PSYCHOMETRIKA—VOL. 48, NO. 3.
SEPTEMBER, 1983
NOTES AND COMMENTS

SIMULATING MULTIVARIATE NONNORMAL DISTRIBUTIONS

C. DAVID VALE
VINCENT A. MAURELLI

ASSESSMENT SYSTEMS CORPORATION

A method for generating multivariate nonnormal distributions with specified intercorrelations and marginal means, variances, skews, and kurtoses is proposed. As an example, the method is applied to the generation of simulated scores on three psychological tests administered to a single group of individuals.

Key words: random numbers, random number generation.

Monte carlo computer simulations are used to model a variety of real-world statistical and psychometric processes. The value of a simulation is closely tied to the fidelity with which it represents the real-world environment that it attempts to model. All monte carlo procedures require the generation of random numbers. Random number generators typically available are limited to generating random numbers with uniform or normal distributions. However, to adequately replicate the real-world environment, random numbers with distinctly nonnormal distributions are often required.

Fleishman [1978] noted that real-world distributions of variables are typically characterized by their first four moments (i.e., mean, variance, skew, and kurtosis). He presented a procedure for generating nonnormal random numbers with these moments specified. His procedure accomplished this by simply taking a linear combination of a random number drawn from a normal distribution and its square and cube. Using this simple transformation and tables (or a system of equations) provided by Fleishman, random numbers with moments matching a real-world data set could be easily generated.

Tadikamalla [1980] criticized Fleishman's procedure for producing distributions of variables for which the exact distribution was unknown and which thus lacked probability-density and cumulative-distribution functions and which, furthermore, could not produce distributions with all possible combinations of skew and kurtosis. Tadikamalla proposed five alternative procedures for generating nonnormal random numbers and compared all six for speed of execution, simplicity of implementation, and generality. Fleishman's procedure was the easiest to implement and executed most quickly. However, it was less general than some of the other procedures because it lacked a distribution function, thus making some theoretical derivations impossible (calculation of percentiles of the population distribution, for example). Also, its inability to generate some extreme combinations of skew and kurtosis made it less useful in certain conditions.

Occasionally a simulation problem is encountered that requires the generation of multivariate nonnormal random numbers (random numbers with specified intercorrelations as well as specified moments). Fleishman's procedure has an advantage over the other procedures in that it can easily be extended to generate multivariate random numbers with specified intercorrelations and univariate means, variances, skews, and kurtoses.

The authors wish to thank David J. Weiss for his critical review of a draft of this manuscript and Kathleen A. Gialluca and Jack T. Litzau for their assistance in verifying equations and computer programs.

Requests for reprints should be sent to C. David Vale, Assessment Systems Corporation, 2233 University Avenue, Suite 310, St. Paul, Minnesota 55114.

0033-3123/83/0900-4010$00.50/0

This paper presents the details of this extension and evaluates the extension through an example.

A Summary of the Univariate Procedure

Fleishman's technique for generating nonnormal random numbers amounts, in the univariate case, to defining a variable, Y, as a linear combination of the first three powers of a standard normal random variable X:

$$Y = a + bX + cX^2 + dX^3. \tag{1}$$

The constants a, b, c, and d are chosen to provide Y with the specified distributional form. To determine the constants, Fleishman expanded (1) to express the first four moments of the nonnormal variable Y in terms of the first fourteen moments of X. Since X is normally distributed, the first fourteen moments are known constants. With considerable algebraic manipulation, Fleishman was able to represent the solution to the constants of (1) as a system of nonlinear equations. For a standard distribution (i.e., mean of zero, variance of one), the constants b, c, and d are found by simultaneously solving (2), (3), and (4) (Fleishman's Equations 11, 17, and 18) where γ_1 is the desired skew and γ_2 is the desired kurtosis:

$$b^2 + 6bd + 2c^2 + 15d^2 - 1 = 0 \tag{2}$$

$$2c(b^2 + 24bd + 105d^2 + 2) - \gamma_1 = 0 \tag{3}$$

$$24[bd + c^2(1 + b^2 + 28bd) + d^2(12 + 48bd + 141c^2 + 225d^2)] - \gamma_2 = 0. \tag{4}$$

The constant a is determined from (5):

$$a = -c. \tag{5}$$

Univariate nonnormal random numbers are then generated by drawing normal random numbers and transforming them using the constants a, b, c, and d in (1).

