The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*), *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

Generating univariate and multivariate nonnormal data

Sunbok Lee
Center for Family Research
University of Georgia
Athens, GA
sunboklee@gmail.com

Abstract. Because the assumption of normality is common in statistics, the robustness of statistical procedures to the violation of the normality assumption is often of interest. When one examines the impact of the violation of the normality assumption, it is important to simulate data from a nonnormal distribution with varying degrees of skewness and kurtosis. Fleishman (1978, *Psychometrika* 43: 521–532) developed a method to simulate data from a univariate distribution with specific values for the skewness and kurtosis. Vale and Maurelli (1983, *Psychometrika* 48: 465–471) extended Fleishman's method to simulate data from a multivariate nonnormal distribution. In this article, I briefly introduce these two methods and present two new commands, `rnonnormal` and `rmvnonnormal`, for simulating data from the univariate and multivariate nonnormal distributions.

Keywords: st0371, rnonnormal, rmvnonnormal, nonnormal data, skewness, kurtosis

1 Introduction

The assumption of normality is common in statistics. For example, in many statistical models, sampling distributions of statistics of interest or error terms are typically assumed to follow normal distributions. However, the normality assumption may not hold in practice. The asymptotic normal approximation of the sampling distributions may not hold when sample sizes are too small to justify the use of the asymptotic theory. Also variables are often somewhat skewed or kurtotic because of outliers, truncations, or floor and ceiling effects, which may threaten the validity of the normal error terms. Therefore, the robustness of statistical procedures to the violation of the underlying normality assumption is often of interest. Two good examples of this are the studies on the robustness of the ordinary least-squares regression (Jarque and Bera 1987) or structural equation modeling (Finney and DiStefano 2006) to the violation of their normality assumptions. When one investigates the impact of nonnormality, it is important to simulate nonnormal data with varying degrees of skewness and kurtosis. Fleishman (1978) proposed a power method in which a nonnormal random variable Y can be obtained from the linear combination of the first three powers of a standard normal random variable X. Vale and Maurelli (1983) extended Fleishman's power method to simulate multivariate nonnormal variables. In the following sections, I briefly discuss these two methods. I then introduce two new commands (`rnonnormal` and `rmvnonnormal`) that

 st0371

implement the two methods. I also conduct a simulation study to verify the accuracy of `rnonnormal` and `rmvnonnormal`.

2 Univariate nonnormal data

Fleishman (1978) proposed a method for generating data from a univariate nonnormal distribution with specific values of the skewness and kurtosis. In Fleishman's method, a random variable Y with desired values for the skewness and kurtosis is defined by

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

where X is a random variable distributed normally with zero mean and unit variance. That is, Y is expressed by the linear combination of the first three powers of a standard normal random variable X. The key to Fleishman's method is to determine the coefficients a, b, c, and d in such a way that the distribution of Y has desired moments of the first four orders, that is, the mean, variance, skewness, and kurtosis. To do this, Fleishman (1978) expressed the first four moments of Y in terms of the first four moments of X. For example, the first moment of Y can be expressed by

$$E(Y) = a + bE(X) + cE\left(X^2\right) + dE\left(X^3\right)$$

Similarly, other higher moments of Y also can be expressed in terms of the first four moments of X. Because X is assumed to follow a standard normal distribution, its first four moments are known constants: $E(X) = 0$, $E(X^2) = 1$, $E(X^3) = 0$, and $E(X^4) = 3$. Therefore, the coefficients a, b, c, and d in (1) can be determined given the first four moments of Y. More specifically, suppose that the desired four moments of Y are $E(Y) = 0$, $E(Y^2) = 1$, $E(Y^3) = \gamma_1$, and $E(Y^4) = \gamma_2 + 3$, where γ_1 and γ_2 represent the specific values of skewness and kurtosis, respectively. Then the coefficients a, b, c, and d can be determined using the following equations:

$$\begin{aligned}
a + c &= 0 \\
b^2 + 6bd + 2c^2 + 15d^2 - 1 &= 0 \\
2c\left(b^2 + 24bd + 105d^2 + 2\right) - \gamma_1 &= 0 \\
24\left\{bd + c^2\left(1 + b^2 + 28bd\right) + d^2\left(12 + 48bd + 141c^2 + 225d^2\right)\right\} - \gamma_2 &= 0
\end{aligned}$$

Here γ_1 and γ_2 are the desired skewness and kurtosis of the nonnormal variable Y. Note that when one obtains the coefficients a, b, c, and d, real solutions for the coefficients do not always exist for all values of γ_1 and γ_2. For example, when $\gamma_1 = 1.5$, γ_2 must be larger than 2.46 to get real solutions for the system of equations above.

