Scientific Research Publishing

Applied Mathematics, 2024, 15, 355-389
https://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Comparison of Block Design Nonparametric Subset Selection Rules Based on Alternative Scoring Rules

Gary C. McDonald, Sajidah Alsaeed

Department of Mathematics and Statistics, Oakland University, Rochester, MI, USA
Email: mcdonald@oakland.edu, salsaeed@oakland.edu

How to cite this paper: McDonald, G.C. and Alsaeed, S. (2024) Comparison of Block Design Nonparametric Subset Selection Rules Based on Alternative Scoring Rules. *Applied Mathematics*, **15**, 355-389.
https://doi.org/10.4236/am.2024.155022

Received: April 24, 2024
Accepted: May 27, 2024
Published: May 30, 2024

Abstract

This article compares the size of selected subsets using nonparametric subset selection rules with two different scoring rules for the observations. The scoring rules are based on the expected values of order statistics of the uniform distribution (yielding rank values) and of the normal distribution (yielding normal score values). The comparison is made using state motor vehicle traffic fatality rates, published in a 2016 article, with fifty-one states (including DC as a state) and over a nineteen-year period (1994 through 2012). The earlier study considered four block design selection rules—two for choosing a subset to contain the "best" population (*i.e.*, state with lowest mean fatality rate) and two for the "worst" population (*i.e.*, highest mean rate) with a probability of correct selection chosen to be 0.90. Two selection rules based on normal scores resulted in selected subset sizes substantially smaller than corresponding rules based on ranks (7 vs. 16 and 3 vs. 12). For two other selection rules, the subsets chosen were very close in size (within one). A comparison is also made using state homicide rates, published in a 2022 article, with fifty states and covering eight years. The results are qualitatively the same as those obtained with the motor vehicle traffic fatality rates.

Keywords

Order Statistics, Rank Scoring Methods, Probability of a Correct Selection, Subset Size, Motor Vehicle Traffic Fatality Rates, Homicide Rates, Asymptotic Distributions

1. Introduction

Nonparametric statistical methods are useful for analyzing data that might not satisfy the distributional assumptions of parametric methods (e.g., see Conover

[1]. 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 (e.g., LaVange and Koch [2]. McDonald [3] [4] developed a class of nonparametric (distribution-free) subset selection rules for block (two-way) design experimental data. These selection procedures are based on scores, *i.e.*, functions of the rank values of the data. Subsequently, there have been many applications of these procedures based on the raw ranks of the data: McDonald [5]; Lorenzen and McDonald [6]; Green, *et al.* [7]; Green and McDonald [8]; McDonald [9]; Wang and McDonald [10]. Gupta and Panchapakesan [11] provide a thorough review of the class of parametric and nonparametric ranking and selection procedures.

The purpose of this article is to explore the effect of applying a scoring function of ranks, rather than the raw ranks, to the data and subsequently applying the selection procedure. Specifically, how is the selected subset of populations affected in terms of the size and the content? This will be done with two specific data sets used in earlier publications. The foundations of the subset selection rules, taken from McDonald [3], are described next.

Let $\pi_1, \cdots, \pi_k$ be $k\,(\geq 2)$ independent populations. Let X_{ij}, $j = 1, \cdots, n$; $i = 1, \cdots, k$ be independent samples of size n from the k populations. Assume the random variables X_{ij} have a continuous cumulative distribution function (CDF) $F_i(x;\theta_i)$, where θ_i's belong to some interval Θ on the real line. Suppose $F_i(x;\theta)$ is a stochastically increasing family of distributions in θ; *i.e.*, if $\theta_1 < \theta_2$, then $F_i(x;\theta_1)$ and $F_i(x;\theta_2)$ are distinct and $F_i(x;\theta_2) \leq F_i(x;\theta_1)$ for all x. Examples of such families of distributions are: (1) any location parameter family, *i.e.*, $F_i(x;\theta) = F_i(x-\theta)$; (2) any scale parameter family, *i.e.*, $F_i(x;\theta) = F_i(x/\theta)$, $\theta > 0$, $x > 0$; any family of distribution functions whose densities possess the monotone likelihood ratio property. Figure 1 illustrates that the normal distribution as a location family with respect

