SIMULATION OF TWIN DATA CONTROLLING POPULATION MEAN, VARIANCE, SKEWNESS AND KURTOSIS

JOE C. CHRISTIAN, JOAN E. BAILEY, MARY M. EVANS, K. W. KANG, JAMES A. NORTON Jr., P.L. YU

Departments of Medical Genetics and Psychiatry, Indiana University School of Medicine, Indianapolis, Indiana, USA

A computer system for simulation of quantitative twin data is being developed. The capability is being built in to simulate distributions with known means, standard deviations, skewness and kurtosis.

INTRODUCTION

Computer programs are being developed to simulate twin sampling experiments. These programs are designed to generate data to test ideas about the analysis of twin data, specifically to test the appropriatness and power of analyses in specific situations. Hopefully this system will provide a mechanism for teaching principles, facilitating communication among investigators and formulating new concepts of twin data analysis.

MATERIALS AND METHODS

Previously published analysis of variance models of twin data have been used as the basis for simulation (Christian et al. 1974). The variance and covariance components used in this model are listed in the Table. Presently all of the variance and covariance components shown in the Table have been simulated except epistatic effects (σ^2_i) and genetic environmental covariance (σ_{ge} and $\sigma_{ge}\star$). The mean, variance, skewness, and kurtosis may be controlled for the distribution of each of these sources of variance. One of the major influences which led to our simulation of twin studies was the need to test the robustness of various methods of analysis to deviations from normality. It therefore was critical to be able to simulate distributions with known skewness and kurtosis. Methods are generally available for generating pseudo-random values from many common distributions such as the uniform, normal, Poisson, etc. However, when these studies were started there was no readily available method of developing a continuous spectrum of distributions that varied by one or more parameters from normality. A method was developed for transforming, by a series of exponents, values from a simulated normal distribution so that the mean, variance and either skewness or kurtosis could be controlled. This method was used in the early computer software development. Ramberg and Schmeiser (1974) and Schmeiser (1971 and 1976) have reported the development of a method which allows generation of a family of continuous distributions in which the first four moments are controlled by the manipulation of four parameters. This method is more rapid than the one we developed in that less computer time is required and has the capability of controlling the mean and variance as well as skewness and kurtosis at the same time. At the present time the method of Ramberg and Schmeiser is being incorporated into our system and tested for accuracy and reproducibility.

The original programs were written in Basic-Plus (Digital Equipment Corp., 1974) be-

This is publication No. 76-35 from the Department of Medical Genetics and was supported in part by the Indiana University Human Genetics Center, PHS GM 21054, NIH Contract 71-2307, Training Grant PHS T01 GM 1056, and a grant from the John A. Hartford Foundation, Inc.

Table. *Analysis of variance model for twin studies*

Source of variation	*df*	Mean squares	Expected value of mean squares
MZ twins:			
Among pairs ..	n_{MZ}—1	M_{AMZ}	$2\sigma_a^2+2\sigma_d^2+2\sigma_i^2+\sigma_e^2+4\sigma_{ge}+C_{MZ}$
Within pairs ..	n_{MZ}	M_{WMZ}	σ_e^2 —C_{MZ}
DZ twins:			
Among pairs ..	n_{DZ}—1	M_{ADZ}	$3/2\sigma_a^2+5/4\sigma_d^2+(1+f)\sigma_i^2+\sigma_e^2+2(\sigma_{ge}+\sigma\bigstar_{ge})+C_{DZ}$
Within pairs ..	n_{DZ}	M_{WDZ}	$1/2\sigma_a^2+3/4\sigma_d^2+(1-f)\sigma_i^2+\sigma_e^2+2(\sigma_{ge}-\sigma\bigstar_{ge})-C_{DZ}$

Note: n_{MZ} = number of MZ twin pairs; n_{DZ} = number of DZ twin pairs; *df* = degrees of freedom; σ^2_a = variance component due to additive genetic effects; σ^2_d = variance component due to dominant genetic effects; σ^2_i = variance component due to epistatic genetic effects; σ^2_e = variance component due to environmental effects; σ_{ge} = covariance between genetic and environmental effects in the same individual; $\sigma\bigstar_{ge}$ = covariance between genetic effects on one member of a twin pair and environmental effects of the other member of that twin pair; C_{MZ} = covariance among environmental effects between pairs of MZ twins; C_{DZ} = covariance among environmental effects between pairs of DZ twins; and f = one minus the fraction of epistatic variance manifest within DZ twin sets (Christian et al. 1974).

cause of system limitations, but are now being translated into Fortran IV (Stuart 1970).

RESULTS AND DISCUSSION

In the early stages a method was developed for simulating twin samples from normally distributed populations which was helpful in testing methods of twin analyses (Christian et al. 1974). A specific example is in a paper describing the theoretically most appropriate test of the difference between the means of MZ and DZ twins (Christian and Norton 1977). Twin sampling experiments were simulated which confirmed that the theoretically most appropriate test of this difference indeed had the predicted error rate and also gave measures of the amount of bias in alternative tests.

Under the best of conditions there are assumptions of the twin model that cannot be adequately tested. It therefore becomes important to assess the direction and magnitude of biases introduced by failures of these assumptions. We believe that simulated sampling experiments will be a useful tool in assessing possible effects in the failure of assumptions and a means of communicating ways of correcting deficiencies in the methodology of twin analyses.

REFERENCES

Christian J.C., Kang K.W., Norton J.A. Jr. 1974. Choice of an estimate of genetic variance from twin data. Am. J. Hum. Genet., 26: 154-161.

Christian J.C., Norton J.A. Jr. 1977. A proposed test of the difference between the means of monozygotic and dizygotic twins. Acta Genet. Med. Gemellol., 26: 49-53.

Digital Equipment Corporation 1974. Basic Plus Language Manual Maynard, Massachusetts.

Ramberg J.S., Schmeiser B.W. 1974. An approximate method for generating asymmetric random variables. Communications of Association for Computing Machinery, 17: 78-82.

Schmeiser B.W. 1971. A general algorithm for generating random variables. Master's Thesis, University of Iowa, Iowa City.

Schmeiser B.W. 1976. Personal Communication. Department of Industrial Engineering and Operations Research, Southern Methodist University, Dallas, Texas 75275.

Stuart F. 1970. Fortran Programming. New York: John Wiley and Sons.

Joe C. Christian, M.D., Department of Medical Genetics, Indiana Univeristy Medical Center, 1100 West Michigan Street, Indianapolis, Indiana 46202, USA.