,1.67,0.88,0.96,0.71,1.1,0.76,1.25,1.13,0.99,1.05,0.89,0.98,1.15,1
.29,
1.5,1.46,0.95,0.86,0.68,0.94,0.65,1.62,1.14,1.79,0.95,1.02,0.71,0.86,1.36,0.92,1.19,1.62,
0.91,1.47,0.99,1.3,0.84,1.7,1.23,1.32,1.29,0.93,0.77,0.94,0.8,1.78,0.99,1.46)
y2012=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,

```
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
MV<-
data.frame(State,y1994,y1995,y1996,y1997,y1998,y1999,y2000,y2001,y2002,y2003,y20
04, y2005,y2006,y2007,y2008,y2009,y2010,y2011,y2012)
MV
#Tied ranks are resolved at random
x1<-rank(y1994,ties.method="random"); x2<-rank(y1995,ties.method="random")
x3<-rank(y1996,ties.method="random");x4<-rank(y1997,ties.method="random")
x5<-rank(y1998,ties.method="random");x6<-rank(y1999,ties.method="random")
x7<-rank(y2000,ties.method="random");x8<-rank(y2001,ties.method="random")
x9<-rank(y2002,ties.method="random");x10<-rank(y2003,ties.method="random")
x11<-rank(y2004,ties.method="random");x12<-rank(y2005,ties.method="random")
x13<-rank(y2006,ties.method="random");x14<-rank(y2007,ties.method="random")
x15<-rank(y2008,ties.method="random");x16<-rank(y2009,ties.method="random")
x17<-rank(y2010,ties.method="random");x18<-rank(y2011,ties.method="random")
x19<-rank(y2012,ties.method="random")
ra<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11, x12,x13,x14,x15,x16,x17,x18,x19)
#ra
ram<-as.matrix(ra)
ram
RkSum<-rep(0,k)
for (i in 1:k){RkSum[i]<-sum(ram[i,])}
#RkSum
StRk<-data.frame(State,RkSum)
#StRk
StRkOrd<-StRk[order(StRk$RkSum),]
#StRkOrd
NorSc<-rep(0,51)
#The following are approximations to exact values given by Harter
#for (i in 1:k){NorSc[i]<-qnorm(i/(k+1))}
#NorSc<-round(NorSc,4)
#The following expected values of normal order stats are taken from
#"Expected Values of Normal Order Statistics," by H. Leon Harter (1961)
#If two states are tied and the rank values are r1 and r2 (r1< #If two states are tied and the
rank values are r1 and r2 (r1<r2).then r1
#is assigned to the state abbreviation;r2 is assigned to the other state.
#similarly if there are three or more states are tied in their MVTFRs
NorSc<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,-1.23003,-1.13162, -1.04312,-
0.96213,-0.88701,-0.81661,-0.75004,-0.68666,-0.62592,-0.56742, -0.51080,-0.45578,-
0.40211,-0.34957,-0.29799,-0.24719,-0.19702,-0.14735, -0.09803,-
0.04896,0,0.04896,0.09803,0.14735,0.19702,0.24719,0.29799,
0.34957,0.40211,0.45578,0.51080,0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,1.34207,1.47409,1.63829,1.86371, 2.25678)
sum(NorSc)
```

```
#NorSc
df94<-data.frame(State,x1)
ns94<-rep(0,51) for (i in 1:51){ns94[i]<-NorSc[x1[i]]}
df94<-data.frame(State,x1,ns94)
#df94 #
df95<-data.frame(State,x2)
ns95<-rep(0,51) for (i in 1:51){ns95[i]<-NorSc[x2[i]]}
df95<-data.frame(State,x2,ns95)
#df95 #
df96<-data.frame(State,x3)
ns96<-rep(0,51) for (i in 1:51){ns96[i]<-NorSc[x3[i]]}
df96<-data.frame(State,x3,ns96)
#df96 #
df97<-data.frame(State,x4)
ns97<-rep(0,51) for (i in 1:51){ns97[i]<-NorSc[x4[i]]}
df97<-data.frame(State,x4,ns97)
#df97 #
df98<-data.frame(State,x5)
ns98<-rep(0,51) for (i in 1:51){ns98[i]<-NorSc[x5[i]]}
df98<-data.frame(State,x5,ns98)
#df98 #
df99<-data.frame(State,x6)
ns99<-rep(0,51) for (i in 1:51){ns99[i]<-NorSc[x6[i]]}
df99<-data.frame(State,x6,ns99)
#df99 #
df00<-data.frame(State,x7)
ns00<-rep(0,51) for (i in 1:51){ns00[i]<-NorSc[x7[i]]}
df00<-data.frame(State,x7,ns00)
#df00 #
df01<-data.frame(State,x8)
ns01<-rep(0,51) for (i in 1:51){ns01[i]<-NorSc[x8[i]]}
df01<-data.frame(State,x8,ns01)
#df01 #
df02<-data.frame(State,x9)
ns02<-rep(0,51) for (i in 1:51){ns02[i]<-NorSc[x9[i]]}
df02<-data.frame(State,x9,ns02)
#df02 #
df03<-data.frame(State,x10)
ns03<-rep(0,51) for (i in 1:51){ns03[i]<-NorSc[x10[i]]}
df03<-data.frame(State,x10,ns03)
#df03 #
df04<-data.frame(State,x11)
ns04<-rep(0,51) for (i in 1:51){ns04[i]<-NorSc[x11[i]]}
df04<-data.frame(State,x11,ns04)
#df04 #
```

```
df05<-data.frame(State,x12)
ns05<-rep(0,51) for (i in 1:51){ns05[i]<-NorSc[x12[i]]}
df05<-data.frame(State,x12,ns05)
#df05 #
df06<-data.frame(State,x13)
ns06<-rep(0,51) for (i in 1:51){ns06[i]<-NorSc[x13[i]]}
df06<-data.frame(State,x13,ns06)
#df06 #
df07<-data.frame(State,x14)
ns07<-rep(0,51) for (i in 1:51){ns07[i]<-NorSc[x14[i]]}
df07<-data.frame(State,x14,ns07)
#df07 #
df08<-data.frame(State,x15)
ns08<-rep(0,51) for (i in 1:51){ns08[i]<-NorSc[x15[i]]}
df08<-data.frame(State,x15,ns08) #
df08 #
df09<-data.frame(State,x16)
ns09<-rep(0,51) for (i in 1:51){ns09[i]<-NorSc[x16[i]]}
df09<-data.frame(State,x16,ns09)
 #df09 #
df10<-data.frame(State,x17)
ns10<-rep(0,51) for (i in 1:51){ns10[i]<-NorSc[x17[i]]}
df10<-data.frame(State,x17,ns10)
#df10 #
df11<-data.frame(State,x18)
ns11<-rep(0,51) for (i in 1:51){ns11[i]<-NorSc[x18[i]]}
df11<-data.frame(State,x18,ns11)
#df11 #
df12<-data.frame(State,x19)
ns12<-rep(0,51) for (i in 1:51){ns12[i]<-NorSc[x19[i]]}
df12<-data.frame(State,x19,ns12)
#df12 #
Ns<-data.frame(ns94,ns95,ns96,ns97,ns98,ns99,ns00,ns01,
ns02,ns03,ns04,ns05,ns06,ns07,ns08,ns09,ns10,ns11,ns12)
scm<-as.matrix(Ns)
#scm
NsSum<-rep(0,k) for (i in 1:k){NsSum[i]<-sum(scm[i,])}
#NsSum
StRkNs<-data.frame(State,NsSum)
#StRkNs #
StNs<-data.frame(State,NsSum)
StNs
StNsOrd<-StNs[order(StNs$NsSum),]
StNsOrd
StRks<-data.frame(State,RkSum)
```

```
StRks
StRksOrd<-StRks[order(StRks$RkSum),]
StRksOrd
#
Summary<-data.frame(StRkOrd,StNsOrd)
Summary
#check on distribution of MVTFRs for one year, 1994
hist(y1994,col='red',freq=FALSE,
main="Histogram of 1994 MVTFRs\n Normal Density Overlay")
low<-min(y1994)-0.1;up<-max(y1994)+0.1
curve(dnorm(x,mean(y1994),sd(y1994),),from=low,to=up,add=TRUE)
```

Appendix B

```
#Regression of exp51 and seq(2:51)
#Differences in the Expected value of normal order stats, n=51
exp51<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,-1.23003,-1.13162, -1.04312,-
0.96213,-0.88701,-0.81661,-0.75004,-0.68666,-0.62592,-0.56742, -0.51080,-0.45578,-
0.40211,-0.34957,-0.29799,-0.24719,-0.19702,-0.14735, -0.09803,-
0.04896,0,0.04896,0.09803,0.14735,0.19702,0.24719,0.29799,
0.34957,0.40211,0.45578,0.51080,0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,1.34207,1.47409,1.63829,1.86371, 2.25678)
sum(exp51)
diff<-rep(0,50)
for (i in 1:50){ diff[i]<-exp51[i+1]-exp51[i] }
diff
x<-seq(1:50)
#plot(x,diff,main="Difference in expected values of normal order stats\n k = 51")
#1st point is E[X(2)]-E[X(1)], 2nd point is E[X(3)]-E[X(2)], etc.
xx<-seq(1:51) par(mfrow=c(1,2)) model<-lm(exp51~xx)
plot(xx,exp51,xlab="Ranks",ylab="Normal Scores",main="Normal Scores vs.\n Rank
Values, k = 51",cex.main=1,pch=16)
abline(model,col="red",lwd=2)
legend("topleft","Reg Line\nR^2 = 0.967",col="red",lty=1,cex=0.8,bg="yellow")
plot(x,diff,xlab="Ranks",ylab="Difference",cex.main=1,
main="Diff in Expected Values of\n Normal Order Stats, k = 51",pch=16)
abline(v=25.5,col="red",lty=2)
```