Generation of Multivariate Random Numbers

When normal distributions are desired, the generation of multivariate random numbers is accomplished quite simply using a matrix decomposition procedure (cf. Kaiser & Dickman, 1962). A principal-components factorization (or any factorization, for that matter) is performed on the population correlation matrix that is to underlie the random numbers. To generate a multivariate random number, one random number is generated for each component, and each random variable is defined as the sum of the products of the variable's component loadings and the random number corresponding to each of the components. When the random numbers input are normal, the resulting random numbers are multivariate normal with population intercorrelations equal to those of the matrix originally decomposed. This procedure is the converse of computing component scores. Here the component scores are known and the variable scores are computed.

Multivariate nonnormal random numbers can be generated using a combination of the matrix decomposition procedure and Fleishman's method. Multivariate normal random numbers with specified intercorrelations are generated and then univariately transformed to the desired shapes. The two processes interact, however, and the nonnormal numbers have intercorrelations different from the normal ones. To generate random numbers with specified intercorrelations and moments, the effect of nonnormalizing on the correlations must be anticipated and counteracted. In the procedure described below, an appropriate intermediate correlation matrix is determined from the desired matrix and

the nonnormalizing transformations. When multivariate normal random numbers generated with the intermediate correlation matrix are nonnormalized, the resulting nonnormal variables have the desired population intercorrelations and moments.

The procedure begins with the specification of the constants necessary for Fleishman's procedure to turn standard normal variables into standard nonnormal variables. For each variable independently, these are given by the solution of (2), (3), (4), and (5). Calculation of the nonnormal variables can be formulated in matrix notation as shown below. First defining two variables X_1 and X_2 as variables drawn from standard normal populations, the vector $\mathbf{x}$ is defined as powers zero through three of one of them, in this case X:

$$\mathbf{x}' = [1, X, X^2, X^3]. \tag{6}$$

The weight vector $\mathbf{w}'$ contains the power function weights a, b, c, and d:

$$\mathbf{w}' = [a, b, c, d]. \tag{7}$$

The nonnormal variable Y is then defined as the product of these two vectors:

$$Y = \mathbf{w}'\mathbf{x}. \tag{8}$$

Letting $r_{Y_1Y_2}$ be the correlation between two nonnormal variables Y_1 and Y_2 corresponding to the normal variables X_1 and X_2, and noting that, since the variables are standardized, the correlation between Y_1 and Y_2 is equal to their expected cross product

$$\begin{aligned} r_{Y_1Y_2} &= E(Y_1\, Y_2) \\ &= E(\mathbf{w}_1'\, \mathbf{x}_1\, \mathbf{x}_2'\, \mathbf{w}_2) \\ &= \mathbf{w}_1'\, \mathsf{R}\mathbf{w}_2, \end{aligned} \tag{9}$$

where R is the expected matrix product of $\mathbf{x}_1$ and $\mathbf{x}_2'$:

$$\mathsf{R} = E(\mathbf{x}_1\, \mathbf{x}_2') = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & \rho_{X_1X_2} & 0 & 3\rho_{X_1X_2} \\ 1 & 0 & 2\rho_{X_1X_2}^2 + 1 & 0 \\ 0 & 3\rho_{X_1X_2} & 0 & 6\rho_{X_1X_2}^3 + 9\rho_{X_1X_2} \end{bmatrix} \tag{10}$$

Returning to scalar algebra, collecting terms, and using the relation between a and c given in (5), a third-degree polynomial in $\rho_{X_1X_2}$, the correlation between the normal variables X_1 and X_2, results:

$$\begin{aligned} r_{Y_1Y_2} = {} & \rho_{X_1X_2}(b_1b_2 + 3b_1\, d_2 + 3d_1b_2 + 9d_1\, d_2) \\ & + \rho_{X_1X_2}^2(2c_1c_2) + \rho_{X_1X_2}^3(6d_1\, d_2). \end{aligned} \tag{11}$$

Solving this polynomial for $\rho_{X_1X_2}$ provides the intermediate correlation between the two variables X_1 and X_2 required to provide the desired post-transformation correlation $r_{Y_1Y_2}$. These correlations can then be assembled into a matrix of intercorrelations, and this matrix can be decomposed to yield multivariate normal random numbers for input to Fleishman's transformation procedure.

