NONPARAMETRIC SELECTION PROCEDURES APPLIED TO STATE TRAFFIC FATALITY RATES

Gary C. McDonald
Mathematics Department
General Motors Research Laboratories

ABSTRACT

This article reviews the practical aspects of several nonparametric subset selection rules useful in block design problems, and discusses advantages and disadvantages of these methods. The populations are assumed stochastically ordered by the parameter of interest. Rules based on ranked observations are given for selecting a subset of populations which contains, with a specified confidence level, the population characterized by the smallest (or largest) parameter value. These procedures are applied to state traffic fatality rates recorded yearly (1960-76). New England states and Middle Atlantic states comprise most of the subset asserted, with a 90% confidence level, to contain the state with the smallest fatality rate; whereas, Southern states, Southwestern states and Rocky Mountain states generally comprise the subset for the state with the largest fatality rate. Note that while this example is not based on simulation data, such data would be analyzed in exactly the same fashion.

I. INTRODUCTION

In this article, the use of several nonparametric subset selection procedures will be discussed and illustrated with a set of traffic fatality data. The procedures are simple to use and robust in the sense that inferences to populations apply under very few model assumptions. The statistical procedures discussed here can be motivated by the following model: each of n independent judges orders k populations according to some specified criterion. That is, each judge assigns a rank of 1 to the population least desirable in his (or her) opinion,..., and a rank of k to that which is most desirable. The "best" population is defined to be that (unknown) one which is in fact most desirable according to this given criterion, and, correspondingly, the "worst" is the one least desirable.

Based on these ranks, several selection procedures for choosing a subset of the k populations so as to guarantee that the best (or worst) is included with a probability no less than P^* ($k^{-1} < P^* < 1$) are discussed and illustrated. This model has wide applicability; e.g., the n judges may be taken as n years on which observations are recorded, or as n replicate simulation runs. In this case the observations recorded for a given run are ranked among themselves. This is the spirit of the example presented in this article -- motor-vehicle traffic fatality rates recorded yearly by state. The subset formulation of decision rules is due to Gupta (1956).

It is possible to pool all the observations and use selection rules based on joint ranks. Several classes of these rules have been developed and studied by Lehmann (1963), Dudewicz (1966), Gupta and McDonald (1970), and Lee and Dudewicz (1974). Also, these methods are discussed in the recent books by Kleijnen (1975) and by Lehmann (1975). Many of the properties and limitations discussed in this article have a direct analog in the joint ranking procedures.

By employing the population-judge model (or block design) rather than using joint ranking methods, relatively large savings in computational time and data storage may be realized. This aspect will be illustrated in the context of a numerical example. As with any block design, the experimenter should have a fixed reasonable interpretation of "judge" before actually proceeding with the data analysis. The model has been investigated by McDonald (1972, 1973) and by Lee and Dudewicz (1974). It has been employed in a multiple comparison context by Kramer (1956); Thompson and Willke (1963); and others. References to the use of this model, as well as several related paired-comparison models, may be found in McDonald (1972).

In Section II of this article several selection rules for choosing either the best or the worst populations are stated formally, along with the model assumptions required to support a statistical inference. Section III provides a guide to existing tables of constants, and associated approximations, which are required to implement the selection rules. In Section IV, these techniques are illustrated with a set of motor-vehicle traffic fatality rates which are indexed by state (the population) and time (the run). Several advantages and disadvantages of these methods are discussed briefly in Section V.

II. FORMULATION OF RULES AND SOME BASIC PROPERTIES

Let $\Pi_1, \ldots, \Pi_k$ be $k (\geq 2)$ independent populations. The associated random variables X_{ij}, $j = 1, \ldots, n$; $i = 1, \ldots, k$, are assumed independent and to have a continuous distribution

$F_j(x;\theta_i)$ where θ_i belongs to some interval Θ on the real line. Our basic model assumption is that $F_j(x;\theta)$ is a stochastically increasing family of distributions for each j; i.e., if θ_1 is less than θ_2, then $F_j(x;\theta_1)$ and $F_j(x;\theta_2)$ are distinct and $F_j(x;\theta_2) \leq F_j(x;\theta_1)$ for all x. This covers, for example, models of the form

$$X_{ij} = \mu + \theta_i + \beta_j + \epsilon_{ij} \quad ,$$

where the error term may have any (not necessarily normal) continuous distribution $G(\cdot)$.

The observations are taken in n blocks which are well specified in advance of the analysis. The subscript j indicates the particular block to which the observation x_{ij} corresponds; the i indicates the population. Let R_{ij} denote the rank of observation x_{ij} among $x_{1j},\ldots, x_{kj}$; i.e., if there are exactly r of the observations x_{mj}, $m = 1,\ldots, k$, less than x_{ij} then $R_{ij} = r + 1$. These ranks are well-defined since $F_j(x;\theta)$ is assumed continuous. The variables R_{ij} take integer values from 1 to k inclusive. Our selection procedures are based on the quantities $T_i \equiv \sum_{j=1}^{n} R_{ij}$, the sum of ranks associated with Π_i, $i = 1,\ldots, k$.

