COMMUN. STATIST.-SIMULA. COMPUTA., 11(3), 311-334 (1982)

A DISTRIBUTION-FREE APPROACH TO INDUCING RANK CORRELATION AMONG INPUT VARIABLES

Ronald L. Iman

Sandia National Laboratories
Albuquerque, NM 87185

W. J. Conover

College of Business Admin.
Texas Tech University
Lubbock, TX 79409

Key Words and Phrases: rank correlation; computer models; multivariate distribution; large-scale computer codes; dependences; distribution free.

ABSTRACT

A method for inducing a desired rank correlation matrix on a multivariate input random variable for use in a simulation study is introduced in this paper. This method is simple to use, is distribution free, preserves the exact form of the marginal distributions on the input variables, and may be used with any type of sampling scheme for which correlation of input variables is a meaningful concept. A Monte Carlo study provides an estimate of the bias and variability associated with the method. Input variables used in a model for study of geologic disposal of radioactive waste provide an example of the usefulness of this procedure. A textbook example shows how the output may be affected by the method presented in this paper.

1. INTRODUCTION

Computer models are used widely to simulate the intricate relationships among variables in economic, social, and physical environments, in order to estimate unknown quantities or predict future events. The ever expanding capability and capacity of computers has allowed the complexity of these models to increase dramatically. It is not uncommon to find models which have perhaps several hundred input variables and may take several hours of computer time to generate a single output observation. Investigation of techniques for selecting input values has led to the development of efficient sampling techniques by McKay, Conover and Beckman (1979). Procedures for looking at the effect of different distributional assumptions on input variables have been examined in Iman and Conover (1980).

While much effort has been expended toward development of new statistical techniques for computer modeling, relatively little attention has been given to the problem of incorporating the dependences that may exist among the input variables. Typically the model input variables are assumed to be independent (Iman, Helton, and Campbell, 1981a, 1981b). A study presently underway at Sandia National Laboratories is examining mechanisms by which radionuclides might escape a waste depository in bedded salt (Campbell and Cranwell, 1980). The assumption of independence among input variables may not be appropriate for the models used in this study. For example, significant correlations are expected to exist between hydraulic properties in the vicinity of the disposal site and the time for circulating groundwater to contact radioactive waste.

One approach to incorporating dependences is to consider linear combinations of independent random variables to achieve a desired correlation structure. In the case of normal random variables and random sampling this approach is well known to produce a multivariate normal input vector. However, if the samples are obtained using a stratified sampling scheme, then this approach will destroy the integrity of the stratified sample. That is, the values obtained from a linear combination will no longer map back into each of the original strata which collectively span the range of each input variable. In addition, the linear combinations of non-normal random variables will adversely affect both the random sample and the stratified sample, as the marginal distributions may no longer resemble the original marginal distributions desired on the input variables.

Another approach to incorporating dependences has been developed by Johnson and Ramberg (1977). By viewing the marginal distributions as transformations of normal distributions, a correlation structure can be imposed as follows. An original normal independently distributed input vector is first transformed to a correlated multivariate normal vector as described above. The appropriate transformation is then used to obtain the desired marginal distributions. However, the means, variances and correlations of the transformed variables are difficult to control. Johnson and Ramberg show how to control these moments in the bivariate case for lognormal and inverse hyperbolic sine distributions, but the mathematics becomes intractable for one other distribution they considered, and appears to be equally difficult for other distributions.

In this paper we present a method based on rank correlations which is intended to induce the desired rank dependence among the input variables. The method has the following desirable properties.

1) It is distribution free. That is, it may be used with equal facility on all types of input distribution functions.
2) It is simple. No unusual mathematical techniques are required to implement the method.
3) It can be applied to any sampling scheme for which correlated input variables could logically be considered, while preserving the intent of the sampling scheme. That is, the same numbers originally selected as input values are retained; only their pairing is affected to achieve the desired rank correlation. This means that in Latin

hypercube sampling the integrity of the intervals is maintained. If some lattice structure is used for selection of values, that same structure is retained.

4) The marginal distributions remain intact.

Our approach is based on the premise that rank correlation is a meaningful way to define dependences among input variables. That is, a correlation coefficient computed on raw data may lose meaning and interpretation with non-normal data or in the presence of outliers. On the other hand, rank correlation coefficients can be quite meaningful in most modeling situations, even when the data are normal.

In Section 2 we explain the proposed method for inducing dependences among the input variables, and provide an example of the method. Section 3 presents the results of a Monte Carlo study. An application is discussed in Section 4. An example is presented in Section 5 to show how output may be affected by use of the method. The final section contains a discussion and summary.

2. THE METHOD

Suppose that a random row vector $\underset{\sim}{X}$ has a correlation matrix $\underset{\sim}{I}$. That is, the elements of $\underset{\sim}{X}$ are uncorrelated. Let $\underset{\sim}{C}$ be the desired correlation matrix of some transformation of $\underset{\sim}{X}$. Because $\underset{\sim}{C}$ is positive definite and symmetric, $\underset{\sim}{C}$ may be written as $\underset{\sim}{C} = \underset{\sim}{P}\underset{\sim}{P}'$ where $\underset{\sim}{P}$ is a lower triangular matrix (Scheuer and Stoller, 1962). Then the transformed vector $\underset{\sim}{X}\underset{\sim}{P}'$ has the desired correlation matrix $\underset{\sim}{C}$. This is the theoretical basis for our method.

Our objective is for the Spearman rank correlation matrix $\underset{\sim}{M}$ of the input vectors to be close to the target rank correlation matrix $\underset{\sim}{C}^*$ supplied by the user, while preserving certain important properties of the input vectors such as marginal distributions and properties of the sampling scheme used to obtain the input vectors. It does not appear possible to find a transformation matrix which results in the target rank correlation matrix, so we use scores {a(i)}, for which the correlation matrix $\underset{\sim}{C}$ and the rank correlation matrix $\underset{\sim}{M}$ will be close to each other after the transformation by $\mathbf{P}$. Thus by setting $\mathbf{C}$ equal to $\mathbf{C}^*$, a transformation matrix $\mathbf{P}$ is obtained which will result in a rank correlation matrix $\mathbf{M}$ close to the desired rank correlation matrix $\mathbf{C}^*$.