Appendix C

```
#State Homicide Rates 2005,2014-2020
#Rank sums for the HRs
#Applied Mathematics,2022,13,585-601
#https://www.scirp.org/journal/am
#k is the number of populations (e.g., states); n is the number of blocks (e.g., years)
k=50;n=8
State=c("AK","AL","AR","AZ","CA","CO","CT","DE","FL","GA","HI","IA","ID","IL",
"IN","”KS”,
```

```
"KY","LA","MA","MD","ME","MI","MN","MO","MS","MT","NC","ND","NE","NH","
NJ","NM","NV",
"NY","OH","OK","OR","PA","RI","SC","SD","TN","TX","UT","VA","VT","WA","WI"
,"WV","WY")
y2005=c(1.93,2.47,2.30,2.41,2.17,1.71,1.59,2.13,2.02,2.19,1.29,1.21,1.59,2.15,2.03,1.72,
1.96,2.77,1.51,2.55,1.24,2.17,1.49,2.21,2.41,1.63,2.25,0.00,1.44,0.00,1.92,2.29,2.27,
1.86,1.99,2.06,1.53,2.09,1.57,2.29,1.53,2.33,2.11,1.42,2.10,0.00,1.67,1.79,1.96,0.00)
y2014=c(1.86,2.31,2.26,1.90,1.84,1.61,1.53,2.13,2.07,2.13,1.37,1.44,1.42,2.07,2.01,1.67,
1.86,2.67,1.32,2.14,1.32,2.09,1.29,2.24,2.65,1.53,1.99,0.00,1.63,0.00,1.81,2.15,2.09,
1.63,1.93,2.13,1.42,1.93,1.44,2.25,1.57,2.11,1.93,1.32,1.76,0.00,1.57,1.55,2.03,1.81)
y2015=c(2.30,2.53,2.23,1.98,1.90,1.69,1.67,2.24,2.09,2.21,1.37,1.44,1.32,2.17,2.05,1.86,
2.02,2.74,1.35,2.54,1.24,2.10,1.51,2.47,2.64,1.74,2.06,1.57,1.74,0.00,1.83,2.30,2.14,1.63,
2.05,2.35,1.63,1.99,1.51,2.46,1.78,2.20,1.99,1.32,1.83,0.00,1.63,1.83,1.83,0.00)
y2016=c(2.21,2.68,2.38,2.09,1.95,1.79,1.49,2.18,2.15,2.29,1.51,1.51,1.29,2.43,2.25,1.95,
2.20,2.90,1.35,2.52,0.00,2.14,1.42,2.50,2.71,1.79,2.23,0.00,1.61,0.00,1.84,2.45,2.23,1.67,
2.11,2.36,1.61,2.05,1.40,2.41,1.86,2.39,2.05,1.44,1.98,0.00,1.53,1.87,2.09,0.00)
y2017=c(2.57,2.78,2.49,2.13,1.92,1.84,1.59,2.17,2.10,2.29,1.44,1.63,1.55,2.41,2.20,2.11,
2.21,2.91,1.47,2.53,0.00,2.09,1.37,2.64,2.76,1.79,2.17,0.00,1.49,0.00,1.76,2.35,2.25,1.55,
2.24,2.35,1.57,2.13,0.00,2.44,1.78,2.39,2.02,1.47,1.96,0.00,1.67,1.69,2.11,0.00)
y2018=c(2.24,2.72,2.42,2.06,1.87,1.86,1.51,2.15,2.13,2.26,1.57,1.49,1.40,2.30,2.23,2.03,
2.06,2.82,1.40,2.44,0.00,2.11,1.40,2.65,2.82,1.78,2.10,1.44,1.29,1.27,1.69,2.59,2.26,1.59,
2.15,2.18,1.44,2.10,0.00,2.53,1.72,2.43,1.96,1.37,1.92,0.00,1.69,1.72,2.02,1.76)
y2019=c(2.59,2.77,2.45,2.03,1.83,1.79,1.57,2.06,2.14,2.31,1.44,1.49,1.24,2.31,2.20,1.89,
2.03,2.93,1.40,2.51,1.27,2.11,1.51,2.59,2.99,1.69,2.18,1.57,1.57,1.51,1.63,2.68,1.98,1.59,
2.13,2.39,1.55,2.06,1.44,2.61,1.67,2.43,2.03,1.47,1.95,0.00,1.59,1.78,2.01,1.81)
y2020=c(2.21,2.89,2.79,2.24,2.06,2.02,1.84,2.50,2.27,2.56,1.61,1.67,1.44,2.63,1.48,2.18,
2.46,3.31,1.49,2.65,1.21,2.38,1.67,2.87,3.35,2.13,2.36,1.81,1.76,0.00,1.79,2.59,2.21,1.86,
2.42,2.41,1.71,2.35,1.55,2.76,2.11,2.66,2.25,1.53,2.10,0.00,1.78,2.06,2.18,1.89)
HR<-data.frame(State,y2005,y2014,y2015,y2016,y2017,y2018,y2019,y2020)
HR
#Tied ranks are resolved at random
x1<-rank(y2005,ties.method="random");x2<-rank(y2014,ties.method="random")
x3<-rank(y2015,ties.method="random");x4<-rank(y2016,ties.method="random")
x5<-rank(y2017,ties.method="random");x6<-rank(y2018,ties.method="random")
x7<-rank(y2019,ties.method="random");x8<-rank(y2020,ties.method="random")
ra<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8)
ram<-as.matrix(ra)
#ram
RkSum<-rep(0,k) for (i in 1:k){RkSum[i]<-sum(ram[i,])}
#RkSum
StRk<-data.frame(State,RkSum)
#StRk
StRkOrd<-StRk[order(StRk$RkSum),]
#StRkOrd
#NorSc taken from Harter, Biometrika (1961)
```

```
NorSc<-c(-2.24907,-1.85487,-1.62863,-1.46374,-1.33109, -1.21846,-1.11948,-1.03042,-
0.94887,-0.87321,-0.80225, -0.73513,-0.67117,-0.60986,-0.55077,-0.49354,-0.43789, -
0.38357,-0.33036,-0.27807,-0.22653,-0.17559,-0.12511, -0.07494,-
0.02496,0.02496,0.07494,0.12511,0.17559,
0.22653,0.27807,0.33036,0.38357,0.43789,0.49354,
0.55077,0.60986,0.67117,0.73513,0.80225,0.87321,
0.94887,1.03042,1.11948,1.21846,1.33109,1.46374, 1.62863,1.85487,2.24907)
#Approx to exact NorSc given above
#NorSc<-rep(0,k) #for (i in 1:k){NorSc[i]<-qnorm(i/(k+1))}
#NorSc<-round(NorSc,4)
#NorSc df05<-data.frame(State,x1)
df05<-df05[order(df05$x1,decreasing=FALSE),]
df05<-cbind(df05,NorSc)
#df05
df05<-df05[order(df05$State,decreasing=FALSE),]
#df05
NS05<-df05$NorSc
#NS05
df14<-data.frame(State,x2)
df14<-df14[order(df14$x2,decreasing=FALSE),]
df14<-cbind(df14,NorSc)
#df14
df14<-df14[order(df14$State,decreasing=FALSE),]
#df14
NS14<-df14$NorSc
#NS14
df15<-data.frame(State,x3)
df15<-df15[order(df15$x3,decreasing=FALSE),]
df15<-cbind(df15,NorSc)
#df15
df15<-df15[order(df15$State,decreasing=FALSE),]
#df15
NS15<-df15$NorSc
#NS15
df16<-data.frame(State,x4)
df16<-df16[order(df16$x4,decreasing=FALSE),]
df16<-cbind(df16,NorSc)
#df16
df16<-df16[order(df16$State,decreasing=FALSE),]
#df16
NS16<-df16$NorSc
#NS16
df17<-data.frame(State,x5)
df17<-df17[order(df17$x5,decreasing=FALSE),]
df17<-cbind(df17,NorSc)
```

```
#df17
df17<-df17[order(df17$State,decreasing=FALSE),]
#df17
NS17<-df17$NorSc
#NS17
df18<-data.frame(State,x6)
df18<-df18[order(df18$x6,decreasing=FALSE),]
df18<-cbind(df18,NorSc)
#df18
df18<-df18[order(df18$State,decreasing=FALSE),]
#df18
NS18<-df18$NorSc
#NS18
df19<-data.frame(State,x7)
df19<-df19[order(df19$x7,decreasing=FALSE),]
df19<-cbind(df19,NorSc)
#df19
df19<-df19[order(df19$State,decreasing=FALSE),]
#df19
NS19<-df19$NorSc
#NS19
df20<-data.frame(State,x8)
df20<-df20[order(df20$x8,decreasing=FALSE),]
df20<-cbind(df20,NorSc)
#df20
df20<-df20[order(df20$State,decreasing=FALSE),]
#df20
NS20<-df20$NorSc
#NS20
Ns<-data.frame(NS05,NS14,NS15,NS16,NS17,NS18,NS19,NS20)
scm<-as.matrix(Ns)
#scm
NsSum<-rep(0,k) for (i in 1:k){NsSum[i]<-sum(scm[i,])}
#NsSum
StRkNs<-data.frame(State,RkSum,NsSum)
#StRkNs #
StNs<-data.frame(State,NsSum)
StNsOrd<-StNs[order(StNs$NsSum),]
#StNsOrd
Summary<-data.frame(StRkOrd,StNsOrd)
Summary
```

Appendix D (b1 and b3)

```
#Nonparametric Block Design Selection Procedure Based on Ranks
#k=no. of population;n=no. of blocks;w=no. of simulations
```

```
#P=min Prob of Correct Selection
k<-50;n<-8;w<-50000;P=0.90
#calculate quantiles of max(T)-Ti for iid populations
rnk<-seq(1:k);T<-rep(0,k);U<-rep(0,k);Q<-rep(0,w);b1<-0;
V<-rep(0,k);C<-rep(0,k);W<-rep(0,k)
for (h in 1:w){
M<-matrix(0,nrow=n,ncol=k)
for (j in 1:n){M[j,]<-sample(rnk,size=k,replace=FALSE)}
M
for (i in 1:k){T[i]<-sum(M[,i])}
T
for (j in 1:k){U[j]<-max(T)-T[j]}
U
Q[h]<-U[k] }
message("k = ",k,", n = ",n,", w = ",w,", P = ",P)
quan<-c(0.50,0.75,0.90,0.95,0.99)
Qile<-quantile(Q,quan)
Qile
Qile<-unname(Qile)
Qile
#table(Q)
if (P==0.50){b1<-Qile[1]}
if (P==0.75){b1<-Qile[2]}
if (P==0.90){b1<-Qile[3]}
if (P==0.95){b1<-Qile[4]}
if (P==0.99){b1<-Qile[5]}
c(P,b1)
#Select population i iff Ti>=max(T)-b1
a<-rep(1,k);b<-rep(5,k)
V<-rep(0,k)
for (h in 1:w){
C<-rep(0,k);x<-rep(0,k);rk<-rep(0,k);T<-rep(0,k)
M<-matrix(0,nrow=n,ncol=k);U<-rep(0,k)
for (j in 1:n){
for (i in 1:k){x[i]<-runif(1,a[i],b[i])}
rk<-rank(x)
M[j,]<-rk }
M
for (i in 1:k){T[i]<-sum(M[,i])}
T
for (i in 1:k){U[i]<-max(T)-T[i]
if (U[i]<=b1){C[i]<-1}
else {C[i]<-0} }
V<-V+C }
message("The probabilities of population selections are")
```

```
V/w
ESS<-sum(V)/w
message("The expected subset size is ",round(ESS,3))
```

Appendix E (b2 and b4)

```
#Nonparametric Block Design Selection Procedure Based on Ranks
#k=no. of population;n=no. of blocks;w=no. of simulations
#P=min Prob of Correct Selection
k<-50;n<-8;w<-50000;P=0.90
#calculate quantiles of Tk for iid populations
rnk<-seq(1:k);T<-rep(0,k);U<-rep(0,k);Q<-rep(0,w);b1<-0;
V<-rep(0,k);C<-rep(0,k);W<-rep(0,k)
for (h in 1:w){
M<-matrix(0,nrow=n,ncol=k)
for (j in 1:n){M[j,]<-sample(rnk,size=k,replace=FALSE)}
for (i in 1:k){T[i]<-sum(M[,i])}
Q[h]<-T[k] }
message("k = ",k,", n = ",n,", w = ",w,", P = ",P)
quan<-c(0.01,0.05,0.10,0.25,0.50)
Qile<-quantile(Q,quan)
Qile
Qile<-unname(Qile)
Qile
#table(Q)
if (P==0.50){b2<-Qile[5]}
if (P==0.75){b2<-Qile[4]}
if (P==0.90){b2<-Qile[3]}
if (P==0.95){b2<-Qile[2]}
if (P==0.99){b2<-Qile[1]}
b4<-(n*(k+1))-b2
message("P = ",P,", b2 = ",b2,", b4 = ",b4)
#Select population i iff Ti>b2
a<-rep(1,k);b<-rep(5,k)
V<-rep(0,k)
for (h in 1:w){
C<-rep(0,k);x<-rep(0,k);rk<-rep(0,k);T<-rep(0,k)
M<-matrix(0,nrow=n,ncol=k);U<-rep(0,k)
for (j in 1:n){
for (i in 1:k){x[i]<-runif(1,a[i],b[i])}
rk<-rank(x)
M[j,]<-rk }
for (i in 1:k){T[i]<-sum(M[,i])}
for (i in 1:k){U[i]<-T[i]
if (U[i]>b2){C[i]<-1}
else {C[i]<-0} }
```

```
V<-V+C }
message("The probabilities of population selections are")
V/w ESS<-sum(V)/w
message("The expected subset size is ",round(ESS,3))
```