Letting $\theta_{[i]}$ denote the i^{th} smallest unknown parameter and recalling that $F_j(x;\theta)$ is stochastically increasing, we have

$$F_j(x;\theta_{[1]}) \geq F_j(x;\theta_{[2]}) \geq \cdots \geq F_j(x;\theta_{[k]}) \quad ,$$

$$\text{all } x, \; j = 1, \ldots, n \; .$$

To accommodate the application to be discussed in Section IV, the population characterized by $\theta_{[1]}$ will be called the best, and that characterized by $\theta_{[k]}$ called the worst. Several subset selection procedures, based on the rank sums, will be reviewed. These procedures have the property that the probability of a "Correct Selection" (CS), i.e., including the best (or worst) population in the selected subset, is bounded below by a specified value P^* ($k^{-1} < P^* < 1$). Formally, for a given rule R, the probability of a correct selection should satisfy the inequality,

$$\inf_{\Omega} P(CS|R) \geq P^* \quad ,$$

where

$$\Omega = \{\underline{\theta} = (\theta_1,\ldots, \theta_k): \theta_i \in \Theta, \; i = 1,\ldots, k\}.$$

In some instances, to be noted later, this guarantee may hold only on a subspace Ω' of Ω.

CHOOSING THE WORST POPULATION

For choosing a subset to contain the worst population, the following two rules are considered:

$$R_1 : \text{Select } \Pi_i \text{ iff } T_i \geq \max_{1\leq j\leq k} T_j - b$$

and

$$R_2 : \text{Select } \Pi_i \text{ iff } T_i > d \quad .$$

The constants b and d are chosen to yield the basic P^* - condition. The constants b and d are calculated assuming the θ_i's are all equal. In the case of R_1 this restricts the inference space as indicated in the following property:

> If Ω represents a slippage configuration, i.e., $\theta_{[1]} = \cdots = \theta_{[k-1]} \leq \theta_{[k]}$, then the probability of a correct selection using rules R_1 and R_2 is minimized when the k populations are identically distributed. For an unrestricted parameter space Ω, the same is true for rule R_2; however, a similar result does not hold for R_1.

CHOOSING THE BEST POPULATION

Corresponding rules for choosing subsets of the k populations which contain the best, with a specified confidence, are:

$$R_1' : \text{Select } \Pi_i \text{ iff } T_i \leq \min_{1\leq j\leq k} T_j + b'$$

and

$$R_2' : \text{Select } \Pi_i \text{ iff } T_i < d' \quad .$$

In this case, the constants b' and d' are chosen to be as small as possible while preserving the basic probabilistic guarantee.

The properties for these rules are direct analogs for R_1 and R_2, respectively. That is, R_1' is justified over a slippage space where $\theta_{[1]} \leq \theta_{[2]} = \cdots = \theta_{[k]}$; and R_2' is applicable over the entire parameter space.

III. DETERMINATION OF CONSTANTS FOR SELECTION RULES

This section provides a guide to tables useful in determining the constants required to implement the selection rules discussed.

WORST POPULATION RULES

The constant b used in rule R_1 is tabulated in McDonald (1973) for $k = 2$, $n = 2(1)20$; $k = 3$, $n = 2(1)8$; $k = 4$, $n = 2(1)5$; $k = 5$, $n = 2,3$. The exact distribution of the appropriate statistic is given and so any admissible value of P^* can be used in this range of k and n values.

For large values of n, the following asymptotic expression can be used to determine b:

$$\tilde{b} = h[n\,k\,(k+1)/6]^{1/2} \quad ,$$

where h satisfies

$$\int_{-\infty}^{\infty} [\Phi(x+h2^{1/2})]^{k-1}\, \varphi(x)dx = P^* \quad .$$

In the above equation $\Phi(\cdot)$ and $\varphi(\cdot)$ refer to the cumulative distribution function and probability density function, respectively, of a standard normal variate. The h quantity appearing in this expression has been tabulated by Gupta (1963); Gupta, et al. (1973); and Milton (1963).

The constant d for use with R_2 is tabulated in McDonald (1973) for k = 2, n = 2(1)15; k = 3, n = 2(1)10; k = 4, n = 2(1)7; k = 5, n = 2(1)5. As in the previous case, the entire distribution is tabulated and so any value of P* can be used.

