5-2016

A Percentile-Based Power Method in SAS: Simulating Multivariate Non-Normal Continuous Distributions

Jennifer Koran
Southern Illinois University Carbondale

Todd C. Headrick
Southern Illinois University Carbondale, headrick@siu.edu

Material in this article is copyrighted and initially published by *JMASM*, Vol. 15, No. 1, 836-847. ISSN 1538 – 9472. JMASM Inc PO Box 48023 Oak Park, MI 48237 ea@jmasm.com.

Journal of Modern Applied Statistical Methods
May 2016, Vol. 15, No. 1, 836-847.

ISSN 1538 − 9472

JMASM Algorithms and Code
A Percentile-Based Power Method in SAS: Simulating Multivariate Non-Normal Continuous Distributions

Jennifer Koran
Southern Illinois University
Carbondale, Illinois

Todd C. Headrick
Southern Illinois University
Carbondale, Illinois

The conventional power method transformation is a moment-matching technique that simulates non-normal distributions with controlled measures of skew and kurtosis. The percentile-based power method is an alternative that uses the percentiles of a distribution in lieu of moments. This article presents a SAS/IML macro that implements the percentile-based power method.

Keywords: Monte Carlo, non-normal, percentiles, polynomial, SAS, simulation

Introduction

Fleishman (1978) introduced the power method, an elegant and convenient approach for simulating non-normal data by applying a third-order polynomial transformation to a normally distributed random variable. Multivariate extensions of this approach were later developed in Vale and Maurelli (1983) and Headrick and Sawilowsky (1999). A limitation of this conventional power method approach is that the third- and fourth-order moments of the distribution(s) must be available in order to implement the method (see Headrick & Sawilowsky, 2000).

Obtaining these higher-order moments is not a problem if individual data are available to the researcher. However, given privacy concerns and data restrictions, especially in education and health care, a researcher may desire to simulate data to match descriptive statistics available from publicly available reports. As third- and fourth-order moments are typically of little interest to the general public, these higher-order moments may not be part of publicly available reports.

Dr. Koran is in the Section on Statistics and Measurement. Email her at: jkoran@siu.edu. Dr. Headrick is in the Section on Statistics and Measurement, and is an Assistant Editor of this journal.

To address this situation, Koran, Headrick, and Kuo (2015) introduced a multivariate power method that uses the percentiles of a distribution in place of moments. This approach relies upon the 10th, 25th, 50th, 75th, and 90th percentiles of each variable and can accommodate either Pearson or Spearman correlations between the variables. Further, Koran, Headrick, and Kuo (2015) showed that the percentile-based power method exhibits substantially lower relative bias than the conventional moment-based power method. Thus, the percentile-based power method is preferable to the conventional moment-based power method even when third- and fourth-order moments are available, as it mimics the characteristics of the original data more accurately. In light of the prior theoretical presentation of the percentile-based power method, this article focuses on the presentation of %simPPM, a SAS/IML (SAS Institute, 2013) macro that implements the percentile-based power method to produce the simulated data file.

The Percentile-based Power Method

This section presents the essential elements in applying the percentile-based power method to simulate non-normal univariate and multivariate data distributions. The third - order power method transformation is expressed as (Headrick, 2010, pp.12 - 13)

$$p(Z) = \sum_{i=1}^{4} c_i Z^{i-1} \tag{1}$$

where $Z \sim$ i.i.d. $N(0, 1)$ random variable with standard normal pdf ($\phi(z)$) and cdf ($\Phi(z)$). For the purposes contained herein, we assume that (1) is a strictly increasing, monotonic function. As such, an inverse function p^{-1} exists. Thus, the cdf of (1) is $F(p(z)) = (\Phi(z))$. Differentiating $F(p(z))$ with respect to z yields the pdf of (1), which is $f(p(z)) = \phi(z)/p'(z)$.

Univariate Non-normal Data Generation

The percentile-based power method begins with five percentile values ($q(x)_u$) that are used to produce estimates of the following location, scale, and shape parameters (Karian & Dudewicz, 2011, pp. 172-173)

$$\hat{\gamma}_1 = q(x)_{0.50} \tag{2}$$

$$\hat{\gamma}_2 = q(x)_{0.90} - q(x)_{0.10} \tag{3}$$

$$\hat{\gamma}_3 = \frac{q(x)_{0.50} - q(x)_{0.10}}{q(x)_{0.90} - q(x)_{0.50}} \tag{4}$$

$$\hat{\gamma}_4 = \frac{q(x)_{0.75} - q(x)_{0.25}}{\hat{\gamma}_2} \tag{5}$$

where (2)-(5) are the (i) median, (ii) inter-decile range, (iii) left-right tail-weight ratio (a skew function), and (iv) tail-weight factor (a kurtosis function), respectively.