Alternative procedures for generating nonnormal data are available, but Fleishman's method is the easiest to implement and can be executed the fastest. Also Fleishman's method can easily be extended to generate multivariate nonnormal data. One of the limitations of Fleishman's method is that the exact distribution produced is unknown, so it lacks probability density and cumulative distribution functions (Vale and Maurelli 1983; Tadikamalla 1980). In Fleishman's method, nonnormal data are generated using the first four moments of a random variable without knowing the exact distribution of the random variable.

3 Multivariate nonnormal data

Vale and Maurelli (1983) extended Fleishman's method and proposed a method for generating multivariate nonnormal random numbers with desired intercorrelations. Their method consists of two steps: 1) random numbers are generated from a multivariate normal distribution with a specific correlation matrix, which is called an intermediate correlation matrix; and 2) the generated multivariate normal random numbers are univariately transformed by using Fleishman's method to produce multivariate nonnormal random numbers with desired intercorrelations. The key to this method is to find an intermediate correlation matrix in the first step such that the multivariate normal random numbers with the intermediate correlation matrix are univariately transformed to produce the multivariate nonnormal random numbers with desired intercorrelations.

3.1 Generating correlated multivariate normal random numbers

To generate multivariate nonnormal random numbers, we start by generating multivariate normal random numbers with a specific intermediate correlation matrix. Specifically, we consider an n-dimensional random vector $\mathbf{X}$ that is defined as

$$\mathbf{X} = \mathbf{AZ} + \mathbf{b}$$

where $\mathbf{Z}$ is the n-dimensional random vector whose elements are random variables following the standard normal distributions, $\mathbf{A}$ is an $n \times n$ matrix, and $\mathbf{b}$ is the mean vector of $\mathbf{X}$. Given the standard random vector $\mathbf{Z}$, the matrix $\mathbf{A}$ needs to be determined such that the resulting random vector $\mathbf{X}$ transformed from $\mathbf{Z}$ can have a specific intermediate correlation matrix. To do this, we express the covariance matrix of $\mathbf{X}$, which is denoted as $\mathbf{C_X}$, in terms of the matrix $\mathbf{A}$:

$$\mathbf{C_X} = \mathrm{Cov}(\mathbf{X}) = \mathrm{Cov}(\mathbf{AZ} + \mathbf{b}) = \mathbf{A}\mathrm{Cov}(\mathbf{Z})\mathbf{A}' = \mathbf{AA}' \tag{2}$$

We can determine the matrix $\mathbf{A}$ by using the eigenvalue decomposition, in which a real or complex matrix $\mathbf{C}$ is decomposed into the product of three other matrices,

$$\mathbf{C} = \mathbf{UDU}'$$

where $\mathbf{U} = (u_1, \ldots, u_n)$ is a real or complex unitary matrix containing eigenvectors of $\mathbf{C}$, and $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ is a diagonal matrix containing eigenvalues of $\mathbf{C}$. When $\mathbf{C_X}$ is a positive semidefinite covariance matrix (that is, every eigenvalue of $\mathbf{C_X}$ is nonnegative), $\mathbf{C_X}$ can be expressed as

$$\mathbf{C_X} = \mathbf{UDU}' = \mathbf{UD}^{\frac{1}{2}}\mathbf{D}^{\frac{1}{2}}\mathbf{U}' = \left(\mathbf{UD}^{\frac{1}{2}}\right)\left(\mathbf{UD}^{\frac{1}{2}}\right)' \tag{3}$$

where $\mathbf{D}^{1/2} = \text{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$. Therefore, by comparing (2) with (3), we see that $\mathbf{A} = \mathbf{U}\mathbf{D}^{1/2}$. In sum, by decomposing the desired covariance matrix $\mathbf{C}$ with the eigenvalue decomposition, we determine the transformation matrix $\mathbf{A}$. Now the random vector $\mathbf{X}$, which is transformed from $\mathbf{Z}$ using $\mathbf{A}$, has the desired covariance matrix.