For large values of n, an asymptotic approximation for d is given by:

$$\tilde{d} = [n(k^2-1)/12]^{1/2}\, \Phi^{-1}(1-P^*) + n(k+1)/2 \quad ,$$

where $\Phi^{-1}(\cdot)$ denotes the inverse of the standard normal cumulative distribution function. A tabulation of the inverse function is contained in Owen (1962), and an efficient FORTRAN computer program to evaluate numerically the inverse normal function is given by Milton and Hotchkiss (1969).

BEST POPULATION RULES

The constant b' required to implement rule R_1' is identically equal to the corresponding b value required for rule R_1 for given values of k, n and P*. Thus, the tables and asymptotic formula apply directly.

The constand d' required for rule R_2' is given by

$$d' = n(k+1) - d \quad ,$$

where d is the corresponding value needed for rule R_2. Thus, the appropriate asymptotic form is given by

$$\tilde{d}' = n(k+1)/2 - [n(k^2-1)/12]^{1/2}\, \Phi^{-1}(1-P^*) \quad .$$

TABLE 1

Motor-Vehicle Traffic Fatalities per Year per 100,000,000 Vehicle Miles

STATE	YEAR 60	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76
ALABAMA	7.0	7.0	6.9	7.5	7.0	7.4	7.0	7.1	7.0	7.0	6.4	6.8	6.0	6.2	4.1	3.9	4.0
ARIZONA	7.8	7.6	7.0	6.6	7.2	6.6	7.4	6.2	6.9	6.5	6.3	5.7	5.5	6.0	4.8	4.2	4.1
ARKANSAS	6.0	5.8	6.2	6.8	7.4	6.8	7.3	6.6	6.9	5.6	5.4	5.7	6.0	5.0	3.9	4.1	3.7
CALIFORNIA	5.3	5.2	5.3	5.0	5.2	4.9	5.0	4.7	4.6	4.5	4.2	3.8	3.9	3.8	3.1	3.2	3.1
COLORADO	5.4	5.8	5.1	5.7	6.1	5.4	5.8	5.5	6.1	5.3	5.2	4.6	4.6	4.2	3.8	3.6	3.7
CONNECTICUT	2.8	2.8	3.1	2.9	2.6	3.0	3.0	3.2	3.0	2.6	2.7	2.9	2.6	2.8	2.2	2.2	2.2
DELAWARE	4.1	2.7	4.0	4.3	5.1	4.4	5.0	5.4	5.8	4.5	5.1	3.7	3.8	3.7	3.3	3.4	3.2
DIST. OF COL.	2.8	2.2	2.5	3.9	4.3	3.7	3.8	5.0	4.7	4.6	4.3	3.3	2.5	2.5	2.6	2.4	2.1
FLORIDA	5.7	5.4	5.6	6.0	5.9	6.1	6.0	5.4	6.1	5.7	5.2	5.0	4.5	4.5	3.7	3.2	3.0
GEORGIA	6.3	6.0	6.2	6.8	6.5	6.5	7.0	6.7	6.9	6.4	6.2	5.7	4.7	5.3	4.4	3.5	3.4
IDAHO	7.1	7.9	8.0	5.3	6.5	7.0	6.8	6.7	7.1	7.6	6.9	6.6	6.6	6.5	6.0	4.9	4.1
ILLINOIS	4.6	4.7	4.7	4.9	5.1	4.9	5.2	5.0	4.8	4.7	4.2	4.2	3.8	3.9	3.4	3.4	3.0
INDIANA	5.3	5.0	5.5	5.8	5.9	5.8	5.9	5.8	5.5	5.5	4.8	4.7	4.2	4.2	3.4	3.0	3.4
IOWA	5.0	5.1	4.8	5.8	6.7	6.1	6.9	6.1	6.3	4.6	4.9	4.4	4.6	4.1	3.6	3.4	4.0
KANSAS	5.1	5.3	5.4	5.2	5.6	5.4	5.7	5.4	5.2	6.0	4.9	4.9	4.6	4.1	3.4	3.3	3.2
KENTUCKY	7.0	6.5	6.9	6.4	6.6	6.2	7.0	6.2	5.9	5.7	5.4	4.8	4.7	4.6	3.3	3.6	3.3
LOUISIANA	7.4	6.7	6.4	8.3	9.5	9.0	8.9	8.3	8.5	7.2	7.1	6.5	6.2	6.0	4.4	4.6	4.4
MAINE	4.3	4.1	4.1	4.5	4.2	4.8	4.6	5.0	4.2	4.5	4.5	4.2	3.8	3.6	3.3	3.3	2.8