Let the number of input variables be denoted by K, and let N be the sample size. Let $\mathbf{R}$ be an NxK matrix whose columns represent K independent permutations of an arbitrary set {a(i)}, i=1,...,N, of N numbers, referred to as "scores." The columns should be checked to insure that there are no perfect rank correlations among the scores. Each row of $\mathbf{R}$, say $\mathbf{R}_i$, has K independent components where each component assumes one of the values a(i), i=1,...,N with equal probability. Then the row vector $\mathbf{R}_i$ has population correlation matrix $\mathbf{I}$.

Let $\mathbf{C}^*$ be the user supplied target rank correlation matrix and set $\mathbf{C} = \mathbf{C}^*$. Let $\mathbf{P}$ be a matrix such that $\mathbf{PP}' = \mathbf{C}$. As suggested earlier, the Cholesky factorization scheme used by Scheuer and Stoller (1962) may be used to obtain a lower triangular matrix $\mathbf{P}$ such that $\mathbf{PP}' = \mathbf{C}$. Multiplication by $\mathbf{P}'$, $\mathbf{R}_i\mathbf{P}'$, results in a vector which has the desired population correlation matrix $\mathbf{C}$. Multiplication of the entire matrix $\mathbf{R}$ by $\mathbf{P}'$, $\mathbf{RP}' = \mathbf{R}^*$, gives a matrix $\mathbf{R}^*$ whose rows have the same multivariate distribution as $\mathbf{R}_i\mathbf{P}'$. The rank correlation matrix $\mathbf{M}$ of $\mathbf{R}^*$ should be close to $\mathbf{C}$.

For the rank correlation matrix of the input values to be approximately equal to $\mathbf{C}$, the values in each column of the NxK input matrix are rearranged so that they will have the same ordering as the corresponding column of $\mathbf{R}^*$. Thus the input values have the same sample rank correlation matrix that $\mathbf{R}^*$ has. A numerical example will now be given to illustrate the method.

Suppose the rank correlation matrix $\mathbf{C}^*$ is desired for 6 input variables. A lower triangular matrix $\mathbf{P}$ such that $\mathbf{PP}' = \mathbf{C} = \mathbf{C}^*$ is found by the Cholesky factorization of $\mathbf{C}$. The scores {a(i)} in this example are the van der Waerden scores $\Phi^{-1}(i/(N+1))$, where Φ^{-1} is the inverse function of the standard normal distribution function. For a sample of size N=15 the matrix $\mathbf{R}$ has a random mix of the van der Waerden scores $\Phi^{-1}(i/16)$, i=1,...,15 in each column.

$$\underset{\sim}{C}^{*} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 75 & -.70 \\ 0 & 0 & 0 & .75 & 1 & -.95 \\ 0 & 0 & 0 & -.70 & -.95 & 1 \end{bmatrix}$$

$$\underset{\sim}{P} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & .75 & .6614 & 0 \\ 0 & 0 & 0 & -.70 & -.6425 & .3117 \end{bmatrix}$$

$$\underset{\sim}{R} = \begin{bmatrix} 1.534 & 1.534 & -1.534 & -1.534 & .489 & -.319 \\ -.887 & -.489 & .887 & -.887 & -.157 & .674 \\ -.489 & .674 & -.489 & 1.150 & 1.534 & -.489 \\ .887 & 0.000 & -.674 & .319 & 0.000 & -1.534 \\ 1.150 & -.319 & .489 & .674 & .157 & 1.150 \\ .157 & -1.534 & -.887 & -.674 & -.319 & .157 \\ -1.150 & -.674 & -.157 & .157 & -1.534 & -.157 \\ 0.000 & -.887 & .157 & -.319 & -.674 & .887 \\ .319 & -.157 & .674 & .887 & .674 & 1.534 \\ -.319 & .157 & -.319 & -1.150 & 1.150 & -.887 \\ -1.534 & .887 & 1.150 & 1.534 & -.489 & -1.150 \\ -.157 & -1.150 & 1.534 & -.157 & -1.150 & -.674 \\ .489 & .489 & -1.150 & .489 & -.887 & 0.000 \\ .674 & .319 & .319 & 0.000 & .887 & .319 \\ -.674 & 1.150 & 0.000 & -.489 & .319 & .489 \end{bmatrix}$$

The sample correlation matrix $\underset{\sim}{T}$ for $\underset{\sim}{R}$ is given for the interest of the reader although it is not used in the method.

$$\underset{\sim}{T} = \begin{bmatrix} 1.0000 & .0969 & -.4667 & -.2335 & .2614 & .1748 \\ .0969 & 1.0000 & -.3129 & .0710 & .4838 & -.2270 \\ -.4667 & -.3129 & 1.0000 & .3377 & -.1970 & .1902 \\ -.2335 & .0710 & .3377 & 1.0000 & -.0412 & -.0298 \\ .2614 & .4838 & -.1970 & -.0412 & 1.0000 & .0522 \\ .1748 & -.2270 & .1902 & -.0298 & .0522 & 1.0000 \end{bmatrix}$$

The matrix $\underset{\sim}{R}^*$ is found as $\underset{\sim}{R}\underset{\sim}{P}'$. The Spearman rank correlation matrix $\underset{\sim}{M}$ of $\underset{\sim}{R}^*$ can be compared with the desired rank correlation matrix $\underset{\sim}{C}^*$. The non-zero target correlations agree closely with the desired values while some of the zero target correlations are rather large, e.g. - .5107. The primary reason for this variation is that any particular realization $\underset{\sim}{r}^*$ of $\underset{\sim}{R}^*$ will have a sample correlation that estimates $\underset{\sim}{C}$. That is, if the sample correlation matrix $\underset{\sim}{T}$ associated with $\underset{\sim}{R}$ is exactly equal to $\underset{\sim}{I}$, then the sample correlation matrix of $\underset{\sim}{R}^*$ would be $\underset{\sim}{C}$, and the rank correlation matrix of $\underset{\sim}{R}^*$ would be approximately equal to $\underset{\sim}{C} = \underset{\sim}{C}^*$.

