Behavior Research Methods, Instruments, & Computers
1995, 27 (3), 400-403

NORMUL: A FORTRAN program for testing multivariate normality

RONALD R. HOLDEN and MICHAEL PARENT
Queen's University, Kingston, Ontario, Canada

NORMUL is a FORTRAN program that provides a test of whether data conform to a multivariate normal distribution. The method involves correlating Mahalanobis distances for observed data with expected chi-square percentile values. This obtained correlation is then tested for significance by empirically evaluating the probability of its belonging to a distribution generated from multivariate normal data.

Multivariate significance tests generally assume that observed data conform to some particular distributional form, usually a multivariate normal distribution. Although many such tests may still be applicable in situations where distributional assumptions are not met, the extent of robustness is generally unknown and cannot be assumed. Furthermore, established robustness may be specific to particular types of errors. For example, although multivariate non-normality has only a small effect on Type I error rates for test statistics in multivariate analysis of variance, non-normality can have substantial effects on Type II error rates (Stevens, 1992, p. 247). NORMUL is a mainframe computer program that provides a test of multivariate normality by examining the degree to which the Mahalanobis distances associated with observed cases adhere to an expected chi-square distribution. The probability for the degree of conformity with the observed data is evaluated relative to that for random multivariate normal data drawn from a population whose covariance matrix is identical to that of the observed data.

Various tests of multivariate normality have been proposed, but few have come into widespread acceptance or are readily available. General multivariate texts (e.g., Johnson & Wichern, 1992; Stevens, 1992) suggest a graphical test of multivariate normality whereby ordered Mahalanobis distances for the observed data are paired and plotted with percentiles of the chi-square distribution having degrees of freedom equal to the number of observed variables. According to Johnson and Wichern, if both NS (the number of cases) and NS − NV (the number of variables) exceed 25, the resultant plot based on multivariate normal data should be a straight line. Following Looney and Gulledge's (1985) formal hypothesis testing of univariate normality for probability plots using the Pearson product-moment correlation coefficient, NORMUL extends the technique to multivariate data.

The Mahalanobis distance for each subject j is

$$d_j^2 = (\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}), \quad j = 1, 2, \ldots, \mathrm{NS},$$

where $\mathbf{x}_j$ is the vector of scores for the jth case, $\bar{\mathbf{x}}$ is the vector of means for the variables, and S is the covariance matrix. Based on Looney and Gulledge's (1985) examination of the power of Blom's (1958) plotting position, the percentile of the chi-square distribution with NV degree of freedom is designated:

$$\chi^2_{\mathrm{NV}}((j-.375)/(\mathrm{NS}+.25)), \quad j = 1, 2, \ldots, \mathrm{NS}.$$

Multivariate normality is then represented by the correlation between the ordered chi-square percentile values and observed Mahalanobis distances:

$$r(\chi^2_{\mathrm{NV}}((j-.375)/(\mathrm{NS}+.25)), d_j^2).$$

Because the distribution of this correlation under the null hypothesis of multivariate normality is unknown, empirical sampling methods are used to evaluate the probability of the observed correlation belonging to the distribution of correlations associated with normally distributed data.

Program Description

NORMUL is a mainframe computer program written in VS FORTRAN, using double-precision subroutines from IMSL Version 10 (International Mathematical and Statistical Libraries, 1987). The user indicates the numbers of variables, cases (or subjects), and random data trials. Observed data are input as a cases by variables matrix for which an appropriate format statement must be tailored in the program.

The program output lists the input data, pairs of ordered chi-square percentiles and corresponding observed Mahalanobis distances, the correlation between these pairs, and the probability of this correlation. This probability is based on the likelihood of the correlation for random multivariate normal data from a distribution of samples having the same covariance matrix as the observed data. In addition, output includes the mean, standard deviation, and standard error of the mean for the correlations based on normal data. Finally, a listing of all correlations for the randomly sampled normal data is provided.

Example

Multivariate non-normal data were simulated by using DATASIM (Bradley, 1988). For 100 cases with four variables, one pair of random variables was constructed to be strongly correlated, with each variable's distribution being positively skewed and leptokurtic, while another pair of random variables was developed to be strongly correlated, with each variable's distribution following an exponential distribution (i.e., extremely negative skewed

Correspondence should be addressed to R. R. Holden, Department of Psychology, Queen's University, Kingston, ON, Canada K7L 3N6 (e-mail: holdenr@qucdn.queensu.ca).

Table 1
Generation of Multivariate Non-Normal Data

DATASIM Code (Seed = 814860)

```
? design none 4, nobs 100
? mu .5 .5 27.1 27.1
? sigma .2 .2 3 3
? decimal 3 3 0 0
? lambda c1–c2 −.886 .1333 .0193 .1588
? lambda c3–c4 .993 −.001081 −.001076 −.00000407
? rho .7 .3 .2 \
      .2 .3 \
        .8 \
```

Resultant Variable Descriptive Statistics

		Standard			Variable Intercorrelations			
Variable	Mean	Deviation	Skewness	Kurtosis	1	2	3	4
1	0.500	0.185	0.784	−0.020		.710	.179	.064
2	0.498	0.174	0.796	1.076			.159	.218
3	27.270	2.741	−1.569	2.467				.799
4	27.150	2.826	−1.586	3.303				

and leptokurtic). Between pairs of variables, correlations were designated as moderate. This simulation might be representative of a pretest–posttest design involving two moderately related dependent variables. DATASIM commands are shown in Table 1, as is descriptive information concerning the resultant simulated multivariate non-normal data. For comparison purposes, a set of multivariate normal data conforming to the same parameters as those of the non-normal data was also generated.