MARYLAND	4.5	3.9	4.7	4.3	4.2	4.5	4.4	4.5	4.6	4.1	3.8	3.6	3.7	3.2	3.1	2.7	2.6
MASSACHUSETTS	3.2	3.1	3.7	3.5	3.7	3.9	4.1	4.0	3.9	3.5	3.5	3.2	3.3	3.4	3.4	3.0	2.8
MICHIGAN	5.0	4.9	4.7	5.2	5.5	5.3	5.3	4.7	5.0	4.9	4.1	3.9	3.9	3.8	3.4	3.1	3.3
MINNESOTA	4.9	4.8	4.5	5.1	5.1	5.1	5.5	5.2	5.3	4.8	4.4	4.4	4.1	4.1	3.5	3.0	3.0
MISSISSIPPI	7.8	6.5	6.9	8.2	7.6	7.4	9.0	9.2	7.3	7.3	7.7	7.8	7.0	6.4	4.7	4.3	4.9
MISSOURI	5.3	4.6	5.0	5.8	6.1	6.1	6.0	5.6	6.0	6.0	5.7	5.2	5.0	4.7	3.5	3.5	3.8
MONTANA	6.6	8.1	6.8	6.1	6.8	7.6	7.1	7.8	7.0	7.6	6.5	6.5	7.3	5.8	5.1	5.2	4.6
NEBRASKA	4.3	4.7	5.6	4.7	6.0	4.8	5.1	5.1	5.0	4.6	4.3	4.9	4.5	3.9	3.5	3.4	3.3
NEVADA	8.3	9.5	6.6	8.6	8.9	7.8	7.3	5.9	7.3	7.5	7.4	7.4	6.7	6.4	5.1	4.9	4.5
NEW HAMPSHIRE	4.2	4.0	4.3	4.4	4.5	4.0	4.1	4.4	4.8	4.7	4.4	4.4	3.5	2.8	3.3	2.9	2.9
NEW JERSEY	2.9	3.0	3.1	3.2	3.4	3.4	3.3	3.3	3.6	3.3	3.2	3.0	2.8	2.8	2.4	2.2	2.1
NEW MEXICO	7.8	6.7	7.9	6.7	6.9	8.2	7.2	7.4	7.4	8.0	7.6	6.7	6.6	6.7	5.7	5.7	5.3
NEW YORK	4.4	4.6	4.7	4.5	5.1	4.8	4.8	4.8	4.9	4.9	4.5	4.4	4.3	4.6	4.0	3.8	3.4
NORTH CAROLINA	6.8	6.5	6.7	7.0	7.6	7.3	7.4	7.1	7.2	6.5	6.0	5.9	5.8	5.3	4.5	4.2	3.9
NORTH DAKOTA	6.1	5.7	6.4	5.7	5.2	4.5	6.0	5.9	5.6	5.0	4.5	5.7	5.1	4.8	3.7	3.8	3.3
OHIO	4.9	4.2	4.5	4.9	4.9	5.1	5.3	5.1	4.8	5.0	4.6	3.9	3.9	3.7	3.0	2.8	2.8
OKLAHOMA	5.9	6.1	5.8	6.1	6.0	5.0	5.5	5.8	5.0	5.3	4.7	4.5	4.1	3.7	3.5	3.4	3.6
OREGON	5.6	5.7	5.4	6.0	5.8	6.6	6.2	5.7	5.4	5.6	5.1	4.8	4.8	4.0	4.4	3.6	3.8
PENNSYLVANIA	4.0	3.7	4.0	4.1	4.1	4.4	4.3	4.5	4.2	4.4	4.0	3.8	3.5	3.7	3.2	3.3	2.8
RHODE ISLAND	1.9	2.1	2.7	2.7	2.8	2.5	2.7	2.5	3.3	3.0	3.0	2.5	2.5	2.4	1.8	1.9	2.1
SOUTH CAROLINA	7.8	7.9	7.8	7.7	8.0	7.3	7.9	7.0	7.0	6.4	6.2	5.8	5.6	4.7	4.4	4.0	3.7
SOUTH DAKOTA	7.4	7.0	7.2	5.9	7.2	6.2	6.5	5.3	5.9	6.8	5.1	5.4	5.8	5.6	4.5	3.9	4.3
TENNESSEE	5.6	4.8	4.9	6.4	6.9	6.8	7.4	7.0	6.4	7.1	6.6	5.5	5.1	4.9	4.1	3.5	3.2
TEXAS	4.9	4.8	4.8	5.8	6.0	5.7	6.2	5.8	5.6	5.4	5.2	5.1	4.8	4.6	3.9	4.1	3.7
UTAH	6.4	5.7	5.3	5.8	6.2	5.7	6.4	5.2	5.2	5.3	5.5	5.1	5.7	5.0	3.1	3.5	3.1
VERMONT	6.3	5.1	6.1	6.1	6.8	7.4	6.0	5.8	5.8	5.8	4.6	5.0	4.7	4.7	4.2	4.3	3.4
VIRGINIA	4.7	5.3	5.7	5.3	5.4	5.2	5.1	5.4	5.0	4.8	4.3	4.0	3.8	3.5	3.1	3.0	2.8
WASHINGTON	4.8	4.8	4.8	4.5	5.0	4.7	5.0	4.9	4.9	4.2	4.2	4.0	3.8	3.3	3.4	3.2	3.3
WEST VIRGINIA	6.0	6.0	6.7	6.6	6.9	6.7	6.9	7.5	6.3	6.2	6.6	5.8	5.3	4.7	4.4	4.6	4.1
WISCONSIN	6.0	5.8	5.9	5.2	5.7	5.4	5.6	5.5	5.3	4.8	4.6	4.4	4.2	4.0	3.3	3.3	3.2
WYOMING	7.4	6.4	5.4	7.4	5.8	7.1	5.8	5.5	5.9	7.6	6.4	5.2	5.8	5.6	5.6	5.8	6.4