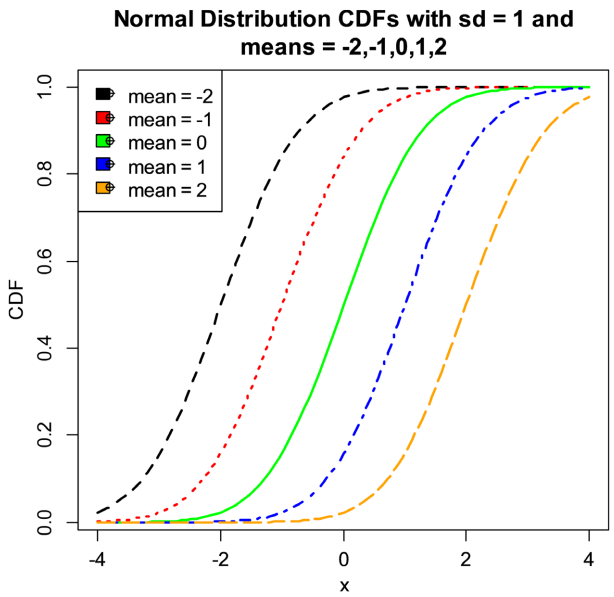

Figure 1. Illustration of stochastic ordering for the normal CDF.

to the mean parameter is stochastically ordered. Note that the CDFs are stacked from top to bottom in inverse order of the mean values.

Let R_{ij} denote the rank of the observation x_{ij} among $x_{1j}, x_{2j}, \cdots, x_{kj}$; *i.e.*, if there are exactly r of the observations $x_{mj}, m = 1, \cdots, k$ 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 $Z(1) \le Z(2) \le \cdots \le Z(k)$ denote an ordered sample of size k from any continuous distribution G, such that $-\infty < a(r) \equiv E[Z(r) \mid G] < \infty$, $r = 1, \cdots, k$. With each of the random variables X_{ij} associate the number $a(R_{ij})$ and define

$$H_i = \sum_{j=1}^{n} a(R_{ij}), \; i = 1, \cdots, k. \tag{1.1}$$

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 $\theta_{[i]}$ denote the i^{th} smallest unknown parameter, it follows that

$$F_j(x;\theta_{[1]}) \ge F_j(x;\theta_{[2]}) \ge \cdots \ge F_j(x;\theta_{[k]}), \; \forall x. \tag{1.2}$$

The population whose associated random variables have the distribution $F_j(x;\theta_{[k]})$ is called the "best" population. In case several populations possess the largest parameter value $\theta_{[k]}$, one of these is tagged at random and called the best. A "Correct Selection" (CS) is said to occur if and only if the best population is included in the selected subset. 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. Formally, for a given selection rule R,

$$\inf_{\Omega} P(CS \mid R) \ge P^*, \tag{1.3}$$

where

$$\Omega = \{\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k) : \theta_i \in \Theta, i = 1, \cdots, k\}. \tag{1.4}$$

The choice of P^* is specified by the analyst and represents the confidence level that the resultant selected subset will contain the best population. The number of populations in the selected subset is a nondecreasing function of P^*.

In a similar fashion, the "worst" population can be defined as that population characterized by the probability distribution $F_j(x;\theta[1])$. Selection procedures can analogously be defined with P^* requirements on the selected subset to contain the worst population as noted in the following Section 2. The assignment of "best" and "worst" is problem specific as will be noted in the applications to follow.

2. Nonparametric (Distribution-Free) Subset Selection Procedures

Four subset selection rules are considered for the analysis of state motor vehicle traffic fatality rates (MVTFRs) as given in McDonald [9]. In this application to be described in Section 3, the populations are states and the blocks are years.

Since low (high) fatality rates are good (bad), the "best" ("worst") state is the one with the smallest (largest) mean fatality rate.

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

R_1: Select π_i iff $H_i \geq \max\left(H_j, j=1,\cdots,k\right) - b_1$.

R_2: Select π_i iff $H_i > b_2$.

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

R_3: Select π_i iff $H_i \leq \min\left(H_j, j=1,\cdots,k\right) + b_3$.

R_4: Select π_i iff $H_i < b_4$.

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^* goal. In cases considered here, these constants are calculated assuming the population parameters are equal and, thus, the distribution of the H statistics are distribution free. As derived in McDonald [3] rules R_1 and R_3 are justified over a slippage space, Ω', where all parameters θ_i are equal with the possible exception of $\theta_{[k]}$ ($\theta[1]$) in case of rule R_1 (R_3); and rules R_2 and R_4 are applicable over the entire parameter space, Ω. That is, the probability of a correct selection will be no less than P^*. 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.