The location, scale, and shape parameters computed from the percentile values are subsequently used, along with the standard normal constants of $z_{0.90} = 1.281\ldots$ and $z_{0.75} = 0.6744\ldots$, to solve for the following four coefficients (Koran, Headrick, & Kuo, 2015, Equations 47-50)

$$c_1 = \hat{\gamma}_1 \tag{6}$$

$$c_2 = \frac{\hat{\gamma}_2\left(\hat{\gamma}_4 z_{0.90}^3 - z_{0.75}^3\right)}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{0.75}^3} \tag{7}$$

$$c_3 = \frac{\hat{\gamma}_2\left(1-\hat{\gamma}_3\right)}{2\left(1+\hat{\gamma}_3\right)z_{0.90}^2} \tag{8}$$

$$c_4 = -\frac{\hat{\gamma}_2\left(\hat{\gamma}_4 z_{0.90} - z_{0.75}\right)}{2z_{0.90}^3 z_{0.75} - 2z_{0.90} z_{.075}^3} \tag{9}$$

However, it is important to check that the solved values of the coefficients form a valid pdf under the conditions described earlier for the power method. In order to produce a valid non-normal pdf, the following criteria must be met (Koran, Headrick, & Kuo, 2015)

$$\begin{aligned} 0 < c_2 < 1 \\ 0 < c_4 < 0.608875 \\ c_3^2 - 3c_2c_4 < 0 \end{aligned} \tag{10}$$

Provided that the criteria in (10) are met, the power method coefficients from (6) - (9) are subsequently substituted into the third-order power method polynomial transformation in (1). The transformed values $p(Z)$ then represent a sample of size n from the univariate distribution defined by the original percentile values $(q(x)_u)$.

Multivariate Non-normal Data Generation

If it is desired to simulate multivariate data that will have a specified correlation structure following the transformation in (1), then standard normal random variables will first have to be simulated with an appropriate intermediate correlation structure. For each pair of variables Z_j, Z_k with $j \neq k$, the intermediate (Pearson) correlation r_{jk} can be determined by solving the following expression for r_{jk} (Headrick, 2010, Equation 4.34, p.114; Koran, Headrick, & Kuo, 2015, Equation 58)

$$\xi_{jk} = \frac{6}{\pi}\left\{\left(\frac{n-2}{n-1}\right)\sin^{-1}\left(\frac{r_{jk}}{2}\right)+\left(\frac{1}{n-1}\right)\sin^{-1}\left(r_{jk}\right)\right\} \tag{11}$$

Where ξ_{jk} is fixed to a specified Spearman correlation.

In many cases a specified Pearson correlation is known instead of a specified Spearman correlation. This requires a different expression to be solved for the intermediate correlation. This new expression uses the following reduced form of Headrick's (2010) Equation 2.59 (p.30)

$$\begin{aligned} E\left[p\left(Z_j\right)p\left(Z_k\right)\right] = &\left(c_{j1}\left(c_{k1}+c_{k3}\right)+c_{k3}\left(c_{k1}+c_{k3}\right)\right) \\ &+r_{jk}\left(c_{j2}c_{k2}+3c_{j4}c_{k2}+3c_{j2}c_{k4}+9c_{j4}c_{k4}\right) \\ &+r_{jk}^2\left(2c_{j3}c_{k3}\right)+r_{jk}^3\left(6c_{j4}c_{k4}\right) \end{aligned} \tag{12}$$

where the fifth-order polynomial expression in Headrick (2010) has been reduced to the appropriate expression for a third-order polynomial and the coefficients $c_{j1} - c_{j4}$ and $c_{k1} - c_{k4}$ are the solved values of (6) - (9) for Z_j and Z_k, respectively.