IV. AN ANALYSIS OF FATALITY RATES

Each year the National Safety Council publishes the motor-vehicle traffic fatality rate (MFR) for each state in the annual editions of Accident Facts. The MFR is the number of motor-vehicle traffic fatalities per 100,000,000 vehicle miles. Basically, we are considering fatalities which occur within one year as a result of an accident involving a motor-vehicle on a trafficway. The death is attributed to the place (state) of the accident. The exact definitions of all the terms used in this context are given in the Accident Facts. The fatality rates are given in Table 1 (to 1 dp) for the contiguous forty-eight states and the District of Columbia for the years 1960 to 1976 inclusive -- so k = 49 and n = 17. The 1960-75 rates were obtained from Accident Facts, annual editions, National Safety Council, Chicago, Ill. The rates for a given year from 1960-74, say t, were obtained from the (t+2) annual edition and are considered "final." The rates for 1975-76 are preliminary estimates. The 1976 estimates, and rates not listed in annual editions, were obtained directly from the National Safety Council.

Our goal is to choose a subset of the 49 states (considering D.C. a state) which can be asserted, with a specified confidence, to contain the state with largest fatality rate (worst population). Likewise, a subset will be chosen for the smallest fatality rate (best population). This objective is often viewed as a screening technique to reduce the number of states to a relatively small number for further study or characterization. For example, a successful characterization contrast of the subset of states chosen for the largest fatality rate with the subset chosen for the smallest may suggest causal factors which play a significant role in determining the MFR for a particular state. I am sure that such scrutiny of the fatality rates has been undertaken carefully in the past. However, the statistical techniques described here provide additional analytic tools to assist this empirical process. Before applying the selection procedures of Section II, the model assumptions will be reviewed and checked briefly.

Let X_{ij} denote the MFR for the i^{th} state and j^{th} year, $i=1,\dots, 49$; $j=1,\dots, 17$. The index i denotes the state in alphabetic order, and the index j denotes the year in increasing order. For example, $X_{11} = 7.0$ and is the MFR for Alabama in 1960; $X_{53} = 5.1$ and is the MFR for Colorado in 1962. The assumptions of Section II are assumed to hold. Simply put, the X_{ij} are assumed independent having a continuous distribution (not necessarily normal), $F_j(x;\theta_i)$, which is stochastically increasing in θ for each fixed j. The θ_i is taken to be the effect of the i^{th} state on the fatality rate. This model applies, for example, in the context of a block design in the absence of an interaction term. Table 2 provides an analysis of variance for the data in Table 1. Both indices, state and year, are highly significant in explaining the variation in the MFR's.

Tukey (1949) developed a test for non-additivity when there is a single observation per cell, as given here. This test on the original data, summarized in Table 2, strongly indicates that an interaction term cannot be deleted from the basic model. However, Tukey also pointed out the influence of the scale of measurement upon the existence or nonexistence of interaction effects. An appropriate choice of a monotonic transformation applied to the data may yield the subsequent Tukey test insignificant suggesting that the interaction is an artifact of the scale of measurement (see Winer (1971)).

The large non-additivity mean square could result from one or more discrepant data points

TABLE 2

ANOVA and Tukey's Test for Traffic Fatality Rate Data

ANALYSIS OF VARIANCE

SOURCE	SUM OF SQUARES	DF	MEAN SQUARES	F-VALUE	PROB(F)
MEAN	21426.276	1			
YEARS	496.833	16	31.052	114.258	0.000
STATES	1053.980	48	21.957	80.795	0.000
RESIDUAL	208.720	768	0.271		
TOTAL	23185.810	833			