Appendix F (d1 and d3 Values)

```
#Nonparametric Block Design Selection Procedure Based on Normal Scores
#k=no. of population;n=no. of blocks;w=no. of simulations
#P=min Prob of Correct Selection
k<-50;n<-8;w<-50000;P=0.90
#calculate quantiles of max(T)-Ti for iid populations
rnk<-seq(1:k);T<-rep(0,k);U<-rep(0,k);Q<-rep(0,w);b1<-0;
V<-rep(0,k);C<-rep(0,k);W<-rep(0,k);nsc<-rep(0,k)
#For approx expected values of normal order statistics use:
#for (i in 1:k){nsc[i]<-qnorm(i/(k+1))}
#for exact expected values of normal order stats read in:
#nsc taken from Harter, Biometrika (1961)
nsc<-c(-2.24907,-1.85487,-1.62863,-1.46374,-1.33109, -1.21846,-1.11948,-1.03042,-
0.94887,-0.87321,-0.80225, -0.73513,-0.67117,-0.60986,-0.55077,-0.49354,-0.43789, -
0.38357,-0.33036,-0.27807,-0.22653,-0.17559,-0.12511, -0.07494,-
0.02496,0.02496,0.07494,0.12511,0.17559,
0.22653,0.27807,0.33036,0.38357,0.43789,0.49354,
0.55077,0.60986,0.67117,0.73513,0.80225,0.87321,
0.94887,1.03042,1.11948,1.21846,1.33109,1.46374, 1.62863,1.85487,2.24907)
for (h in 1:w){
M<-matrix(0,nrow=n,ncol=k)
for (j in 1:n){M[j,]<-sample(nsc,size=k,replace=FALSE)}
M
for (i in 1:k){T[i]<-sum(M[,i])}
T
for (j in 1:k){U[j]<-max(T)-T[j]}
U
Q[h]<-U[k] }
message("k = ",k,", n = ",n,", w = ",w,", P = ",P)
quan<-c(0.50,0.75,0.90,0.95,0.99)
Qile<-quantile(Q,quan)
Qile
Qile<-unname(Qile)
Qile
#table(Q)
if (P==0.50){b1<-Qile[1]}
if (P==0.75){b1<-Qile[2]}
if (P==0.90){b1<-Qile[3]}
if (P==0.95){b1<-Qile[4]}
if (P==0.99){b1<-Qile[5]}
```

```
c(P,b1)
#Select population i iff Ti>=max(T)-b1
a<-rep(1,k);b<-rep(2,k)
V<-rep(0,k)
for (h in 1:w){
C<-rep(0,k);W<-rep(0,k);x<-rep(0,k);rk<-rep(0,k);S<-rep(0,k)
ns<-rep(0,k)
M<-matrix(0,nrow=n,ncol=k)
for (j in 1:n){
for (i in 1:k){x[i]<-runif(1,a[i],b[i])}
rk<-rank(x) for (i in 1:k){ns[i]<-nsc[rk[i]]}
M[j,]<-ns }
M
for (i in 1:k){S[i]<-sum(M[,i])}
S
for (i in 1:k){W[i]<-max(S)-S[i]
if (W[i]<=b1){C[i]<-1}
else {C[i]<-0} }
V<-V+C }
message("The probabilities of population selections are")
V/w
ESS<-sum(V)/w
message("The expected subset size is ",round(ESS,3))
```

Appendix G (d2 and d4 Values)

```
#Nonparametric Block Design Selection Procedure Based on Normal Scores
#k=no. of population;n=no. of blocks;w=no. of simulations
#P=min Prob of Correct Selection
k<-50;n<-8;w<-50000;P=0.90
#calculate quantiles of max(T)-Ti for iid populations
rnk<-seq(1:k);T<-rep(0,k);U<-rep(0,k);Q<-rep(0,w);b1<-0;
V<-rep(0,k);C<-rep(0,k);W<-rep(0,k);nsc<-rep(0,k)
#For approx expected values of normal order statistics use:
#for (i in 1:k){nsc[i]<-qnorm(i/(k+1))}
#for exact expected values of normal order stats read in:
#nsc taken from Harter, Biometrika (1961)
nsc<-c(-2.24907,-1.85487,-1.62863,-1.46374,-1.33109, -1.21846,-1.11948,-1.03042,-
0.94887,-0.87321,-0.80225, -0.73513,-0.67117,-0.60986,-0.55077,-0.49354,-0.43789, -
0.38357,-0.33036,-0.27807,-0.22653,-0.17559,-0.12511, -0.07494,-
0.02496,0.02496,0.07494,0.12511,0.17559,
0.22653,0.27807,0.33036,0.38357,0.43789,0.49354,
0.55077,0.60986,0.67117,0.73513,0.80225,0.87321,
0.94887,1.03042,1.11948,1.21846,1.33109,1.46374, 1.62863,1.85487,2.24907)
for (h in 1:w){
M<-matrix(0,nrow=n,ncol=k)
```

```
for (j in 1:n){M[j,]<-sample(nsc,size=k,replace=FALSE)}
for (i in 1:k){T[i]<-sum(M[,i])}
Q[h]<-T[k] }
message("k = ",k,", n = ",n,", w = ",w,", P = ",P)
quan<-c(0.01,0.05,0.10,0.25,0.50)
Qile<-quantile(Q,quan)
Qile
Qile<-unname(Qile)
Qile
#table(Q) if (P==0.50){d2<-Qile[5]}
if (P==0.75){d2<-Qile[4]}
if (P==0.90){d2<-Qile[3]}
if (P==0.95){d2<-Qile[2]}
if (P==0.99){d2<-Qile[1]}
d4<--d2
c(P,d2)
message("P = ",P,", d2 = ",d2,", d4 = ",d4)
#Select population i iff Ti>d2
a<-rep(1,k);b<-rep(5,k)
V<-rep(0,k)
for (h in 1:w){
C<-rep(0,k);W<-rep(0,k);x<-rep(0,k);rk<-rep(0,k);S<-rep(0,k)
ns<-rep(0,k)
M<-matrix(0,nrow=n,ncol=k)
for (j in 1:n){
for (i in 1:k){x[i]<-runif(1,a[i],b[i])}
rk<-rank(x)
for (i in 1:k){ns[i]<-nsc[rk[i]]}
M[j,]<-ns }
for (i in 1:k){S[i]<-sum(M[,i])}
for (i in 1:k){W[i]<-S[i]
if (W[i]>d2){C[i]<-1}
else {C[i]<-0} }
V<-V+C }
message("The probabilities of population selections are")
V/w
ESS<-sum(V)/w
message("The expected subset size is ",round(ESS,3))
```

Appendix H

```
#k = 7, n = 2
#generates all the permutations of 1:7
k<-7;n=2
f<-factorial(k)
f2<-f^2
```

```
D<-matrix(0,nrow=f,ncol=k)
E<-matrix(0,nrow=f,ncol=k)
C<-matrix(0,nrow=f2,ncol=k)
F<-matrix(0,nrow=f2,ncol=k)
permutations <- function(n){
if(n==1){ return(matrix(1))
} else {
sp <- permutations(n-1)
p <- nrow(sp)
A <- matrix(nrow=n*p,ncol=n)
for(i in 1:n){
A[(i-1)*p+1:p,] <- cbind(i,sp+(sp>=i))
}
return(A)
}
}
D<-permutations(k)
dim(D)
head(D,5)
tail(D,5)
a<-c(rep(0,f));b<-c(rep(0,f))
for (i in 1:f){a[i]<-(i-1)*f+1
b[i]<-i*f }
#for (i in a[1]:b[1]){C[i,]<-D[1,]+D[i,]}
#for (i in a[2]:b[2]){C[i,]<-D[2,]+D[i-f,]}
#for (i in a[3]:b[3]){C[i,]<-D[3,]+D[i-2*f,]}
#for (i in a[4]:b[4]){C[i,]<-D[4,]+D[i-3*f,]}
#for (i in a[5]:b[5]){C[i,]<-D[5,]+D[i-4*f,]}
####
#for (i in a[f]:b[f]){C[i,]<-D[f,]+D[i-(f-1)*f,]} ####
for (i in 1:f){
for (j in 1:f){ C[(i-1)*f+j,]<-D[i,]+D[j,] } }
dim(C)
head(C,5)
tail(C,5)
S<-c(rep(0,f2))
for(i in 1:f2){S[i]<-max(C[i,])-C[i,1]}
head(S,5)
tail(S,5)
table(S)
df<-data.frame(table(S))
df
Pr<-df$Freq/f2
Pr<-round(Pr,5)
CDF<-cumsum(Pr)
```

```
df1<-data.frame(df,Pr,CDF) df1
############################
v3<-c(rep(0,f2))
for(i in 1:f2){v3[i]<-max(C[i,])-6}
message("max(Ti)-d for k = 7, n = 2, d = 6 and
P* = 0.74904")
head(v3,5)
tail(v3,5)
v2<-c(rep(0,f2))
for(i in 1:f2){v2[i]<-max(C[i,])-8}
message("max(Ti)-d for k = 7, n = 2, d = 8 and
P* = 0.906")
head(v2,5)
tail(v2,5)
K<-c(rep(0,f2))
for (i in 1:f2){
if (C[i,1]>=v3[i]){K[i]<-1}
if (C[i,2]>=v3[i]){K[i]<-K[i]+1}
if (C[i,3]>=v3[i]){K[i]<-K[i]+1}
if (C[i,4]>=v3[i]){K[i]<-K[i]+1}
if (C[i,5]>=v3[i]){K[i]<-K[i]+1}
if (C[i,6]>=v3[i]){K[i]<-K[i]+1}
if (C[i,7]>=v3[i]){K[i]<-K[i]+1}
}
length(K)
message("number of pops chosen with k = 7, n = 2,
d = 6, and P* = 0.74904 for each of the ",f2," rank sums")
head(K,5)
tail(K,5)
L<-c(rep(0,f2))
for (i in 1:f2){
if (C[i,1]>=v2[i]){L[i]<-1}
if (C[i,2]>=v2[i]){L[i]<-L[i]+1}
if (C[i,3]>=v2[i]){L[i]<-L[i]+1}
if (C[i,4]>=v2[i]){L[i]<-L[i]+1}
if (C[i,5]>=v2[i]){L[i]<-L[i]+1}
if (C[i,6]>=v2[i]){L[i]<-L[i]+1} if (C[i,7]>=v2[i]){L[i]<-L[i]+1}
}
length(L)
message("number of pops chosen with k = 7, n = 2,
d = 8, and P* = 0.906 for each of the ",f2," rank sums")
head(L,5)
tail(L,5)
#########################################################
#Now replace ranks by normal scores (k=7) given by ns
```

```
ns<-c(-1.35218,-0.75737,-0.35271,0,0.35271,0.75737,1.35218)
sum(ns)
E<-matrix(0,nrow=f,ncol=k)
for (i in 1:f){ for (j in 1:k){
for (m in 1:k){
if (D[i,j]==m){E[i,j]<-ns[m]}
}
}
}
dim(E)
for (i in 1:f){
for (j in 1:f){
F[(i-1)*f+j,]<-E[i,]+E[j,]
}
}
dim(F)
head(F,5)
tail(F,5)
U<-c(rep(0,f2))
for (i in 1:f2){U[i]<-max(F[i,])-F[i,1]}
U<-round(U,5)
length(U)
head(U,5)
tail(U,5)
table(U)
df3<-data.frame(table(U))
Pro<-df3$Freq/f2
Pro<-round(Pro,5)
CDF1<-cumsum(Pro)
df4<-data.frame(df3,Pro,CDF1)
df4[40:length(Pro),]
#################################
v5<-c(rep(0,f2))
for (i in 1:f2){v5[i]<-max(F[i,])-2.70436}
message("max(Si)-d for k = 7, n = 2, d = 2.70436
and P* = 0.75324")
head(v5,5)
tail(v5,5)
v4<-c(rep(0,f2))
for (i in 1:f2){v4[i]<-max(F[i,])-3.62429}
message("max(Si)-d for k = 7, n = 2, d = 3.62429
and P* = 0.90840")
head(v4,5)
tail(v4,5)
K1<-c(rep(0,f2))
```

```
for (i in 1:f2){
if (F[i,1]>=v5[i]){K1[i]<-1}
if (F[i,2]>=v5[i]){K1[i]<-K1[i]+1}
if (F[i,3]>=v5[i]){K1[i]<-K1[i]+1}
if (F[i,4]>=v5[i]){K1[i]<-K1[i]+1}
if (F[i,5]>=v5[i]){K1[i]<-K1[i]+1}
if (F[i,6]>=v5[i]){K1[i]<-K1[i]+1}
if (F[i,7]>=v5[i]){K1[i]<-K1[i]+1}
}
length(K1)
message("number of pops chosen with k = 7, n = 2,
d = 6, and P* = 0.74904 for each of the ",f2," norm scores")
head(K1,5)
tail(K1,5)
L1<-c(rep(0,f2))
for (i in 1:f2){
if (F[i,1]>=v4[i]){L1[i]<-1}
if (F[i,2]>=v4[i]){L1[i]<-L1[i]+1}
if (F[i,3]>=v4[i]){L1[i]<-L1[i]+1}
if (F[i,4]>=v4[i]){L1[i]<-L1[i]+1}
if (F[i,5]>=v4[i]){L1[i]<-L1[i]+1}
if (F[i,6]>=v4[i]){L1[i]<-L1[i]+1}
if (F[i,7]>=v4[i]){L1[i]<-L1[i]+1}
}
length(L1)
message("number of pops chosen with k = 7, n = 2,
d = 8, and P* = 0.906 for each of the ",f2," norm scores")
head(L1,5)
tail(L1,5)
###################################
#K-K1 < 0 means fewer pops chosen using rank scores at P* = 0.75
#L-L1 < 0 means fewer pops chosen using rank scores at P* = 0.90
Z<-K-K1
max(Z)
min(Z)
W<-L-L1
max(W)
min(W)
table(Z)
table(W)
#note that Z[142] = -2
#C[142,]=c(2,5,5,11,10,12,11)
#C[142,]=c(1,2,3,4,5,6,7)+c(1,3,2,7,5,6,4)
#rank sums = c(2,5,5,11,10,12,11)
#norm scores = c(-2.70436,-1.11008,-1.11008,1.35218,0.70542,
```

```
# 1.51474,1.35218)
#Rank procedure chooses popi if Ti>=max(T)-6=6 so 4 chosen
#Norm score procedure choose if Si>=max(S)-2.70436=-1.18962 so 6 chosen
#Rank procedure chooses 2 fewer than Norm score procedure
```

Chapter 6

Appendix I.