It only remains to generate the NxK matrix of input vectors, according to any desired method or distribution, as if the K input random variables were independent of each other. Then the values of the variable in each column are arranged so they have the same order (rank) as the corresponding column in $\underset{\sim}{R}^*$. Thus the sample Spearman rank correlation of the input vectors will be the same as the sample Spearman rank correlation of $\underset{\sim}{R}^*$, given by $\underset{\sim}{M}$ for this example. Also, the identity of the original marginal distributions on the input variables has been maintained, as the procedure explained above merely provides a means for pairing the variables and does not change the numbers themselves.

$$\underset{\sim}{R}^* = \begin{bmatrix} 1.534 & 1.534 & -1.534 & -1.534 & -.827 & .660 \\ -.887 & -.489 & .887 & -.887 & -.769 & .932 \\ -.489 & .674 & -.489 & 1.150 & 1.877 & -1.943 \\ .887 & 0.000 & -.674 & .319 & .239 & -.701 \\ 1.150 & -.319 & .489 & .674 & .609 & -.214 \\ .157 & -1.534 & -.887 & -.674 & -.717 & .726 \\ -1.150 & -.674 & -.157 & .157 & -.897 & .827 \\ 0.000 & -.887 & .157 & -.319 & -.685 & .933 \\ .319 & -.157 & .674 & .887 & 1.111 & -.576 \\ -.319 & .157 & -.319 & -1.150 & -.102 & -.210 \\ -1.534 & .887 & 1.150 & 1.534 & .827 & -1.118 \\ -.157 & -1.150 & 1.534 & -.157 & -.878 & .639 \\ .489 & .489 & -1.150 & .489 & -.220 & .228 \\ .674 & .319 & .319 & 0.000 & .587 & -.471 \\ -.674 & 1.150 & 0.000 & -.489 & -.156 & .290 \end{bmatrix}$$

$$\underset{\sim}{M} = \begin{bmatrix} 1.0000 & .0607 & -.4036 & -.0821 & .0964 & -.1179 \\ .0607 & 1.0000 & -.2857 & .1321 & .4107 & -.5107 \\ -.4036 & -.2857 & 1.0000 & .2714 & .1429 & -.0571 \\ -.0821 & .1321 & .2714 & 1.0000 & .6714 & -.7036 \\ .0964 & .4107 & .1429 & .6714 & 1.0000 & -.8679 \\ -.1179 & -.5107 & -.0571 & -.7036 & -.8679 & 1.0000 \end{bmatrix}$$

Note that the transformation matrix $\underset{\sim}{P}$ depends only on $\underset{\sim}{C}$. As a result, random variation in the sample correlation matrix of $\underset{\sim}{R}$ carries through the transformation, so that the sample correlation matrix of $\underset{\sim}{R}\underset{\sim}{P}'$ may not be close enough to $\underset{\sim}{C}$ for all applications of this procedure. This concern led to the development of a variance reduction technique in which the transformation matrix is adjusted so that the final sample correlation matrix will be much closer to $\underset{\sim}{C}$. This variance reduction technique considerably decreases the variability of the sample correlation matrix and could be used with the method in situations where it is desired to have the uncorrelated variables nearly orthogonal. Another application of this variation of the method would be associated with time consuming computer models where only a limited number of computer runs can be made and the user wants the actual rank correlation matrix of the input variables to be very close to that which he specifies.

In order to avoid the problem associated with $\underset{\sim}{R}$ not necessarily having a sample correlation matrix equal to $\underset{\sim}{I}$, and thus the sample correlation matrix of $\underset{\sim}{R}^*$ not being close enough to $\underset{\sim}{C}$ to satisfy the user, a matrix $\underset{\sim}{S}$ is found such that $\underset{\sim}{STS'} = \underset{\sim}{C}$ where $\underset{\sim}{T}$ is the sample correlation matrix associated with $\underset{\sim}{R}$. Consider only realizations of $\underset{\sim}{R}$ which have distinct (non-identical) columns, so that $\underset{\sim}{T}$ is positive definite and symmetric. The Cholesky factorization may be used to find a lower triangular matrix $\underset{\sim}{Q}$ such that $\underset{\sim}{T} = \underset{\sim}{QQ'}$. This along with the fact that $\underset{\sim}{C} = \underset{\sim}{PP'}$ allows the equation involving $\underset{\sim}{S}$ to be rewritten as $\underset{\sim}{SQQ'S'} = \underset{\sim}{PP'}$ for which one solution is $\underset{\sim}{SQ} = \underset{\sim}{P}$ or $\underset{\sim}{S} = \underset{\sim}{PQ}^{-1}$. Note that $\underset{\sim}{S}$ is also lower triangular. The matrix $\underset{\sim}{R}_B^* = \underset{\sim}{RS'}$ has a correlation matrix exactly equal to $\underset{\sim}{C}$.

Continuing with the above example a lower triangular matrix $\underset{\sim}{Q}$ is found by the Cholesky factorization such that $\underset{\sim}{QQ'} = \underset{\sim}{T}$. The matrix $\underset{\sim}{S}$ is found as $\underset{\sim}{PQ}^{-1}$. This method defines $\underset{\sim}{R}_B^*$ as $\underset{\sim}{RS'}$. Finally, the Spearman rank correlation matrix $\underset{\sim}{M}_B$ of $\underset{\sim}{R}_B^*$ can be found and compared with $\underset{\sim}{M}$ and the desired rank correlation matrix $\underset{\sim}{C}^*$. This comparison shows the non-zero target correlations again to be in close agreement with the desired values while the zero target correlations are as a whole much closer to zero than appeared in the matrix $\underset{\sim}{M}$. In the next section we present a short Monte Carlo study to compare the method and the variation more closely and to see what bias may be involved with van der Waerden scores and with this particular matrix $\underset{\sim}{C}^*$.

