

AD-A044 890

PURDUE UNIV LAFAYETTE IND DEPT OF STATISTICS
SUBSET SELECTION PROCEDURES WITH SPECIAL REFERENCE TO THE ANALY--ETC(U)
SEP 77 S S GUPTA, J C HSU
MIMEOGRAPH SER-511

N00014-75-C-0455

F/G 12/1

NL

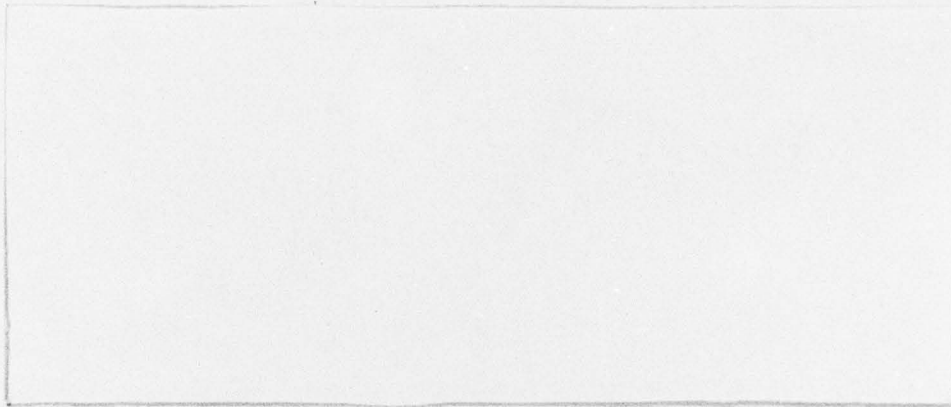
UNCLASSIFIED

1 of 1
ADA044890

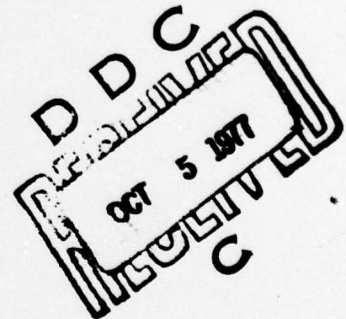


END
DATE
FILMED
11-77
DDC

AD A 044890



PURDUE UNIVERSITY



DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

AD No. _____
DDC FILE COPY

DISTRIBUTION STATEMENT A
Approved for release;
Distribution unlimited

6

SUBSET SELECTION PROCEDURES WITH SPECIAL REFERENCE TO THE ANALYSIS OF TWO-WAY LAYOUT: APPLICATION TO MOTOR-VEHICLE FATALITY DATA.

by

10

Shanti S./Gupta and Jason C./Hsu
Purdue University

9

Technical rept.,

DDC
OCT 5 1977
C

Department of Statistics
Division of Mathematical Sciences

14 Mimeograph Ser. 511

11 Sept. 77

12 Sp.

15 *This research was supported by the Office of Naval Research contract
N00014-75-C-0455 at Purdue University.

1473

291 730 ✓

LB

SUBSET SELECTION PROCEDURES WITH SPECIAL REFERENCE TO
THE ANALYSIS OF TWO-WAY LAYOUT: APPLICATION TO
MOTOR-VEHICLE FATALITY DATA

by

Shanti S. Gupta and Jason C. Hsu
Purdue University

ABSTRACT

In this paper, the origin of selection and ranking problems is discussed. Then the two basic approaches to the selection problem - the indifference zone approach and the subset selection approach - are reviewed briefly. As an application, Gupta's subset selection procedure is applied to motor-vehicle fatality data which fits into a two-way layout.

I. INTRODUCTION AND ORIGIN OF THE PROBLEM

A common problem faced by an experimenter is one of comparing several categories or populations. These may be, for example, different varieties of a grain, different competing manufacturing processes for an industrial product, different drugs (treatments) for a specific disease, or different alternatives under which a simulated system is run. In other words, we have k (≥ 2) populations and each population is characterized by the value of a parameter of interest θ_i , which may be, in the example of drugs, an appropriate measure of the effectiveness of a drug. The classical approach to this problem is to test the hypothesis $H_0: \theta_1 = \dots = \theta_k$, where $\theta_1, \dots, \theta_k$ are the values of the parameter for these populations. In the case of normal populations with means $\theta_1, \dots, \theta_k$ and a common variance σ^2 , the test can be carried out using the F-ratio of the analysis of variance.