TUKEY'S TEST FOR NON-ADDITIVITY

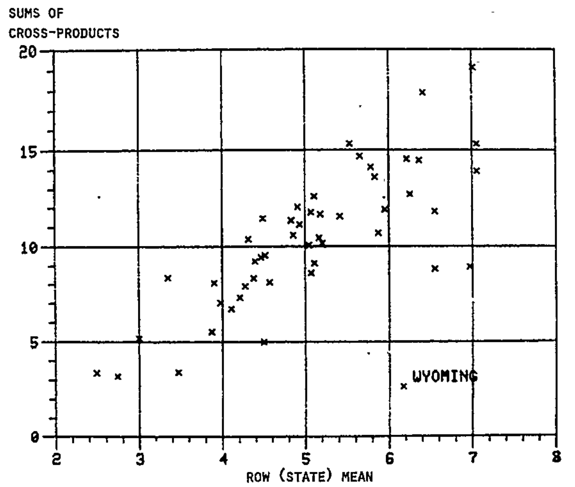
SOURCE	SUM OF SQUARES	DF	MEAN SQUARES	F-VALUE	PROB(F)
RESIDUAL	208.720	768	0.271		
NON-ADDITVITY	31.871	1	31.871	138.229	0.000
BALANCE	176.848	767	0.230		

and/or from analysis in an inappropriate scale of measurement. Tukey suggests a plot of the sums of cross-products versus row means, i.e., $\sum_{j=1}^{17} X_{ij}(\bar{X}_{\cdot j}-\bar{X}_{\cdot\cdot})$ vs. $\bar{X}_{i\cdot}$, as a tool to examine the influence of these factors. A single discrepant observation tends to be reflected by one point high or low and the others distributed around a nearly horizontal regression line. An analysis in an inappropriate scale tends to be reflected by a slanting regression line.

The cross-product plot for the fatality rate data is given in Figure 1. A slanting regression

FIGURE 1

Sums of Cross-products for Traffic Fatality Rates



line is readily apparent, as is one low point corresponding to the state of Wyoming. This low point is due primarily to the increasing large fatality rates for Wyoming over the last several years. Wyoming is the only entry for which the fatality rate has been nondecreasing since 1973. The strong slanting nature of the plot strongly suggests a change of scale.

Table 3 provides an analysis of variance and Tukey's test for non-additivity for the transformed data $Y_{ij} = \ell n(X_{ij}-1)$. This table indicates that factors, state and year, remain significant while the interaction term is insignificant. Thus, a block design with no interaction does not appear unreasonable. Since the rank procedures are invariant to order preserving transformations, the analysis can proceed on the original data. It should be noted that the checks appearing in Tables 2 and 3 do assume a normal distribution for the error terms to substantiate the F-tests; however, the rank procedures to be applied do not impose such an assumption.

Since the values of k (=49) and n (=17) are large for this application, the constants required to implement the selection rules are determined by the asymptotic formulae given in Section III. Taking P* = .90, the h-solution to

$$\int_{-\infty}^{\infty} [\Phi(x+h2^{1/2})]^{48} \varphi(x)dx = .90 \quad ,$$

as given in Table 1 of Gupta et al. (1973), is h = 2.5816. Thus,

$$\tilde{b} = (2.5816)\ [17(49)(50)/6]^{1/2} \doteq 215.09 \quad .$$

Also, $\Phi^{-1}(1-P^*) = \Phi^{-1}(.1) = -1.28155$, as given in Owen (1962), and so

$$\tilde{d} = [17(49^2-1)/12]^{1/2}\ (-1.28155) + 17(50)/2 \doteq 350.27 \quad .$$

TABLE 3

ANOVA and Tukey's Test for Transformed Traffic Fatality Rate Data

ANALYSIS OF VARIANCE

SOURCE	SUM OF SQUARES	DF	MEAN SQUARES	F-VALUE	PROB(F)
MEAN	1477.5496	1			
YEARS	39.0298	16	2.4393	161.0992	0.000
STATES	81.5632	48	1.6992	112.2200	0.000
RESIDUAL	11.6290	768	0.0151		
TOTAL	1609.7717	833			

TUKEY'S TEST FOR NON-ADDITIVITY

SOURCE	SUM OF SQUARES	DF	MEAN SQUARES	F-VALUE	PROB(F)
RESIDUAL	11.6290	768	0.0151		
NON-ADDITVITY	0.0004	1	0.0004	0.0281	0.861
BALANCE	11.6286	767	0.0151		

The remaining needed value is simply

$$\tilde{d}' = 17(50) - \tilde{d} \doteq 499.73 \quad .$$

Now, the two selection rules for choosing a subset containing the worst population, i.e., the state with the largest fatality rate, are given by:

R_1 : Select the $i^{\underline{th}}$ state iff $T_i \geq \max_{1\leq j\leq 49} T_j - 215.09$,

and

R_2 : Select the $i^{\underline{th}}$ state iff $T_i > 350.27$.

The corresponding rules for selecting the subset for the best population, the state with the smallest fatality rate, are

R_1' : Select the $i^{\underline{th}}$ state iff $T_i \leq \min_{1\leq j\leq 49} T_j + 215.09$,

and

R_2' : Select the $i^{\underline{th}}$ state iff $T_i < 499.73$.