$$\underset{\sim}{Q} = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ .0969 & .9953 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ -.4667 & -.2689 & .8425 & 0.0000 & 0.0000 & 0.0000 \\ -.2335 & .0941 & .3015 & .9196 & 0.0000 & 0.0000 \\ .2614 & .4606 & .0580 & -.0446 & .8451 & 0.0000 \\ .1748 & -.2451 & .2443 & -.0431 & .1223 & .9126 \end{bmatrix}$$

$$\underset{\sim}{S} = \begin{bmatrix} 1.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ -.0974 & 1.0047 & 0.0000 & 0.0000 & 0.0000 & 0.0000 \\ .5228 & .3207 & 1.1869 & 0.0000 & 0.0000 & 0.0000 \\ .0924 & -.2079 & -.3891 & 1.0874 & 0.0000 & 0.0000 \\ -.1207 & -.5400 & -.3593 & .8535 & .7827 & 0.0000 \\ .0217 & .5971 & .2374 & -.7844 & -.8097 & .3415 \end{bmatrix}$$

$$\underset{\sim}{R}_B^{\star} = \begin{bmatrix} 1.5340 & 1.3919 & -.5267 & -1.2483 & -1.3888 & 1.2836 \\ -.8870 & -.4050 & .4322 & -1.2900 & -.8275 & .9524 \\ -.4890 & .7248 & -.6199 & 1.2554 & 2.0529 & -2.0354 \\ .8870 & -.0864 & -.3362 & .6911 & .4074 & -.9149 \\ 1.1500 & -.4325 & 1.0794 & .7153 & .5559 & -.3125 \\ .1570 & -1.5565 & -1.4627 & -.0543 & .3032 & -.2826 \\ -1.1500 & -.5652 & -1.0038 & .2656 & -.5075 & .6006 \\ 0.0000 & -.8912 & -.0981 & -.2235 & -.3772 & .6065 \\ .3190 & -.1888 & .9164 & .7644 & 1.0887 & -.6444 \\ -.3190 & .1888 & -.4951 & -1.1885 & -.0131 & -.3209 \\ -1.5340 & 1.0405 & .8474 & .8943 & .2194 & -.4308 \\ -.1570 & -1.1402 & 1.3698 & -.5430 & -.9453 & .4982 \\ .4890 & .4437 & -.9524 & .9228 & -.1867 & .3642 \\ .6740 & .2549 & .8333 & -.1281 & .3260 & -.3284 \\ -.6740 & 1.2211 & .0164 & -.8332 & -.7073 & .9644 \end{bmatrix}$$

$$\underset{\sim}{M}_B = \begin{bmatrix} 1.0000 & .0214 & .0464 & .0357 & .2179 & -.0786 \\ .0214 & 1.0000 & -.0643 & .0821 & -.0143 & -.0500 \\ .0464 & -.0643 & 1.0000 & -.0786 & .0536 & -.1143 \\ .0357 & .0821 & -.0786 & 1.0000 & .7286 & -.7036 \\ .2179 & -.0143 & .0536 & .7286 & 1.0000 & -.8893 \\ -.0786 & -.0500 & -.1143 & -.7036 & -.8893 & 1.0000 \end{bmatrix}$$

3. MONTE CARLO RESULTS

The brief Monte Carlo study reported in this section examines the sampling behavior of the Spearman sample rank correlation matrices M and M_B of the previous section, using the van der Waerden scores as before. The target correlation matrix is $\underset{\sim}{C}^*$ of Section 2 and sample sizes considered are N = 15, 25, 50 and 100. The Monte Carlo results are based on 100 repetitions, with the method reported in Table 1 and the variation of the method in Table 2. The Monte Carlo results in these tables show that the bias, if any, is small. The observed mean rank correlations for 100 repetitions are close to the desired values in almost every case, i.e., within one or two standard deviations ($s_{\bar{x}} = s/\sqrt{100}$). The estimates improve, that is, the bias and the standard deviations get smaller, as N gets larger, which one might expect. The variance estimates in both tables decrease at a rate close to N; however, the variance estimates for the variation of the method are roughly 12 to 15 times smaller than those for the method itself.

4. AN APPLICATION

This section presents an application of the method to a model used to estimate the risk associated with geologic disposal of radioactive waste. Input to this model includes time to groundwater contact with radioactive waste which is correlated with other input variables such as hydraulic, thermal, and mechanical properties of several rock types near the waste depository. Thus it is necessary to define a target correlation structure between properties of the rock units near the depository and the time to groundwater contact with radioactive waste. In this example, 15 variables are defined including the ones just mentioned. Thus the desired correlation matrix is a 15 x 15 symmetric matrix which must be positive definite. The nonzero target correlations are indicated in parentheses in Table 3, along with the actual rank correlation structure generated using the variation of the method of Section 2 with N = 100.

Table 1. Means and Standard Deviations of Rank Correlations Associated with the Spearman Matrix $\underset{\sim}{M}$ Based on 100 Monte Carlo Runs. The Target Rank Correlation Matrix is $\underset{\sim}{C}^*$ and Results are Recorded by Matrix Position and Sample Size.

		N = 15		N = 25		N = 50		N = 100	
i,j	Desired Correlation	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
1,2	0.0	-.0084	.2872	.0196	.1997	.0070	.1336	-.0059	.0986
1,3	0.0	-.0097	.2562	-.0188	.2020	-.0022	.1425	.0053	.1091
1,4	0.0	-.0270	.2557	-.0302	.2054	.0108	.1543	.0040	.1019
1,5	0.0	-.0248	.2614	.0016	.2291	.0155	.1436	.0062	.1056
1,6	0.0	.0273	.2678	-.0005	.2342	-.0093	.1491	-.0049	.1046
2,3	0.0	.0182	.3127	-.0050	.1900	.0096	.1296	.0013	.0952
2,4	0.0	-.0402	.2334	-.0025	.2052	.0008	.1489	-.0128	.1089
2,5	0.0	-.0318	.2140	-.0308	.2021	-.0124	.1398	-.0065	.1072
2,6	0.0	.0292	.2322	.0266	.2000	.0090	.1422	.0082	.1048
3,4	0.0	-.0270	.2517	-.0317	.2039	-.0031	.1409	.0106	.0933
3,5	0.0	-.0038	.2522	-.0138	.1972	-.0010	.1490	-.0010	.0929
3,6	0.0	-.0178	.2692	.0010	.2022	-.0002	.1479	-.0028	.0919
4,5	0.75	.7424	.0880	.7321	.0720	.7406	.0499	.7328	.0356
4,6	-0.70	-.6991	.1182	-.6692	.0961	-.6894	.0649	-.6811	.0489
5,6	-0.95	-.9262	.0365	-.9280	.0294	-.9386	.0146	-.9400	.0121