The above classical approach is inadequate and unrealistic in the sense that it often cannot answer the experimenter's real questions, such as, how to identify the best category? Often in practice, after the hypothesis $H_0: \theta_1 = \dots = \theta_k$ has been rejected, one of the multiple-comparison procedures designed for making inferences concerning all pairwise differences of θ_i or all linear contrasts of θ_i is employed, and based on its outcome some purported 'best' set of populations is chosen. But this method of choosing a 'best' set of populations is indirect and does not control any error rate relevant to the problem, for example, the probability of an incorrect selection.

II. TWO APPROACHES TO THE SELECTION PROBLEM

The formulation of a k -sample problem as a selection and ranking problem enables the experimenter to answer his natural questions regarding the best category. There are two basic approaches to the problem of selection. The first approach is what is known as the indifference zone approach introduced by Bechhofer in [1]. The second approach is the subset selection approach introduced by Gupta in [6].

In order to explain the two approaches, consider the problem of selecting the population with the largest mean from k normal populations with unknown means μ_i , $i = 1, \dots, k$, and a common known variance σ^2 . Let \bar{x}_i , $i = 1, \dots, k$, denote the sample means of independent samples of size n from these populations. The 'natural' procedure is to select the population that yields the largest \bar{x}_i . The experimenter would, of course, want a guarantee that this procedure will pick the population with the largest μ_i with a probability not less than a specified level P^* . For the problem to be meaningful, P^* should be between $1/k$ and 1. Since we do not know the true configuration of the μ_i , we look for the least favorable configuration (LFC) for which the probability of a correct selection (PCS) is at a minimum. Without restrictions on the μ_i , $i=1, \dots, k$, the LFC is given by $\mu_1 = \dots = \mu_k$ for which the probability guarantee cannot be met, whatever the sample size n .

A natural modification is to insist on the minimum probability guarantee whenever the best population is sufficiently superior to the next best. In other words, the experimenter specifies a positive constant Δ^* and requires PCS to be at least P^* whenever $\mu_{[k]} - \mu_{[k-1]} \geq \Delta^*$, where $\mu_{[1]} \leq \dots \leq \mu_{[k]}$ denote the ordered means. So the minimization of PCS is over the part Ω_{Δ^*} of the parameter space in which $\mu_{[k]} - \mu_{[k-1]} \geq \Delta^*$. The complement of Ω_{Δ^*} is called the indifference zone for the obvious reason. The problem is to determine the minimum sample size n required in order to achieve $PCS > P^*$ for the LFC. This approach is known as the indifference zone approach.

Subset Selection Procedures ... Continued

In the subset selection approach, the goal is to select a non-empty subset of the populations so as to include the best population. Here the size of the selected subset is not fixed in advance, but rather is determined by the observations themselves. For our example of normal populations with unknown means μ_1, \dots, μ_k and common known variance σ^2 , the rule proposed by Gupta in [6] selects the population that yields \bar{x}_i if and only if $\bar{x}_i \geq \max_{1 \leq j \leq k} \bar{x}_j - d_1 \sigma / \sqrt{n}$, where $d_1 = d_1(k, P^*) > 0$ is determined so that the PCS is at least P^* . The constant d_1 is determined by

$$\int_{-\infty}^{\infty} \phi^{k-1}(t+d_1) d\phi(t) = P^*$$

where ϕ is the cumulative distribution function of a standard normal variable. Tables for d_1 for selected values of k and P^* are available in Gupta, Nagel, and Panchapakesan [9]. In the case where σ^2 is common but unknown, Gupta's procedure is to select the population that yields \bar{x}_i if and only if $\bar{x}_i \geq \max_{1 \leq j \leq k} \bar{x}_j - d_2 s / \sqrt{n}$ where s^2 is the usual pooled estimate of σ^2 based on $v = k(n-1)$ degrees of freedom. $d_2 = d_2(k, v, P^*)$ is chosen to satisfy the P^* condition and is determined by

$$\int_0^{\infty} \int_{-\infty}^{\infty} \phi^{k-1}(t+d_2 u) d\phi(t) dQ_v(u) = P^*$$

where ϕ is as before and Q_v is the distribution function of x_v / \sqrt{v} . For selected values of P^* , k and v the values of d_2 were tabulated by Gupta and Sobel in [11].