2.1. Calculation of Selection Rules 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 $a(r) \equiv E\left[Z(r)\right] = r/(n+1)$. The selection procedures can then be stated in terms of ranks and rank sums. Thus,

$$R_1\text{: Select } \pi_i \text{ iff } T_i \geq \max\left(T_j, j=1,\cdots,k\right) - b_1 \tag{2.1}$$

$$R_2\text{: Select } \pi_i \text{ iff } T_i > b_2. \tag{2.2}$$

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

$$R_3\text{: Select } \pi_i \text{ iff } T_i \leq \min\left(T_j, j=1,\cdots,k\right) + b_3 \tag{2.3}$$

$$R_4\text{: Select } \pi_i \text{ iff } T_i < b_4, \tag{2.4}$$

where $T_i = \sum_{j=1}^{n} R_{ij}$, $i = 1,\cdots,k$.

The calculation of these constants is treated in McDonald [4] [9] [12] for both small and large samples. For the purposes of this article, the asymptotic values are used. The value of b_1 to meet the P^* requirement is the solution to

$$\int_{-\infty}^{\infty}\left[\phi\left(x + cb_1\right)\right]^{k-1}\varphi(x)\,\mathrm{d}x = P^*, \tag{2.5}$$

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

$$c = c(n,k) = \left[12/nk(k+1)\right]^{1/2}. \tag{2.6}$$

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

$$b_2 = \left[n\left(k^2 - 1\right)/12 \right]^{1/2} \phi^{-1}\left(1 - P^*\right) + n\left(k+1\right)/2, \tag{2.7}$$

where $\phi^{-1}(\cdot)$ is the inverse standard normal CDF, and

$$b_4 = n\left(k+1\right) - b_2. \tag{2.8}$$

The selection rules defined in (2.1) through (2.4) are based on rank sums. These arise from the expected values of the order statistics from a standard uniform distribution. This article addresses the question of how the choice of the distribution of the order statistics affects the performance of the subset selection procedures by choosing an alternate distribution for G as in the next section. Conover [1] compares the nonparametric Kruskal-Wallis test based on rank scores to that based on normal scores. He concludes that the asymptotic relative efficiency may be greater or less than one depending on the particular situation. This further motivates such assessments of performance characteristics for nonparametric subset selection procedures.

2.2. Calculation of Selection Rules Constants, *G* = Standard Normal Distribution

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 than 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.

With the choice of G to be the normal distribution with mean = 0 and standard deviation = 1, the score of X_{ij}, call it $a(R_{ij})$, is the expected value of the i^{th} order statistic drawn from a sample of size k from the standard normal distribution. Extensive tabulations (to 5 dp) of expected values of normal order statistics are given by Harter [13] for sample sizes k = 2(1)100(25)250(50)400. Birnbaum and Dudman [14] also provide tabulations of these expected values along with corresponding calculations from the logistic distribution. The selection procedures can then be stated in terms of scores and score sums. Thus,

$$\mathrm{Q}_1\text{: Select } \pi_i \text{ iff } \; S_i \geq \max\left(S_j,\, j = 1,\cdots,k\right) - d_1 \tag{2.9}$$

$$\mathrm{Q}_2\text{: Select } \pi_i \text{ iff } \; S_i > d_2. \tag{2.10}$$

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

$$\mathrm{Q}_3\text{: Select } \pi_i \text{ iff } \; S_i \leq \min\left(S_j,\, j = 1,\cdots,k\right) + d_3 \tag{2.11}$$

$$Q_4\text{: Select } \pi_i \text{ iff } \; S_i < d_4, \tag{2.12}$$

where $S_i = \sum_{j=1}^{n} a\left(R_{ij}\right)$, $i = 1, \cdots, k$.