A Numerical Example

An example of the use of this technique results from a simulation of test equating in which scores on three word-analogy tests—one easy, one of medium difficulty, and one

TABLE 1

Characteristics of the Original Data

	Moments				Intercorrelations		
	Mean	Variance	Skew	Kurtosis	Easy	Medium	Difficult
Easy	13.6000	19.2502	-.5485	-.2103	1.0000		
Medium	9.0319	21.3287	.3386	-.9035	.7787	1.0000	
Difficult	5.2340	12.5621	1.0283	.9272	.6159	.6892	1.0000

difficult—were simulated. The score distributions, taken from scores of real people on three real tests, are shown in Table 1.

The mean scores ranged from 13.6 down to 5.2, becoming lower as the tests became progressively more difficult. Variances of the easy and medium difficulty tests were near 20, but the difficult test's variance was near 12. The skew changed from negative to positive with increasing test difficulty. No trends were obvious in the kurtosis, which ranged from about $-.9$ to .9. Correlations among the three tests ranged from .62 to .78.

Generation of the random numbers began by determining the power function constants from (2) through (5) for each of three variables so that each would have moments corresponding to one of the three test scores. The constants for these variables are shown in Table 2.

These constants were then applied to (11) separately for each variable pair to determine an intermediate correlation. The intermediate correlations were assembled into the matrix shown in Table 3. Correlations in this matrix were all higher than those in the original matrix. Differences ranged from .032 to .064.

The intermediate matrix was factored using a principal components decomposition, and three normally distributed standard variables were produced with population correlations equal to those of the intermediate matrix. These three variables were then nonnormalized by applying the previously determined power function constants to (1).

Fifty samples of 2000 random numbers were drawn for each of the three variables. The normal random numbers were generated by applying the Box-Muller transformation [Box & Muller, 1958] to uniformly distributed random numbers generated using a multiple multiplicative congruential algorithm [Wichman & Hill, 1982]. After each multiple of 100 numbers was drawn, the average biases of the 50 means, variances, skews, kurtoses, and intercorrelations were computed. The root-mean-square errors of each of these sample values around the population values were also computed.

TABLE 2

Power Function Constants

	a	b	c	d
Easy	.1148	1.0899	-.1148	-.0357
Medium	-.1014	1.2443	.1014	-.0939
Difficult	-.2107	1.0398	.2107	-.0293

TABLE 3

Intermediate Intercorrelation Matrix

	Easy	Medium	Difficult
Easy	1.0000		
Medium	.8279	1.0000	
Difficult	.6802	.7212	1.0000

Table 4 presents the average biases of the moments and intercorrelations for the three simulated tests at sample sizes of 2000. Bias is defined here as the sample value minus the population value. The entries in Table 4 are all means of 50 replications. Standard errors of these mean biases are given in parentheses for the means, variances, and intercorrelations. Standard errors for skew and kurtosis could not be calculated because the required higher-order moments of the population distribution were unknown. All biases were well within two standard errors of zero and most were within one, for the parameters with known standard errors. Although standard errors of the skews and kurtoses were unknown, the biases were all very near zero. This suggests that the procedure produced the specified population distributions.

Figure 1 shows root-mean-square errors (RMSE) for the four univariate moments plotted as a function of sample size. The relative errors of the means of the three tests ordered as would be expected; the easy and medium tests, which had approximately equal variances, had essentially equivalent RMSE curves while the difficult test, which had a smaller variance, had a consistently smaller RMSE. All three tests had similar variance RMSE curves; although one might expect the difficult test to have a smaller RMSE because of its smaller variance, the leptokurtosis of its distribution cancels this effect out (cf. Lindgren, 1976, p. 216). RMSE curves for the three tests ranked consistently at all sample sizes for skew and kurtosis; the difficult test had the largest error and the medium test had the smallest error. In general, all four of these RMSE curves appeared to converge toward zero as the sample size increased. This tends to confirm that Fleishman's procedure, as implemented here, works properly. These curves also provide information regarding the sampling errors of the four moments for three distributions of distinctly