Ordering Probabilities for Three Independent Normal Variates

```
#General Normal Ordering for k=3 and n=2 #Compute the Pr(X1<X2<X3) where Xi's
are independent normal variates with
#means m[i] and standard deviations s[i].
#order 1 is (X1<X2<X3); order 2 is (X1<X3<X2); order 3 is (X2<X1<X3);
#order 4 is (X2<X3<X1); order 5 is (X3<X1<X2); order 6 is (X3<X2<X1).
#Input the values of the means and standard deviations in next line
m=c(0,0,0);s=c(1,1,1)
order<-NULL; error<-NULL
Prob1<-NULL;Prob2<-NULL;Prob3<-NULL;Prob4<-NULL;Prob5<-
NULL;Prob6<-NULL
integrand1<-function(x){
((pnorm((s[2]*x+m[2]-m[1])/s[1]))*(pnorm((-x*s[2]+m[3]-
m[2])/s[3]))*dnorm(x,mean=0,sd=1))
}
int1<-integrate(integrand1,lower=-Inf,upper=Inf)
order[1]<-int1$value
order[1]
error[1]<-int1$abs.error
integrand2<-function(x){
((pnorm((s[3]*x+m[3]-m[1])/s[1]))*(pnorm((-x*s[3]+m[2]-
m[3])/s[2]))*dnorm(x,mean=0,sd=1))
}
int2<-integrate(integrand2,lower=-Inf,upper=Inf)
order[2]<-int2$value
order[2]
error[2]<-int2$abs.error
integrand3<-function(x){
((pnorm((s[1]*x+m[1]-m[2])/s[2]))*(pnorm((-x*s[1]+m[3]-
m[1])/s[3]))*dnorm(x,mean=0,sd=1))
}
int3<-integrate(integrand3,lower=-Inf,upper=Inf)
order[3]<-int3$value
order[3]
error[3]<-int3$abs.error
integrand4<-function(x){
```

```
((pnorm((s[3]*x+m[3]-m[2])/s[2]))*(pnorm((-x*s[3]+m[1]-
m[3])/s[1]))*dnorm(x,mean=0,sd=1))
}
int4<-integrate(integrand4,lower=-Inf,upper=Inf)
order[4]<-int4$value
order[4]
error[4]<-int4$abs.error
integrand5<-function(x){
((pnorm((s[1]*x+m[1]-m[3])/s[3]))*(pnorm((-x*s[1]+m[2]-
m[1])/s[2]))*dnorm(x,mean=0,sd=1))
}
int5<-integrate(integrand5,lower=-Inf,upper=Inf)
order[5]<-int5$value
order[5]
error[5]<-int5$abs.error
integrand6<-function(x){
((pnorm((s[2]*x+m[2]-m[3])/s[3]))*(pnorm((-x*s[2]+m[1]-
m[2])/s[1]))*dnorm(x,mean=0,sd=1))
}
int6<-integrate(integrand6,lower=-Inf,upper=Inf)
order[6]<-int6$value
order[6]
error[6]<-int6$abs.error
df<-data.frame(order,error)
df
Total<-sum(order)
Total
#ProbX is the probability that order X & order[i] occur in 2 samples (blocks)
for (I in 1:6){
Prob1[i]<-order[1]*order[i]
}
for (I in 1:6){
Prob2[i]<-order[2]*order[i]
}
for (I in 1:6){
Prob3[i]<-order[3]*order[i]
}
for (I in 1:6){
Prob4[i]<-order[4]*order[i]
}
for (I in 1:6){
Prob5[i]<-order[5]*order[i]
}
for (I in 1:6){
Prob6[i]<-order[6]*order[i]
```

```
}
df1<-data.frame(Prob1,Prob2,Prob3,Prob4,Prob5,Prob6)
round(df1,5)
sum(Prob1,Prob2,Prob3,Prob4,Prob5,Prob6)
#PrX is the probability that Tmax-T3 = X for X=0,1,2,3,4
Pr0<-Prob1[1]+Prob1[2]+Prob1[3]+Prob1[4]+Prob1[6]+Prob2[1]+
Prob2[3]+Prob2[4]+Prob3[1]+Prob3[2]+Prob3[3]+
Prob3[4]+Prob3[5]+Prob4[1]+Prob4[2]+Prob4[3]+
Prob5[3]+Prob6[1]
Pr1<-Prob1[5]+Prob3[6]+Prob5[1]+Prob6[3]
Pr2<-Prob2[2]+Prob2[6]+Prob4[4]+Prob4[5]+Prob5[4]+Prob6[2]
Pr3<-Prob2[5]+Prob4[6]+Prob5[2]+Prob5[6]+Prob6[4]+Prob6[5]
Pr4<-Prob5[5]+Prob6[6]
df2<-data.frame(Pr0,Pr1,Pr2,Pr3,Pr4)
round(df2,5)
#cumdfX is the cdf of Tmax-T3 for X=0,1,2,3,4
cumdf0<-Pr0
cumdf1<-cumdf0+Pr1
cumdf2<-cumdf1+Pr2
cumdf3<-cumdf2+Pr3
cumdf4<-cumdf3+Pr4
df3<-data.frame(cumdf0,cumdf1,cumdf2,cumdf3,cumdf4)
round(df3,5)
```

Appendix J.
Pr(CS) for Three Independent Normal Populations with Slippage Configuration (Figure 1 Left Panel)

```
#norm_ord_computations using normal ordering.txt
#slippage parameter configurations
#mu=(0,0,delta)
require(splines)
delta<-seq(from=0,to=5.2,by=0.1)
delta
length(delta)
Pr0<-c(0.5,0.53756,0.57477,0.61131,0.64687,0.68115,0.71391,
0.74494,0.77405,0.80114,0.82612,0.84895,0.86964,0.88822,
0.90478,0.91942,0.93224,0.94338,0.95299,0.96121,0.96819,
0.97407,0.97899,0.98308,0.98645,0.98922,0.99146,0.99328,
0.99474,0.9959,0.99683,0.99756,0.99813,0.99857,0.99892,
0.99918,0.99939,0.99954,0.99966,0.99975,0.99982,0.99987,
0.9999,0.99993,0.99995,0.99996,0.99997,0.99998,0.99999,
0.99999,0.99999,1,1)
length(Pr0)
plot(delta,Pr0)
fit0<-smooth.spline(delta,Pr0)
```

```
lines(fit0,lwd=2,col=”red”)
Pr1<-c(0.61111,0.64818,0.68394,0.71811,0.75042,0.78069,0.80877,
0.83454,0.85797,0.87906,0.89785,0.91444,0.92892,0.94145,
0.95218,0.96127,0.96891,0.97525,0.98047,0.98473,0.98816,
0.9909,0.99307,0.99477,0.99609,0.9971,0.99787,0.99845,
0.99888,0.9992,0.99943,0.9996,0.99972,0.99981,0.99987,
0.99991,0.99994,0.99996,0.99997,0.99999,0.99999,0.99999,
1,1,1,1,1,1,1,1,1,1)
length(Pr1)
par(ask=TRUE)
plot(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col=”blue”)
Pr2<-c(0.77778,0.80502,0.83024,0.85338,0.87438,0.89328,0.91009,
0.92491,0.93784,0.949,0.95853,0.9666,0.97334,0.97892,
0.9835,0.9872,0.99017,0.99252,0.99437,0.9958,0.9969,
0.99774,0.99836,0.99883,0.99917,0.99942,0.9996,0.99972,
0.99981,0.99987,0.99992,0.99994,0.99996,0.99998,0.99999,
0.99999,0.99999,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
length(Pr2)
par(ask=TRUE)
plot(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col=”green”)
Pr3<-c(0.94444,0.9533,0.9611,0.96789,0.97374,0.97873,0.98293,
0.98644,0.98933,0.99169,0.99359,0.99511,0.99631,0.99724,
0.99796,0.99851,0.99892,0.99922,0.99945,0.99962,0.99973,
0.99982,0.99988,0.99992,0.99995,0.99996,0.99998,0.99999,
0.99999,0.99999,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1)
length(Pr3)
par(ask=TRUE)
plot(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col=”purple”)
Pr4<-c(rep(1,53))
length(Pr4)
par(ask=TRUE)
plot(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col=”orange”)
par(ask=TRUE)
plot(delta,Pr0,ylab=”P(CS|d)”)
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col=”red”)
```

```
title("Pr(T3>=max(T1,T2,T3)-d) = Pr(CS|d)
for 3 Normal Populations—slippage config & 2 blocks")
points(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
points(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
points(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
points(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col="orange")
legend(2.5,0.85,legend=c("d = 0","d = 1","d = 2","d = 3","d = 4"),
col=c("red","blue","green","purple","orange"),pch=c("o","o","o","o","o"),
lty=c(1,2,3,4,5),ncol=1)
```

Appendix K. Pr(CS) for Three Independent Normal Populations with Equi-Spaced Configuration (Figure 1 Right Panel)

```
#norm_ord_computations using normal ordering.txt
#equi-spaced parameter configurations
#mu=(0,delta,2*delta)
require(splines)
delta<-seq(from=0,to=4.2,by=0.1)
delta
length(delta)
Pr0<-c(0.5,0.55563,0.60899,0.65921,0.70571,0.74812,0.78629,
0.82026,0.85017,0.87625,0.89875,0.91797,0.93421,0.94777,
0.95895,0.96806,0.97539,0.98122,0.98579,0.98933,0.99205,
0.99412,0.99568,0.99684,0.99771,0.99835,0.99881,0.99915,
0.9994,0.99958,0.9997,0.9998,0.99986,0.9999,0.99993,
0.99996,0.99997,0.99998,0.99999,0.99999,0.99999,1,1)
length(Pr0)
plot(delta,Pr0)
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col="red")
Pr1<-c(0.61111,0.66562,0.71566,0.76058,0.80009,0.83419,0.8632,
0.88756,0.90786,0.92467,0.93856,0.95001,0.95946,0.96725,
0.97367,0.97895,0.98328,0.98681,0.98967,0.99197,0.99381,
0.99527,0.99641,0.9973,0.99799,0.99851,0.99891,0.99921,
0.99943,0.99959,0.99971,0.9998,0.99986,0.9999,0.99993,
0.99996,0.99997,0.99998,0.99999,0.99999,0.99999,1,1)
length(Pr1)
par(ask=TRUE)
```

```
plot(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
Pr2<-c(0.77778,0.81764,0.85252,0.88246,0.90774,0.92868,0.94573,
0.95938,0.97011,0.97839,0.98466,0.98932,0.9927,0.99512,
0.9968,0.99795,0.99871,0.99921,0.99953,0.99972,0.99984,
0.99991,0.99995,0.99997,0.99999,0.99999,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1)
length(Pr2)
par(ask=TRUE)
plot(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
Pr3<-c(0.94444,0.95735,0.96797,0.97649,0.9832,0.98829,0.99207,
0.99478,0.99667,0.99794,0.99876,0.99928,0.9996,0.99978,
0.99989,0.99994,0.99997,0.99999,0.99999,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
length(Pr3)
par(ask=TRUE)
plot(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
Pr4<-c(rep(1,43))
length(Pr4)
par(ask=TRUE)
plot(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col="orange")
```

```
par(ask=TRUE)
plot(delta,Pr0,ylab="P(CS|d)")
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col="red")
title("Pr(T3>=max(T1,T2,T3)-d) = Pr(CS|d)
for 3 equi-spaced Normal Populations & 2 blocks")
points(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
points(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
points(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
points(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
```

```
lines(fit4,lwd=2,col="orange")
legend(2.5,0.85,legend=c("d = 0","d = 1","d = 2","d = 3","d = 4"),
col=c("red","blue","green","purple","orange"),pch=c("o","o","o","o","o"),
lty=c(1,2,3,4,5),ncol=1)
```