It should be pointed out that the two approaches, namely, indifference zone and subset selection, differ in that the former requires specification of two constants P^* and Δ^* to select a fixed number t , say, of populations; the later (subset selection) requires only one constant, namely, P^* to be specified and selects a random size subset depending on the outcome of the experiment.

Performance of subset selection procedures can be discussed in terms of true probability of a correct selection, expected subset size, expected proportion selected, and other similar quantities. A number of performance studies have been carried out, see, for example, Gupta [7] and Deely and Gupta [4]. Gupta and Panchapakesan [10] gave a comprehensive account of the relevant work in the area up to that time. Since then progress has been made in several directions. Dudewicz and Dalal [5] considered, among other things, two stage procedures for the normal means problem with unknown and unequal variances. Gupta and Huang [8] also considered the normal means problem with unequal variances. In his Ph.D. thesis [2], Berger considered the minimaxity

and admissability of subset selection procedures. Two recent Monte Carlo studies by Chernoff and Yahav [3] and Hsu [12] showed that for the normal means problem discussed, the class of Gupta's normal means procedures are nearly optimal in the sense that with respect to normal priors, their integrated risks are close to those of Bayes procedures.

In the next section we shall illustrate the use of subset procedures by applying the method just described to traffic fatality data.

III. AN ANALYSIS OF MOTOR-VEHICLE FATALITY DATA

In McDonald [13], the use of nonparametric subset selection procedures is illustrated by the application of these procedures to a set of traffic fatality data. For comparison purposes, we shall use the same data set. We are indebted to Dr. McDonald for allowing us to use this data and for suggesting the useful transformation used subsequently.

The traffic fatality data used in McDonald [13] are motor-vehicle traffic fatality rates (MFR) for the forty-eight contiguous states and the District of Columbia for the years 1960 to 1976. See Table 1 of McDonald [13]. It would be of interest to select out those states that have MFR much higher or lower than average. Further investigation of these states might identify factors related to MFR. We shall illustrate the use of subset selection procedures by selecting a set of 'best' populations and a set of 'worst' populations.

Let X_{ij} denote the MFR for the i th state and the 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. Our goal is roughly to select the states having the lowest (highest) "average" MFR. For an appropriate model we consider the two-way layout:

$$X_{ij} = m + a_i + b_j + (ab)_{ij} + e_{ij}, \quad i=1, \dots, 49, \quad j=1, \dots, 17 \quad (1)$$

where

$$\sum_{i=1}^{49} a_i = 0, \quad \sum_{j=1}^{17} b_j = 0, \quad \sum_{i=1}^{49} (ab)_{ij} = 0, \quad \sum_{j=1}^{17} (ab)_{ij} = 0$$

and

e_{ij} are independently distributed with means 0.

Our goal, stated in terms of the model (1), is as follows:

Goal 1: Select the states having the smallest (largest) $m+a_i$.

Note that our model is the fixed-effect model. The factor 'year' is not considered to be a random factor since it can be observed from the data that from around 1968, there has been a general decreasing trend in the fatality rate.

Using traditional analysis of variance techniques, one would first test the hypothesis $H: (ab)_{ij} = 0$ for all i, j . If this hypothesis is accepted, one would proceed to test whether each of the main effects is significant. However, our goal is the stated Goal 1. We are not particularly interested in whether the main effects are significant.