3.2 Determining intermediate correlations

Given the multivariate normal random numbers with specified intercorrelations, multivariate nonnormal random numbers can be obtained by applying Fleishman's (1978) power transformation method univariately. One problem with the two-step procedure is that Fleishman's transformation changes the prespecified intercorrelations in the multivariate normal random numbers. As a result, the transformed multivariate nonnormal random numbers no longer have the prespecified intercorrelations. Therefore, the intercorrelations in generating multivariate normal random numbers must be specified such that the transformed nonnormal random numbers have the desired intercorrelations. Vale and Maurelli (1983) derived a formula describing the relationship between the correlation for two normal random variables and the correlation for transformed nonnormal random variables. Suppose that Y_1 and Y_2 are nonnormal random variables that are transformed from the standard normal random variables X_1 and X_2 by using Fleishman's transformation as follows:

$$Y_1 = a_1 + b_1 X_1 + c_1 X_1^2 + d_1 X_1^3$$

$$Y_2 = a_2 + b_2 X_2 + c_2 X_2^2 + d_2 X_2^3$$

Then, the correlation $r_{Y_1 Y_2}$ between Y_1 and Y_2 can be expressed in terms of the correlation $\rho_{X_1 X_2}$ between X_1 and X_2 with the following equation:

$$r_{Y_1 Y_2} = \rho_{X_1 X_2}(b_1 b_2 + 3 b_1 d_2 + 3 d_1 b_2 + 9 d_1 d_2) + \rho^2_{X_1 X_2}(2 c_1 c_2) + \rho^3_{X_1 X_2}(6 d_1 d_2) \quad (4)$$

Given the desired correlation $r_{Y_1 Y_2}$ for the nonnormal random variables, we can determine the intermediate correlation $\rho_{X_1 X_2}$ for the normal random variables by solving (4). Then, with the intermediate intercorrelations, we can obtain the multivariate nonnormal random numbers with desired moments by applying Fleishman's method univariately.

4 rnonnormal and rmvnonnormal commands

4.1 Syntax

The commands `rnonnormal` and `rmvnonnormal` generate univariate and multivariate nonnormal random numbers with specified skewness and kurtosis by implementing the power methods developed by Fleishman (1978) and Vale and Maurelli (1983). The syntax of each command is as follows:

`rnonnormal, n(`#`) skewness(`#`) kurtosis(`#`)`

`rmvnonnormal, n(`#`) skewness(`*vectorname*`) kurtosis(`*vectorname*`)`
`correlation(`*matname*`)`

Options for rnonnormal

`n(`#`)` specifies the sample size of univariate nonnormal random numbers. `n()` is required.

`skewness(`#`)` specifies the skewness of univariate nonnormal random numbers. `skewness()` is required.

`kurtosis(`#`)` specifies the kurtosis of univariate nonnormal random numbers. `kurtosis()` is required.

Options for rmvnonnormal

`n(`#`)` specifies the sample size of multivariate nonnormal random numbers. `n()` is required.

`skewness(`*vectorname*`)` specifies the vector with skewness of each random variable in multivariate nonnormal random variables, where k is the dimension of the vector *vectorname*. `skewness()` is required.

`kurtosis(`*vectorname*`)` specifies the vector with kurtosis of each random variable in multivariate nonnormal random variables, where k is the dimension of the vector *vectorname*. `kurtosis()` is required.

`correlation(`*matname*`)` specifies the matrix of intercorrelations among multivariate nonnormal random variables, where k is the number of rows and columns of the matrix *matname*. `correlation()` is required.

4.2 Stored results

rnonnormal stores the following in r():

Scalars			
r(a)	a in Fleishman's equation	r(b)	b in Fleishman's equation
r(c)	c in Fleishman's equation	r(d)	d in Fleishman's equation
r(skew)	sample skewness	r(kurt)	sample kurtosis
r(sd)	sample standard deviation	r(mean)	sample mean
Matrices			
r(Y)	$n \times 1$ random numbers		

rmvnonnormal stores the following in r():

Matrices			
r(table)	descriptive statistics	r(Y)	$n \times k$ random numbers

4.3 Examples

In this section, I present examples using the commands rnonnormal and rmvnonnormal. I provide seeds for random-number generation for replicability. I also provide sample statistics such as sample skewness, kurtosis, and correlations. The sample statistics can deviate from prespecified values. However, the averages of the sample statistics across replications would be very close to prespecified values for sufficiently large samples, as will be shown by the simulation study in the next section.