Appendix L. Ordering Probabilities for Three Independent Bi-Uniform Variates

```
#Bi-Uniform for nonparametric block design, k=3,n=2
#See McDonald (1972), Sankhya, Series A, Vol. 34, Part 1, pp.53-64
#The distribution is a mixture of two uniform densities
#Population 1: Uniform(-c,0) and uniform(1,1+a), c>-0 and a>-0
#Population 2: Uniform(-c+delta1,delta1) and uniform(1+delta1,1+a+delta1),
#c>0, a>0, delta1>=0
#Populatiom 3: Uniform(-c+delta2,delta2) and uniform(1+delta2,1+a+delta2),
#c>0, a>0, delta2>=delta1
#set c, a, delta1, delta2 and number of random draws (nsim)
c<-5;a<-1;delta1<-0;delta2<-0;nsim<-200000
y1=NULL;y2=NULL;y3=NULL
z1=NULL;z2=NULL;z3=NULL
p1=NULL;p2=NULL;p3=NULL
for (I in 1:nsim){
y1[i]<-runif(1,-c,0)
z1[i]<-runif(1,1,1+a)
p1[i]<-rbinom(1,1,0.5)
x1[i]<-p1[i]*y1[i] + (1-p1[i])*z1[i]
}
for (j in 1:nsim){
y2[j]<-runif(1,-c+delta1,delta1)
z2[j]<-runif(1,1+delta1,1+a+delta1)
p2[j]<-rbinom(1,1,0.5)
x2[j]<-p2[j]*y2[j] + (1-p2[j])*z2[j]
}
for (k in 1:nsim){
y3[k]<-runif(1,-c+delta2,delta2)
z3[k]<-runif(1,1+delta2,1+a+delta2)
p3[k]<-rbinom(1,1,0.5)
x3[k]<-p3[k]*y3[k] + (1-p3[k])*z3[k]
}
```

```
df<-data.frame(x1,x2,x3)
head(df,5)
tail(df,5)
avg<-c(mean(x1),mean(x2),mean(x3))
std<-c(sd(x1),sd(x2),sd(x3))
var<-std^2
```

```
avg
std
```

```
r1=NULL;r2=NULL;rk1=NULL;rk2=NULL;add=NULL
d=NULL
n1<-floor((nsim+1)/2)
for (I in 1:n1){
j<-2*i-1
r1<-c(x1[j],x2[j],x3[j])
r2<-c(x1[j+1],x2[j+1],x3[j+1])
rk1<-rank(r1)
rk2<-rank(r2)
add<-rk1+rk2
d[i]<-max(add)-add[3]
}
```

```
pd<-table(d)
pd
sum(pd)
den<-pd/sum(pd)
den
cdf<-cumsum(den)
cdf
summary<-data.frame(den,cdf)
summary
```

Appendix M. Pr(CS) for Three Independent Bi-Uniform Populations with Slippage Configuration (Figure2 Left Panel)

```
#Computations using Counterexample Distribution k=3,n=2
#slippage parameter configurations
#c=a=1; 200,000 simulations
# theta=(0,0,delta)
require(splines)
delta<-seq(from=0,to=1.0,by=0.1)
delta
length(delta)
```

```
Pr0<-c(0.50035,0.56392,0.61917,0.67020,0.70652,
0.74557,0.76886,0.78770,0.80021,0.81007,0.81374)
length(Pr0)
plot(delta,Pr0)
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col="red")
Pr1<-c(0.61072,0.67262,0.72432,0.77131,0.80291,
0.83586,0.85383,0.86842,0.88060,0.88930,0.89132)
length(Pr1)
```

```
par(ask=TRUE)
plot(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
Pr2<-c(0.77821,0.82438,0.85743,0.88818,0.90752,
0.92597,0.93527,0.94186,0.94859,0.95162,0.95389)
length(Pr2)
par(ask=TRUE)
plot(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
Pr3<-c(0.94404,0.95971,0.96915,0.97726,0.98219,
0.98690,0.98899,0.99011,0.99115,0.99182,0.99229)
length(Pr3)
par(ask=TRUE)
plot(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
Pr4<-c(rep(1,11))
length(Pr4)
par(ask=TRUE)
plot(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col="orange")
```

```
par(ask=TRUE)
plot(delta,Pr0,ylab="P(CS|d)",xlim=c(0,1),ylim=c(0.5,1))
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col="red")
title("Pr(T3>=max(T1,T2,T3)-d) = Pr(CS|d)
for 3 slippage Counterexample Populations & 2 blocks")
points(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
points(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
points(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
points(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col="orange")
legend(0.8,0.7,legend=c("d = 0","d = 1","d = 2","d = 3","d = 4"),
col=c("red","blue","green","purple","orange"),pch=c("o","o","o","o","o"),
lty=c(1,2,3,4,5),ncol=1)
```

Appendix N. Pr(CS) for Three Independent Bi-Uniform Populations with Equi-Spaced Configuration (Figure 2 Right Panel)

```
#Computations using Counterexample Distribution k=3,n=2
#equi-spaced parameter configurations
#c=a=1; 200,000 simulations
# theta=(0,delta,2*delta)
require(splines)
delta<-seq(from=0,to=1.0,by=0.1)
delta
length(delta)
```

```
Pr0<-c(0.50035,0.59002,0.66110,0.71370,0.74518,
0.76439,0.76724,0.78088,0.80236,0.82969,0.86351)
length(Pr0)
plot(delta,Pr0)
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col="red")
Pr1<-c(0.61072,0.69698,0.76194,0.81093,0.84024,
0.86286,0.87304,0.88545,0.90072,0.91396,0.92584)
length(Pr1)
par(ask=TRUE)
plot(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
Pr2<-c(0.77821,0.84163,0.88157,0.91152,0.92619,
0.93778,0.94237,0.94811,0.95687,0.96347,0.97230)
length(Pr2)
par(ask=TRUE)
plot(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
Pr3<-c(0.94404,0.96501,0.97582,0.98380,0.98613,
0.98695,0.98671,0.98726,0.98959,0.99269,0.99564)
length(Pr3)
par(ask=TRUE)
plot(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
Pr4<-c(rep(1,11))
length(Pr4)
par(ask=TRUE)
plot(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col="orange")
```

```
par(ask=TRUE)
plot(delta,Pr0,ylab="P(CS|d)",xlim=c(0,1),ylim=c(0.5,1))
fit0<-smooth.spline(delta,Pr0)
lines(fit0,lwd=2,col="red")
title("Pr(T3>=max(T1,T2,T3)-d) = Pr(CS|d)
for 3 equi-spaced Bi-Uniform Populations & 2 blocks")
points(delta,Pr1)
fit1<-smooth.spline(delta,Pr1)
lines(fit1,lwd=2,col="blue")
points(delta,Pr2)
fit2<-smooth.spline(delta,Pr2)
lines(fit2,lwd=2,col="green")
points(delta,Pr3)
fit3<-smooth.spline(delta,Pr3)
lines(fit3,lwd=2,col="purple")
points(delta,Pr4)
fit4<-smooth.spline(delta,Pr4)
lines(fit4,lwd=2,col="orange")
legend(0.8,0.7,legend=c("d = 0","d = 1","d = 2","d = 3","d = 4"),
col=c("red","blue","green","purple","orange"),pch=c("o","o","o","o","o"),
lty=c(1,2,3,4,5),ncol=1)
```

Appendix O. Calculation of A and Pε to Satisfy Equation (6.3.1)

```
#For a given k and epsilon, solves for A and P
#Specify k, epsilon, and an appropriate seq for A
k<-3;epsilon<-0.06066
integrand<-function(x,A=0){((pnorm(x+A))^(k-1))*dnorm(x)}
int<-function(A){integrate(integrand,lower=-Inf,upper=Inf,A=A)$value}
A = seq(0,10,by=0.5)
len<-length(A)
pstar<-sapply(A,int)
df<-data.frame(A,pstar)
ipstar<-qnorm(pstar)
bound<-ipstar*sqrt(2)*(1+epsilon)
diff<-bound-A
print("If diff>0, then A<bound and counterexample inequality holds")
df1<-data.frame(A,pstar,ipstar,bound,diff)
df1
plot(A,bound,main=paste("bound vs. A, k = ",k,", epsilon = ",epsilon,
"\nline is A = bound"), ylim=c(bound[1],bound[len]))
abline(a = 0,b = 1)
```