The expression in (12) for using a specified Pearson correlation to solve for the intermediate (Pearson) correlation also uses the following expressions for computing the mean *m* and variance *v* from the percentile-based power method constant coefficients in (6) - (9) as

$$m_j = c_{j1} + c_{j3} \tag{13}$$

and

$$v_j = c_{j2}^2 + 2c_{j3}^3 + 6c_{j2}c_{j4} + 15c_{j4}^2 \tag{14}$$

Expressions analogous to (13) and (14) may be used for computing m_k and v_k, respectively. Thus, when a Pearson correlation is specified, the intermediate (Pearson) correlation r_{jk} can be found by substituting the expressions from (12), (13), and (14) into the following expression and solving for r_{jk}

$$\rho_{jk} = \frac{E\left[p\left(Z_j\right) p\left(Z_k\right)\right] - m_j m_k}{\sqrt{v_j v_k}} \tag{15}$$

Where ρ_{jk} is fixed to a specified Pearson correlation.

To apply the multivariate percentile-based power method, first compute the intermediate correlation(s) from either specified Spearman or Pearson correlation(s) in (11) or (15), respectively. Then, simulate standard normal random variables based on a Cholesky factorization of the intermediate correlation(s). Finally, apply the transformation in (1) to the individual simulated standard normal random variables. The resulting transformed values then represent a sample from the multivariate distribution with the specified correlation structure.

The Percentile-based Power Method in SAS/IML

The SAS/IML macro %simPPM implements the procedures described in the previous section. This %simPPM macro can be accessed by your program by including the following lines

```
filename simppm "directory of file simPPM";
%include simppm(simPPM) / nosource2;
```

Calling %simPPM uses the following pieces of information: 1) the number of variables, 2) the file path and name of an external ASCII file with the percentiles, 3) the file path and name of an external ASCII file with the specified correlations (multivariate method only), 4) an indication of whether the specified correlations are Pearson or Spearman (1 for Pearson, 2 for Spearman; multivariate method only), 5) the desired sample size, 6) a random number seed (optional), and 7) the file path and name for the ASCII output file containing the simulated data.

Data in the percentiles and correlations files should be space delimited, and there cannot be any missing values. The percentiles are laid out such that there are five rows, the 10th percentile in the first row, the 25th percentile in the second row, the 50th percentile in the third row, the 75th percentile in the fourth row, and the 90th percentile in the fifth row. There are as many columns in the percentiles file as there are variables to be simulated. The correlations file has as many rows and columns as there are variables to be simulated, with the variables appearing in the same order as in the percentiles file, and the correlations should be arranged in a full symmetric matrix with ones on the diagonal. Examples of the layout for the percentiles and correlations files are shown in the appendix and explained further in the next section.

Examples

Univariate Example

Suppose we wish to simulate the Idaho Standards Achievement Test (ISAT) mathematics scale scores for 25 third grade students. The 2011 scale score to percentile rank conversion tables for the ISAT are publicly available (Stoneberg, 2011). We arranged the 10th, 25th, 50th, 75th, and 90th percentiles in the ASCII file ex1percentiles.txt for analysis as shown in the appendix. With the ex1percentiles.txt file saved in the folder C:\SAS\, the %simPPM call for this example is:

```
%simPPM(1, C:\SAS\ex1percentiles.txt, , , 25, 54321, C:\SAS\ex1simdata.txt)
```

The "1" indicates that there is only one variable to be simulated. The "C:\SAS\ex1percentiles.txt" gives the file path and name of the ASCII file containing the percentiles. The next two arguments are left blank, as these are only needed for multivariate applications of the percentile-based power method. The "25" indicates that we would like a sample size of 25 for the simulated data.