▷ **Example**

rnonnormal

```
. set seed 777
. rnonnormal, n(1000) skewness(1.5) kurtosis(3.75)
. return list
scalars:
                r(a) =  -.2210276210126192
                r(b) =  .8658862035231392
                r(c) =  .2210276210126192
                r(d) =  .0272206991580893
             r(kurt) =  3.612271257758691
             r(skew) =  1.452691091582093
               r(sd) =  1.027291300889708
             r(mean) =  .0202128245923377
matrices:
                r(Y) :  1000 x 1
```

◁

▷ **Example**

rmvnonnormal

```
. set seed 735
. matrix C = (1,0.3\0.3,1)
. matrix S = (1.5,2)
. matrix K = (3.5,4)
. rmvnonnormal, n(1000) skewness(S) kurtosis(K) correlation(C)
. return list
  matrices:
              r(table) :  2 x 4
                  r(Y) :  1000 x 2
. matrix list r(table)
r(table)[2,4]
          mean          sd    skewness    kurtosis
Y1  -.02194593   .99951995   1.4408197   2.9355151
Y2   .00421734   1.0344233   1.7600212   3.7932998
. correlate _all
(correlate command was used to check sample correlation matrix)
             |       Y1       Y2
-------------+------------------
          Y1 |   1.0000
          Y2 |   0.3372   1.0000
```

◁

5 Simulation study

In this section, I demonstrate the accuracy of `rnonnormal` and `rmvnonnormal` in terms of obtaining Fleishman's (1978) coefficients and recovering the skewness and kurtosis.

The accuracy of obtaining Fleishman's coefficients

Fleishman (1978) provided a table containing coefficients of the power transformation for various values of skewness and kurtosis. Some selected coefficients in Fleishman's table are listed in table 1 and compared with the values from `rnonnormal`. For each condition of skewness and kurtosis in table 1, Fleishman's coefficients and the ones from `rnonnormal` are exactly the same up to 14 decimal points. Fleishman's coefficients from `rnonnormal` can be checked in the stored results of `rnonnormal`.

Table 1. Fleishman's coefficients

skewness	kurtosis	source	b	c	d
1.50	3.75	F	0.86588620352314	0.22102762101262	0.02722069915809
		R	0.86588620352314	0.22102762101262	0.02722069915809
1.00	2.00	F	0.90475830311225	0.14721081863342	0.02386092280190
		R	0.90475830311225	0.14721081863342	0.02386092280190
0.50	3.25	F	0.78088173005011	0.05749287097856	0.06735271683459
		R	0.78088173005011	0.05749287097856	0.06735271683459
0.00	3.75	F	0.74802080799221	0.00000000000000	0.07787271610187
		R	0.74802080799221	0.00000000000000	0.07787271610187

F = values from Fleishman (1978), R = values from `rnonnormal`

The accuracy of recovering skewness and kurtosis

In this section, I evaluate the accuracy of `rnonnormal` and `rmvnonnormal` in terms of recovering the prespecified values for the skewness and kurtosis. Following Fleishman (1978), to test `rnonnormal`, I set the values for the skewness at 0, 0.25, 0.5, 0.75, 1, and 1.25 and the values for the kurtosis at -1, 0, 1, 2, 3, and 4. I also set the sample sizes for nonnormal random numbers at 10, 25, 50, 100, 200, 1,000, and 2,000 to examine the impact of sample size on the accuracy of recovery. Each simulation condition is replicated 3,000 times. The average of the sample skewness and kurtosis across replications is calculated using the `simulate` command. Note that some simulation conditions such as `skewness()` $= 1$ and `kurtosis()` $= -1$ are omitted from the results because, as previously mentioned, Fleishman's coefficients do not always exist for all values of the skewness and kurtosis. The simulation results for `rnonnormal` are plotted in figures 1 and 2. In the figures, the x axis indicates the true values for the skewness and kurtosis, and the y axis indicates the estimated values of the skewness and kurtosis. Therefore, simulation results that are close to the $y = x$ line indicate that the true and estimated values are close to each other. For all combinations of skewness and kurtosis, the averages of the estimated values across 3,000 replications are quite close to the true values when the sample sizes are larger than 1,000. This can also be checked from figure 5, in which the errors between true and estimated values for some cases are plotted.