Table 2. Means and Standard Deviations of Rank Correlations Associated with the Spearman Matrix $\underset{\sim}{M}_B$ Based on 100 Monte Carlo Runs. The Target Rank Correlation Matrix is $\underset{\sim}{C}^*$ and Results are Recorded by Matrix Position and Sample Size.

		N = 15		N = 25		N = 50		N = 100	
i,j	Desired Correlation	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s	$\bar{x}$	s
1,2	0.0	.0032	.0834	.0009	.0556	-.0012	.0418	.0010	.0271
1,3	0.0	-.0011	.0734	-.0080	.0513	-.0050	.0366	-.0003	.0262
1,4	0.0	-.0148	.0786	-.0081	.0474	.0019	.0347	-.0036	.0261
1,5	0.0	-.0139	.0647	.0104	.0536	.0028	.0331	-.0031	.0294
1,6	0.0	.0099	.0666	-.0115	.0452	-.0033	.0378	.0023	.0314
2,3	0.0	.0149	.0832	-.0003	.0565	-.0008	.0430	.0020	.0294
2,4	0.0	-.0075	.0774	.0084	.0564	-.0007	.0368	.0021	.0272
2,5	0.0	-.0085	.0666	.0086	.0508	-.0046	.0372	.0008	.0273
2,6	0.0	.0081	.0687	.0009	.0564	.0064	.0369	-.0023	.0267
3,4	0.0	-.0057	.0792	-.0044	.0594	.0005	.0388	-.0000	.0255
3,5	0.0	.0062	.0801	-.0028	.0599	.0003	.0394	-.0005	.0314
3,6	0.0	.0010	.0876	-.0002	.0575	.0000	.0386	-.0026	.0313
4,5	0.75	.7216	.0603	.7288	.0404	.7319	.0289	.7344	.0171
4,6	-0.70	-.6769	.0719	-.6656	.0536	-.6795	.0337	-.6830	.0192
5,6	-0.95	-.9232	.0309	-.9288	.0264	-.9382	.0124	-.9410	.0076

Table 3. Actual Rank Correlations Generated from a Sample of Size N = 100 for the Input Variables Used in a Model to Estimate the Risk Associated with Geologic Isolation of Radioactive Waste. Numbers in Parentheses Represent the Nonzero Target Correlations.

Variable Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	-.0477													
3	-.0140	.0234												
4	-.0135	-.0380	-.0123											
5	.3102 (.3000)	-.2990 (-.3000)	.0358	-.0042										
6	.0208	.0283	-.0111	-.0217	-.0447									
7	.0245	-.0274	.0102	.0381	.0267	-.0208								
8	.0342	.0159	-.0232	.0289	.0192	.0246	.0099							
9	.0291	-.0224	.0003	.0218	-.0008	-.0036	-.0088	-.0005						
10	.0168	.0413	.0010	.0146	.0357	.0104	-.0116	.0150	-.0499					
11	-.0287	.6895 (.7000)	.0201	.0025	-.0095	.0190	-.0099	.0303	-.0233	.0217				
12	.4438 (.4500)	.4738 (.5000)	.0035	.0111	-.0195	-.0078	.0101	.0006	-.0083	.0139	-.0227			
13	-.3719 (-.3500)	.0070	-.0350	.0309	-.0286	-.0180	-.0236	.0357	.0197	-.0173	-.0101	-.0174		
14	.0223	-.3017 (-.3500)	.0051	.0058	.0221	-.0239	.0292	-.0026	-.0315	.0360	.0083	.0205	-.0279	
15	-.0399	.0294	-.0310	-.0184	-.0518	.0461	-.0183	-.0221	-.0064	.0358	.0313	-.0071	.0088	-.0236

Variable Number

Examination of the entries in Table 3 shows excellent agreement with the target correlation matrix even though no attempt was made to improve on the entries by considering other matrices resulting from a new 100 x 15 matrix of scores $\underset{\sim}{R}$. That is, the user of this method is free to generate as many rank correlation matrices as desired before beginning the actual computer model runs, but for this example, we considered only one such matrix. It is worth noting that the largest difference between the sample correlations and the target correlations is .0518, out of 105 pairs of variables.

5. AN EXAMPLE SHOWING HOW OUTPUT IS AFFECTED

Thus far in this paper the emphasis has been on methodology for making the distribution of the input variables in a simulation study resemble more closely the desired multivariate input distribution, by matching, in some sense, the correlation matrix. Intuitively it seems reasonable to expect that the output from such a simulation study would also resemble more closely the true output, more closely that is than if this method had not been used as inputs. However reasonable such a result might seem, it is not easy (perhaps not possible) to show such a result analytically. Therefore a brief textbook example is used to see if such a modification of the input does in fact result in improved output - improved from the standpoint of being closer to the true answer than if independent input random variables had been used.

For this textbook example a four component random variable from a multivariate normal distribution with $\underset{\sim}{\mu}' = (1,2,2,3)$ and

$$\underset{\sim}{\Sigma} = \begin{bmatrix} 1 & .8 & .3 & .6 \\ .8 & 1 & .4 & .9 \\ .3 & .4 & 1 & .7 \\ .6 & .9 & .7 & 1 \end{bmatrix}$$

was used as input to the function

$$Y = X_1 + X_2X_3 - X_2\ln|X_1| + \exp(X_4/4)$$

The multivariate normal distribution was used because it is about the only multivariate distribution that can be handled in a simulation study, without using approximate methods such as introduced in this paper. The particular values for $\underset{\sim}{\mu}$, $\underset{\sim}{\Sigma}$ and the function Y were arbitrarily selected to create an example.