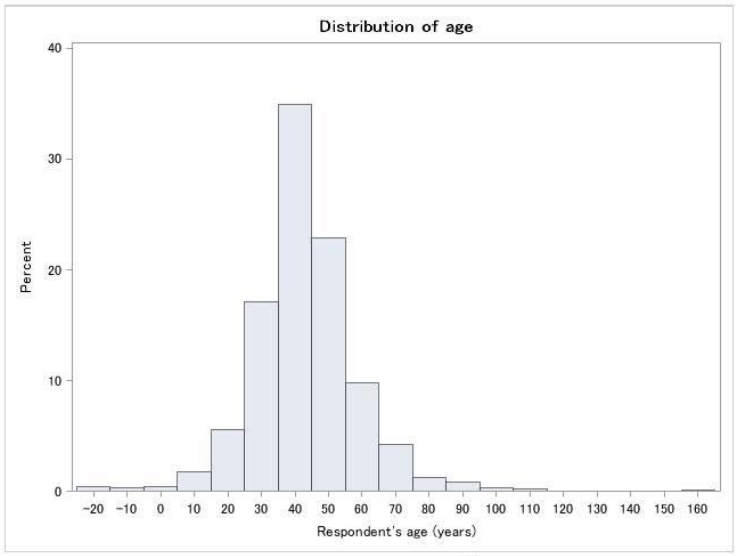
The "54321" is a random number seed to use in generating the data. Alternately, the random number seed may be left blank. If the random number seed is not specified, SAS/IML will choose a random number seed, but the generated data will not be able to be replicated exactly because the random number seed will not be known. It is recommended that the user specify a random number seed. The "C:\SAS\ex1simdata.txt" indicates the file path and name of the ASCII file where the simulated data are to be stored. When the macro call is submitted successfully to SAS/IML, the percentile power method coefficients are produced as shown in the appendix. As this is a small example with only 25 cases, the contents of ex1simdata.txt after submitting the macro call are also shown in the appendix.

Multivariate Example

For the next example, percentiles and Pearson correlations were obtained from n = 527 respondents in the 2012 General Social Survey (Smith, Marsden, Hout, & Kim, 2013) who provided responses to two items. The first item was the respondent's age (AGE). The second item asked, "Approximately how much money or the cash equivalent of property have you contributed in each of the fields listed in the past 12 months? b. Education" (TOTEDUC) (Smith, Marsden, Hout, & Kim, 2013). Suppose we desired to simulate 1000 responses to these two survey items.

The 10th, 25th, 50th, 75th, and 90th percentiles were arranged for the AGE and TOTEDUC variables in the ASCII file ex2percentiles.txt for analysis as shown in the appendix, and the specified Pearson correlation matrix for the AGE and TOTEDUC variables in the ASCII file ex2correlations.txt for analysis as shown in the appendix. With the ex2percentiles.txt and ex2correlations.txt files saved in the folder C:\SAS\, the %simPPM call for this example is:

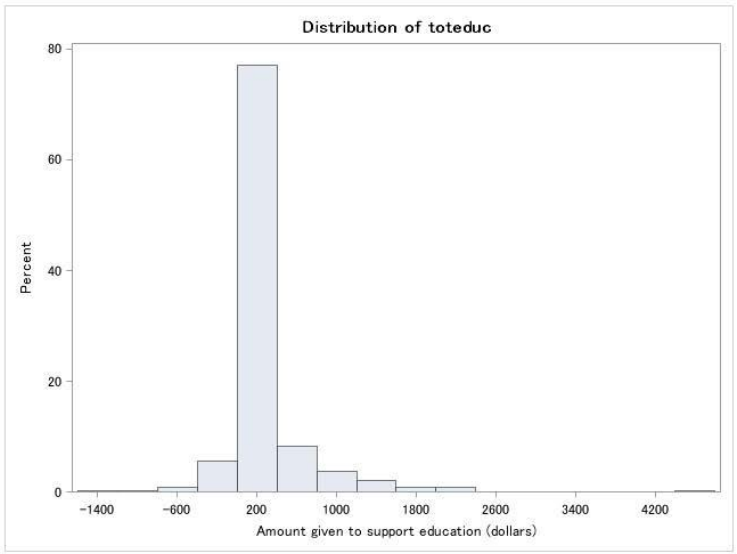
```
%simPPM(2, C:\SAS\ex2percentiles.txt, C:\SAS\ex2correlations.txt, 1, 1000,
7654321, C:\SAS\ex2simdata.txt)
```



Percentiles

10th	26.700
25th	34.787
50th	41.984
75th	50.998
90th	60.617

A



Percentiles

10th	34.588
25th	74.442
50th	95.034
75th	237.418
90th	637.174

B

Figure 1. Distributions of 1000 cases of simulated data for two variables. Panel A depicts the variable AGE, and Panel B depicts TOTEDUC.

The "2" indicates that there are two variables to be simulated. The "C:\SAS\ex2percentiles.txt" gives the file path and name of the ASCII file containing the percentiles. The "C:\SAS\ex2correlations.txt" gives the file path and name of the ASCII file containing the specified correlation matrix. The "1" indicates that specified Pearson correlations are being provided in ex2correlations.txt (alternately a "2" here would indicate specified Spearman correlations). The "1000" indicates that we would like a sample size of 1000 for the simulated data. The "7654321" is a random number seed to use in generating the data. The "C:\SAS\ex2simdata.txt" indicates the file path and name of the ASCII file where the simulated data are to be stored. When the macro call is submitted successfully to SAS/IML, the percentile power method coefficients and intermediate correlation matrix are produced as shown in the appendix.