To apply these selection procedures, the state MFR's, for each year, must be ranked. The state with the lowest rate receives a rank of 1,..., and the state with the highest rate receives a rank of 49. For the year 1960, Rhode Island has rank 1 and Nevada has rank 49; however, a difficulty arises in assigning the rank 2. Connecticut and the District of Columbia both have rates of 2.8 (to 1 dp) and are equal contenders for the ranks of 2 and 3. When ties occur such as this, each of the tied states are assigned the average rank; e.g., both Connecticut and the District of Columbia are assigned a rank of 2.5 for the year 1960. After ranking the states for each year, the rank sum for a state is computed by summing the ranks across years. The rank sum for the $i^{\underline{th}}$ state is denoted by T_i and the collection of rank sums is given in Table 4.

The presence of tied observations results in a difficulty common to many rank techniques. In Section II, the observations were assumed to have a continuous distribution implying that tied observations do not occur and that a complete ranking of the populations is available from each judge. The distribution theory underlying the determination of constants in Section III depends on this implication of no ties. The presence of ties substantially complicates the distribution theory, and exact results are virtually nonexistent. In many instances, this difficulty can be avoided by calculating (or measuring) the observations to additional significant digits so as to eliminate ties. In the motor-vehicle fatality rate example this could be done by using the fatality and mileage data to compute rates to additional decimal places. However, since this example employs the asymptotic results, and the number of observations tied for a specific rank within a given year is small compared to the number of states, the resulting error should be small using the easily obtained rates. The treatment of ties with rank methods is discussed at some length by Hájek (1969) and Lehmann (1975).

TABLE 4

State Rank Sums and Selected Subsets with P* = .90

		Selected Subset for†			
		Best		Worst	
State	Rank Sum	R_1'	R_2'	R_1	R_2
Rhode Island	23.50	*	*		
Connecticut	39.50	*	*		
New Jersey	55.00	*	*		
Dist. of Col.	91.00	*	*		
Massachusetts	103.50	*	*		
Maryland	126.50	*	*		
Pennslyvania	133.00	*	*		
New Hampshire	160.00	*	*		
Maine	185.50	*	*		
Washington	220.00	*	*		
Ohio	224.00	*	*		
Delaware	226.00	*	*		
California	239.50		*		
Illinois	244.50		*		
Virginia	268.00		*		
Michigan	275.50		*		
Minnesota	287.50		*		
New York	303.50		*		
Nebraska	311.00		*		
Wisconsin	368.50		*		*
Kansas	382.50		*		*
Indiana	401.50		*		*
Oklahoma	413.00		*		*
Florida	443.00		*		*
Iowa	444.00		*		*
Utah	452.50		*		*
Colorado	453.00		*		*
N. Dakota	461.00		*		*
Texas	463.00		*		*
Missouri	477.50		*		*
Oregon	484.00		*		*
Vermont	534.00				*
Kentucky	537.50				*
Tennessee	567.00			*	*
Georgia	597.00			*	*
Arkansas	618.00			*	*
S. Dakota	619.00			*	*
Wyoming	635.50			*	*
W. Virginia	643.50			*	*
Arizona	691.50			*	*
N. Carolina	693.00			*	*
S. Carolina	694.50			*	*
Alabama	702.00			*	*
Idaho	713.00			*	*
Montana	733.50			*	*
Louisiana	757.50			*	*
Nevada	772.00			*	*
Mississippi	777.50			*	*
New Mexico	778.50			*	*

†States chosen by selection rule are indicated with *.

The results of applying the selection procedures R_1, R_2, R_1' and R_2' are contained in Table 4. Since $\max_{1\leq j\leq 49} T_j = 778.50$, the rule R_1 selects all those states for which $T_i \geq 778.50 - 215.09 = 563.41$.

There are sixteen such states with a sufficiently large rank sum. The other rules are applied similarly.

The statistical conclusions of this analysis can be summarized by referring to the inference property stated in Section II. The thirty states selected using rule R_2 can be asserted, with 90% confidence, to contain the state which is characterized by the largest fatality rate. The same confidence statement applies to the sixteen states selected using rule R_1 if, in fact, forty-eight of the states have the same fatality rate and one (unknown) state has a rate at least as large as this common value. In order to justify the inference over the unrestricted parameter space of fatality rates, twice as many states are selected, in this example, than are required for an inference over the slippage configuration. Similar inference remarks apply to the subsets selected using rules R_2' and R_1' and asserted to contain the state with the smallest fatality rate.

The subset of sixteen states determined by R_1 to have the worst state contains primarily Southern states, Southwestern states and Rocky Mountain states. On the other hand, the subset of twelve states selected by R_1' to contain the best state consists primarily of New England states and Middle Atlantic states; however, the states of Ohio and Washington are included in this group also.