Output was considered to be the distribution function of Y and the first four moments of Y. The "true" answers were obtained by taking a random sample of size 1000 from the multivariate normal distribution and examining the output. This was compared with the output obtained using the following four cases:

1. A Latin hypercube sample of size 50 was obtained. That is, the univariate normal distributions N(1,1), N(2,1), N(2,1) and N(3,1) were used to obtain 50 observations from each, independently of the others. The variation of the method in Section 2 was used with these observations to induce a Spearman rank correlation matrix resembling $\underset{\sim}{\Sigma}$.
2. Random samples were used in place of Latin hypercube samples in case 1.
3. The same values used in case 1 above were randomly mixed to remove correlation; that is, an uncorrelated Latin hypercube sample was created by generating a random pairing of the values used in case 1.
4. The same values used in case 2 were paired randomly, as in case 3, to obtain uncorrelated random variables.

These samples of size 50 were replicated 10 times to see how the procedures compared with the standard.

Table 4 contains a summary of the estimates of the first four moments of Y. The population values are compared with the values obtained using the four cases. The average of the ten replications, and the standard deviation computed over the ten reps, is given. Note that the use of the method in Section 2 (cases 1 and 2) results in closer estimates of three of the four moments, on the average, and that the standard deviations of the estimates associated with use of the method are smaller in all four cases. In particular, the more interesting estimates are the first two

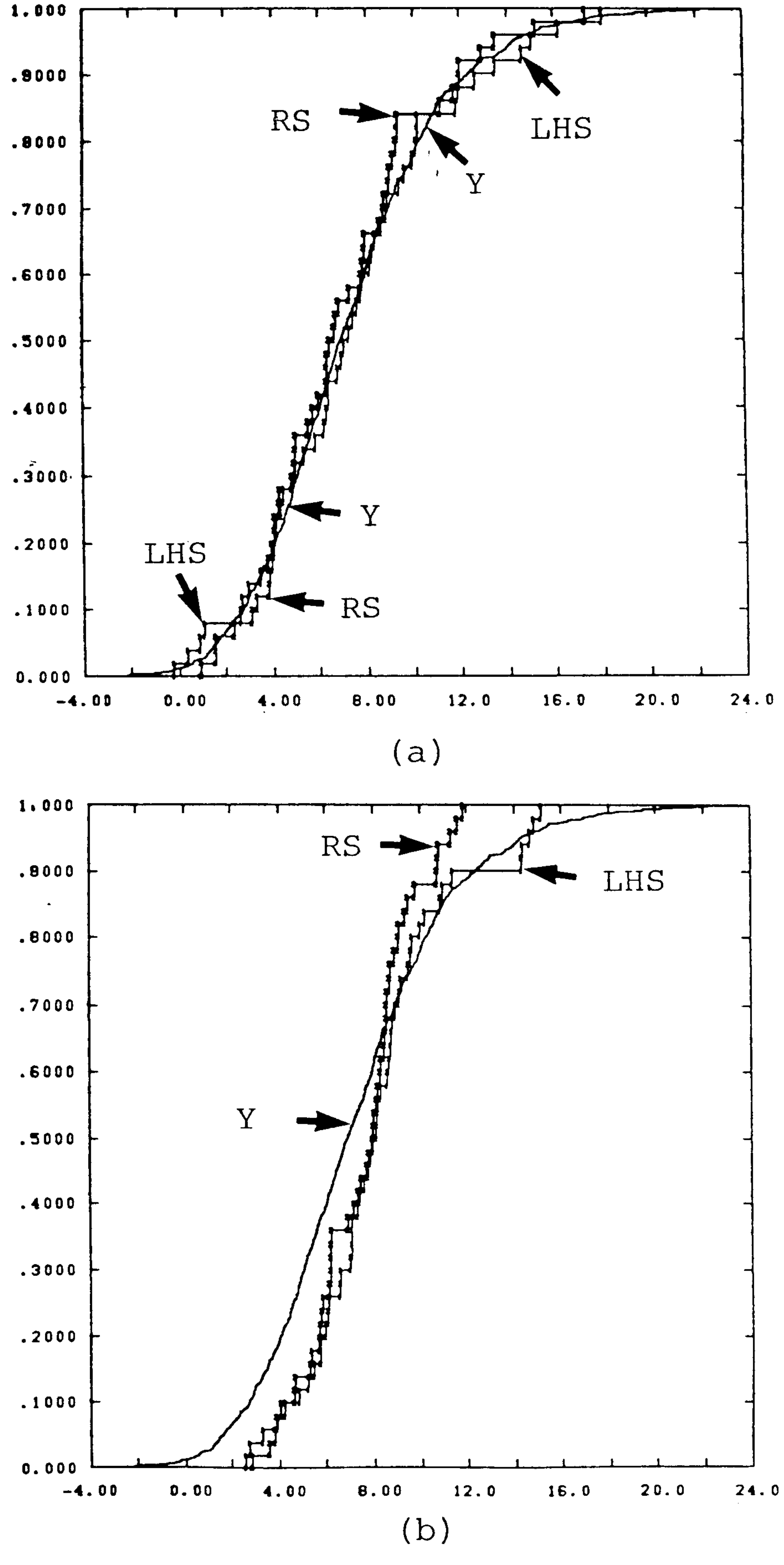

Figure 1. Population Distribution Function (Y) for Textbook Example Along with Estimates for it Based on Latin Hypercube Samples (LHS) and Random Samples (RS) of Size 50 Using (a) Correlated Input and (b) Independent Input.