Chapter 5

Appendix P. EXPANDED MVTFRs DATA

```
#MVTFR Ranks 1994_2022
#Rank sums for the MVTFRs
#k is the number of populations (e.g., states); n is the number of blocks (e.g., years)
k<-51;n<-29
State=c("AL","AK","AZ","AR","CA","CO","CT","DE","DC","FL","GA","HI","ID","IL"
,"IN","IA",
"KS","KY","LA","ME","MD","MA","MI","MN","MS","MO","MT","NE","NV","NH","
NJ","NM","NY","NC",
"ND","OH","OK","OR","PA","RI","SC","SD","TN","TX","UT","VT","VA","WA","WV
","WI","WY")
y1994=c(2.21,2.05,2.33,2.44,1.56,1.74,1.14,1.59,2.00,2.2,1.72,1.54,2.15,1.68,1.59,1.86,1
.79,
1.95,2.25,1.51,1.47,0.94,1.67,1.49,2.77,1.9,2.22,1.75,2.26,1.13,1.26,2.18,1.49,1.99,1.39,
1.4,1.93,1.68,1.56,0.89,2.27,2.02,2.23,1.79,1.9,1.25,1.38,1.35,2.08,1.42,2.15)
y1995=c(2.2,2.11,2.61,2.37,1.52,1.84,1.13,1.61,1.67,2.19,1.74,1.64,2.13,1.68,1.49,2.03,1
.76,
2.07,2.31,1.49,1.5,0.92,1.79,1.35,2.94,1.87,2.28,1.61,2.24,1.11,1.27,2.29,1.46,1.9,1.13,
1.35,1.74,1.91,1.57,1.00,2.28,2.06,2.24,1.76,1.73,1.71,1.29,1.33,2.16,1.45,2.41)
y1996=c(2.23,1.97,2.36,2.21,1.43,1.71,1.1,1.51,1.59,2.12,1.76,1.84,1.99,1.53,1.49,1.73,1
.89,
1.98,2.37,1.32,1.32,0.83,1.67,1.3,2.65,1.88,2.12,1.8,2.18,1.22,1.31,2.25,1.34,1.89,1.26,
1.35,1.96,1.73,1.52,0.97,2.34,2.24,2.12,2.02,1.64,1.38,1.23,1.44,1.97,1.44,1.94)
y1997=c(2.23,1.76,2.19,2.35,1.32,1.62,1.19,1.79,1.8,2.08,1.69,1.65,2.01,1.41,1.36,1.67,1
.82,
1.97,2.44,1.45,1.31,0.87,1.58,1.22,2.73,1.89,2.82,1.77,2.13,1.12,1.23,2.21,1.37,1.81,1.47,
1.39,2.02,1.62,1.59,1.06,2.18,1.86,2.02,1.77,1.79,1.48,1.4,1.32,2.08,1.33,1.81)
y1998=c(1.94,1.55,2.17,2.2,1.2,1.6,1.12,1.4,1.63,2.05,1.63,1.5,1.97,1.38,1.42,1.55,1.82,1
.91,
2.3,1.42,1.25,0.78,1.46,1.31,2.77,1.81,2.47,1.79,2.19,1.11,1.15,1.91,1.23,1.87,1.25,1.36,1
.8,
1.61,1.48,0.93,2.34,2.04,1.94,1.74,1.65,1.58,1.29,1.27,1.9,1.26,1.92)
y1999=c(2.03,1.74,2.18,2.07,1.19,1.54,1.01,1.18,1.18,2.06,1.52,1.21,1.99,1.42,1.46,1.68,
1.95,
1.75,2.28,1.28,1.2,0.8,1.44,1.22,2.66,1.64,2.24,1.64,2.01,1.18,1.11,2.05,1.26,1.71,1.64,1.
36,
1.74,1.19,1.52,1.06,2.41,1.82,2.01,1.67,1.63,1.38,1.19,1.21,2.08,1.31,2.42)
y2000=c(1.76,2.3,2.11,2.24,1.22,1.63,1.11,1.49,1.37,1.99,1.47,1.55,2.04,1.38,1.25,1.51,1
.64,
1.75,2.3,1.19,1.17,0.82,1.41,1.19,2.67,1.72,2.4,1.53,1.83,1.05,1.08,1.9,1.13,1.74,1.19,1.2
9,
1.5,1.33,1.49,0.96,2.34,2.05,1.99,1.72,1.65,1.12,1.24,1.18,2.14,1.4,1.88)
y2001=c(1.75,1.89,2.12,2.08,1.27,1.73,1.03,1.58,1.81,1.77,1.53,1.61,1.84,1.37,1.27,1.49,
1.75,
1.83,2.2,1.33,1.27,0.9,1.34,1.06,2.18,1.62,2.3,1.36,1.72,1.15,1.08,2.00,1.2,1.67,1.45,1.29,
1.57,1.42,1.49,1.01,2.27,2.00,1.85,1.73,1.24,1.17,1.27,1.21,1.91,1.33,2.16)
```

```
y2002=c(1.8,1.82,2.18,2.13,1.27,1.71,1.04,1.4,1.33,1.76,1.41,1.34,1.86,1.35,1.09,1.31,1.
78,
1.95,2.09,1.47,1.23,0.86,1.28,1.2,2.43,1.77,2.59,1.64,2.12,1.01,1.1,1.97,1.15,1.7,1.32,1.3
1,
1.62,1.26,1.54,1.03,2.23,2.12,1.73,1.73,1.34,0.98,1.18,1.2,2.19,1.37,1.95)
y2003=c(1.71,1.98,2.07,2.09,1.31,1.48,0.95,1.57,1.87,1.71,1.47,1.43,2.05,1.36,1.15,1.42,
1.64,
1.99,2.13,1.39,1.19,0.86,1.27,1.18,2.33,1.81,2.41,1.54,1.91,0.98,1.05,1.92,1.11,1.66,1.41,
1.17,
1.47,1.46,1.48,1.24,2.01,2.38,1.73,1.71,1.29,0.83,1.23,1.09,1.96,1.42,1.79)
y2004=c(1.95,2.02,2.01,2.22,1.25,1.45,0.93,1.44,1.15,1.65,1.44,1.46,1.77,1.24,1.3,1.23,1
.57,
2.04,2.08,1.3,1.16,0.87,1.12,1.00,2.28,1.64,2.04,1.32,1.95,1.26,0.99,2.18,1.08,1.64,1.32,1
.15,
1.67,1.28,1.38,0.98,2.11,2.24,1.89,1.6,1.2,1.25,1.17,1.02,2.02,1.31,1.77)
y2005=c(1.92,1.45,1.97,2.05,1.32,1.26,0.88,1.4,1.29,1.75,1.52,1.39,1.85,1.27,1.31,1.45,1
.44,
2.08,2.14,1.13,1.09,0.8,1.09,0.98,2.32,1.83,2.26,1.43,2.06,1.24,1.01,2.04,1.03,1.53,1.62,
1.2,1.71,1.38,1.5,1.05,2.21,2.22,1.79,1.5,1.12,0.95,1.18,1.17,1.82,1.36,1.88)
y2006=c(1.99,1.49,2.07,2.01,1.29,1.1,0.98,1.57,1.02,1.65,1.49,1.58,1.76,1.17,1.27,1.4,1.
55,
1.91,2.17,1.25,1.16,0.78,1.04,0.87,2.2,1.59,2.34,1.39,1.97,0.93,1.02,1.88,1.03,1.53,1.41,
1.11,1.57,1.35,1.41,0.98,2.08,2.08,1.82,1.48,1.11,1.11,1.19,1.12,1.96,1.22,2.07)
y2007=c(1.81,1.59,1.7,1.96,1.22,1.14,0.92,1.23,1.22,1.56,1.46,1.33,1.6,1.16,1.23,1.43,1.
38,
1.8,2.19,1.22,1.09,0.79,1.04,0.89,2.04,1.43,2.45,1.32,1.68,0.96,0.95,1.54,0.97,1.62,1.42,
1.13,1.61,1.31,1.37,0.8,2.11,1.62,1.7,1.42,1.11,0.86,1.25,1.0,2.1,1.27,1.6)
y2008=c(1.63,1.27,1.52,1.81,1.05,1.15,0.95,1.35,0.94,1.5,1.37,1.04,1.52,0.98,1.11,1.34,1
.29,
1.74,2.03,1.06,1.07,0.67,0.96,0.78,1.79,1.41,2.12,1.09,1.56,1.06,0.8,1.39,0.92,1.4,1.33,1.
1,
1.55,1.24,1.36,0.79,1.86,1.35,1.5,1.48,1.06,1.0,1.0,0.94,1.82,1.05,1.68)
y2009=c(1.38,1.3,1.31,1.8,0.95,1.01,0.71,1.28,0.8,1.3,1.18,1.09,1.46,0.86,0.9,1.19,1.31,1
.67,
1.84,1.1,0.99,0.62,0.9,0.74,1.73,1.27,2.01,1.15,1.19,0.85,0.8,1.39,0.87,1.28,1.72,0.92,1.5
7,
1.11,1.22,1.01,1.82,1.48,1.4,1.35,0.93,0.97,0.94,0.87,1.82,0.96,1.4)
y2010=c(1.34,1.17,1.27,1.7,0.84,1.96,1.02,1.13,0.67,1.25,1.12,1.13,1.32,0.88,1.0,1.24,1.
44,
1.58,1.59,1.11,0.88,0.64,0.97,0.73,1.61,1.16,1.69,0.98,1.16,0.98,0.76,1.38,0.92,1.29,1.27,
0.97,1.4,0.94,1.32,0.81,1.65,1.58,1.47,1.29,0.95,0.98,0.9,0.8,1.64,0.96,1.66)
y2011=c(1.38,1.57,1.39,1.67,0.88,0.96,0.71,1.1,0.76,1.25,1.13,0.99,1.05,0.89,0.98,1.15,1
.29,
1.5,1.46,0.95,0.86,0.68,0.94,0.65,1.62,1.14,1.79,0.95,1.02,0.71,0.86,1.36,0.92,1.19,1.62,
0.91,1.47,0.99,1.3,0.84,1.7,1.23,1.32,1.29,0.93,0.77,0.94,0.8,1.78,0.99,1.46)
```