The calculation of the constants d_1, …, d_4 follow the same lines of derivation as for the respective constants used with the uniform distribution order statistics in Section 2.1. The asymptotic value of the value of d_1 to meet the P^* requirement is the solution to

$$\int_{-\infty}^{\infty} \left[\phi\left(x + hd_1\right)\right]^{k-1} \varphi\left(x\right) \mathrm{d}x = P^*, \tag{2.13}$$

where

$$h = h\left(n,k\right) = \left[\left(k-1\right)/\left(n \cdot ssq\right)\right]^{1/2}, \tag{2.14}$$

and $ssq = \sum_{i=1}^{k} \left[a\left(R_{ij}\right)\right]^2$. The value of $d_3 = d_1$. The value of d_2 and d_4 are given by

$$d_2 = \left[n \cdot ssq/k\right]^{1/2} \phi^{-1}\left(1 - P^*\right), \tag{2.15}$$

and

$$d_4 = -d_2. \tag{2.16}$$

3. Description of State Motor Vehicle Traffic Fatality Rates (MVTFRs)

The state MVTFRs per year analyzed in McDonald [9] are used here to illustrate the impact that the two rank scoring rules described in Section 2 have on the selected subsets using selection procedures R_1, …, R_4 and Q_1, …, Q_4. The data are given in **Appendix A** (to 2 dp) of the cited reference and contained in the R-code of **Appendix A** of this article. The fatality rates are given for 51 states (taking the District of Columbia as a state) for the years 1994, …, 2012. The two letter abbreviation for states is given as the variable "State" and the fatality rates for the respective years are given in the variables "y1994", …, "y2012" in the order of the states specified in "State". Thus $k = 51$ populations (states) and $n = 19$ blocks (years) comprise the data set. The National Highway Traffic Safety Administration (NHTSA) publishes the MVTFRs for all U.S. states each year in the Fatality Analysis Reporting System (FARS). The data can be accessed through the government website: www-fars.nhtsa.dot.gov. The fatality rate per year for each state is expressed as the number of fatalities per 100 million vehicle miles of travel (VMT).

The cited [9] reference notes the possibility of interaction between the populations and blocks based on the Tukey [15] one degree-of-freedom test. However, raising the fatality rates to the power 0.3 indicates no significant evidence of interaction, and use of a two-way additive model for the transformed rates is plausible. That is,

$$X_{ij}^{0.3} = \mu + \theta_i + \beta_j + \epsilon_{ij}, \tag{3.1}$$

where θ_i indicates the particular state effect, β_j indicates the year effect, and ϵ_{ij} the random error. The distribution of the transformed MVTFRs will be stochas-

tically ordered in θ as it is a location parameter. 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 reference provides a more detailed discussion of the data and the form of the assumed additive model.

4. Applications to the MVTFRs Data

To apply the selection rules to the MVTFRs data set, the constants b_1, …, b_4 and d_1, …. , d_4 need to be obtained. The values for b_1 and d_1 are based on determining the values of $c \cdot b_1$ and $h \cdot d_1$, based on (2.5), (2.6) and (2.13), (2.14) for given values of k, n, and P^*. These two products are equal and the constants b_1 and d_1 are obtained by dividing the product by c and h respectively. The common value of $c \cdot b_1$ and $h \cdot d_1$, call it w, is easily obtained by noting that the integral expressing in (2.5) is an increasing function of w and using a R-code such as

```
w<-3.5
fucn<-function(x){(pnorm(x+w))^50*dnorm(x)}
integrate(fucn,lower = -Inf,upper = Inf)
```

and successive interval halving to converge on $w = 3.666$ for $k = 51$, $n = 19$, and $P^* = 0.90$. The resultant constants for implementing the eight subset selection procedures are given in Table 1 (to 2 dp).

Execution of the R-code in **Appendix A** yields the state rank sums and the state normal score sums given in Table 2. 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. The "averaging" option was used in McDonald [9] and so there are slight differences between results given in the **Appendix B** of that reference and Table 2 given here.

With Table 1 and Table 2, the selection rules given in Sections 2.1 and 2.2 can be applied to the state MVTFRs specified in **Appendix A**. Using $P^* = 0.90$, selection rule R_1 can be now stated as

$$R_1: \text{Select } \pi_i \text{ iff } T_i \geq \max\left(T_j, j = 1, \cdots, k\right) - b_1 = 930 - 237.53 = 692.47, \quad (4.1)$$

and 16 states are thus included in the chosen subset. Using R_2, all states with

Table 1. 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

Table 2. Rank Sums and Normal Score Sums for MVTFR data, $k = 51$, $n = 19$.

State	Rank Sum	State	Rank Sum	State	NS Sum	State	NS Sum
MA	23	PA	488	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
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		

Table 3. Number of states chosen by selection rules with $P^* = 0.90$, $k = 51$, and $n = 19$.