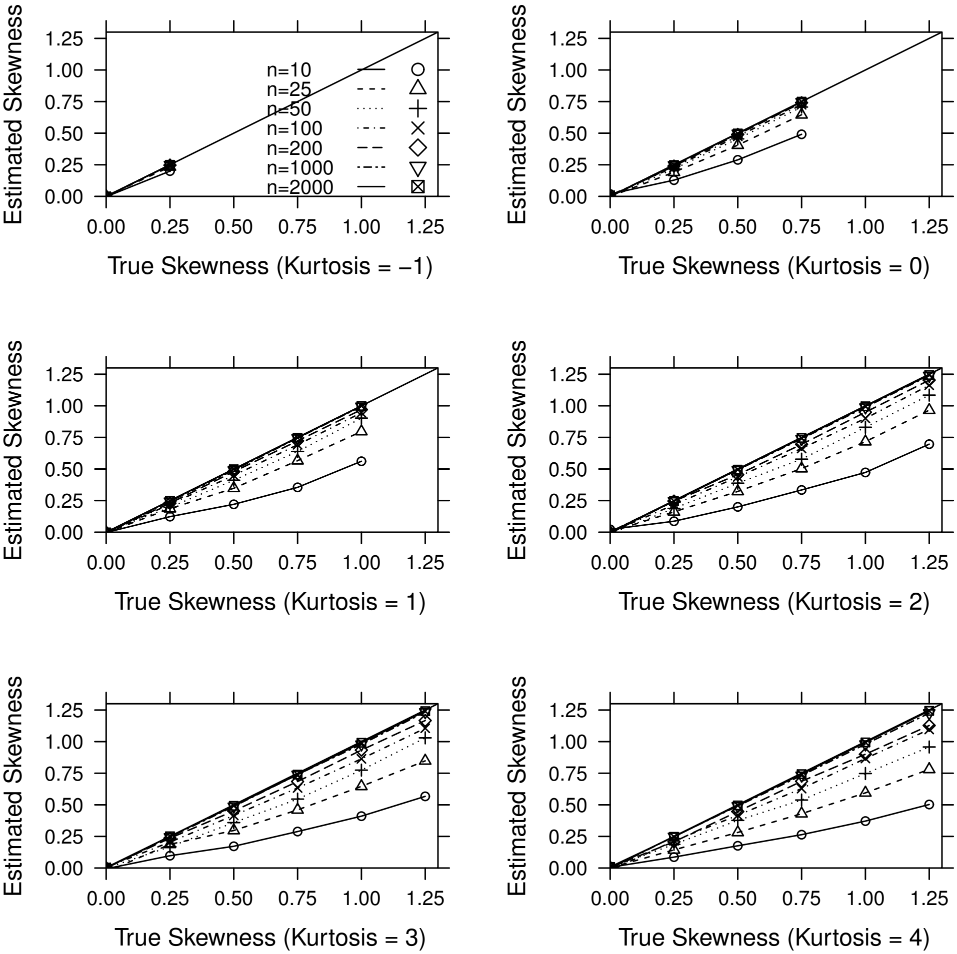

Figure 1. Estimated skewness from `rnonnormal`

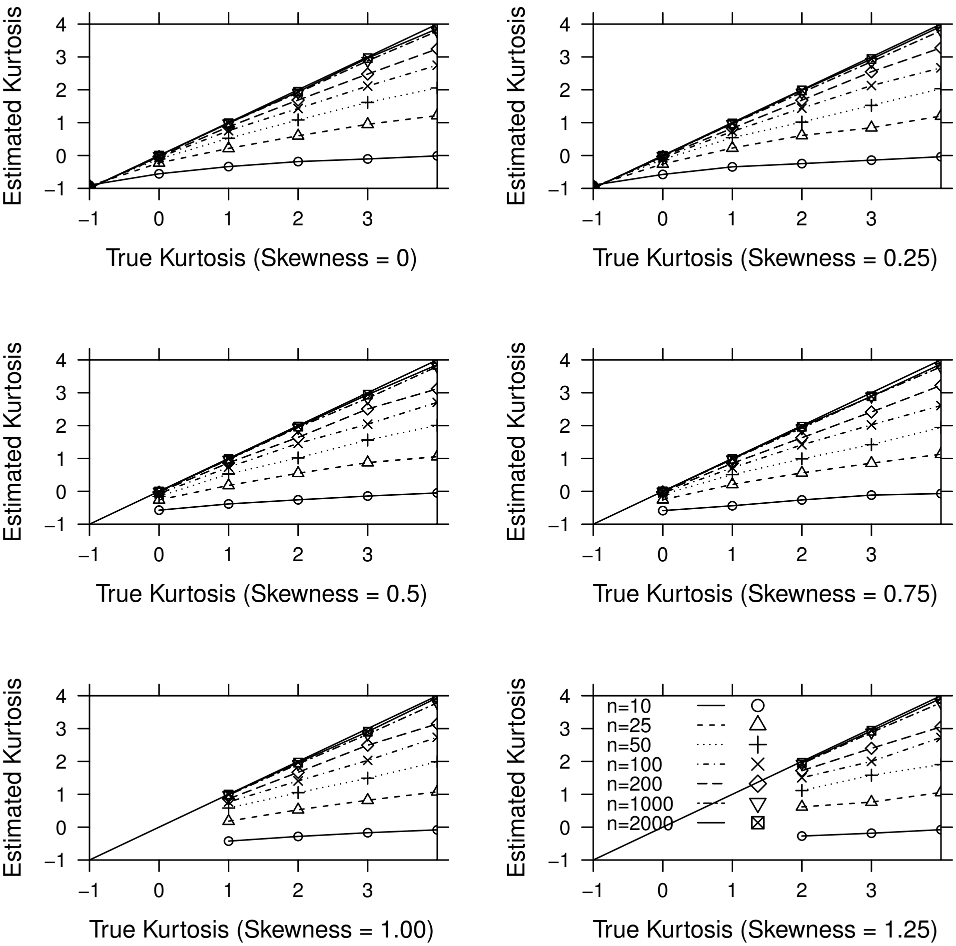

Figure 2. Estimated kurtosis from `rnonnormal`

To test the accuracy of `rmvnonnormal`, I use the command to generate bivariate non-normal random samples. The correlation between 2 nonnormal random variables is set at 0.3. The values for the skewness and kurtosis for 1 nonnormal random variable are fixed to 1.5 and 3.0, respectively. The values for the skewness and kurtosis for the other nonnormal random variable are manipulated to have the same values as in the simulation for `rnonnormal`. The simulation results for `rmvnonnormal` are plotted in figures 3 and 4, and the results are similar to those from `rnonnormal`. The averages of the estimated values across 3,000 replications are quite close to the true values when the sample sizes are larger than 1,000, which also can be observed in figure 5.