TABLE 4

Average Bias at a Sample Size of 2000

	Moments				Intercorrelations		
	Mean	Variance	Skew	Kurtosis	Easy	Medium	Difficult
Easy	-.0061 (.0139)	.0381 (.0814)	-.0058	.0016	----		
Medium	.0111 (.0146)	-.0103 (.0706)	.0016	.0060	.0006 (.0012)	----	
Difficult	-.0144 (.0112)	-.0928 (.0679)	.0012	.0124	-.0009 (.0020)	-.0015 (.0017)	----

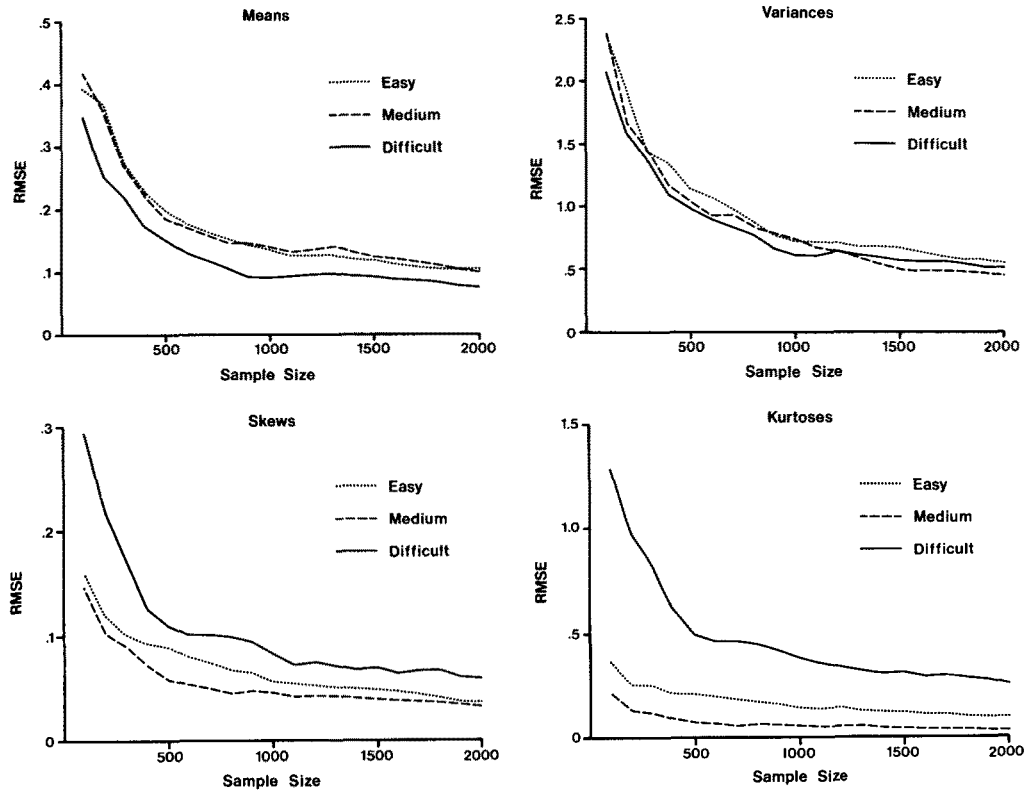

FIGURE 1.
Root mean square errors of univariate sample moments as a function of sample size.

different shapes at a variety of sample sizes. This may be particularly useful for the skew and kurtosis since formulas for their theoretical errors are not readily available.

Figure 2 shows root-mean-square errors for the intercorrelations among the three tests plotted as a function of sample size. The correlation between the easy and medium tests consistently had the smallest RMSE and the correlation between the medium and difficult tests had the largest RMSE. The correlation between the easy and difficult tests, the tests with the most radically different distributions, consistently had the medium RMSE. Like the RMSE curves of the univariate moments, these curves generally appeared to converge toward zero as sample size increased. This, combined with the univariate data, suggests that the multivariate random number generation procedure described in this paper works as expected.