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

rank sums exceeding 411.77 are selected yielding a subset containing 30 states. Following the two examples just given, Table 3 provides the number of selected states in the subsets chosen by the four rules using rank sums and the four rules using normal scores.

Clearly 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 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 Q_2 and R2 (Q_4 and R_4) are within one of each other (are equal).

The correlation between the rank values and the normal score values is 0.983. The R-code of **Appendix B** produces **Figure 2**. The left displays the normal scores vs. the rank scores along with the least squares regression line (Reg Line). The linear fit looks quite good with the exception of the two or three end points on both sides. A notable difference in the rank values and the normal score values is the spacing between successive values. The spacing between any two successive values of the uniform order statistics is $1/(k+1)$, and so the difference in rank values is one, a constant. For the normal scores the spacing for extreme values is much larger than the other spacings. **Figure 2** (right) displays the differences between the successive expected values of the order statistics from the standard normal distribution for $k=51$, *i.e.*, $\text{diff}[x]=a(x+1)-a(x)$, $x=1,\cdots,50$ (see **Appendix B** for the R-code). For example, for $x=1$, $\text{diff}[1]=a(2)-a(1)=0.39307$, the maximum spacing value shared with diff[50]. The minimum spacing value is diff[25] = diff[26] = 0.04896. Thus the more extreme values carry a substantially larger differential weight than the more moderate values, and the spacings are symmetric about the point 25.5 as indicated by the vertical line in the right panel.

5. Applications to the State Homicide Rates (SHRs) Data

The data set, as analyzed by Wang and McDonald [10] is given in the R-code of

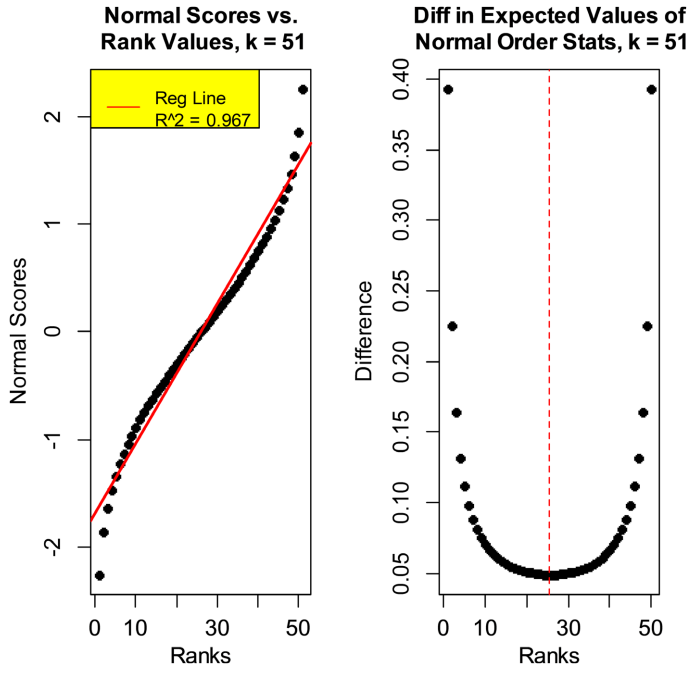

Figure 2. Comparison of normal scores and ranks for $k=51$.

Appendix C. It consists of state homicide rates (*i.e.*, homicides per 100,000 residents) for the years 2005, 2014-2020. An indicated rate of 0.00 is not actually zero since only 2 decimal points were retained for the data. The SHRs are obtained from the Center for Disease Control and Prevention (CDC) at https://www.cdc.gov/nchs/pressroom/sosmap/homicide_mortality/homicide.htm. The CDC website does not contain data for the years 2006 through 2013.

The 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. Since this transformation is monotone, analysis can be done directly with the rates without the power transformation as is done in Section 4. The rank sums and the normal score sums, calculated with the R-code of **Appendix C**, are given in Table 4.

Table 4. Rank sums and normal score sums for SHR data, $k = 50$, $n = 8$.

State	Rank Sum	State	Rank 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
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
RI	62	OH	262	MA	−8.66497	OH	3.00196
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
CA	202	LA	398	CA	−0.06694	LA	17.20416

Applying the selection rules to the SHRs data set, the constants b_1, …, b_4 and d_1, …, d_4 need to be obtained as in Section 4 and given in Table 5 (to 2 dp). 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 Section 5 since n is relatively small. For comparison, the asymptotic values of the selection constants are given in the last row of Table 5 in italics and are seen to be quite close to the simulated values. 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.