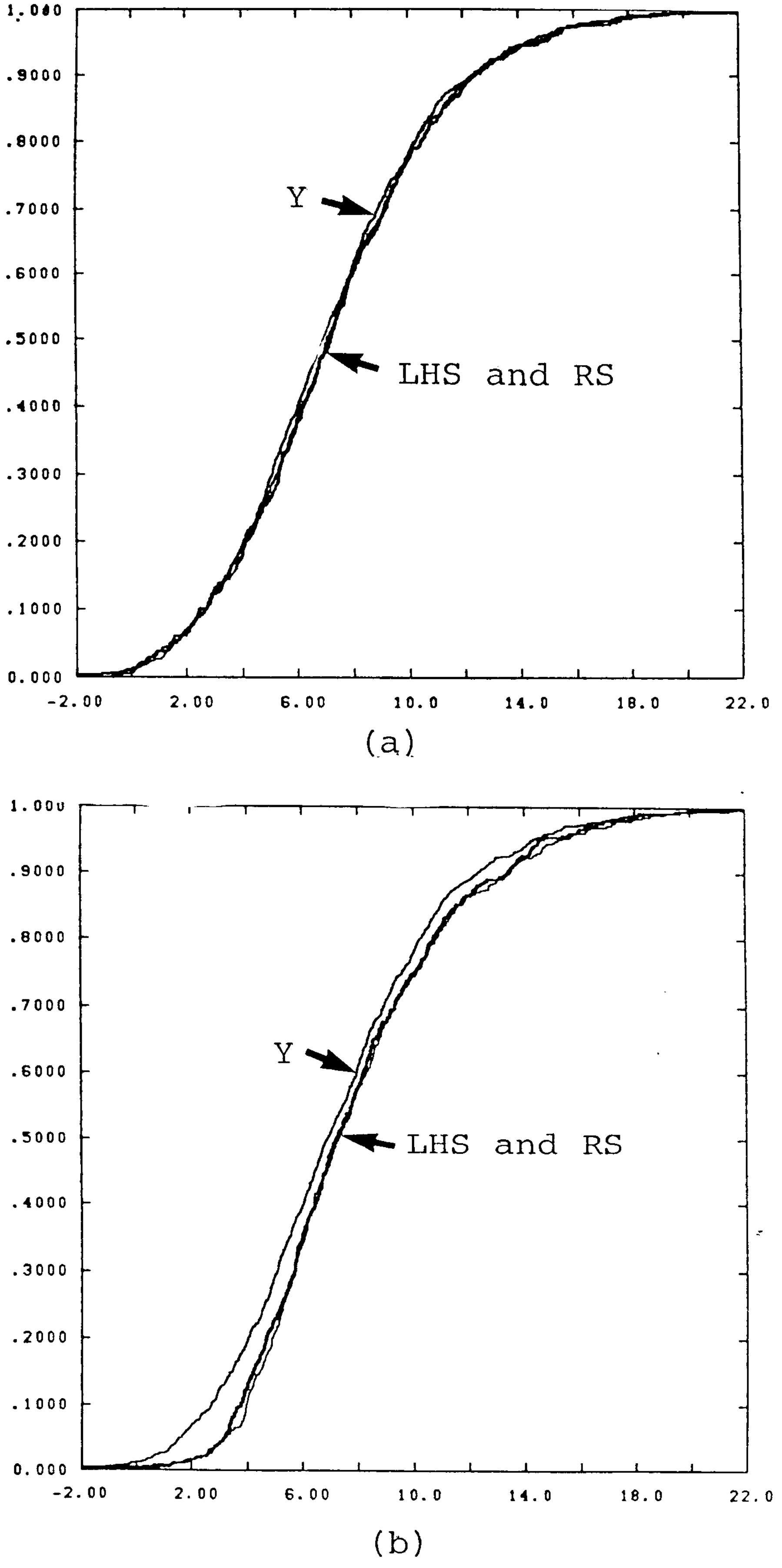

Figure 2. Population Distribution Function (Y) for Textbook Example Along with Estimates for it Based on 10 Latin Hypercube Samples (LHS) and 10 Random Samples (RS) Each of Size 50 Using (a) Correlated Input and (b) Independent Input.

TABLE 4

Summary of Moments Over Reps

		μ	s	M_3	M_4
Population Values		7.29	3.91	40.85	965.9
Case		Estimates of Moments and Their Standard Deviations			
1	LHS	7.39(.06)	3.91(.21)	30.96(17.86)	787.4(334.4)
2	RS	7.50(.60)	3.92(.47)	35.48(25.23)	878.0(652.9)
3	LHS	7.98(.65)	3.70(.50)	43.00(25.47)	725.9(535.7)
4	RS	7.92(.78)	3.62(.72)	42.15(38.62)	728.7(609.4)

moments, which are considerably improved using the method of Section 2.

The c.d.f. of the output is estimated from the empirical c.d.f.'s in Figures 1 and 2 using the first replication. Figure 1(a) compared the "true" c.d.f. with cases 1 and 2. Both cases appear to follow the true curve closely. On the other hand, cases 3 and 4, as depicted in Figure 1(b), appear to underestimate the true c.d.f. for small values of Y. To see if this was a chance occurrence the ten replications were considered again and the average c.d.f.'s were plotted in Figure 2. The same pattern indicated in Figures 1(a) and (b) shows up in Figures 2(a) and (b). That is, results based on the method in Section 2 give a better estimate of the c.d.f. than is obtained using independent input variables, for both random sampling and Latin hypercube sampling - at least for this simple example. Additionally the c.d.f.s in Figure 2(b) are outside of the 95% Kolmogorov bound of .060.

6. SUMMARY AND DISCUSSION

A method for pairing observations on independent random variables in order to induce a desired rank correlation structure is given in this paper. This method, unlike methods based on linear combinations of random variables, preserves the exact marginal distributions, may be used with any distributions, is simple to use, and may be applied to any sampling scheme for which correlated variables could logically be considered. Monte Carlo studies and the application in a computer model with many variables indicate that the expected value of the rank correlation matrix obtained using this procedure is very close to the desired form. We should point out that if the sample rank correlation obtained is not satisfactory to the user, nothing prevents the prospective user from generating several matrices of scores, computing the Spearman rank correlation matrix for each one, and choosing that matrix R that provides the most preferred rank correlations. This approach would permit a pairing of values of input variables that would yield rank correlations as close to the desired structure as the user thinks is necessary. It is worth noting that even if the desired correlation matrix is $\underset{\sim}{I}$, the variation of the method in Section 2 will produce a sample rank correlation matrix which more closely resembles orthogonal input than one would have using a strictly random input.