In order to achieve our goal, intuitively we need to have good estimates of the a_i 's. We also need to have estimates of the variances of these estimates. For the latter it is generally necessary to have $(ab)_{ij} = 0$ for all i, j . Therefore, Tukey's test for additivity was run to test the hypothesis $H_0: (ab)_{ij} = 0$ for all i, j . Unfortunately the test rejected the null hypothesis. However, it is often possible to transform the data so that the interaction term for the transformed data is statistically insignificant. For this MFR data, the monotone transformation $Y_{ij} = \ln(X_{ij}-1)$ appears to be such a transformation. Tukey's test on the transformed data showed no significance against the hypothesis of no interaction. (For the analysis that led to the choice of this transformation, see McDonald [13].) Thus, for the transformed data the following model (2) appears to be reasonable:

$$(2) \quad Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}, \quad i=1, \dots, 49, \quad j=1, \dots, 17$$

where

$$\sum_{i=1}^{49} \alpha_i = 0, \quad \sum_{j=1}^{17} \beta_j = 0,$$

and

ϵ_{ij} are independently distributed with means 0. To investigate further, tests were run on the sample residuals $y_{ij} - y_{i.} - y_{.j} + y_{..}$ where $y_{i.} = \sum_{j=1}^{17} y_{ij}/17$,

$$y_{.j} = \sum_{i=1}^{49} y_{ij}/49 \quad \text{and} \quad y_{..} = \sum_{i=1}^{49} \sum_{j=1}^{17} y_{ij}/(49 \times 17).$$

Test of homogeneity of variance of the residuals showed no significance. Against the hypothesis that the residuals are normally distributed, the two-tailed Kolmogorov-Smirnov test rejects at the 5% level but not at the 1% level. See Table 1. Plotting of the residuals on normal probability paper further reveals that the distribution of the residuals during the first seven years has a slightly longer left-hand tail than the normal distribution while during the last ten years it is essentially normal. Thus it appears not unreasonable to assume that ϵ_{ij} are independent $N(0, \sigma^2)$ for some σ^2 .

TABLE 1

Kolmogorov-Smirnov Goodness of Fit Test

Test Dist.=Normal(mean=.0000, std. dev.=.5963)

cases	max. diff.	K-S Z	2-tailed P
833	-.0181	.5222	.9480

Our original goal is to select those states i that have the lowest (highest) $E(X_{i.})$ where $X_{i.} =$

$$\sum_{j=1}^{17} X_{ij}/17. \quad \text{Under the model (2) and the assumption}$$

that ϵ_{ij} are independent $N(0, \sigma^2)$, it can be shown that the relative ordering among $E(Y_{i.})$ where $Y_{i.} =$

$$\sum_{j=1}^{17} Y_{ij}/17 = \sum_{j=1}^{17} \ln(X_{ij}-1)/17$$

is the same as the relative ordering among $E(X_{i.})$. This amounts to saying that the 'transformed' mean fatality rates have the same relative ordering as the original untransformed mean fatality rates. Hence, we can restate our original goal (Goal 1) in terms of the quantities in the model (2) as

Goal 2: Select the states having the smallest (largest) $\mu + \alpha_i$.

Before we apply Gupta's normal means procedure to the transformed data, some comments are in order. We have stated earlier that from the Monte Carlo studies of Chernoff and Yahav [3] and Hsu [12] we know that in the normal case Gupta's procedure performs well. Hence for Goal 2, Gupta's procedure will have good performance. But the transformation changed the scale of measurement for the means and substantially changed the variances of the relevant quantities. One might question whether a procedure good for Goal 2 is necessarily good for Goal 1. Those Monte Carlo studies showed that Gupta's procedure is good for a variety of loss functions corresponding to the general goal of selecting a subset of good (bad) populations. Hence in terms of Goal 1, applying Gupta's procedure to the transformed data should give good results.