Table 6 provides the number of selected states in the subsets chosen by the four rules using rank sums and the four rules using normal scores. As noted for the MVTFRs (Table 3), clearly 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.

6. Summary and Conclusions

As observed here, 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 [6]). In general for the rank sums, $n(k+1)/2 \le \max(T_j, j = 1, \cdots, k) \le n \cdot k$, so for $k = 51$ and $n = 19$, $494 \le \max(T_j) \le 969$. For $P^* = 0.90$, $b_1 = 237.53$, so $256.47 \le \max(T_j) - b_1 \le 731.47$. With $b_1 = 237.53$ and $\max(T_j) = 930$, then rule R_1 selects all states such that $T_i \ge 692.47$. For rule R_2 the determination of b_2 as seen in (2.7) does not depend on the ranks. It depends only on k, n, and P^*. So here $b_2 = 411.77$ and thus R_2 chooses all states

Table 5. 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
$b_1 = 148$	$b_2 = 151$	$b_3 = 148$	$b_4 = 257$	$d_1 = 10.13$	$d_2 = -3.54$	$d_3 = 10.13$	$d_4 = 3.54$
150.85	*151.7*	*150.85*	*256.3*	*10.18*	*−3.53*	*10.18*	*3.53*

Table 6. Number of states chosen by selection rules with $P^* = 0.90$, $k = 50$, and $n = 8$.

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

such that $T_i > 411.77$. So which rule places more populations 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.

With the traffic fatality rates considered here, rule Q_1 placed seven states in the selected subset and rule R_1 placed sixteen states in the selected subset. So which of these two rules to use in practice? From Figure 2, 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 3 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.

The same statements would apply to the choice between Q_3 and R_3. Clearly there is more work to be done in this area of statistical inference. This article 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. Is it possible for R_1 to place fewer populations in the selected subset than rule Q_1?

While the examples given in Sections 4 and 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

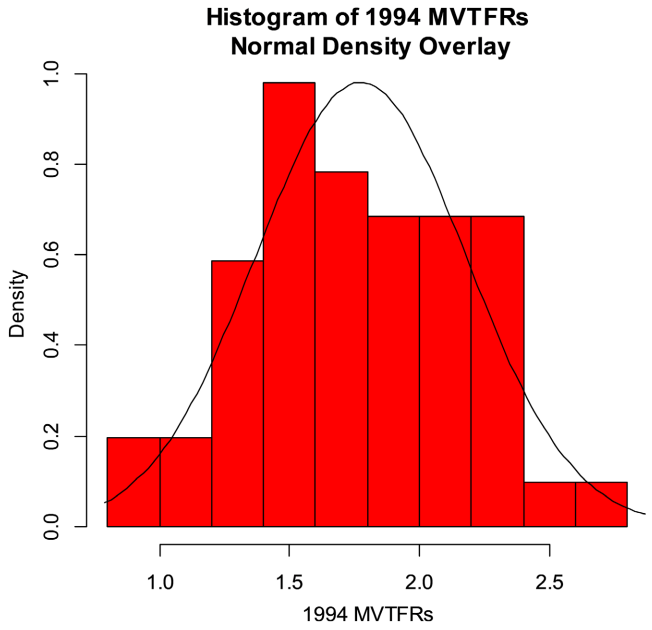

Figure 3. Distribution of 1994 motor vehicle traffic fatality rates for states.

Table 7. Ranked data for $k = 7$, $n = 2$, and selected populations noted in red, $P^* = 0.75$.

Population	π_1	π_2	π_3	π_4	π_5	π_6	π_7
Block 1 ranks	1	2	3	4	5	6	7
Block 2 ranks	1	3	2	7	5	6	4
T_i	2	5	5	11	10	12	11
S_i	−2.70436	−1.11008	−1.11008	1.35218	0.70542	1.51474	1.35218

Table 8. The Number of Configurations (N) for Which the Number of Populations Chosen by R_1 Less 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

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 7, then using the selection rules given in (2.1) and (2.9) rule R_1 chooses four populations and rule, Q_1 chooses six populations.

With $k = 7$, there are seven factorial (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 8. Negative Δ-values indicate that subset selections using R_1 result in fewer chosen population that does that using Q_1. The specific configuration given in Table 7 is one of the 40,320 given in Table 8 under $\Delta = -2$. Of the total number of possible configurations, 705,600 (or approximately 2.8 percent) yield smaller subset sizes chosen by R_1 compared to that of Q_1.