In Figure 1, Mahalanobis distances for the sets of simulated non-normal and normal data are plotted relative to corresponding chi-square percentiles associated with four degrees of freedom (i.e., the number of variables). Although such plots may be "eyeballed" for linearity, inferences drawn are subject to the vagaries of individual observers and, hence, may be subjective, unreliable, or invalid.

Examination of the simulated non-normal data through NORMUL is presented in an abridged form in Table 2. Although the correlation between the ordered chi-square percentiles and Mahalanobis distances for the simulated non-normal data exceeds .9645, the probability of this value for multivariate normal data is less than .008. Consequently, on the basis of a conventional desired Type I error rate of .05, this would result in a rejection of the null hypothesis that the simulated data are multivariate normal. In contrast, for the generated multivariate normal data, a correlation between the ordered chi-square percentiles and Mahalanobis distances of .9960, $p > .83$, would not lead to a rejection of the null hypothesis of multivariate normality.

Limitations

NORMUL is a mainframe program requiring a FORTRAN compiler and access to IMSL subroutines. Users are cautioned that even moderate numbers of variables (e.g., 4) and random trials (e.g., 1,000) can require a nontrivial amount of computer time.

The statistical power associated with using NORMUL remains to be empirically evaluated. Preliminary testing with 1,000 sets of multivariate non-normal data generated with the use of the DATASIM code in Table 1 indicates a power of .92 for a Type I error rate of .05. Nevertheless, further research on the power for varying types of non-normal distributions and on the comparative merits of this approach to other test statistics is clearly desirable.

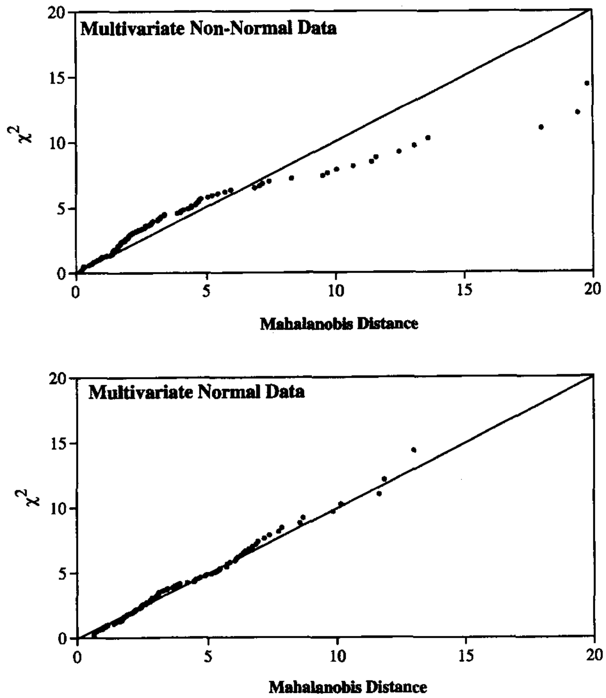

Figure 1. Plots of paired ordered chi-square percentile values and observed Mahalanobis distances for non-normally and normally distributed data.

Table 2
Abridged NORMUL Output

```
INPUT DATA

0.707      0.760      29.000      30.000

0.317      0.458      28.000      28.000

  .          .           .           .

  .          .           .           .

  .          .           .           .

0.374      0.491      29.000      30.000

0.487      0.420      29.000      28.000

CHI             D
SQUARES         SQUARES
ORDERED         ORDERED
 0.232           0.226

 0.384           0.279

   .               .

   .               .

   .               .

11.034          17.978

12.159          19.417

14.359          19.782

ORDERED CHI-SQUARES AND D-SQUARES CORRELATION = 0.9645396
                                        PROBABILITY < 0.008000

FOR 1000 TRIALS OF RANDOM MULTIVARIATE NORMAL DATA, THE DISTRIBUTION OF
CORRELATIONS HAS THE FOLLOWING CHARACTERISTICS:

                                 MEAN = 0.9912724
                   STANDARD DEVIATION = 0.0063720
               STANDARD ERROR OF MEAN = 0.0002015

ORDERED RANDOM MULTIVARIATE NORMAL CORRELATIONS ARE:

0.9339883  0.9417402  0.9510506  0.9591396  0.9634058  0.9639581  0.9640943  0.9648002

0.9651556  0.9659209  0.9662724  0.9663322  0.9668088  0.9685802  0.9688609  0.9690301

    .          .          .          .          .          .          .          .

    .          .          .          .          .          .          .          .

    .          .          .          .          .          .          .          .

0.9978267  0.9978484  0.9978566  0.9978723  0.9978893  0.9979016  0.9980077  0.9980126

0.9981002  0.9981382  0.9981840  0.9982246  0.9982527  0.9982543  0.9982786  0.9984903
```

Availability

The program NORMUL, example data, and output are available at no charge either through e-mail (holdenr@qucdn.queensu.ca) or by sending a DOS-formatted floppy disk to R. R. Holden, Department of Psychology, Queen's University, Kingston, ON Canada K7L 3N6.

REFERENCES

Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York: Wiley.

Bradley, D. R. (1988). *DATASIM* [Computer program]. Lewiston, ME: Desktop Press.

International Mathematical and Statistical Libraries (1987). *International Mathematical and Statistical Libraries reference manual* (10th ed.). Houston: Author.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Looney, S. W., & Gulledge, T. R., Jr. (1985). Use of the correlation coefficient with normal probability plots. *American Statistician*, **39**, 75-79.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

(Manuscript received February 21, 1994;
revision accepted for publication July 20, 1994.)