The file ex2simdata.txt after submitting the macro call contains 1000 cases with two variables. The correlation between AGE and TOTEDUC is 0.0847 in the simulated data. Descriptive statistics and graphs summarizing the distribution of the simulated data are shown in Figure 1.

Note that a small proportion of the cases of AGE are invalid (< 18 years), and a small proportion of the cases of TOTEDUC are invalid (< $0). The distribution may be truncated and cases associated with these invalid values removed from the simulated data at the user's discretion.

Conclusion

The SAS/IML macro %simPPM simulates data to match univariate and multivariate non-normal distributions defined by percentiles and correlations. Unlike other power method approaches used for simulating non-normal data, %simPPM does not require the user to have direct access to conventional measures of skew and kurtosis, making this an ideal approach for simulating data to match distribution information available in public reports without direct access to individually identifiable data.

References

Fleishman, A. I. (1978). A method for simulating non-normal distributions. *Psychometrika, 43*(4), 521-532. doi:10.1007/BF02293811

Headrick, T. C. (2010). *Statistical simulation: Power method polynomials and other transformations*. Boca Raton, FL: CRC Press.

Headrick, T. C., & Sawilowsky, S. S. (1999). Simulating correlated multivariate nonnormal distributions: Extending the Fleishman power method. *Psychometrika*, *64*, 25-35. doi:10.1007/BF02294317

Headrick, T. C., & Sawilowsky, S. S. (2000). Weighted simplex procedures for determining boundary points and constants for the univariate and multivariate power methods. *Journal of Educational and Behavioral Statistics*, *25*(4), 417-436. doi:10.3102/10769986025004417

Karian, Z. A., & Dudewicz, E. J. (2011). *Handbook of fitting statistical distributions with R*. Boca Raton FL: CRC Press.

Koran, J., Headrick, T. C., & Kuo, T. C. (2015). Simulating univariate and multivariate nonnormal distributions through the method of percentiles. *Multivariate Behavioral Research*, *50*(2), 216-232. doi:10.1080/00273171.2014.963194

SAS Institute (2013). *Base SAS 9.4 procedures guide: Statistical procedures*. Cary, NC: author. http://www.sas.com/.

Smith, T. W., Marsden, P. V., Hout, M. & Kim, J. (2013). General Social Survey 1972-2012 (SDA 4.0) [Data file].Retrieved from http://sda.berkeley.edu/archive.htm.

Stoneberg, B. (2011). Idaho Standards Achievement Tests Scale Score to Percentile Rank Conversion Tables Spring 2011. Boise, ID:

Idaho State Department of Education Assessment Division.

Vale, C. D., & Maurelli, V. A. (1983). Simulating multivariate nonnormal distributions. *Psychometrika*, *48*(3), 465-471. doi:10.1007/BF02293687

Appendix

Example Input and Output

Input and Output for Univariate Example

Input: The contents of the ASCII file ex1percentiles.txt are:
188.5
197.0
205.8
216.3
228.1

Output: The contents of the SAS output window are:
Percentile Power Method coefficients
C
205.8
13.869234
1.5221864
0.9625014

Output: The contents of ex1simdata.txt file are:
10.609710187
2.6051849323
75.05922126
29.634033258
62.861351473
80.326129351
99.236151941
83.577624847
100.16942799
81.800333938
52.766922741
110.68001386
43.654002567
74.415441414
52.073124328
77.869219855

43.866586956
126.84710015
80.419015418
58.74947208
112.64205437
30.900259549
96.584691093
105.65723356
44.858063646

Input and Output for Multivariate Example

Input: The contents of the ASCII file ex2percentiles.txt are:
27.176 15.500
33.667 25.350
41.444 90.339
48.901 180.737
59.500 600.529

Input: The contents of the ASCII file ex2correlations.txt are:
1 .10
.10 1

Output: The contents of the SAS output window are:
Percentile Power Method coefficients
C
41.444 90.339
10.78791 71.871855
1.1532084 132.53707
1.1102007 95.214843
Intermediate Pearson correlations
PI
1 0.1322937
0.1322937 1