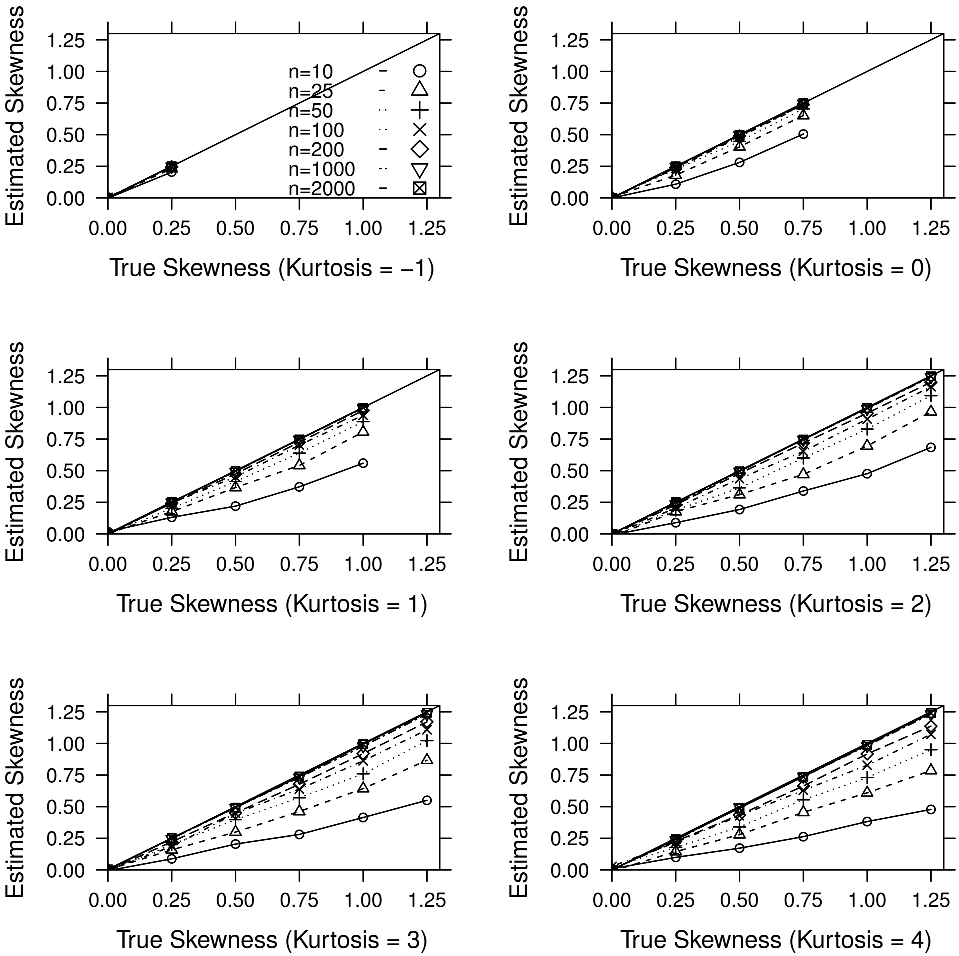

Figure 3. Estimated skewness from `rmvnonnormal`

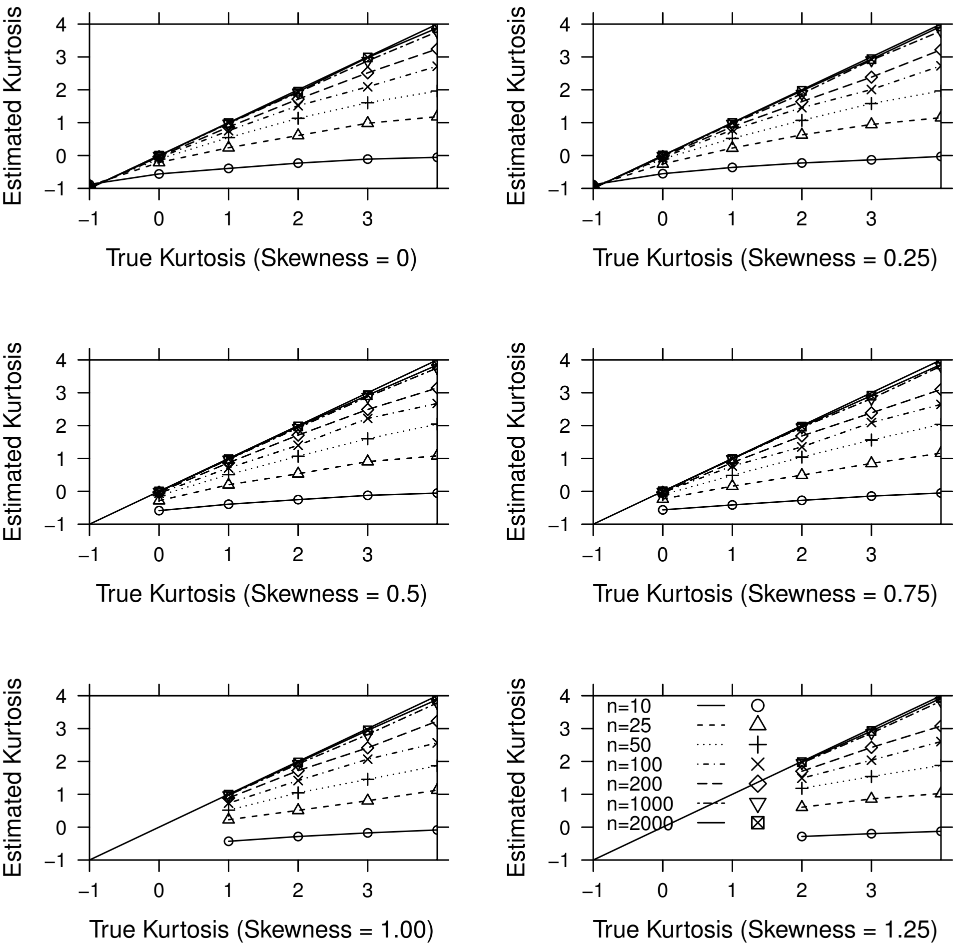

Figure 4. Estimated kurtosis from `rmvnonnormal`

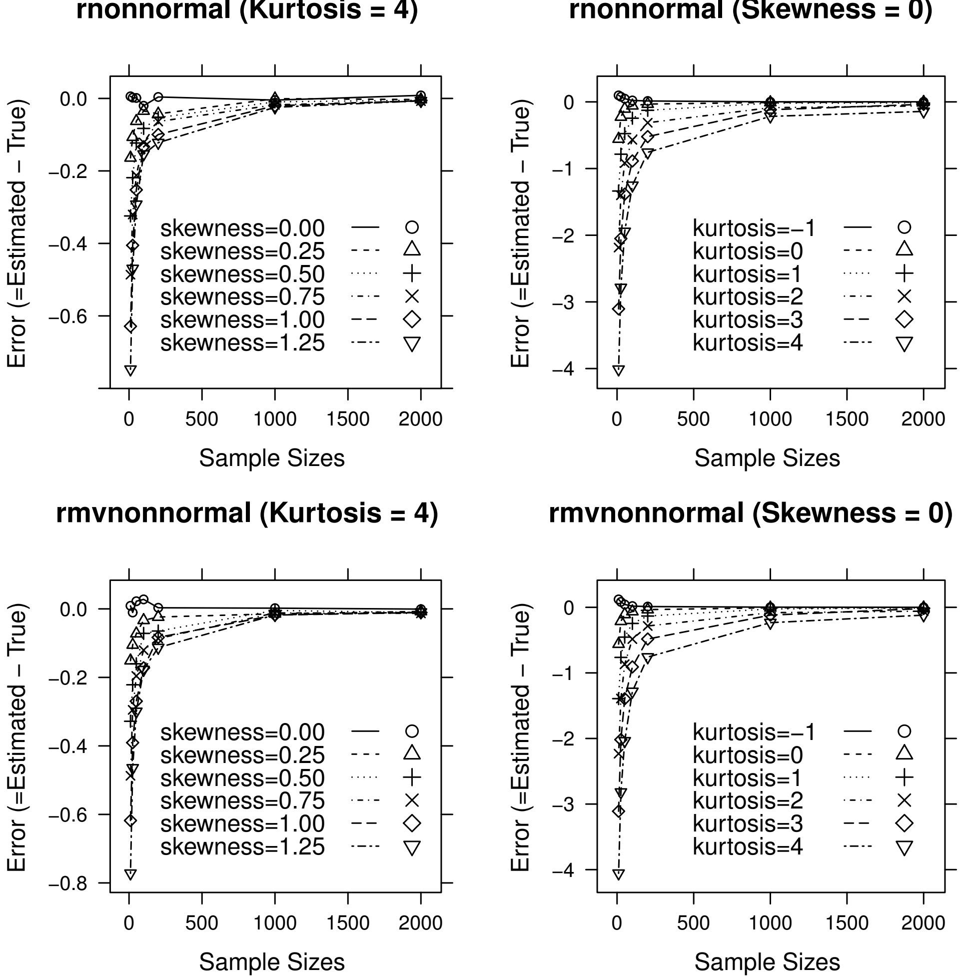

Figure 5. Errors of `rnonnormal` and `rmvnonnormal` for some cases

6 Conclusion

In this article, I briefly introduced two methods for simulating univariate and multivariate nonnormal data. Fleishman (1978) proposed a method for simulating univariate nonnormal data by transforming random numbers sampled from a standard normal distribution. Vale and Maurelli (1983) extended Fleishman's univariate method to simulate multivariate nonnormal data with specified values for the skewness, kurtosis, and intercorrelations. I then introduced two new commands, `rnonnormal` and `rmvnonnormal`, and demonstrated their accuracy by implementing them in example simulation studies. According to the simulation results, on average, `rnonnormal` and `rmvnonnormal` recover the true values for the skewness and kurtosis when the sample sizes are larger than 1,000.

7 References

Finney, S. J., and C. DiStefano. 2006. Non-normal and categorical data in structural equation modeling. In *Structural Equation Modeling: A Second Course*, ed. G. R. Hancock and R. O. Mueller, 269–314. Greenwich, CT: Information Age Publishing.

Fleishman, A. I. 1978. A method for simulating non-normal distributions. *Psychometrika* 43: 521–532.

Jarque, C. M., and A. K. Bera. 1987. A test for normality of observations and regression residuals. *International Statistical Review* 55: 163–172.

Tadikamalla, P. R. 1980. On simulating non-normal distributions. *Psychometrika* 45: 273–279.

Vale, C. D., and V. A. Maurelli. 1983. Simulating multivariate nonnormal distributions. *Psychometrika* 48: 465–471.

About the author

Sunbok Lee received his PhD in Educational Psychology with a quantitative methods major from the University of Georgia, where he is now a postdoctoral associate in the Center for Family Research.