Although this paper used van der Waerden scores in the examples, we used other scores in order to see what the relative merits of several types of scores might be. If ranks are used as scores, i.e., $a(i) = i$, the variation as expressed in Tables 2 and 3 is smaller than that obtained using van der Waerden scores. However, pairwise plots of the input variables did not appear as "natural," in our opinion, as when van der Waerden scores were used. That is, the resulting bivariate scatter diagrams formed elliptical patterns when van der Waerden scores were used, but appeared to be pinched in the middle and spread out in the tails when ranks were used. The use of random normal deviates instead of scores (a different set for each sample) did not change the bivariate plots noticeably, but resulted in the highest standard deviations of the sample rank correlations of the three types of scores. The intuitive appeal of van der Waerden scores is that they resemble values of normal random variables, for which the relationship between correlation of the ranks and correlation of the data, even after a linear transformation, is well behaved with a correlation approaching $\sqrt{3/\pi}$ as the sample size increases without bound.

In any type of computer modeling involving random sampling of the input variables, whether it is simple random sampling, stratified sampling, or Latin hypercube sampling, the validity of the model output depends to a great extent on how closely the sampled joint distribution of the input variables agrees with the true joint distribution. That is, if a correlation structure exists among the input variables, but the actual sampling takes place as if the input variables were independent, the theoretical properties of the statistics formed from the output may no longer be valid. Estimators intended to be unbiased or consistent may not be. The procedure presented in this paper can be expected to bring the joint distribution of the input variables closer to the true joint distribution than would be attained under the assumption of independence. It should be recognized, however, that by matching marginal distributions and the correlation matrix, one does not match the entire joint distribution function of the input variables, and therefore there is no guarantee that the output will be any closer to the true form than if this methodology were not used at all. That is, the unmatched characteristics of the input distribution may be the dominating characteristics for some aspects of the output. A brief example was used to help alleviate fears of this happening. However, other examples may be invented which possibly show this methodology to be ineffective in improving the output.

While it is true that there is much more to a multivariate input distribution than a mere collection of marginal distributions and a covariance matrix, it is usually not possible to obtain more rigid specifications than those. In fact, it is more usual to find only the marginal input distributions specified, with the correlation matrix defaulted to the identity matrix for simplicity. Since the immediate objective of the simulation study is to come as close to realism as possible, the methods in this paper should be used whenever correlation is appropriate. Of course, if more complete information about the multivariate input distribution is available, it should be used in the sampling scheme if methods are available for incorporating that information.

A recent technical report by Iman, Davenport, and Zeigler (1980) at Sandia National Laboratories provides a user's manual and computer listings for implementing the methods presented in this paper. A copy of this report can be obtained from the first listed author on this paper. In addition the variation of the method in Section 2 has been used in a paper by Iman and Davenport (1982) to provide bivariate plots of correlated random variables with various combinations of marginal distributions and rank correlations.

ACKNOWLEDGMENTS

Funding for this research effort was provided by the Nuclear Regulatory Commission and is part of a project at Sandia National Laboratories to develop the methodology associated with geologic disposal of radioactive waste. The authors wish to express their appreciation to Diane Zeigler and Michael J. Shortencarier at Sandia National Laboratories for their programming assistance. The authors would like to thank James M. Davenport of Texas Tech University for his suggestions which led to the use of general scores in the matrix $\underset{\sim}{R}$ rather than just ranks. The suggestion leading to the variation of the method given in Section 2 was made by R. G. Easterling of Sandia National Laboratories.

BIBLIOGRAPHY

Campbell, J. E. and Cranwell, R. M. (1980). Risk Methodology for Geologic Disposal of Radioactive Waste: A Model for Incorporating Feedback Effects in Salt Dissolution Processes. Tech. Report, SAND80-0067, Sandia National Laboratories, Albuquerque, NM.

Iman, R. L. and Conover, W. J. (1980). Small Sample Sensitivity Analysis Techniques for Computer Models, With an Application to Risk Assessment. Communications in Statistics, A9(17), 1749-1842. Rejoinder to Comments, 1863-1874.

Iman, R. L. and Davenport, J. M. (1982). Rank Correlation Plots for Use With Correlated Input Variables. Communications in Statistics, Vol. B11, No. 3.

Iman, R. L., Davenport, J. M., and Zeigler, D. K. (1980). Latin Hypercube Sampling (A Program User's Guide). Tech. Report SAND79-1473, Sandia National Laboratories, Albuquerque, NM.

Iman, R. L., Helton, J. C., and Campbell, J. E. (1981a). An Approach to Sensitivity Analysis of Computer Models: Part I - Introduction, Input Variable Selection and Preliminary Variable Assessment. Journal of Quality Technology, 13(3), 174-183.

Iman, R. L., Helton, J. C., and Campbell, J. E. (1981b). An Approach to Sensitivity Analysis of Computer Models: Part II - Ranking of Input Variables, Response Surface Validation, Distribution Effect and Technique Synopsis. Journal of Quality Technology, 13(4), 232-240.

Johnson, Mark E. and Ramberg, John S. (1977). Transformations of the Multivariate Normal Distribution With Applications to Simulation. Tech. Report LA-UR-77-2595, Los Alamos Scientific Laboratory, New Mexico.

McKay, M. D., Conover, W. J., and Beckman, R. J. (1979). A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. Technometrics, 21, 239-245.

Scheuer, E. M. and Stoller, D. S. (1962). On the Generation of Normal Random Vectors. Technometrics, 4, 278-281.

NOTE ADDED BY EDITOR:

This paper was presented at the Joint Statistical Meetings in Detroit, Michigan on August 11, 1981. The Section on Physical and Engineering Sciences of the American Statistical Association chose this paper as runner up as the most outstanding presentation at these meetings.

Received January, 1981; Revised November, 1981.

Recommended by Wm. R. Schucany, Southern Methodist University, Dallas, Texas

Refereed Anonymously.