V. SUMMARY DISCUSSION

The advantages and disadvantages of using a selection procedure based on ranks are parallel to those of any statistical rank method. Primary advantages are computational ease, valid inference based on weak assumptions and the knowledge of an ordinal relationship for the populations (i.e., actual numerical observations are not required). The first of these advantages can be substantial in large scale repetitive computation. For example, to provide for a future analysis of the fatality rate data, all that need be stored is the rank sums, given in Table 4, of forty-eight states. The sum for the remaining state is determined since the rank sums for k populations and n judges must sum to $n\,k(k+1)/2$. When the 1977 fatality rates become available, these can be ranked and the individual state rank sums updated easily.

Disadvantages of these methods include restricted inference, unknown operating characteristics (including optimal choice of score functions) and computationally difficult distributions. The restricted inference is applicable to the selection rules R_1 and R_1'. Little is known about characteristics, such as expected subset size, of these procedures outside of the probability of a correct selection. In our example, the selection rule R_2 selected considerably more states than the rule R_1. How much of this can be explained as a property of the techniques versus a sample phenomenon is not known. Finally, the distribution theory associated with these rank procedures can be quite complicated. Some exact results are available and asymptotic results can be applied easily. However, tied observations result in substantial complications and no results are available currently to handle this problem. When possible, the experimenter should avoid these difficulties by refining the measurements or calculations to avoid ties.

Selection procedures based on ranks are developed sufficiently well to play a useful role in the analysis and interpretation of data. These techniques, however, are in the embryo state within the user community. Increased usage of these methods in data analysis and simulation studies, and the appropriate documentation of the findings, will encourage further research and development of these methods to understand and abate the disadvantages, and to expand the advantages, briefly discussed here.

ACKNOWLEDGMENTS

The author gratefully acknowledges the assistance of Mrs. Leona Henry with the data analysis aspects of this paper.

REFERENCES

1. Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. (25), 16-29.

2. Dudewicz, E. J. (1966). The efficiency of a nonparametric selection procedure: largest location parameter case, Technical Report #14, Dept. of Ind. Eng. and Oper. Res., Cornell University, Ithaca, N. Y.

3. Gupta, S. S. (1956). On a decision rule for a problem in ranking means, Mimeograph Series #150, Institute of Statistics, University of North Carolina, Chapel Hill, N. C.

4. Gupta, S. S. (1963). Probability integrals of multivariate normal and multivariate t. Ann. Math. Statist. (34), 794-828.

5. Gupta, S. S. and McDonald, G. C. (1970). On some classes of selection procedures based on ranks, in Nonparametric Techniques in Statistical Inference, M. L. Puri (ed.), Cambridge University Press, London, 491-514.

6. Gupta, S. S., Nagel, K. and Panchapakesan, S. (1973). On the order statistics from equally correlated normal random variables. Biometrika (60), 403-413.

7. Hâjek, J. Nonparametric Statistics. Holden-Day, Inc., San Francisco, 1969.

8. Kleijnen, J. P. C. Statistical Techniques in Simulation, Part II. Marcel Dekker, Inc., New York, 1975.

9. Kramer, A. (1956). A quick rank test for significance of differences in multiple comparisons. Fd. Tech. (10), 391-392.

10. Lee, Y. J. and Dudewicz, E. J. (1974). Nonparametric ranking and selection procedures, Technical Report #105, Department of Statistics, The Ohio State University, Columbus, Ohio.

11. Lehmann, E. L. (1963). A class of selection procedures based on ranks. Math. Annalen (150), 268-275.

12. Lehmann, E. L. Nonparametrics: Statistical Methods Based on Ranks. Holden-Day, Inc., San Francisco, 1975.

13. McDonald, G. C. (1972). Some multiple comparison selection procedures based on ranks. Sankhyā, Series A, (34), 53-64.

14. McDonald, G. C. (1973). The distribution of some rank statistics with applications in block design selection problems. Sankhyā, Series A, (35), 187-204.

15. Milton, R. C. (1963). Tables of the equally correlated multivariate normal probability integral, Technical Report #27, Department of Statistics, University of Minnesota, Minneapolis, Minn.

16. Milton, R. C. and Hotchkiss, R. (1969). Computer evaluation of the normal and inverse normal distribution functions. Technometrics (11), 817-822.

17. Owen, D. B. Handbook of Statistical Tables. Addison-Wesley Publishing Company, Inc., Reading, Mass., 1962 .

18. Thompson, W. A. and Willke, T. A. (1963). On an extreme rank sum test for outliers. Biometrika (50), 375-383.

19. Tukey, J. W. (1949). One degree of freedom for non-additivity. Biometrics (5), 232-242.

20. Winer, B. J. Statistical Principles in Experimental Design, 2nd ed., McGraw-Hill, Inc., New York, 1971 .