To apply Gupta's normal means procedure we shall estimate each $\mu + \alpha_i$ by $y_{i.}$ and σ^2 by $s^2 = \Sigma(y_{ij} - y_{i.} - y_{.j} + y_{..})^2 / (48 \times 16)$. Table 2 lists $y_{i.}$ and the corresponding states in ascending order of $y_{i.}$. The calculated value of s^2 is also given. From Gupta, Nagel and Panchapakesan [9], the d_2 values corresponding to $P^* = .90$ in 3.651. Therefore, to select a subset of states such that with probability .90 the state with the best (lowest) true MFR is included, we select those states with $y_{i.} \leq y_{38} + d_2 s / \sqrt{n} = .360 + 0.109 = 0.469$. Only Rhode Island is selected. To select a subset of states such that with probability .90 the state with the true worst (highest) MFR is included, we select those states with $y_{i.} \geq y_{30} - d_2 s / \sqrt{n} = 1.777 - 0.109 = 1.668$. Six states are selected. See Table 2. For $P^* = .99$, for the set of 'best' populations, again only Rhode Island is selected. For the set of 'worst' populations, ten states are selected.

Let us compare Gupta's normal means procedure with McDonald's rank sum procedure (described in detail in McDonald [13]). For $P^* = .90$, the normal means procedure selects six states as 'worst' populations while the rank sum procedure R_1 selects ten. This is not surprising since more assumptions are made in applying the normal means procedure, hence

Subset Selection Procedures ... Continued

one is able to obtain stronger results. The rank sum procedure R_1 selects twelve states as 'best' populations while the normal means procedure selects only one. This may seem mildly surprising but a careful examination of the basic MFR data readily reveals the reason. From Table 1 of McDonald [13] one sees that the MFR for Rhode Island is consistently much smaller than average. This causes the normal means procedure to select that state alone. The rank sum procedure is based on relative ranks only. It is designed so that the information concerning the magnitude of the differences in the sample is ignored. Hence there is no drastic reduction in the number of states selected for the rank sum procedure.

IV. CONCLUSION

In the last twenty-five years, research in the area of selection and ranking procedures has progressed steadily. These procedures clearly have great potential for application in simulation studies and in other areas. They have not been used more perhaps because it calls for giving up the ingrained habit of testing of hypothesis on the part of applied statisticians. In view of the fact that some optimality properties of these procedures are becoming known, the time is right for making an effort in applying these procedures in practice.

TABLE 2

Selection of States in Terms of MFR

i state	y_i	$p^* = .90$ $p^* = .99$		
		Gupta	McDonald	Gupta
38 Rhode Island	.360	*	*	*
6 Connecticut	.538		*	
29 New Jersey	.667		*	
8 Dist. of Col.	.775		*	
20 Massachusetts	.899		*	
19 Maryland	1.037		*	
37 Pennsylvania	1.045		*	
28 New Hampshire	1.065		*	
18 Maine	1.114		*	
7 Delaware	1.130		*	
34 Ohio	1.163		*	
46 Washington	1.167		*	
4 California	1.195			
12 Illinois	1.198			
45 Virginia	1.210			
21 Michigan	1.218			
22 Minnesota	1.231			
31 New York	1.245			
26 Nebraska	1.251			
48 Wisconsin	1.310			
15 Kansas	1.323			
13 Indiana	1.333			
35 Oklahoma	1.339			
14 Iowa	1.374			
5 Colorado	1.378			
9 Florida	1.380			
33 North Dakota	1.389			
42 Texas	1.390			

Table 2 ... Continued

i state	y_i	$p^* = .90$ $p^* = .99$		
		Gupta	McDonald	Gupta
43 Utah	1.395			
24 Missouri	1.403			
36 Oregon	1.415			
44 Vermont	1.458			
16 Kentucky	1.467			
41 Tennessee	1.495			*
10 Georgia	1.533			*
3 Arkansas	1.546			*
40 South Dakota	1.561			*
47 West Virginia	1.580			*
32 North Carolina	1.625			*
49 Wyoming	1.636			*
2 Arizona	1.636			*
39 South Carolina	1.648			*
1 Alabama	1.651			*
25 Montana	1.691	*	*	*
11 Idaho	1.698	*	*	*
17 Louisiana	1.758	*	*	*
23 Mississippi	1.773	*	*	*
27 Nevada	1.775	*	*	*
30 New Mexico	1.777	*	*	*

$s^2 = 0.0152$ $s = 0.123$ * denote selected state

ACKNOWLEDGEMENT

The authors are grateful to Dr. Gary C. McDonald, Assistant Head, Department of Mathematics, General Motors Research Laboratories, for providing in addition to the MFR data an advance copy of his paper, and for helpful discussions. Thanks are also due to Professor Edward J. Dudewicz, Statistics Department, Ohio State University for reading a first draft of this paper and making helpful suggestions. This work was supported in part by the Office of Naval Research Contract N00014-75-C-0455 at Purdue University.