y2012=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
#####
y2013=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2014=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2015=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2016=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2017=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2018=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2019=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2020=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2021=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)
y2022=c(1.33,1.23,1.37,1.65,0.88,1.01,0.75,1.24,0.42,1.27,1.11,1.25,1.13,0.91,0.99,1.16,
1.32,
1.58,1.54,1.16,0.89,0.62,0.99,0.69,1.51,1.21,1.72,1.1,1.07,0.84,0.79,1.43,0.91,1.23,1.69,
1.0,1.48,1.01,1.32,0.82,1.76,1.46,1.42,1.43,0.82,1.07,0.96,0.78,1.76,1.04,1.33)

```
#####
MV<-
data.frame(State,y1994,y1995,y1996,y1997,y1998,y1999,y2000,y2001,y2002,y2003,y20
04,
y2005,y2006,y2007,y2008,y2009,y2010,y2011,y2012,y2013,y2014,y2015,y2016,y2017,
y2018,y2019,
y2020,y2021,y2022)
MV
#Tied ranks are resolved at random
x1<-rank(y1994,ties.method="random");x2<-rank(y1995,ties.method="random")
x3<-rank(y1996,ties.method="random");x4<-rank(y1997,ties.method="random")
x5<-rank(y1998,ties.method="random");x6<-rank(y1999,ties.method="random")
x7<-rank(y2000,ties.method="random");x8<-rank(y2001,ties.method="random")
x9<-rank(y2002,ties.method="random");x10<-rank(y2003,ties.method="random")
x11<-rank(y2004,ties.method="random");x12<-rank(y2005,ties.method="random")
x13<-rank(y2006,ties.method="random");x14<-rank(y2007,ties.method="random")
x15<-rank(y2008,ties.method="random");x16<-rank(y2009,ties.method="random")
x17<-rank(y2010,ties.method="random");x18<-rank(y2011,ties.method="random")
x19<-rank(y2012,ties.method="random");x20<-rank(y2013,ties.method="random")
x21<-rank(y2014,ties.method="random");x22<-rank(y2015,ties.method="random")
x23<-rank(y2016,ties.method="random");x24<-rank(y2017,ties.method="random")
x25<-rank(y2018,ties.method="random");x26<-rank(y2019,ties.method="random")
x27<-rank(y2020,ties.method="random");x28<-rank(y2021,ties.method="random")
x29<-rank(y2022,ties.method="random")
ra<-data.frame(x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12,x13,x14,x15,x16,x17,
x18,x19,x20,x21,x22,x23,x24,x25,x26,x27,x28,x29)
#ra
ram<-as.matrix(ra)
ram
RkSum<-rep(0,k)
for (i in 1:k){RkSum[i]<-sum(ram[i,])}
#RkSum
StRk<-data.frame(State,RkSum)
#StRk
StRkOrd<-StRk[order(StRk$RkSum),]
#StRkOrd
NorSc<-rep(0,51)
#The following are approximations to exact values given by Harter
#for (i in 1:k){NorSc[i]<-qnorm(i/(k+1))}
#NorSc<-round(NorSc,4)
#The following expected values of normal order stats are taken from
#"Expected Values of Normal Order Statistics," by H. Leon Harter (1961)
#If two states are tied and the rank values are r1 and r2 (r1<r2), then r1
#is assigned to the state that is lower in alphabetical order (using the
#two letter state abbreviation; r2 is assigned to the other state.
```

```
#Similarly if three are more states are tied in their MVTFRs.
NorSc<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,-1.23003,-1.13162,
-1.04312,-0.96213,-0.88701,-0.81661,-0.75004,-0.68666,-0.62592,-0.56742,
-0.51080,-0.45578,-0.40211,-0.34957,-0.29799,-0.24719,-0.19702,-0.14735,
-0.09803,-0.04896,0,0.04896,0.09803,0.14735,0.19702,0.24719,0.29799,
0.34957,0.40211,0.45578,0.51080,0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,1.34207,1.47409,1.63829,1.86371,
2.25678)
sum(NorSc)
#NorSc
df94<-data.frame(State,x1)
ns94<-rep(0,51)
for (i in 1:51){ns94[i]<-NorSc[x1[i]]}
df94<-data.frame(State,x1,ns94)
#df94
#
df95<-data.frame(State,x2)
ns95<-rep(0,51)
for (i in 1:51){ns95[i]<-NorSc[x2[i]]}
df95<-data.frame(State,x2,ns95)
#df95
#
df96<-data.frame(State,x3)
ns96<-rep(0,51)
for (i in 1:51){ns96[i]<-NorSc[x3[i]]}
df96<-data.frame(State,x3,ns96)
#df96
#
df97<-data.frame(State,x4)
ns97<-rep(0,51)
for (i in 1:51){ns97[i]<-NorSc[x4[i]]}
df97<-data.frame(State,x4,ns97)
#df97
#
df98<-data.frame(State,x5)
ns98<-rep(0,51)
for (i in 1:51){ns98[i]<-NorSc[x5[i]]}
df98<-data.frame(State,x5,ns98)
#df98
#
df99<-data.frame(State,x6)
ns99<-rep(0,51)
for (i in 1:51){ns99[i]<-NorSc[x6[i]]}
df99<-data.frame(State,x6,ns99)
#df99
```

```
#
df00<-data.frame(State,x7)
ns00<-rep(0,51)
for (i in 1:51){ns00[i]<-NorSc[x7[i]]}
df00<-data.frame(State,x7,ns00)
#df00
#
df01<-data.frame(State,x8)
ns01<-rep(0,51)
for (i in 1:51){ns01[i]<-NorSc[x8[i]]}
df01<-data.frame(State,x8,ns01)
#df01
#
df02<-data.frame(State,x9)
ns02<-rep(0,51)
for (i in 1:51){ns02[i]<-NorSc[x9[i]]}
df02<-data.frame(State,x9,ns02)
#df02
#
df03<-data.frame(State,x10)
ns03<-rep(0,51)
for (i in 1:51){ns03[i]<-NorSc[x10[i]]}
df03<-data.frame(State,x10,ns03)
#df03
#
df04<-data.frame(State,x11)
ns04<-rep(0,51)
for (i in 1:51){ns04[i]<-NorSc[x11[i]]}
df04<-data.frame(State,x11,ns04)
#df04
#
df05<-data.frame(State,x12)
ns05<-rep(0,51)
for (i in 1:51){ns05[i]<-NorSc[x12[i]]}
df05<-data.frame(State,x12,ns05)
#df05
#
df06<-data.frame(State,x13)
ns06<-rep(0,51)
for (i in 1:51){ns06[i]<-NorSc[x13[i]]}
df06<-data.frame(State,x13,ns06)
#df06
#
df07<-data.frame(State,x14)
ns07<-rep(0,51)
```

```
for (i in 1:51){ns07[i]<-NorSc[x14[i]]}
df07<-data.frame(State,x14,ns07)
#df07
#
df08<-data.frame(State,x15)
ns08<-rep(0,51)
for (i in 1:51){ns08[i]<-NorSc[x15[i]]}
df08<-data.frame(State,x15,ns08)
#df08
#
df09<-data.frame(State,x16)
ns09<-rep(0,51)
for (i in 1:51){ns09[i]<-NorSc[x16[i]]}
df09<-data.frame(State,x16,ns09)
#df09
#
df10<-data.frame(State,x17)
ns10<-rep(0,51)
for (i in 1:51){ns10[i]<-NorSc[x17[i]]}
df10<-data.frame(State,x17,ns10)
#df10
#
df11<-data.frame(State,x18)
ns11<-rep(0,51)
for (i in 1:51){ns11[i]<-NorSc[x18[i]]}
df11<-data.frame(State,x18,ns11)
#df11
#
df12<-data.frame(State,x19)
ns12<-rep(0,51)
for (i in 1:51){ns12[i]<-NorSc[x19[i]]}
df12<-data.frame(State,x19,ns12)
#df12
#
df13<-data.frame(State,x20)
ns13<-rep(0,51)
for (i in 1:51){ns13[i]<-NorSc[x20[i]]}
df13<-data.frame(State,x20,ns13)
#df13
#
df14<-data.frame(State,x21)
ns14<-rep(0,51)
for (i in 1:51){ns14[i]<-NorSc[x21[i]]}
df14<-data.frame(State,x21,ns14)
#df14
```

```
#
df15<-data.frame(State,x22)
ns15<-rep(0,51)
for (i in 1:51){ns15[i]<-NorSc[x22[i]]}
df15<-data.frame(State,x22,ns15)
#df15
#
df16<-data.frame(State,x23)
ns16<-rep(0,51)
for (i in 1:51){ns16[i]<-NorSc[x23[i]]}
df16<-data.frame(State,x23,ns16)
#df16
#
df17<-data.frame(State,x24)
ns17<-rep(0,51)
for (i in 1:51){ns17[i]<-NorSc[x24[i]]}
df17<-data.frame(State,x24,ns17)
#df17
#
df18<-data.frame(State,x25)
ns18<-rep(0,51)
for (i in 1:51){ns18[i]<-NorSc[x25[i]]}
df18<-data.frame(State,x25,ns18)
#df18
#
df19<-data.frame(State,x26)
ns19<-rep(0,51)
for (i in 1:51){ns19[i]<-NorSc[x26[i]]}
df19<-data.frame(State,x26,ns19)
#df19
#
df20<-data.frame(State,x27)
ns20<-rep(0,51)
for (i in 1:51){ns20[i]<-NorSc[x27[i]]}
df20<-data.frame(State,x27,ns20)
#df20
#
df21<-data.frame(State,x28)
ns21<-rep(0,51)
for (i in 1:51){ns21[i]<-NorSc[x28[i]]}
df21<-data.frame(State,x28,ns21)
#df21
#
df22<-data.frame(State,x29)
ns22<-rep(0,51)
```

```
for (i in 1:51){ns22[i]<-NorSc[x29[i]]}
df22<-data.frame(State,x29,ns22)
#df22
#
Ns<-data.frame(ns94,ns95,ns96,ns97,ns98,ns99,ns00,ns01,
ns02,ns03,ns04,ns05,ns06,ns07,ns08,ns09,ns10,ns11,ns12)
scm<-as.matrix(Ns)
#scm
NsSum<-rep(0,k)
for (i in 1:k){NsSum[i]<-sum(scm[i,])}
#NsSum
StRkNs<-data.frame(State,NsSum)
#StRkNs
#
StNs<-data.frame(State,NsSum)
StNs
StNsOrd<-StNs[order(StNs$NsSum),]
StNsOrd
StRks<-data.frame(State,RkSum)
StRks
StRksOrd<-StRks[order(StRks$RkSum),]
StRksOrd
#
Summary<-data.frame(StRkOrd,StNsOrd)
Summary
#check on distribution of MVTFRs for one year, 1994
hist(y1994,col='red',freq=FALSE,xlab="1994 MVTFRs",
main="Histogram of 1994 MVTFRs\n Normal Density Overlay")
low<-min(y1994)-0.1;up<-max(y1994)+0.1
curve(dnorm(x,mean(y1994),sd(y1994),),from=low,to=up,add=TRUE)
#
#check on distribution of MVTFRs for one year, 2012
#hist(y2012,col='red',freq=FALSE,xlab="2012 MVTFRs",ylim=c(0,1.4),
#main="Histogram of 2012 MVTFRs\n Normal Density Overlay")
#low<-min(y2012)-0.2;up<-max(y2012)+0.2
#curve(dnorm(x,mean(y2012),sd(y2012),),from=low,to=up,add=TRUE)
```

App. A.expanded.txt
Displaying App. A.expanded.txt.

Appendix Q: constants calculations for different values of P* (0.75,0.90,0.95,0.99)

```
################################
# CALCULATIONs RELATED TO WHEN P*=0.75
################################
```

```
# Find the d's constants#
# find d1 by simulation #
# N simulations#
k=51; n=29; N=25000
size<-k*n
m<-rep(0,size)
M<-matrix(m,nrow=n,ncol=k,byrow=TRUE)
S<-rep(0,k)
row.seq<-rep(0,k)
#for (i in 1:k){
#row.seq[i]<-qnorm(i/(k+1))
#}
#row.seq<-seq(1:k)
#normal.score.51 taken from H. Leon Harter (1961), Biometrika
norm.score.51<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,
-1.23003,-1.13162,-1.04312,-0.96213,-0.88701,
-0.81661,-0.75004,-0.68666,-0.62592,-0.56742,
-0.51080,-0.45578,-0.40211,-0.34957,-0.29799,
-0.24719,-0.19702,-0.14735,-0.09803,-0.04896,0.00000,
0.04896,0.09803,0.14735,0.19702,0.24719,
0.29799,0.34957,0.40211,0.45578,0.51080,
0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,
1.34207,1.47409,1.63829,1.86371,2.25678)
row.seq<-norm.score.51
for (m in 1:N){
for (i in 1:n){
M[i,]<-sample(row.seq,replace=FALSE)
}
M
for (j in 1:k){
S[j]<-sum(M[,j])
}
S
T[m]<-max(S)-S[k]
}
#T
#sort(unique(T))
quan<-c(0.50,0.75,0.90,0.95,0.97,0.99)
round(quantile(T,quan),4)
norm.sc.sq<-(norm.score.51)^2
ssq<-sum(norm.sc.sq)
h<-((k-1)/(n*ssq))^0.5
h
d1=15.7484
```

```
h*d1
Pstar<-0.75
d2<-((n*ssq)/k)^0.5*qnorm(1-Pstar)
d2
d3=d1
d3
d4=-d2
d4

w<-h*d1
w

############################################################
# Find the b's constants#
#find the constant b1 and d1 based on determining the values of c.b1 and h.d1
#which they share a common value call it w#


# find b1 #
#k populations; n blocks#
w<-2.971838 # P* is 75
k=51
n=29
func<-function(x){(pnorm(x+w))^50*dnorm(x)}
integrate(func,lower=-Inf,upper=Inf)
c<-(12/(n*k*(k+1)))^0.5
# to find b1 now, all we need is to divide w by c#
b1<-w/c
b1
Pstar<-0.75
b2<-((n*(k^2-1))/12)^0.5*qnorm(1-Pstar)+(n*(k+1))/2
b2
b3=b1
b3
b4<-n*(k+1)-b2
b4

################################
# CALCULATIONs RELATED TO WHEN P*=0.90
################################
# Find the d's constants#
# find d1 by simulation #
# N simulations#
k=51; n=29; N=25000
size<-k*n
```

```
m<-rep(0,size)
M<-matrix(m,nrow=n,ncol=k,byrow=TRUE)
S<-rep(0,k)
row.seq<-rep(0,k)
#for (i in 1:k){
#row.seq[i]<-qnorm(i/(k+1))
#}
#row.seq<-seq(1:k)
#normal.score.51 taken from H. Leon Harter (1961), Biometrika
norm.score.51<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,
-1.23003,-1.13162,-1.04312,-0.96213,-0.88701,
-0.81661,-0.75004,-0.68666,-0.62592,-0.56742,
-0.51080,-0.45578,-0.40211,-0.34957,-0.29799,
-0.24719,-0.19702,-0.14735,-0.09803,-0.04896,0.00000,
0.04896,0.09803,0.14735,0.19702,0.24719,
0.29799,0.34957,0.40211,0.45578,0.51080,
0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,
1.34207,1.47409,1.63829,1.86371,2.25678)
row.seq<-norm.score.51
for (m in 1:N){
for (i in 1:n){
M[i,]<-sample(row.seq,replace=FALSE)
}
M
for (j in 1:k){
S[j]<-sum(M[,j])
}
S
T[m]<-max(S)-S[k]
}
#T
#sort(unique(T))
quan<-c(0.50,0.75,0.90,0.95,0.97,0.99)
round(quantile(T,quan),4)
norm.sc.sq<-(norm.score.51)^2
ssq<-sum(norm.sc.sq)
h<-((k-1)/(n*ssq))^0.5
h
d1=19.4745
h*d1
Pstar<-0.90
d2<-((n*ssq)/k)^0.5*qnorm(1-Pstar)
d2
d3=d1
```

```
d3
d4=-d2
d4

w<-h*d1
w

#############################################################
# Find the b's constants#
#find the constant b1 and d1 based on determining the values of c.b1 and h.d1
#which they share a common value call it w#

# find b1 #
#k populations; n blocks#
w<-3.674981 # P* is 90
k=51
n=29
func<-function(x){(pnorm(x+w))^50*dnorm(x)}
integrate(func,lower=-Inf,upper=Inf)
c<-(12/(n*k*(k+1)))^0.5
# to find b1 now, all we need is to divide w by c#
b1<-w/c
b1
Pstar<-0.90
b2<-((n*(k^2-1))/12)^0.5*qnorm(1-Pstar)+(n*(k+1))/2
b2
b3=b1
b3
b4<-n*(k+1)-b2
b4
################################
# CALCULATIONs RELATED TO WHEN P*=0.95
################################
# Find the d's constants#
# find d1 by simulation #
# N simulations#
k=51; n=29; N=25000
size<-k*n
m<-rep(0,size)
M<-matrix(m,nrow=n,ncol=k,byrow=TRUE)
S<-rep(0,k)
row.seq<-rep(0,k)
#for (i in 1:k){
#row.seq[i]<-qnorm(i/(k+1))
#}
```

```
#row.seq<-seq(1:k)
#normal.score.51 taken from H. Leon Harter (1961), Biometrika
norm.score.51<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,
-1.23003,-1.13162,-1.04312,-0.96213,-0.88701,
-0.81661,-0.75004,-0.68666,-0.62592,-0.56742,
-0.51080,-0.45578,-0.40211,-0.34957,-0.29799,
-0.24719,-0.19702,-0.14735,-0.09803,-0.04896,0.00000,
0.04896,0.09803,0.14735,0.19702,0.24719,
0.29799,0.34957,0.40211,0.45578,0.51080,
0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,
1.34207,1.47409,1.63829,1.86371,2.25678)
row.seq<-norm.score.51
for (m in 1:N){
for (i in 1:n){
M[i,]<-sample(row.seq,replace=FALSE)
}
M
for (j in 1:k){
S[j]<-sum(M[,j])
}
S
T[m]<-max(S)-S[k]
}
#T
#sort(unique(T))
quan<-c(0.50,0.75,0.90,0.95,0.97,0.99)
round(quantile(T,quan),4)
norm.sc.sq<-(norm.score.51)^2
ssq<-sum(norm.sc.sq)
h<-((k-1)/(n*ssq))^0.5
h
d1=21.5202
h*d1
Pstar<-0.95
d2<-((n*ssq)/k)^0.5*qnorm(1-Pstar)
d2
d3=d1
d3
d4=-d2
d4

w<-h*d1
w
```

```
############################################################
# Find the b's constants#
#find the constant b1 and d1 based on determining the values of c.b1 and h.d1
#which they share a common value call it w#


# find b1 #
#k populations; n blocks#
w<-4.061019 # P* is 95
k=51
n=29
func<-function(x){(pnorm(x+w))^50*dnorm(x)}
integrate(func,lower=-Inf,upper=Inf)
c<-(12/(n*k*(k+1)))^0.5
# to find b1 now, all we need is to divide w by c#
b1<-w/c
b1
Pstar<-0.95
b2<-((n*(k^2-1))/12)^0.5*qnorm(1-Pstar)+(n*(k+1))/2
b2
b3=b1
b3
b4<-n*(k+1)-b2
b4


################################
# CALCULATIONs RELATED TO WHEN P*=0.99
################################
# Find the d's constants#
# find d1 by simulation #
# N simulations#
k=51; n=29; N=25000
size<-k*n
m<-rep(0,size)
M<-matrix(m,nrow=n,ncol=k,byrow=TRUE)
S<-rep(0,k)
row.seq<-rep(0,k)
#for (i in 1:k){
#row.seq[i]<-qnorm(i/(k+1))
#}
#row.seq<-seq(1:k)
#normal.score.51 taken from H. Leon Harter (1961), Biometrika
norm.score.51<-c(-2.25678,-1.86371,-1.63829,-1.47409,-1.34207,
-1.23003,-1.13162,-1.04312,-0.96213,-0.88701,
```

```
-0.81661,-0.75004,-0.68666,-0.62592,-0.56742,
-0.51080,-0.45578,-0.40211,-0.34957,-0.29799,
-0.24719,-0.19702,-0.14735,-0.09803,-0.04896,0.00000,
0.04896,0.09803,0.14735,0.19702,0.24719,
0.29799,0.34957,0.40211,0.45578,0.51080,
0.56742,0.62592,0.68666,0.75004,0.81661,
0.88701,0.96213,1.04312,1.13162,1.23003,
1.34207,1.47409,1.63829,1.86371,2.25678)
row.seq<-norm.score.51
for (m in 1:N){
for (i in 1:n){
M[i,]<-sample(row.seq,replace=FALSE)
}
M
for (j in 1:k){
S[j]<-sum(M[,j])
}
S
T[m]<-max(S)-S[k]
}
#T
#sort(unique(T))
quan<-c(0.50,0.75,0.90,0.95,0.97,0.99)
round(quantile(T,quan),4)
norm.sc.sq<-(norm.score.51)^2
ssq<-sum(norm.sc.sq)
h<-((k-1)/(n*ssq))^0.5
h
d1=25.5102
h*d1
Pstar<-0.99
d2<-((n*ssq)/k)^0.5*qnorm(1-Pstar)
d2
d3=d1
d3
d4=-d2
d4

w<-h*d1
w

############################################################
# Find the b's constants#
#find the constant b1 and d1 based on determining the values of c.b1 and h.d1
#which they share a common value call it w#
```

```
# find b1 #
#k populations; n blocks#
w<-4.813962 # P* is 99
k=51
n=29
func<-function(x){(pnorm(x+w))^50*dnorm(x)}
integrate(func,lower=-Inf,upper=Inf)
c<-(12/(n*k*(k+1)))^0.5
# to find b1 now, all we need is to divide w by c#
b1<-w/c
b1
Pstar<-0.99
b2<-((n*(k^2-1))/12)^0.5*qnorm(1-Pstar)+(n*(k+1))/2
b2
b3=b1
b3
b4<-n*(k+1)-b2
b4
```