Given the findings in this article, what should be done in practice? 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, for a fixed value of P*. Then use the results that yield the smaller subset size for the given probability of correct selection criteria.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Conover, W.J. (1999) Practical Nonparametric Statistic. 3rd Edition, John Wiley & Sons, Inc., New York.

[2] LaVange, L.M. and Koch, G.G. (2006) Rank Score Tests. *Circulation*, **114**, 2528-2533.

https://doi.org/10.1161/CIRCULATIONAHA.106.613638

[3] McDonald, G.C. (1972) Some Multiple Comparison Selection Procedures Based on Ranks. *Sankhya*: *The Indian Journal of Statistics Series A*, **34**, 53-64.

[4] McDonald, G.C. (1973) The Distribution of Some Rank Statistics with Applications in Block Design Selection Problems. *Sankhya*: *The Indian Journal of Statistics Series A*, **35**, 187-204.

[5] McDonald, G.C. (1979) Nonparametric Selection Procedures Applied to State Traffic Fatality Rates. *Technometrics*, **21**, 515-523. https://doi.org/10.1080/00401706.1979.10489822

[6] Lorenzen, T.J. and McDonald, G.C. (1984) A Nonparametric Analysis of Urban, Rural, and Interstate Traffic Fatality Rates. In: Santner, T.J. and Tamhane, A.C., Eds., *Design of Experiments-Ranking and Selection*, Marcel Dekker, New York, 143-178.

[7] Green, J., McDonald, G.C. and Rao, N. (2006) Using Selection Procedures to Analyze State Traffic Fatality Rates. *American Journal of Mathematical and Management Sciences*, **26**, 387-416. https://doi.org/10.1080/01966324.2006.10737680

[8] Green, J. and McDonald, G.C. (2009) Nonparametric Subset Selection Procedures: Applications and Properties. *American Journal of Mathematical and Management Sciences*, **29**, 413-436. https://doi.org/10.1080/01966324.2009.10737766

[9] McDonald, G.C. (2016) Applications of Subset Selection Procedures and Bayesian Ranking Methods in Analysis of Traffic Fatality Data. *WIREs Computational Statistics*, **8**, 222-237. https://doi.org/10.1002/wics.1385

[10] Wang, A.Q. and McDonald, G.C. (2022) Analysis of State Homicide Rates Using Statistical Ranking and Selection Procedures. *Applied Mathematics*, **13**, 585-601. https://doi.org/10.4236/am.2022.137037

[11] Gupta, S.S. and Panchapakesan, S. (1979) Multiple Decision Procedures. John Wiley & Sons, Inc. Republished in the Classics in Applied Mathematics Series, No. 44 (2002), Society for Industrial and Applied Mathematics, Philadelphia. https://epubs.siam.org/doi/pdf/10.1137/1.9780898719161.fm

[12] McDonald, G.C. (2021) Computing Probabilities for Rank Statistics Used with Block Design Nonparametric Subset Selection Rules. *American Journal of Mathematical and Management Sciences*, **41**, 38-50. https://doi.org/10.1080/01966324.2021.1910885

[13] Harter, H.L. (1961) Expected Values of Normal Order Statistics. *Biometrika*, **48**, 151-165. https://doi.org/10.1093/biomet/48.1-2.151

[14] Birnbaum, A. and Dudman, J. (1963) Logistic Order Statistics. *Annals of Mathematical Statistics*, **34**, 658-663. https://doi.org/10.1214/aoms/1177704178

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

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","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,
1.75,2.3,1.19,1.17,0.82,1.41,1.19,2.67,1.72,2.4,1.53,1.83,1.05,1.08,1.9,1.13,1.74,1.19,1.29,
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,
1.95,2.09,1.47,1.23,0.86,1.28,1.2,2.43,1.77,2.59,1.64,2.12,1.01,1.1,1.97,1.15,1.7,1.32,1.31,
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,
1.84,1.1,0.99,0.62,0.9,0.74,1.73,1.27,2.01,1.15,1.19,0.85,0.8,1.39,0.87,1.28,1.72,0.92,1.57,
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,y2004,
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<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
    #
    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 (b_1 and b_3)

```
#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 (b_2 and b_4)

```
#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 (d_2 and d_4 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
```