BIBLIOGRAPHY

1. Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. Ann. Math. Statist. 25, 16-39.
2. Berger, R. (1977). Minimax, admissible, and gamma-minimax multiple decision rules. Mimeo. Ser. No. 489, Dept. of Statistics, Purdue University, W. Lafayette, IN.
3. Chernoff, H. and Yahav, J. A. (1977). On selecting a set of good populations. Statistical Decision Theory and Related Topics II. (ed. S. S. Gupta and D. S. Moore), 37-55. Academic Press, NY.
4. Deely, J. J. and Gupta, S. S. (1968). On the properties of subset selection procedures. Sankhya Ser. A 30, 37-50.
5. Dudewicz, E. J. and Datta, S. R. (1975). Allocations of observations in ranking and selection with unequal variances. Sankhya Ser. B 37, 28-78.

ACC. SYSTEM Per

NOIS Write Section

DOC Buff Section

UNANNOUNCED

JUS I LOCATION

BY

DISTRIBUTION/AVAILABILITY NOTES

Dist v v

A

6. Gupta, S. S. (1956). On a decision rule for a problem in ranking means. Mimeo Ser. No. 150, Inst. of Statistics, University of North Carolina, Chapel Hill, NC.
7. Gupta, S. S. (1965). On some multiple decision (selection and ranking) rules. Technometrics 7, 225-245.
8. Gupta, S. S. and Huang, D. Y. (1976). Subset selection procedures for the means and variances of normal populations: Unequal sample size case. Sankya 38.
9. Gupta, S. S., Nagel, K. and Panchapakesan, S. (1973). On the order statistic from equally correlated normal random variables. Biometrika 60, 403-413.
10. Gupta, S. S. and Panchapakesan, S. (1972). On multiple decision (subset selection) procedures. Jour. of Math. and Phy. Sci. 6, 1-71.
11. Gupta, S. S. and Sobel, M. (1957). On a statistic which arises in selection and ranking problems. Ann. Math. Statist. 28, 956-967.
12. Hsu, J. C. (1977). On some decision-theoretic contributions to the problem of subset selection. Mimeo. Ser. No. 491, Dept. of Statistics, Purdue University, W. Lafayette, IN.
13. McDonald, G. C. (1977). An application of non-parametric selection procedures to an analysis of motor-vehicle traffic fatality rates. These proceedings of the 1977 Winter Simulation Conference.

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Mimeograph Series #511 ✓	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Subset Selection Procedures with Special Reference to the Analysis of Two-Way Layout: Application to Motor-Vehicle Fatality Data	5. TYPE OF REPORT & PERIOD COVERED Technical	
	6. PERFORMING ORG. REPORT NUMBER Mimeo. Series #511	
7. AUTHOR(s) Shanti S. Gupta and Jason C. Hsu	8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0455 ✓	
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Statistics ✓ Purdue University W. Lafayette, IN 47907	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC	12. REPORT DATE September 1977	
	13. NUMBER OF PAGES 5	
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report) Unclassified	
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE	
16. DISTRIBUTION STATEMENT (of this Report) Approved for public release; distribution unlimited.		
17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number) Subset selection, ranking and selection, two-way layout, motor-vehicle fatality rate.		
20. ABSTRACT (Continue on reverse side if necessary and identify by block number) In this paper, the origin of selection and ranking problems is discussed. Then the two basic approaches to the selection problem - the indifference zone approach and the subset selection approach - are reviewed briefly. As an application, Gupta's subset selection procedure is applied to motor-vehicle traffic fatality data which fits into a two-way layout.		