REFERENCES

Chick, Stephen E., and Koichiro Inoue. "New two-stage and sequential procedures for selecting the best simulated system." *Operations Research* 49.5 (2001): 732-743.

Chick, Stephen E., and Peter Frazier. "Sequential sampling with economics of selection procedures." *Management Science* 58.3 (2012): 550-569.

Conover, William Jay. *Practical nonparametric statistics*. Vol. 350. john wiley & sons, 1999.

Dunnett, Charles W., and Milton Sobel. "A bivariate generalization of Student's t-distribution, with tables for certain special cases." *Biometrika* 41.1-2 (1954): 153-169.

Elkadry, Alaa, and Gary C. McDonald. "Analyzing continuous randomized response data with an indifference-zone selection procedure." *Communications in Statistics-Theory and Methods* 47.14 (2018): 3508-3522.

Fan, Weiwei, L. Jeff Hong, and Barry L. Nelson. "Indifference-zone-free selection of the best." *Operations Research* 64.6 (2016): 1499-1514.

Gibbons, Jean D., Ingram Olkin, and Milton Sobel. "An introduction to ranking and selection." *The American Statistician* 33.4 (1979): 185-195.

Gibbons, Jean Dickinson, Ingram Olkin, and Milton Sobel. *Selecting and ordering populations: A new statistical methodology*. Society for Industrial and Applied Mathematics, 1999.

Green, Jaclyn, Gary C. McDonald, and Nirupama Rao. "Using selection procedures to analyze state traffic fatality rates." *American Journal of Mathematical and Management Sciences* 26.3-4 (2006): 387-416.

Green, Jaclyn, and Gary C. McDonald. "Nonparametric subset selection procedures: applications and properties." *American Journal of Mathematical and Management Sciences* 29.3-4 (2009): 413-436.

Gupta, Shanti S., and G. C. MacDonald. *Nonparametric Procedures in Multiple Decisions: Ranking and Selection Procedures*. Purdue University. Department of Statistics, 1980.

Gupta, Shanti S., and Gary C. McDonald. "On some classes of selection procedures based on ranks." *Nonparametric techniques in statistical inference* (1970): 491-514.

Gupta, Shanti S., and Subramanian Panchapakesan. *Multiple decision procedures: theory and methodology of selecting and ranking populations*. Society for Industrial and Applied Mathematics, 2002.

Gupta, Shanti S., and S. Panchapakesan. "Subset selection procedures: review and assessment." *American Journal of Mathematical and Management Sciences* 5.3-4 (1985): 235-311.
GUPTA, SHANTIS, and Takashi Matsui. "On the least favorable configuration of a selection procedure based on ranks." (1987).

Gupta, Shanti S. "On a selection and ranking procedure for gamma populations." *Annals of the Institute of Statistical Mathematics* 14.1 (1962): 199-212.

Gupta, Shanti S. "On some multiple decision (selection and ranking) rules." *Technometrics* 7.2 (1965): 225-245.

Gupta, Shanti S. "Selection and ranking procedures: A brief introduction." *Communications in Statistics-Theory and Methods* 6.11 (1977): 993-1001.

Harter, H. Leon. "Expected values of normal order statistics." *Biometrika* 48.1/2 (1961): 151-165.

Hong, L. Jeff, Weiwei Fan, and Jun Luo. "Review on ranking and selection: A new perspective." *Frontiers of Engineering Management* 8.3 (2021): 321-343.

Kim, Seong-Hee, and Barry L. Nelson. "A fully sequential procedure for indifference-zone selection in simulation." *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 11.3 (2001): 251-273.

Kuehl, Robert O. "Design of experiments: statistical principles of research design and analysis." *(No Title)* (2000).

Lee, Young Jack, and Edward J. Dudewicz. "On the least favorable configuration of a selection procedure based on rank sums: counterexample to a postulate of Matsui." *Journal of the Japan Statistical Society, Japanese Issue* 9.2 (1980): 65-69.

Lorenzen, Thomas J., and Gary C. McDonald. "A nonparametric analysis of urban, rural, and interstate traffic fatality rates." *Design of Experiments—Ranking and Selection. New York: Marcel Dekker* (1984): 143-178.

Matsui, Takashi. "Selection problems based on ranked data." *Hiroshima mathematical journal* 28.1 (1998): 55-93.

McDonald, Gary C., and Sajidah Alsaeed. "Comparison of Block Design Nonparametric Subset Selection Rules Based on Alternative Scoring Rules." *Applied Mathematics* 15.5 (2024): 355-389.

McDonald, Gary C., and Sajidah Alsaeed. "On Performance Characteristics of a Nonparametric Subset Selection Procedure with a Small Randomized Block Experimental Design." *Applied Mathematics* 15.9 (2024): 630-650.

McDonald, Gary C. "Applications of subset selection procedures and Bayesian ranking methods in analysis of traffic fatality data." *Wiley Interdisciplinary Reviews: Computational Statistics* 8.6 (2016): 222-237.

McDonald, Gary C. "Characteristics of block design selection procedures and a counterexample." *Sankhyā: The Indian Journal of Statistics, Series B* (1985): 47-55.

McDonald, Gary C. "Computing Probabilties for Rank Statistics Used with Block Design Nonparametric Subset Selection Rules." *American Journal of Mathematical and Management Sciences* 41.1 (2022): 38-50.

McDonald, Gary C. "Nonparametric selection procedures applied to state traffic fatality rates." *Technometrics* 21.4 (1979): 515-523.

McDonald, Gary C. "Some multiple comparison selection procedures based on ranks." *Sankhyā: The Indian Journal of Statistics, Series A* (1972): 53-64.

McDonald, Gary C. "The distribution of some rank statistics with applications in block design selection problems." *Sankhyā: The Indian Journal of Statistics, Series A* (1973): 187-204.

Mosteller, Frederick. "A k-sample slippage test for an extreme population." *Selected Papers of Frederick Mosteller* (2006): 101-109.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). Applied linear statistical models.

Pan, Guohua. "Subset selection with additional order information." *Biometrics* (1996): 1363-1374.

Rinott, Yosef. "On two-stage selection procedures and related probability-inequalities." *Communications in Statistics-Theory and methods* 7.8 (1978): 799-811.

Rizvi, M. Haseeb, and George G. Woodworth. "On selection procedures based on ranks: counterexamples concerning least favorable configurations." *The Annals of Mathematical Statistics* 41.6 (1970): 1942-1951.

Santner, Thomas J. "A Restricted Subset Selection Approach to Ranking and Selection Problems." *Ann. Statist.* 3.1 (1975): 334-349.21

Sidak, Zbynek, Pranab K. Sen, and Jaroslav Hajek. *Theory of rank tests*. Elsevier, 1999.

Tukey, J.W. (1949) One Degree-of-Freedom for Non-Additivity. Biometrics, 5, 232-242. https://doi.org/10.2307/3001938

Winer, B. J., Brown, D. R., & Michels, K. M. (1971). *Statistical principles in experimental design* (Vol. 2, p. 596). New York: Mcgraw-hill.