Summary and Conclusions

This paper has suggested a method of generating multivariate nonnormal random numbers based on an extension of Fleishman's [1978] power function method for the univariate case. As demonstrated in a numerical example, the procedure appears to produce random numbers with intercorrelations and univariate moments near specified values. Although the shortcomings of Fleishman's method pointed out by Tadikamalla [1980] also apply to this method, it does provide a means of generating multivariate nonnormal random numbers with specified moments. Such a simple extension of any of the other methods suggested by Tadikamalla is not so obvious.

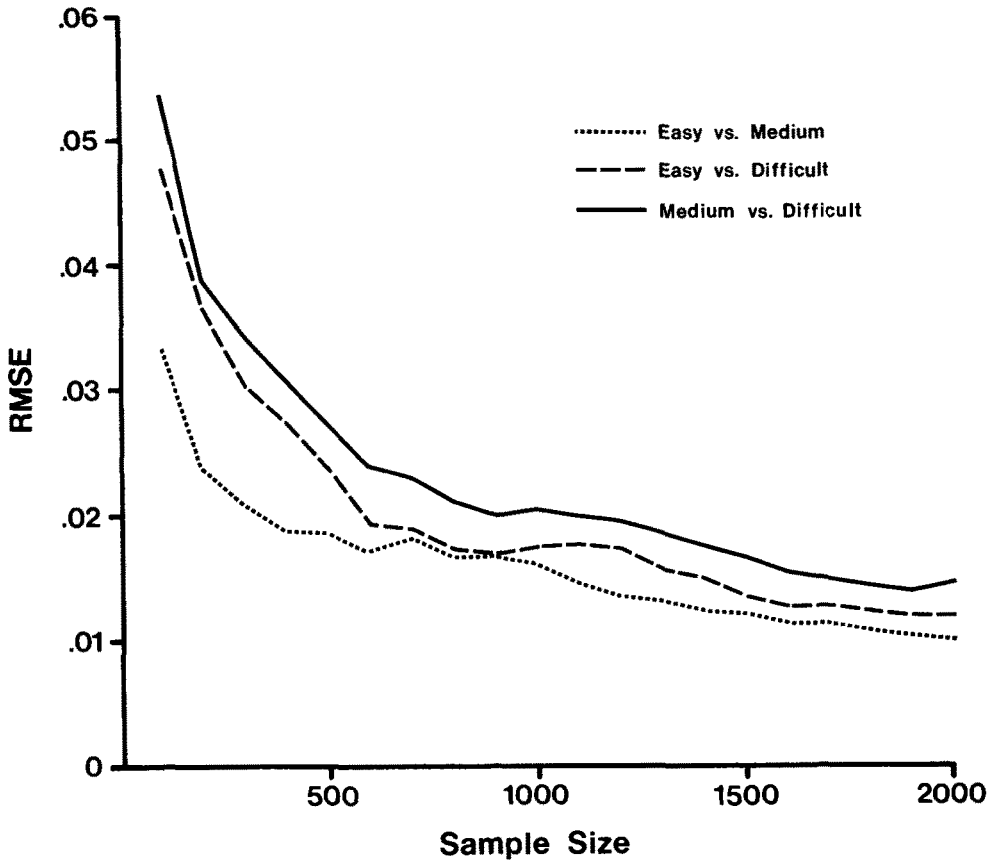

FIGURE 2.
Root mean square errors of sample intercorrelations as a function of sample size.

REFERENCES

Box, G. E. P., & Muller, M. E. A note on the generation of random normal deviates. *Annals of Mathematical Statistics*, 1958, *29*, 610–611.

Fleishman, A. I. A method for simulating nonnormal distributions. *Psychometrika*, 1978, *43*, 521–532.

Kaiser, H. F., & Dickman, K. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika*, 1962, *27*, 179–182.

Lindgren, B. W. *Statistical theory*. New York: Macmillan, 1976.

Tadikamalla, P. R. On simulating nonnormal distributions. *Psychometrika*, 1980, *45*, 273–279.

Wichman, B. A., & Hill, I. D. An efficient and portable pseudo-random number generator. *Applied Statistics*, 1982, *31*, 188–190.

Manuscript received 7/21/81
Final version received 1/13/83