1971

A note on generating multivariate data with desired means, variances, and covariances

Capra, J.R.; Elster, R.S.

Capra, J. R., and R. S. Elster. "A note on generating multivariate data with desired means, variances, and covariances." Educational and Psychological Measurement 31.3 (1971): 749-752.

https://hdl.handle.net/10945/60666

EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1971, 31, 749–752.

A NOTE ON GENERATING MULTIVARIATE DATA WITH DESIRED MEANS, VARIANCES, AND COVARIANCES

J. R. CAPRA AND R. S. ELSTER
Naval Postgraduate School
Monterey, California

THE problem to be discussed involves creating a set of n observations on p variables, with the p variables having specified means, variances, and covariances. The method which will be presented differs from those previously given by Kaiser and Dickman (1962) and Wherry, Naylor, Wherry, and Fallis (1965), in that the procedure does not use the models of principal component or factor analysis.

Derivation and Procedure

Let A be a p by n matrix of n independent observations on p variables. If the p variables are independent, with means of zero and unit variances, then, assuming the normal model A is distributed as a sample from a multivariate normal population with a mean vector of 0 and a variance-covariance matrix equal to the identity matrix. More succinctly, A is distributed as $N\ (0, I)$.

Anderson (1958, p. 21) showed that if one transforms A in the following way:

$$Z = CA, \tag{1}$$

then Z is distributed as $N(0,\ CIC')$ or $N(0,\ CC')$, where C is a p by p matrix used to transform A. Given a specified correlation matrix R, the problem is to decompose R such that $CC' = R$, where C will be a lower triangular matrix since R is symmetric.

If one can derive C and if one has specified the correlation matrix R, then a transformation exists which, when applied to A, will

give a set of observations on p variables with means of zero, unit variances, and the specified correlations among them. It is then a simple task to apply linear transformations in order to achieve the desired means and variances.

The numerical technique for deriving C from R uses Crout factorization (Kunz, 1957, pp. 226–229). The following recursion, which is easily programmed, allows C to be derived (Odell and Feiveson, 1966):

$$
\begin{aligned}
C_{i1} &= R_{i1}/\sqrt{R_{11}}, \qquad 1 \leq i \leq p \\
C_{ii} &= \sqrt{R_{ii} - \sum_{k=1}^{i-1} C_{ik}^2}, \qquad 1 < i \leq p \\
C_{ij} &= \left[R_{ij} - \sum_{k=1}^{i-1} C_{ik}C_{jk}\right] \Big/ C_{jj}, \qquad 1 < j < i \leq p \\
C_{ij} &= 0, \qquad i < j \leq p.
\end{aligned}
\tag{2}
$$

Ordinarily, a researcher will establish the n observations on each of the p variables of A by using a random number generator. The assumption made is that the random number generator will yield variables having zero means, unit variances, and zero intercorrelations. Because random number generators yield these characteristics only in the limit, a researcher may wish initially to adjust A such that in fact the p variables do have zero means, unit variances, and zero intercorrelations. (Of course, this operation will somewhat distort the randomness of the sample.) Nevertheless, to allow this adjustment, if the user judges it to be desirable, the following option is available.

Let X be a $p \times n$ matrix obtained from the random number generator, the sample mean vector being given by $\bar{x}$ and the sample variance-coveriance matrix being given by M. The goal is to transform X into a set of data with sample means of zero and a sample variance-covariance matrix of I. By definition,

$$XX' = M.$$

Consequently, what is needed is a matrix, D, such that

$$DXX'D' = DMD' = I.$$

This matrix could then be used to transform the matrix X into one having a sample variance-covariance matrix I:

$$Y = DX.$$

Now, since

$$DMD' = I,$$

then

$$M = D^{-1}D'^{-1},$$

which means we can obtain D^{-1} by Crout factorization and D itself by simple matrix inversion.

The sample means of this new set of observations are given by

$$\bar{y} = D\bar{x},$$

where $\bar{y}$ is a p component vector. If Y_i refers to the ith column of Y and $\bar{y}_i$ refers to the ith element of the $\bar{y}$ vector, a matrix with zero means can be obtained by subtracting $\bar{y}_i$ from each element in Y_i, for each i, from 1 to p.

Required Input Data

All that is required of the user are the correlation matrix, the desired mean and variance for each variable, and the sample size (number of observations on each variable) which he desires to have generated. Of course, the user should insure that the correlation matrix is nonsingular and positive semidefinite.

Output from the Program

The program will generate a multivariate sample from a population with the specified means, variances, and covariances. Sample means, variances, and correlation coefficients are also computed.

Summary

A method is shown for creating a set of n observations on p variables, with the p variables having specified means, variances, and covariances. This method differs from previous techniques in that it uses Crout factorization to develop the desired variance-covariance matrix instead of using the methods of component or factor analysis. Because the procedure assumes that it begins with p variables having zero means, unit variances, and zero intercorrelations, a procedure is also given for transforming the original data so that they fulfill these conditions.

REFERENCES

Anderson, T. W. *An introduction to multivariate statistical analysis.* New York: Wiley, 1958.

Kaiser, H. F. and Dickman, K. Sample and population score matrices and sample correlation matrices from an arbitrary population correlation matrix. *Psychometrika,* 1962, 27, 178–182.

Kunz, K. S. *Numerical analysis.* McGraw-Hill, 1957.

Odell, P. L. and Feiveson, A. H. A numerical procedure to generate a sample covariance matrix. *American Statistical Association Journal,* 1966, 61, 199–203.

Wherry, R. J., Sr., Naylor, J. C., Wherry, R. J., Jr., and Fallis, R. F. Generating multiple samples of multivariate data with arbitrary population parameters. *Psychometrika,* 1965, 30, 303–313.